Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 857

Maximilian Alexander Coenen

Probabilistic Pose Estimation and 3D Reconstruction of Vehicles from Stereo Images

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5269-7

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover
ISSN 0174-1454, Nr. 362, Hannover 2020

Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 857

Probabilistic Pose Estimation and 3D Reconstruction of Vehicles from Stereo Images

Von der Fakultät für Bauingenieurwesen und Geodäsie

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

Vorgelegt von

Maximilian Alexander Coenen, M.Sc.

Geboren am 23.04.1987 in Köln

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5269-7

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover
ISSN 0174-1454, Nr. 362, Hannover 2020

Adresse der DGK:

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften (DGK)
Alfons-Goppel-Straße 11 ● D – 80 539 München
Telefon +49 – 331 – 288 1685 ● Telefax +49 – 331 – 288 1759
E-Mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission:

Vorsitzender: Prof. Dr.-Ing. habil. Christian Heipke

Referent: apl. Prof. Dr. techn. Franz Rottensteiner

Korreferenten: apl. Prof. Dr.-Ing. Claus Brenner
Prof. Dr.-Ing. Olaf Hellwich (TU Berlin)

Tag der mündlichen Prüfung: 25.05.2020

ISSN 0065-5325 ISBN 978-3-7696-5269-7

Abstract

The pose estimation and reconstruction of 3D objects from images is one of the major problems that are addressed in computer vision and photogrammetry. The understanding of a 3D scene and the 3D reconstruction of specific objects are prerequisites for many highly relevant applications of computer vision such as mobile robotics and autonomous driving. To deal with the inverse problem of reconstructing 3D objects from their 2D projections, a common strategy is to incorporate prior object knowledge into the reconstruction approach by establishing a 3D model and aligning it to the 2D image plane. However, current approaches are limited due to inadequate shape priors and the insufficiency of the derived image observations for a reliable association and alignment with the 3D model. The goal of this thesis is to infer valuable observations from the images and to show how 3D object reconstruction can profit from a more sophisticated shape prior and from a combined incorporation of the different observation types.

To achieve this goal, this thesis presents three major contributions for the particular task of 3D vehicle reconstruction from street-level stereo images. First, a subcategory-aware deformable vehicle model is introduced that makes use of a prediction of the vehicle type for a more appropriate regularisation of the vehicle shape. Second, a Convolutional Neural Network (CNN) is proposed which extracts observations from an image. In particular, the CNN is used to derive a prediction of the vehicle orientation and type, which are introduced as prior information for model fitting. Furthermore, the CNN extracts vehicle keypoints and wireframes, which are well-suited for model association and model fitting. Third, the task of pose estimation and reconstruction is addressed by a versatile probabilistic model. Suitable parametrisations and formulations of likelihood and prior terms are introduced for a joint consideration of the derived observations and prior information in the probabilistic objective function. As the objective function is non-convex and discontinuous, a proper customized strategy based on stochastic sampling is proposed for inference, yielding convincing results for the estimated poses and shapes of the vehicles.

To evaluate the performance and to investigate the strengths and limitations of the proposed method, extensive experiments are conducted using two challenging real-world data sets: the publicly available KITTI benchmark and the ICSENS data set, which was created in the scope of this thesis. On both data sets, the benefit of the developed shape prior and of each of the individual components of the probabilistic model can be shown. The proposed method yields vehicle pose estimates with a median error of up to 27 cm for the position and up to 1.7° for the orientation on the data sets. A comparison to state-of-the-art methods for vehicle pose estimation shows that the proposed approach performs on par or better, confirming the suitability of the developed model and inference procedure.

Keywords detection, 3D reconstruction, pose estimation, CNN, active shape model, probabilistic model, Monte Carlo sampling

Kurzfassung

Die Bestimmung der Pose und die 3D Rekonstruktion der Form von Objekten aus Bildinformation ist eine der zentralen Aufgaben im Bereich Computer Vision und Photogrammetrie. Eine Rekonstruktion der Umgebung und bestimmter Objekte ist Voraussetzung für aktuell relevante Anwendungen, bspw. im Bereich der Robotik oder des autonomen Fahrens. Eine übliche Strategie, um mit dem schlecht konditionierten Problem der Bestimmung von 3D Objekten aus 2D Bildern umzugehen, ist die Einführung von a priori Wissen über die Objekte in die Rekonstruktion. Hierzu wird ein 3D Modell des Objektes angenommen und so in die Daten eingepasst, dass es mit den Bildbeobachtungen übereinstimmt. Bestehende Repräsentationen von a priori Modellen sind jedoch unzulänglich und die Beobachtungen, die aus den Bildern abgeleitet werden, reichen nicht für eine verlässliche Zuordnung und Einpassung der Modelle aus. Das Ziel dieser Arbeit ist es, geeignete Beobachtungen aus den Bildern zu generieren und zu zeigen, wie die 3D Rekonstruktion von der Nutzung einer verbessertern Objektrepräsentation und einer gemeinsamen Berücksichtigung mehrerer Beobachtungstypen profitieren kann. Vor diesem Hintergrund werden in dieser Arbeit vorranging drei wissenschaftliche Beiträge präsentiert, die zum Ziel der Rekonstruktion von Fahrzeugen aus Stereobildern entwickelt wurden. Zum Einen wird ein neuartiges differenziertes Fahrzeugmodell vorgestellt, welches eine vorausgehende Prädiktion des Fahrzeugtyps nutzt, um eine bessere Repräsentation des beobachteten Fahrzeugs zu erhalten. Des Weiteren wird ein CNN präsentiert, welches genutzt wird, um semantische Beobachtungen, wie charakteristische Landmarken oder Kanten der Fahrzeuge, die sich als sehr nützlich für die Modelleinpassung erweisen, aus den Bildern abzuleiten. Das CNN wird außerdem dazu genutzt, um den Fahrzeugtyp sowie die Fahrzeugorientierung zu prädizieren, welche als a priori Wissen in die Modelleinpassung mit einfließen. Zuletzt wird ein umfassendes probabilistisches Modell für die 3D Rekonstruktion entwickelt, welches alle extrahierten Informationen berücksichtigt. Hierzu werden passende Parametrisierungen und Formulierungen der Likelihood-Funktionen sowie der a priori Terme vorgestellt. Ein auf stochastischem Sampling bestehendes Optimierungsverfahren wird für die Inferenz der nicht-konvexen und unstetigen Zielfunktion entwickelt.

Umfangreiche Experimente werden anhand von zwei anspruchsvollen Datensätzen durchgeführt um die Qualität sowie Stärken und Limitierungen des vorgestellten Verfahrens zu analysieren. Der Nutzen, der durch das vorgestellte Fahrzeugmodell sowie durch die einzelnen Komponenten des probabilistischen Modells erzielt wird, kann auf beiden Datensätzen, dem KITTI sowie dem IC-SENS Datensatz, einem im Rahmen dieser Arbeit erzeugten Datensatz, nachgewiesen werden. Ein Vergleich der erzielten Ergebnisse zu verwandten Arbeiten bestätigt die Eignung des entwickelten Verfahrens für die Posenbestimmung und die Rekonstruktion von Fahrzeugen.

Schlagworte Detektion, 3D Rekonstruktion, Posen Bestimmung, CNN, Active Shape Model, Probabilistisches Modell, Monte Carlo Sampling

Nomenclature

Acronyms

ASM	Active Shape Model
CE	Cross-Entropy
CNN	Convolutional Neural Network
DPM	Deformable Part Model
DoF	Degree of Freedom
FCN	Fully Convolutional Network
FPN	Feature Pyramid Network
HoG	Histogram of oriented Gradients
IoU	Intersection over Union
MAD	Median Absolute Deviation
MAP	Maximum a Posterior
MCM	Monte Carlo Methods
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
OA	Overall Accuracy
PCA	Principal Component Analysis
RCNN	Region Convolutional Neural Network
RMS	Root Mean Square
SfM	Structure from Motion
SVM	Support Vector Machine

SGD	Stochastic Gradient Descent
TSDF	Truncated Signed Distance Function

Coordinate Systems

${}^{M}C$	The camera model coordinate system
${}^{\Omega}C$	The ground plane coordinate system
${}^{A}C$	The vehicle body coordinate system
Q	Rotation matrix describing the rotation between the coordinate systems ${}^{M}C$ and ${}^{\Omega}C$
T	Vector describing the translation between the coordinate systems ${}^{M}C$ and ${}^{\Omega}C$

Detection

$\mathcal{B}$	Image bounding box
$\mathcal{V}$	Set of detected vehicles
Ω	3D ground plane
Φ	Probabilistic free-space grid map
$\mathbf{X}_v$	The set of 3D points belonging to vehicle v
$\mathbf{X}_\Omega$	The set of 3D points belonging to the ground plane Ω
$\mathbf{X}_{Obj}$	The set of 3D points belonging to an arbitrary object
v	Individual vehicle detection $v \in \mathcal{V}$
σ_x	Depth uncertainty of a stereo triangulated 3D point
ρ_g	Probability of grid cell $g \in \Phi$ for being free space
$\mathbf{o}^{det}$	Observation vector derived from the vehicle detection with $\mathbf{o}^{det} = (\mathbf{X}_v, {}^{l}\mathcal{B}, {}^{r}\mathcal{B})$

Active Shape Model

$\mathcal{K}$	Set of keypoints/vertices of the ASM
$\mathcal{K}_A$	Set of *appearance keypoints* of the ASM with $\mathcal{K}_A \subset \mathcal{K}$
$\mathcal{T}$	Set of different vehicle types

M_{Δ}	Triangulated surface of the ASM
$M_{\mathcal{W}}$	Wireframe of the ASM
$M(\gamma)$	Active shape model deformed according to the shape parameter vector γ
$M(\mathbf{s})$	Active shape model deformed and transformed according to the state vector $\mathbf{s}$
ν_n	A training exemplar to learn the ASM
γ	Shape vector containing the shape parameters defining a deformed ASM
σ_s	Square root of the s'th eigenvalue of the ASM
$\mathbf{e}_s$	The s'th eigenvector of the ASM
$\mathbf{m}$	Mean shape of the ASM

CNN

$\Pi_{\mathcal{T}}$	Probability distribution for the vehicle type
Π_{ϑ}	Probability distribution for the vehicle viewpoint ϑ
$\mathcal{H}_{\mathcal{K}}$	Keypoint probability maps
$\mathcal{H}_{\mathcal{W}}$	Wireframe probability maps
$\mathcal{L}_{(\cdot)}$	Loss function
ϑ	Vehicle viewpoint
$\mathbf{o}^{cnn}$	Observation vector derived from the CNN with $\mathbf{o}^{cnn} = ({}^{l}\mathcal{H}_{\mathcal{K}}, {}^{l}\mathcal{H}_{\mathcal{W}}, {}^{r}\mathcal{H}_{\mathcal{K}}, {}^{r}\mathcal{H}_{\mathcal{W}})$

Probabilistic model

$\mathbf{s}$	State vector $\mathbf{s} = (\mathbf{t}, \theta, \gamma)$
$\mathbf{t}$	Vehicle position on the groundplane Ω
θ	Vehicle heading, i.e. the rotation angle about the normal vector of the ground plane Ω
$\mathbf{o}$	Observation vector $\mathbf{o} = (\mathbf{o}^{det}, \mathbf{o}^{cnn})$ associated to every vehicle detection v
$p(\mathbf{X}_v \vert \mathbf{s})$	The 3D likelihood
$p(\mathcal{H}_{\mathcal{K}} \vert \mathbf{s})$	The keypoint likelihood
$p(\mathcal{H}_{\mathcal{W}} \vert \mathbf{s})$	The wireframe likelihood

$p(\mathbf{t})$	The position prior
$p(\theta)$	The orientation prior
$p(\gamma)$	The shape prior

Inference

N_j	Number of offspring particles of iteration j
n_p	Overall number of particles
n_b	Number of best scoring particles
$\mathbf{s}_j$	Set of particles of iteration j
w_j	Importance weights of the particles of iteration j

Contents

1 Introduction

> *"In many cases it is sufficient to know or assume that an object owns a kind of certain regularities in order to interpret the correct 3D shape from its perspective image."*
>
> *Hermann von Helmholtz, 1867*

The image based reconstruction of three-dimensional (3D) scenes and objects is a major topic of interest in computer vision and photogrammetry. While humans are able to understand and reason about 3D geometry from a short glimpse of the scene and are even capable of predicting the extends and shape of partially occluded or hidden objects, this task is very challenging for vision algorithms. The perspective projection from 3D to the 2D image plane leaves many ambiguities about 3D objects, causing their reconstruction and the retrieval of their pose and shape to be ill-posed and difficult to solve. Nonetheless, 3D scene understanding and 3D reconstruction of specific target objects has a great relevance for several disciplines such as medicine (Angelopoulou et al., 2015), augmented reality (Yang et al., 2013), robotics (Du et al., 2019), and autonomous driving (Janai et al., 2020). The motivation of this thesis is founded by applications related to the latter discipline. The highly dynamic nature of street environments is one of the biggest challenges for autonomous driving applications. The precise reconstruction of moving objects, especially of other cars, is fundamental to ensure safe navigation and to enable applications such as interactive motion planning and collaborative positioning. To this end, cameras provide a cost-effective solution to deliver perceptive data of a vehicle's surroundings.

Given this background, this thesis presents a method for precise vehicle reconstruction from street-level stereo images. The poses of other vehicles w.r.t. the observing vehicle, i.e. their relative positions and orientations in 3D, can directly be derived from the reconstructions. Throughout this thesis the following definition of pose and pose related notions are used:

The notion of a 6DoF (6 degrees of freedom) object pose contains the position of the object's reference point in form of its 3D coordinates in the reference coordinate frame as well as the 3D orientation of the object in the same coordinate frame. In this context, the 3D orientation comprises three rotation angles which describe the rotation of the object body frame, which is rigidly attached to the vehicle and centred at its reference point, w.r.t. the reference coordinate frame. In addition to the orientation, the notion of an object's viewpoint is introduced. In 3D, the viewpoint describes the two rotation angles representing the elevation and azimuth of the object body frame *w.r.t. to a 3D ray, e.g. an image ray, intersecting the object reference point. The notion of a 3DoF (3 degrees of freedom) object pose in a 2D reference frame consists of two coordinates representing the position and one angle, representing the orientation. In accordance to that, a viewpoint in 2D is also represented by one angle representing the azimuth.*

Geometrically, the image-based reconstruction of 3D objects from their 2D projections is an inverse problem, suffering from the ambiguous mapping inherited by the perspective projection, causing the task to be ill-posed. To address the class of ill-posed problems, a common approach is to introduce suitable constraints, e.g. derived from prior knowledge about the 3D shape, a procedure which is also called *regularisation* (Poggio et al., 1985). Thus, an essential strategy in work on image based object reconstruction is to establish a 3D model and align it to the 2D image plane. To relax the requirement for precisely known object models, parametrised deformable shape priors can be formulated, increasing the set of free parameters by the parameters defining the shape. For object reconstruction usually entities such as keypoints, edges, or contours of the deformable 3D model are aligned to their corresponding counterparts localised in the image. Often, the alignment is performed by minimising the reprojection error between projections of the 3D model entities and the associated 2D image detections (Pavlakos et al., 2017; Leotta and Mundy, 2009). However, this procedure is susceptible to outliers and erroneous detections. Establishing a sufficient number of correct and sufficiently precise correspondences between model and image entities is a difficult problem. Existing approaches use model edges and contours to align them with image edges, usually derived from gradient information (Leotta and Mundy, 2009; Ramnath et al., 2014; Ortiz-Cayon et al., 2016). However, these approaches are highly sensitive to illumination, reflections, contrast, object colour and model initialisation, because these factors can cause erroneous edge-to-edge correspondences, thus prohibiting a correct model alignment. Other approaches utilise keypoint detections to be used for model alignment (Pavlakos et al., 2017; Murthy et al., 2017b; Chabot et al., 2017). However, compared to edges, keypoints are less stable since they deliver less geometric constraints compared to edges. In addition, feature representations in the image domain lead to the effect that already small localisation errors in 2D are likely to cause large errors in 3D space. Stereo-image based approaches use 3D points reconstructed from stereo correspondences to align the 3D shape model (Engelmann et al., 2016). However, the 3D points lack the information about their exact counterpart on the model and are usually associated to the closest point on the model surface, thus demanding for a good model initialisation. Often, these difficulties define the limitations of current state-of-the art methods for model based object reconstruction as they cause such approaches to be highly sensitive to, for instance, noisy feature detection, illumination conditions, model initialisation, and/or imprecise keypoint localisation.

Following the path of model based object reconstruction, these limitations are addressed in this thesis on the basis of stereo images, firstly by training a Convolutional Neural Network (CNN) for the extraction of vehicle entities, such as keypoint and wireframes, as well as for the derivation of prior knowledge, and secondly by formulating a comprehensive probabilistic model for vehicle fitting, building upon the outputs of the CNN and stereo information. A stereoscopic setup leads to less flexibility concerning data acquisition but helps to constrain the ill-posed problem of 3D reasoning. Research objectives raised in the context of this thesis are related to:

- a contrast and illumination insensitive extraction of suitable features for the model alignment. In this context, a CNN is exploited to predict semantic vehicle keypoints and semantic vehicle wireframe edges, the latter being proposed as new feature for model fitting in this thesis.
- deriving proper state priors for object pose and shape using a CNN to establish meaningful constraints for regularising the ill-posed problem. In this context, a deep learning based prediction of the vehicle viewpoint and the vehicle type is proposed as prior information for the vehicle's orientation and its shape, thus serving as direct observation of the target parameters,
- a probabilistic fusion of different alignment strategies both, in 2D and 3D, exploiting synergistic effects anticipated from the complementary and/or redundant nature of the fused data,
- gaining robustness against noisy, false or imprecise feature detections by formulating a probabilistic model while properly incorporating observation and detection uncertainties and
- establishing an optimisation procedure that is insensitive to initialisation errors and local optima.

1.1 Contributions

This thesis presents a methodology for the detection and 3D reconstruction of vehicles from stereoscopic images to finally obtain precise estimates for the vehicle's pose and shape. Based on initially detected vehicles, the contributions of this thesis are:

- **A new subcategory-aware deformable vehicle model to be used as shape prior.** In extension to existing work, e.g. (Zia et al., 2013), where a deformable shape model is always learned for the entire class *vehicle*, this work presents a shape model which also learns individual modes for different vehicle subcategories. Thus, the proposed model allows a more detailed shape regularisation if an a priori prediction of the vehicle type is available. The presented shape prior leads to better constraints on the vehicle shape and evidentially also enhances the results of pose estimation.
- **A new multi-task CNN for the extraction of suitable features and a priori estimates.** The CNN simultaneously detects vehicle keypoints and vehicle wireframe edges, the latter being proposed as new feature which is in contrast to standard gradient based edge

extraction which highly depends on illumination conditions and contrast. Furthermore, as the CNN is trained to distinguish between semantically different wireframe edges belonging to different parts/sides of the vehicle, better prospects for model-to-image-edge correspondences are obtained, too. To achieve this goal, a new loss function for the tasks of keypoint and wireframe prediction is introduced. The CNN also delivers a probability distribution for the vehicle's orientation as well as the vehicle's subcategory. For the orientation estimation, a new hierarchical classifier and a hierarchical class structure are proposed. Suitable state priors for orientation and shape are derived from the predicted probability distributions.

- **A comprehensive probabilistic model for vehicle reconstruction.** The probabilistic formulation combines multiple observation likelihoods, based on the keypoint and wireframe probability maps predicted by the CNN, as well as on stereo-reconstructed 3D points. Building the model directly on the basis of the raw keypoint and wireframe probability maps makes the explicit association of image and model entities obsolete and reduces the impact of both, false as well as imprecise keypoint/wireframe localisations. Furthermore, state prior terms for each group of parameters, namely position, orientation and shape, are incorporated as regularisation terms based on derived scene knowledge and the probability distributions derived from the CNN. The probabilistic model is complemented with a Monte-Carlo based inference procedure similar to (Zia et al., 2013), that is tailored to reduce the sensitivity w.r.t. parameter initialisation and local optima of the probabilistic formulation.

1.2 Thesis outline

The thesis is structured as follows. In Chapter 2, fundamental concepts for understanding the presented methodology are given. Chapter 3 puts this work into the context of related work on image based object reconstruction and pose estimation with a focus on vehicles as objects of interest. A detailed explanation of the methodological contributions of this thesis is provided in Chapter 4. The setup of the experiments to evaluate the method is described in Chapter 5, and their results are shown and discussed in Chapter 6. A conclusion and outlook on future work are given in Chapter 7.

2 Basics

This chapter presents mathematical and conceptual foundations that are relevant for the methodology presented in this thesis. Sec. 2.1 provides an overview on the functionality of Convolutional Neural Networks (CNN) and presents basic architectures that are used or adapted within this thesis. In Sec. 2.2, the Active Shape Model (ASM), used here to represent a deformable vehicle model, is described, while Sec. 2.3 introduces Monte Carlo techniques for parameter optimisation.

2.1 Convolutional Neural Networks

For several years, Convolutional Neural Networks (CNN) (LeCun et al., 1990) have shown enormous success in solving computer vision problems of various kinds, including image classification (Simonyan and Zisserman, 2015), object detection (Ren et al., 2015), semantic segmentation (Long et al., 2015) and instance segmentation (He et al., 2017). The basic principles and components of CNNs are outlined in this section. Furthermore, examples of selected specific CNN architectures which serve as a basis for the methodology developed in this thesis are presented.

In the context of image analysis, usually the input to the neural network is a digital (single or multi-channel) image. Generally speaking, a CNN is composed of a series of convolutional layers, followed by a non-linear activation function and pooling layers. Fig. 2.1 shows an example of this structure. In contrast to the general multi-layer perceptron (MLP) (Goodfellow et al., 2016), in which each neuron computes a linear mapping, i.e. a weighted sum, of the input and the input topology is therefore lost, CNNs preserve the spatial structure of the input image and weights are shared in the individual layers. Mathematically, this procedure leads to the utilisation of convolution filters in the convolutional layers, whose spatial size defines the receptive field of the filter and whose depth corresponds to the number channels of the respective input layer (cf. Fig. 2.1). Optionally, an additive bias is applied in every convolution, introducing one parameter to each convolution filter in addition to the filter weights. The output of the convolution filters is passed through a non-linear activation function to enable the modelling of non-linear mappings by a CNN. Examples for commonly applied activation functions are the Sigmoid function (Han and Moraga, 1995) with

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

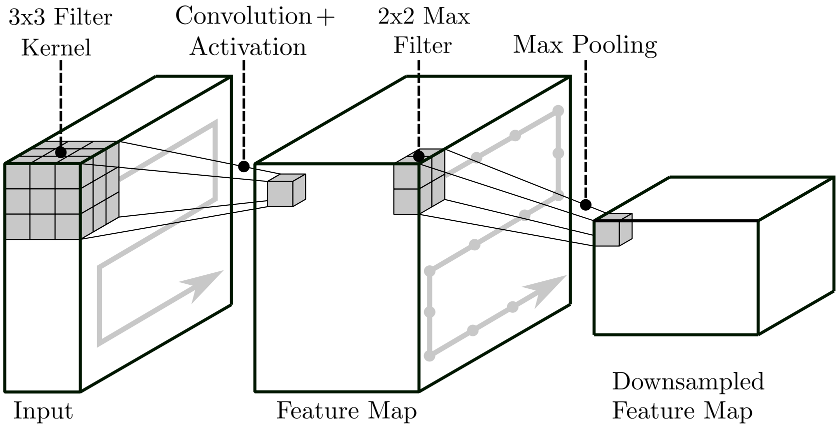


Figure 2.1: Schematic overview of a convolutional layer, followed by a max pooling layer. Convolution filters (here: spatial kernel size of 3x3) followed by an activation function are used to create the feature maps. The kernel weights are shared for all positions of the input. In this example, a 2x2 max pooling filter with stride 2, indicated by the dotted array, is used to create the spatially downsampled feature map.

or the Rectified Liner Unit (ReLU) function (Hahnloser et al., 2000) with

$$\mathrm{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } x > 0. \end{cases} \tag{2.2}$$

Within one convolutional layer, multiple local filter kernels are applied to all positions of the input to produce intermediate outputs which are often referred to as feature maps (cf. Fig. 2.1). The depth of the feature maps corresponds to the number of applied filter kernels. By sharing the weights of the filter kernels at all positions of the input, an enormous reduction of weight parameters that have to be learned is achieved. As shown in Fig. 2.1, the pooling layers are used to reduce the size of the feature maps and to increase the receptive field of the subsequent filters by downsampling the feature maps while preserving the spatial structure of features in the feature map. The most frequently used technique for downsampling is *Max Pooling* in which a max filter is applied to the input map. For downsampling the input, the step size in which the filter is shifted over the input, which is referred to as *stride*, has to be chosen according to the desired downsampling factor. The series of convolutional layers incl. the activation function and pooling layers is usually repeatedly applied in a CNN. Furthermore, other than as shown in Fig. 2.1, convolutional layers are often followed by additional convolutional layers before pooling is applied.

The output of a CNN can vary and depends on the task which is to be solved by the network. For instance, in the case of image classification (Simonyan and Zisserman, 2015), at least one fully connected layer is introduced at the end of the network which connects all its input nodes to a number of output nodes that corresponds to the number of classes C. These neurons produce the raw class scores s_j with $j \in [1, C]$ as real numbers ranging from $[-\infty, +\infty]$ for each class. To normalise the raw scores such that the final scores can be interpreted as probabilities $\hat{y}_j$ for each of the classes such that $\sum \hat{y}_j = 1$, the softmax function is used (Bishop, 2006). It determines the probability $\hat{y}_k$ for class $k \in C$ by

$$\hat{y}_k = \frac{\exp(s_k)}{\sum_j \exp(s_j)}. \tag{2.3}$$

Fully convolutional networks (Long et al., 2015) and encoder-decoder networks (Ronneberger et al., 2015) differ from classification networks in that they typically do not contain fully connected layers. Such networks are used e.g. for semantic segmentation, where the task is to associate a class label to each pixel of the input image. To this end, the intermediate downsampled feature maps obtained after several convolution and pooling operations are upsampled again until they reach the target resolution by applying suitable upsampling techniques, such as e.g. bilinear upsampling or strided transposed convolutions (sometimes also called *deconvolution*) (Zeiler et al., 2010). The kernels of a strided transposed convolution introduce weight parameters which have to be learned during training, too (Long et al., 2015).

2.1.1 Training

Learning all network parameters $\mathbf{w}$, namely all weights and biases of the filter kernels, is called training and is typically performed in a supervised manner, which means that the availability of training samples with corresponding reference data is required. The particular nature of how a training sample and its reference is defined depends on the task that is to be solved by the network. In general, training is performed by minimising a loss function $\mathcal{L}(\mathbf{w})$ which provides a measure of how good the result of the network fits to the known reference. The definition of the loss function also depends on the task. Examples of commonly utilised loss functions are given later. Starting from a (random) initialisation of the weights, the optimisation is performed iteratively using stochastic mini-batch gradient descent (SGD) (Goodfellow et al., 2016).

Initialisation: One way to initialise a CNN is to use random weights drawn from a Gaussian distribution. However, with fixed standard deviations, deeper models expose difficulties to converge (Simonyan and Zisserman, 2015). To overcome this problem, the *Xavier* (Glorot and Bengio, 2010) or the *He* (He et al., 2015) initialisation methods have been proposed, which choose a dynamic definition of the standard deviation in dependency on the number of input units to the convolution filters.

SGD: In the SGD procedure, a subset of the training samples, forming a mini-batch, are selected and propagated through the network, which is called a forward pass. The loss is calculated according to the resulting output. The gradients $\nabla\mathcal{L}(\mathbf{w})$ of the loss w.r.t. to the parameters are computed in an efficient way by exploiting the chain rule, a procedure called back-propagation (Rumelhart et al., 1986). The new weights $\mathbf{w}_{\text{new}}$ are derived from those of the previous iteration $\mathbf{w}_{\text{old}}$ according to

$$\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} - \eta \nabla \mathcal{L}(\mathbf{w}). \tag{2.4}$$

The learning rate η is a hyperparameter that defines the step size in which the weights are updated. However, plain SGD faces several problems such as local minima, noisy gradients, or regions in the loss curve in which the loss changes quickly in one direction and slowly in another. To overcome those problems, SGD with momentum is applied, which introduces running averages of gradients or variant forms of that to the weight update step. The *Adam* optimizer is one example of SGD with momentum which is frequently used to train CNNs and is also applied in this thesis. For a detailed exposition of this method it is referred to (Kingma and Ba, 2015).

Loss functions: The training of multi-class classification networks requires training data including a reference class for each training sample. A frequently used loss for classification tasks is the cross-entropy loss. With n training samples per mini-batch and C different classes to distinguish, the cross-entropy (CE) loss $\mathcal{L}_{\text{CE}}$ is calculated according to

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{C} t_{ij}\log(\hat{y}_{ij}). \tag{2.5}$$

In Eq. 2.5, $\hat{y}_{ij}$ denotes the softmax output for the i^{th} sample of the batch for the j^{th} class and $t_{ij} \in 0, 1$ is a binary variable which indicates whether the sample belongs to the j^{th} class or not. The CE loss is used in this work for the training of the classification branches of the proposed CNN.

A standard loss for training a network for regression tasks, i.e. for tasks that aim at predicting a floating value, is the mean squared error (MSE) loss $\mathcal{L}_{MSE}$ with

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{2.6}$$

In this case, y_i corresponds to the reference value of the i^{th} sample while $\hat{y}_i$ equals the predicted output of the network. Note that for fully convolutional networks which produce pixel-wise outputs both loss functions can be applied to calculate the loss for the output of each pixel. For pixel-wise classification tasks the output has multiple channels, one for each class. In this work, the standard MSE loss is adapted by proposing a customized loss for the prediction of probability maps which contain a pixel-wise probability for the appearance of vehicle keypoints and wireframe edges, respectively.

Transfer learning: Often, it is not necessary to learn all network weights from scratch for every new task. It is common practice to use a pre-trained network, i.e. a network that was trained for a specific task on a specific data set, for fine-tuning or as a feature extractor for a different task and/or data set. These strategies are summarised under the term *transfer learning* in the literature (Yosinski et al., 2014). When used as feature extractor, the weights of the network are usually fixed and a new classifier is learned on top of the features according to a new task. Fine-tuning refers to a strategy in which all or only some layers (typically the last few layers) of the pre-trained network are retrained together with the classification head for the new task or dataset (Donahue et al., 2014). This procedure is justified by the hierarchical structure of a CNN: while the feature maps produced by the earlier layers are more generic, independent from the data or the task, the feature maps learned at later stages represent more specific features that are more sensitive to the task and data they are trained for (Yosinski et al., 2014). This is why fine-tuning is often employed in some of the top layers only while freezing the previous layers. However, it is also possible to fine-tune all network weights using the pre-trained weights only as potentially better initialisation compared to a random initialisation. Usually smaller learning rates are applied for fine-tuning tasks as it is expected that the pre-trained weights are already relatively good. Adapting pre-trained networks for new tasks or data sets is especially helpful when only a comparably small amount of training data is available.

2.1.2 CNN Architectures

In this section, several CNN architectures which are used or adapted within this thesis are presented.

VGG19

VGG19 (Simonyan and Zisserman, 2015) is a CNN architecture proposed for the task of image classification. Originally, the network was trained to classify images containing a single object into one of 1000 object classes (Russakovsky et al., 2015). An overview of the architecture is shown in Fig. 2.2. The network consists of 19 layers with weights (16 convolutional and three fully connected layers) and five pooling layers in total. The convolutional layers use d filter kernels with a spatial size of 3x3 and the ReLU as non-linear activation function. The exact numbers d of filters applied in the different layers are given in Fig. 2.2. Max pooling with kernel size of 2x2 and stride 2 is used. Three fully connected layers are applied at the end of the network: the first two layers have 4096 nodes each, while the third layer (the class layer) contains 1000 nodes, one per class. The activation function of this classification layer is the softmax function. For training, three channel colour images of size 224x224 with known object class were used (Russakovsky et al., 2015) and the cross-entropy loss function was applied. More details can be found in (Simonyan and Zisserman, 2015).

The great performance achieved by the VGG19 on the *ImageNet* challenge (Russakovsky et al., 2015) is the reason why the pre-trained network released by the authors is often used as feature

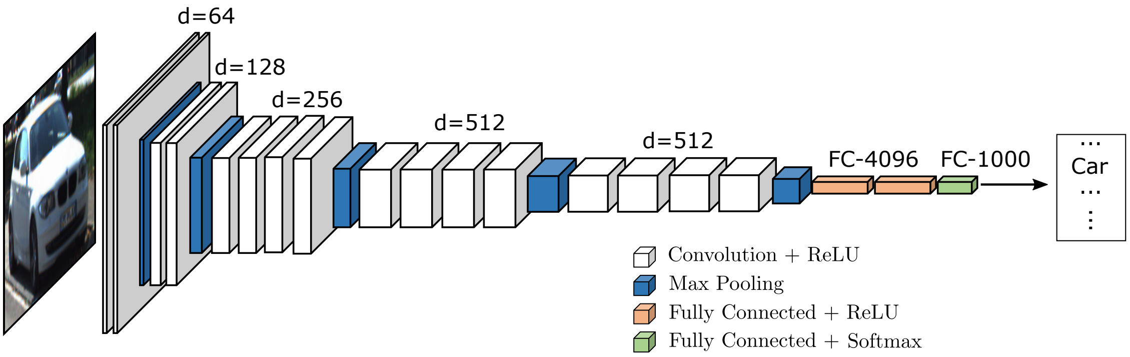


Figure 2.2: Architecture of the original VGG19 classification network (Simonyan and Zisserman, 2015). The input is an RGB image. After several convolutional and max-pooling layers, three fully connected layers are used to produce a probability for each of 1000 classes.

extractor or as pre-trained network for fine-tuning in various subsequent tasks. The architecture of VGG19 together with its pre-trained weights serves as a basis for the multi-task CNN which is developed in this thesis and which is described in Chapter 4.

U-Net

Building upon (Long et al., 2015), the U-Net is a fully convolutional network, originally used for semantic segmentation (Ronneberger et al., 2015). U-Nets or U-Net-like architectures are widely used for tasks in which the output is of the same spatial extent as the input. To this end, the architecture of U-Net, which can be seen in Fig. 2.3, follows a symmetric encoder-decoder structure.

The encoder part follows the typical architecture of a convolutional network and consists of repeated convolutional layers using a number of d filters of size 3x3, each followed by a ReLU activation function. Also, four max pooling layers with filter size 2x2 and stride 2 are applied to decrease the spatial resolution of the input and the subsequent feature maps by a factor of 0.5. The decoder part is symmetric to the encoder part but replaces the max pooling layers by 2x2 upsampling operations (in this case strided transposed convolutions are used), which spatially upsample the feature maps by a factor 2. Skip connections, also known as bypass connections (He et al., 2016), are used between corresponding pairs of blocks of the encoder and the decoder in order to concatenate feature maps of the encoder with feature maps of the same size of the decoder. This allows the subsequent convolutions to take place with awareness of the original pixels or feature maps, respectively, and leads to better segmentation results at object boundaries. U-Net does not contain any fully connected layers. Instead, the final layer performs convolutions with a 1x1 kernel to map the final feature map to the desired number of classes. A 1x1 convolution corresponds to a linear function, its output is forwarded to a non-linear activation function to produce the final predictions which correspond to the label maps. In the original paper (Ronneberger et al., 2015),

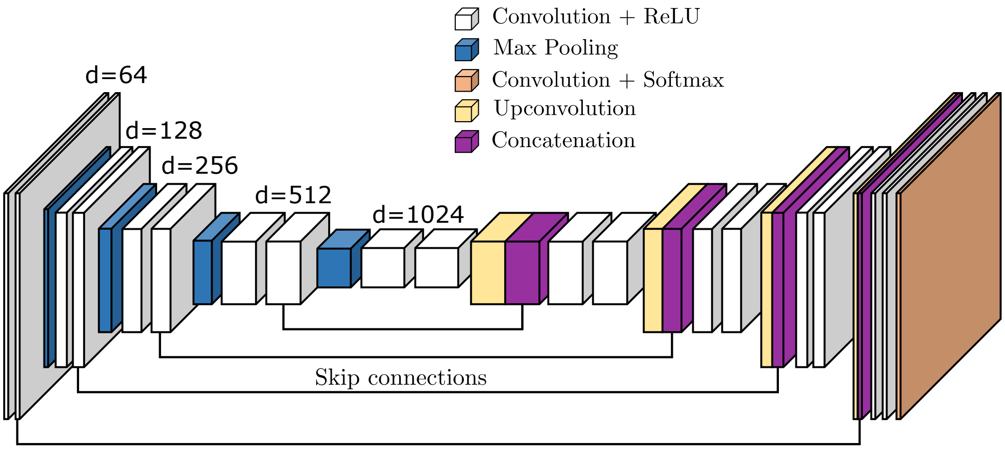


Figure 2.3: Architecture of the symmetric U-Net segmentation network (Ronneberger et al., 2015). It consists of an encoder part, which performs several convolutional and max-pooling operations, and a decoder part, which performs a series of convolutions and upsampling operations. Skip connections are established between corresponding parts of the encoder and the decoder.

the softmax function is used as activation function, combined with a pixel-wise cross-entropy loss to train the network to distinguish between two classes (foreground and background). Consequently, the training procedure requires a pixel-wise reference, i.e. a label image of the same size as the input.

In this thesis, the ideas and principles of the U-Net architecture are adapted for the computation of the vehicle keypoint and wireframe heatmaps (cf. Chapter 4).

Mask RCNN

The Mask RCNN (He et al., 2017) is a CNN architecture for joint object detection and instance segmentation. In addition to an object bounding box and the object's semantic class the network outputs a high-quality segmentation mask for each object. To this end, the architecture builds on the *Faster RCNN* network for object detection proposed in (Ren et al., 2015), extending it with an additional output branch to produce the instance segmentation masks. A high-level overview of the Mask RCNN architecture is shown in Fig. 2.4. The first part of the architecture consists of a convolutional neural network, processing the input image and producing a feature map. More specifically, a ResNet (He et al., 2016) architecture in conjunction with a Feature Pyramid Network (FPN) (Lin et al., 2017) is used to form the backbone. A region proposal network, which is in essence adapted from (Ren et al., 2015), extracts regions of interests (RoI) by proposing a set of rectangular object candidates in the feature map. A small, fixed sized feature map is created for each RoI in the feature map using a bilinear interpolation based technique, referred to as *RoIAlign*

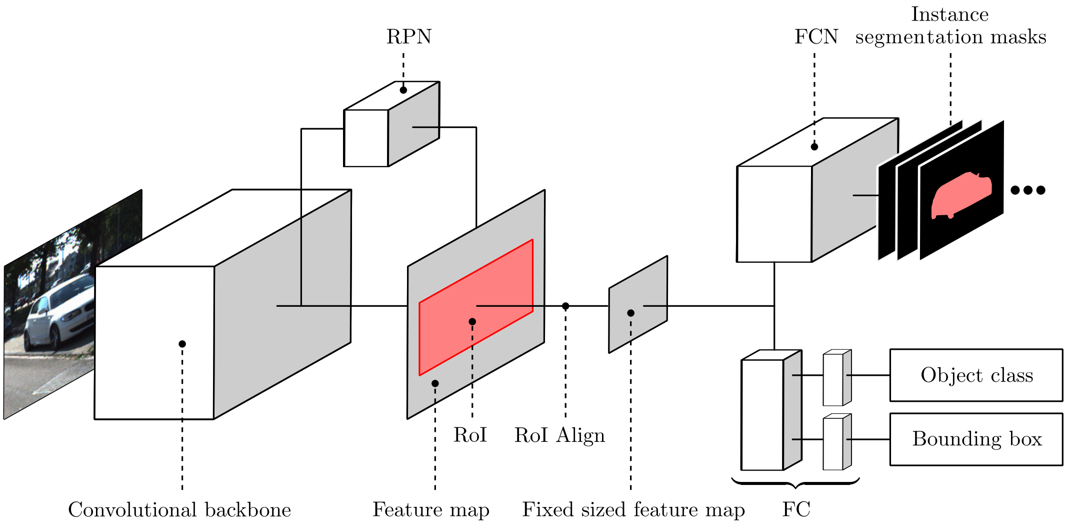


Figure 2.4: High-level overview of the Mask RCNN architecture (He et al., 2017). A three channel input image is first processed by the convolutional backbone network. A Region Proposal Network (RPN) is trained to extract object bounding boxes as regions of interest (RoI). The network output consists of a class label and a bounding box, produced by fully connected (FC) layers, as well as instance segmentation masks, produced by a fully convolutional network (FCN) head, for each detected object and each class.

(He et al., 2017). The head of the Mask RCNN predicts an object class and regresses an object bounding box for each RoI. In parallel, instance segmentation masks are additionally produced for each RoI and object class. In order to predict the object class of each RoI, the fixed size feature map is fed into a series of fully connected layers to predict the probability of the RoI belonging to each of a set of object classes using the softmax function. Furthermore, a FC layer is used to regress a real numbered bounding box position for the object. A FCN head is used to predict the instance segmentation masks from the same feature map using a pixel-wise sigmoid function as the activation function of the last layer.

For the training of the Mask RCNN a multi-task loss $\mathcal{L}_{\text{maskRCNN}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{mask}}$ is defined. Here, the classification loss $\mathcal{L}_{\text{class}}$ corresponds to cross-entropy loss while the bounding box loss $\mathcal{L}_{\text{bbox}}$ makes use of a L1 norm of the difference of the regressed bounding boxes and the reference bounding boxes. The mask loss $\mathcal{L}_{\text{mask}}$ is defined as the average binary cross-entropy loss over the pixels of the bounding box. Consequently, training the Mask RCNN requires training images with annotated object bounding boxes, object class labels and an annotated instance segmentation mask for each object.

The method proposed in this thesis makes use of the Mask RCNN for the detection and precise delineation of vehicles in the input images (cf. Chapter 4). The detected vehicles serve as input for the developed reconstruction procedure.

2.2 Active Shape Model

The reconstruction of 3D objects from (stereo-)images, especially of objects with variable shapes, is an ill-posed problem. The use of shape priors for reconstruction constrains the problem and thus simplifies its solution. When dealing with variable object shapes, non-rigid but deformable shape priors are required, which are able to adapt to all expected possible shapes but always deliver globally plausible object instances and, thus, constrain the overall object geometry.

One technique for deformable shape modelling is the Active Shape Model (ASM) (Cootes et al., 1995), which is learned from a set of N distinct training models of an object class with a well-defined topology to capture the object's intra-class variability. While originally applied to 2D shapes, the ASM can be used to derive a deformable shape prior in any dimension. Similar to (Zia et al., 2013), in this work a 3D ASM is learned as a deformable representation of vehicles. This mathematical explanation of the ASM focuses on learning the shape representation in 3D space.

One 3D training exemplar ν_n, with $n = [1, N]$ consists of a set of keypoints $k \in [1, K]$, represented by their 3D vertex locations with $\nu_n = [x_n^1, y_n^1, z_n^1, ..., x_n^K, y_n^K, z_n^K]^T$. In this context, corresponding keypoints of different exemplars must represent the same physical point of the object. To learn the ASM, the 3D coordinates of exemplar vertices need to be represented in a commonly defined coordinate system, which has the same longitudinal axis and reference point as origin (e.g. the exemplar's centre of mass) among all exemplars. The ASM computation is based on Principal Component Analysis (PCA) applied to the covariance matrix $\Sigma_{\nu\nu}$ of the keypoints of the training exemplars, resulting in the eigenvalues σ_s^2 and eigenvectors $\mathbf{e}_s$ of that matrix.

The computation of the covariance matrix requires the mean shape $\mathbf{m}$ which is calculated according to

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \nu_n. \tag{2.7}$$

The covariance matrix is computed by

$$\Sigma_{\nu\nu} = \frac{1}{N-1} V V^T, \tag{2.8}$$

where V is a matrix collecting the keypoint coordinates of the exemplars reduced by the mean shape:

$$V = [\nu_1 - \mathbf{m},\ \nu_2 - \mathbf{m},\ ...,\ \nu_N - \mathbf{m}]. \tag{2.9}$$

The eigenvectors $\mathbf{e}_s$ of the covariance matrix $\Sigma_{\nu\nu}$ encode the directions of the principal shape deformations given the training exemplars which represent the object class. The square roots of the corresponding eigenvalues σ_s^2 thus represent the standard deviations of the deformation in the direction of the respective eigenvectors. A deformed object model $M(\gamma)$ can be created by a linear combination of the mean shape and the eigenvectors, the latter being weighted by their standard deviations and scaled by the *shape parameters* γ_s which are used to control the amount of deformation in each principal direction with:

$$M(\gamma) = \mathbf{m} + \sum_{s=1}^{3K} \gamma_s \ \sigma_s \ \mathbf{e}_s. \tag{2.10}$$

Defining the shape parameters γ_s in the range of $[-3, 3]$ allows to synthesise shape models covering 99.7% (three times the standard deviation) of the variations being present in the training set. In practice, to reduce the dimensionality of the shape parameter space, often only the first n_s eigenvectors corresponding to the largest eigenvalues are considered in the ASM computation, because the eigenvectors corresponding to the small eigenvalues usually represent minor shape variations and thus can be neglected (Tsin et al., 2009).

2.3 Monte Carlo based optimisation

Many tasks in the field of computer vision and machine learning involve the estimation of unknown quantities, in the following denoted as state variables $\mathbf{s}$, from real-world data, i.e. from some given observations $\mathbf{o}$. A probabilistic model of the task allows to estimate the unknown variables from the variables that are observed. When prior knowledge about the target phenomenon can be modelled or is available, e.g. from expert knowledge, the probabilistic model of the problem can be based on a Bayesian formulation. The Bayesian formulation utilises the prior distributions for the unknown quantities and likelihood functions relating these quantities to the observations (Doucet et al., 2001). The inference of the unknown state is based on the maximum a posterior (MAP) criterion, i.e. on maximising the posterior distribution obtained from the Bayes theorem

$$p(\mathbf{s}|\mathbf{o}) = \frac{p(\mathbf{o}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{o})} \rightarrow \max, \tag{2.11}$$

with the *likelihood* term $p(\mathbf{o}|\mathbf{s})$, the *prior* probability $p(\mathbf{s})$ and the *evidence* $\mathrm{p}(\mathbf{o})$. As $p(\mathbf{o})$ is independent from the state variables and usually is only required as a normalisation factor to be able to interpret the posterior as a probability, the MAP estimate can be obtained without knowledge about the distribution $p(\mathbf{o})$, such that $p(\mathbf{s}|\mathbf{o}) \propto p(\mathbf{o}|\mathbf{s})p(\mathbf{s})$. For computational reasons, instead of maximising the posterior, often the negative logarithm of the posterior is minimised to derive the optimal values $\hat{\mathbf{s}}$ for the unknown state variables. As a consequence, the state optimisation is based on minimising an objective function E, such that

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmin}} \ E(\mathbf{s}, \mathbf{o}, \sigma_{\mathbf{o}}), \tag{2.12}$$

with

$$E(\mathbf{s}, \mathbf{o}, \sigma_{\mathbf{o}}) = -\log p(\mathbf{o}|\mathbf{s}) - \log p(\mathbf{s}) \tag{2.13}$$

In Eqs. 2.12 and 2.13, the parameters $\sigma_{\mathbf{o}}$ are used to adapt the model to the characteristics of specific data and can, for instance, be represented by observation uncertainties. In the probabilistic formulation, these uncertainties are usually considered in the likelihood terms. Standard methods for local optimisation of non-linear functions include first-order approaches based on the Jacobian of the objective or second-order approaches based on the Hessian of the objective. However, for these approaches to reach the global optimum, the objective function has to fulfil some restrictions, such as being (quasi-)convex in the continuous domain (Boyd and Vandenberghe, 2004).

The deformable 3D model fitting approach developed in this thesis is based on a non-linear, non-convex and discontinuous objective function which involves elements that are not Gaussian and thus precludes first or second-order solutions for optimisation (Doucet et al., 2001). To such problems, Monte Carlo Methods (MCM) can be applied to approximate the optimal state parameters $\hat{\mathbf{s}}$. A detailed description of such methods can be found in (Bishop, 2006). MCM belong to simulation-based methods and have the advantage of not requiring any assumptions about the linearity, convexity or statistical distribution on the model (Doucet et al., 2001). Furthermore, they are easy to implement even for the optimisation of complicated objective functions, and finally, they can be parallelised, which in principle favours their applicability to problems with real-time requirements (Kroese et al., 2014).

The central idea of MCM is to find the optimum of a potentially high-dimensional function by generating a set of random samples $\mathbf{s}^i$, with $i \in [1, n_p]$, which are called *particles* (Doucet et al., 2001). If the number n_p of particles is large enough, such that the parameter space is densely represented by the samples, the optimum state can be estimated given the samples by $\hat{\mathbf{s}} = \operatorname{argmin}_{\mathbf{s}^i} E(\mathbf{s}^i)$. However, in a high-dimensional parameter space and/or a parameter space with potentially large parameter ranges, a dense sampling of the parameters for the particle based optimisation is computationally intractable. To overcome this problem, sequential techniques can be used, in which both, particle sampling and state estimation are performed alternatively in an iterative process. Fig. 2.5 shows the scheme of a sequential MCM approach which is applied for the parameter estimation in this thesis, exemplarily depicted for a one-dimensional objective function E and for one iteration. After initialising the first set of particles $\mathbf{s}_0^i$, each iteration $j = 1, ..., n_{it}$ consists of of two steps, the weight computation for each of the initial particles and a resampling step, in which new particles are drawn for the next iteration based on the preceding particles and their corresponding weights. The initial particles $\mathbf{s}_0^i$ are drawn from the prior distribution $p(\mathbf{s}_0)$. When no prior information is available but the range of parameter values is known, usually an uniform distribution is applied for $p(\mathbf{s}_0)$ to draw the particles from. In each iteration j, a normalised weight w_j^i with $\sum_i w_j^i = 1$ is associated to each particle $\mathbf{s}_j^i$. One way of estimating the weights is to define them according to the particle's score obtained from the objective function, such that $w_j^i \propto E(\mathbf{s}_j^i)^{-1}$. The key idea of the sequential MCM is to eliminate particles receiving low weights and consequently, to multiply particles having high weights in the resampling step. To this end, a number of offspring N_j^i is

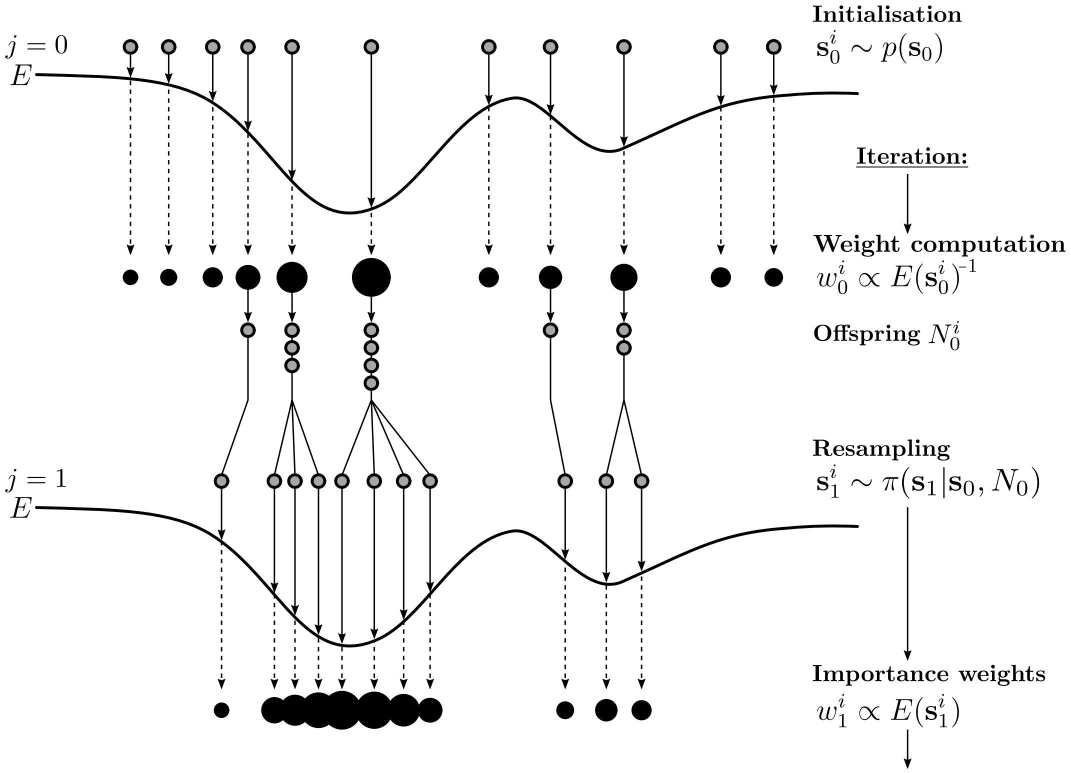

$j = 0$
E
Initialisation
$\mathbf{s}_0^i \sim p(\mathbf{s}_0)$
Iteration:
Weight computation
$w_0^i \propto E(\mathbf{s}_0^i)^{-1}$
Offspring N_0^i
$j = 1$
E
Resampling
$\mathbf{s}_1^i \sim \pi(\mathbf{s}_1|\mathbf{s}_0, N_0)$
Importance weights
$w_1^i \propto E(\mathbf{s}_1^i)$

Figure 2.5: Schematic procedure of the MCM based optimisation for one dimension (Doucet et al., 2001). In a first step, the particles $\mathbf{s}_0^i$ are initialised. In each iteration j, a weight w_j^i is calculated and associated to each particle (a larger size of the filled black circles corresponds to a higher weight). Corresponding to the weights, a number of offspring is calculated for each particle. In a resampling step, new particles are drawn based on the preceding particles and according to the number of offspring. For a detailed explanation the reader is referred to the main text.

associated to each particle (Doucet et al., 2001). In a general setting, the offspring can be determined by $N_j^i = w_j^i \cdot n_p$ with $\sum_i N_j^i = n_p$, such that the number of offspring is larger for particles with higher weights and vice versa. Note that the numbers of offspring need to be integer values such that proper rounding operations are required for the offspring calculation just mentioned. The resampling procedure applied in this thesis differs from this approach in that a user-defined number for the n_b best scoring particles is chosen. These particles are defined as seed particles for the subsequent iteration by setting the weights for these particles to $\frac{n_p}{n_b}$ and the weights of all other particles to 0. As a consequence, the same amount of offspring results for the best n_b particles, while the remaining particles are eliminated. To generate the offspring particles for the subsequent iteration based on the preceding particles and the associated number of offspring, generally speaking, an arbitrary transition function $\pi(\mathbf{s}_j|\mathbf{s}_{j-1}, N_{j-1})$ is used. In this thesis, this function corresponds to drawing the offspring particles from a uniform distribution centred at the preceding seed particles. By decreasing the range of the applied uniform distribution in each iteration, the particles iteratively move towards the global minimum of the objective function E, as insinuated in Fig. 2.5. A variant of the described procedure has already been used for 3D object modelling in (Zia et al., 2013), where it is referred to as *stochastic hill-climbing* method. The iterative MCM procedure as described above enables an efficient exploration of the search space while preserving its applicability to arbitrary (e.g. multi-modal, non-convex, discontinuous) objective functions.

3 State of the art

To put the contributions of this thesis into context, this chapter provides an overview of related work on object reconstruction and object pose and shape recovery. The focus will be on methods that deal with objects, data and applications related to autonomous driving with *vehicles* as objects of main interest. In this chapter, an distinction is made between data driven approaches (Sec. 3.1) and model driven approaches (Sec. 3.2). Sec. 3.3 discusses the current state-of-the-art and summarises open questions.

On an abstract level, **data driven approaches** for object pose and shape recovery make use of the input data, mostly a single image, to derive an intermediate representation of the data, usually in the form of some kind of feature representation, on the basis of which the target parameters for pose and shape are inferred directly. Often, these data driven methods are coupled with approaches for object detection, such that the pose and shape estimation is a by-product of the detection or is built as an extension on top of an object detector. In contrast to that, in **model driven approaches** the detection and state estimation of the object are usually treated as decoupled tasks. Arbitrary object detectors are typically applied in a first step to find instances of the target objects. In a second step, a flexible shape representation of the target object based on prior knowledge is fitted to the observations of the detected instances to determine the object's pose and shape.

3.1 Data driven approaches

Methods which extract a feature representation of the input data and train a classifier or regressor based on this feature representation to infer the target parameters for object pose and/or shape are categorised as data driven approaches in this overview. Traditionally, the feature representation is based on suitable handcrafted features, while recently almost exclusively CNNs are used for both, feature extraction and parameter estimation.

This section discusses related work on data-driven vehicle pose and shape estimation. First, approaches are reviewed which aim at estimating the vehicle viewpoint, disregarding the vehicle's location in 3D (cf. Sec. 3.1.1). Sec. 3.1.2 gives an overview about existing methods on 3D pose estimation which, however, do not reason about a representation of the vehicle's shape. Finally, in Sec. 3.1.3 data-driven methods are reviewed that infer a joint representation for 3D object pose and shape.

3.1.1 Viewpoint prediction

A coarse estimation of the vehicle viewpoint from single images is delivered already by a number of early vehicle detection approaches. To overcome the huge variety of visual appearances of vehicles due to their intra-class variability or changes in the camera viewpoint and to handle occlusions, part-based detection approaches were introduced by Leibe et al. (2006) and Felzenszwalb et al. (2010). They divide the object into several distinctive parts and learn a detector for each part. Usually a global part configuration, considering the topology of the individual parts, is learned for the detection of the entire object. Pepik et al. (2015) additionally learn the viewpoint dependent arrangement of the parts in the global configuration to associate each detection already with the estimation of one of eight discrete viewpoint classes. However, the discretisation of the viewpoint into a small set of classes only allows the derivation of coarse viewpoint categories rather than the definite viewpoint angle.

Other approaches explicitly train viewpoint specific classifiers to determine a vehicle's viewpoint. In these approaches, the viewpoint typically also consists of a discrete number of viewpoint-bins and a classifier is trained to predict the correct bin. In (Ozuysal et al., 2009), a classifier is learned for a number of 16 vehicle azimuth classes based on SIFT-like features. Similarly, Tulsiani and Malik (2015) train a CNN to predict a class for both, azimuth and elevation angle bins, respectively. In (Su et al., 2015) a CNN is trained to predict a class for azimuth, elevation, and in-plane rotation (around the optical axis) viewpoint angles. To achieve more fine-grained results for the viewpoint, the authors distinguish 360 viewpoint classes of the detected object. More fine-grained discretisations call for extensive amounts of training data. Thus, in (Su et al., 2015) synthetic data generated from model renderings are created to train the network. However, training on synthetic data usually leads to a noticeable drop of performance when the classifier is applied to real-world data. Summarising, most of the mentioned approaches only give a coarse indication about the viewpoint in a discretised form and the vehicle's position in 3D, and its shape is disregarded completely. Furthermore, the problem remains that a compromise between the classifier's complexity and the level of detail of the viewpoint estimation has to be found in all cases. Striving for more precise viewpoint representations requires more training data.

3.1.2 3D pose prediction

One step towards capturing 3D object information from images is taken by early approaches which internally enrich a part-based detector by linking 3D object knowledge to the parts and transferring this information to the objects after detection. To that end, the increasing amount of freely available CAD data is often exploited. For instance, Liebelt and Schmid (2010) treat appearance and geometry as separate learning tasks. They train an appearance based detector for object parts from real images and learn 3D geometry for each part from synthetic 3D models. Linking the detected parts with their associated 3D geometry allows an approximate estimation of 3D pose. Similarly, Pepik et al. (2012) adapt the Deformable Part Model (DPM) of (Felzenszwalb et al., 2010). They add 3D information from CAD models to the deformable parts and incorporate 3D constraints to enforce part correspondences. Thomas et al. (2007) enrich the Implicit Shape Model (ISM) of Leibe et al. (2006) by adding depth information from training images to the ISM and transfer the 3D information to the test images, which allows the estimation of coarse 3D pose information. Hödlmoser et al. (2013) tackle pose estimation in 3D space by defining a number of pose classes where each pose class corresponds to a different combination of viewpoint and distance classes. Consequently, the overall number of pose classes corresponds to the product of the number of distinguished viewpoint and distance bins. They train a Random Forest classifier (Breiman, 2001) to distinguish between the pose classes based on Histogram of oriented Gradients (HoG) features (Dalal and Triggs, 2005). However, due to the class-wise representation of the pose parameters, a precise and continuous pose estimation is not provided by the approaches mentioned so far.

With the emergence of CNNs the prediction of 3D object pose from images has experienced a huge boost and work on 3D object detection has expanded significantly over the last few years. A naive approach for 6DoF pose estimation would be to train a CNN to regress the target parameters from detections in the 2D image. However, regressing the pose, especially the translation, from a single image without prior knowledge is an ambiguous task. This is why typically other solutions are found in the literature, which e.g. incorporate prior scene knowledge or find alternative representations for the target pose parameters. Frequently, oriented 3D bounding boxes are estimated for the objects which implicitly contain the information about the object's pose, i.e. their position and orientation.

One line of work follows the two-step procedure that was already successfully applied by Region Convolutional Neural Network (RCNN) approaches for 2D object detection (Ren et al., 2015), by generating 3D object proposals in a first step, which are passed through a CNN to generate the final detections in a second step. In this context, Chen et al. (2015) make use of stereo image data and the 3D information derived from it to introduce geometric priors, such as object height and point density, and to reason about free-space in order to derive 3D bounding box proposals for street-level objects. Their follow-up work (Chen et al., 2016) replaces the geometric priors by priors based on scene-context and object shape, which are derived from semantic segmentation and instance segmentation using monocular images. While these methods have shown good results, they are computational expensive due to the generation and the processing of the large number of object proposals initialised in 3D.

Another line of work builds upon the success of existing work on 2D object detection for the task of 3D bounding box estimation. Ku et al. (2019) make use of 2D object detections in the image to infer 3D bounding box proposals by leveraging the relation between the 2D bounding box and an estimated object height. However, small errors in the 2D bounding box estimates or the height estimates of the 3D bounding box are likely to cause large errors in the position estimation of the object in 3D space. In (Xiang et al., 2018), a CNN is trained to localise the object centre in the image and to regress the object's orientation and distance to finally derive the object's 3D pose. However, predicting object distances, i.e. the absolute scale from single images is an ill-posed problem and therefore causes ambiguous solutions. Given 2D detections delivered by an object detector, Mousavian et al. (2017) propose a CNN to regress the object extents and orientation from single images instead of regressing the 3D translation in object space. The fact that the perspective projection of the 3D bounding box should fit to the 2D image bounding box is used to infer the absolute translation of the object from the regressed object extents and orientation. Similar to this, Tekin et al. (2018) and Grabner et al. (2018) propose to train a CNN for the prediction of the 2D image locations of the projected 3D bounding box vertices to estimate the 6DoF pose of objects with known size via spatial resection. However, these approaches are highly sensitive to errors in the 2D predictions and inaccuracies of the regressed parameters.

Xu and Chen (2018) and Ma et al. (2019) resort to the large body of work that focuses on depth estimation from single images, e.g. (Godard et al., 2017), and incorporate a module for monocular depth estimation in their approach for 3D object detection. However, while approaches for monocular depth estimation work impressively well when the network is trained and tested on the same domain, these methods exhibit difficulties w.r.t. their transferability to other domains.

To summarise, while the approaches mentioned in this section have shown impressive results w.r.t estimating an object's oriented 3D bounding box when trained and tested on the same data domain, doubts can be entertained about their generalisation capability. Estimating 3D dimensions or 3D pose from single images is highly ambiguous and therefore causes large average 3D position errors which can amount to up to several meters (Mousavian et al., 2017). Furthermore, classifiers learned on the statistics and objects of a training domain are likely to experience a significant drop of performance when applied to a different data domain. Besides, the mentioned approaches entirely neglect the reasoning about the object's shape by only representing objects as 3D boxes. However, the fine-grained estimation of an object's shape can be important for several applications and tasks, such as e.g. the (re-)identification of objects (Tang et al., 2019).

3.1.3 3D pose and shape prediction

In this section, methods are reviewed that do not only reason about an object's pose, but also give a prediction of a parametric or non-parametric representation of the object's shape, and thus, strive for 3D reconstruction of pose and shape at the same time.

Xiang et al. (2015) train a sliding window based detector for what they call 3D voxel patterns (3DVP) of vehicles. Support vector machines (SVM) (Cortes and Vapnik, 1995) are used for this purpose. 3DVP are learned by clustering similar 3D voxel exemplars, each of which contains a voxelised 3D model of a vehicle as well as a 2D image and its 2D segmentation mask to encode visibility patterns resulting from occlusion or truncation. In the follow-up work (Xiang et al., 2017), the detection of 3DVP is incorporated in a RCNN network resulting in a better performance. A softmax output layer of the RCNN network is used to classify the detection bounding boxes as one of the different 3DVPs. By associating a 3DVP to each detection in the image and transferring the 3D information implicitly contained in the 3DVP to the detections, a pose and shape aware object reconstruction is obtained. However, the level of detail achieved for the reconstruction heavily depends on the number of 3DVP clusters and the finite number of voxel representations that are considered to learn the clusters. These restrictions are likely to cause the reconstructions to be quite generalised.

In (Wang et al., 2018), object shape is represented by a 3D location field, a 3-channel representation in which every image pixel belonging to the object is mapped to its corresponding location on the surface of the 3D model. The Mask-RCNN architecture is adapted and trained to detect vehicles, regress their pose and output the 3D location field instead of the original binary segmentation mask. However, this non-parametric shape representation only delivers information for the visible parts of the objects.

Recent work on object reconstruction delivers the 3D object pose together with a set of parameters defining the shape of an object given an parametric shape representation. Kundu et al. (2018) learn a ten-dimensional shape representation for vehicles by applying PCA to a training set of voxelised vehicle models. A RCNN is trained to detect vehicles and to regress the ten-dimensional shape vector in addition to the 3D pose parameters, thus obtaining a complete vehicle reconstruction in 3D space. To avoid the expensive annotation of ground-truth shapes in 3D, a render-and-compare loss is applied to train the network for pose and shape, which measures the similarity of the backprojected silhouette resulting from the predicted pose and shape parameters and a reference image segmentation mask. However, measuring pose and shape discrepancies based on a perspective 2D silhouette is highly ambiguous. Furthermore, this loss is unfavourable in case of object occlusions in the image. Instead of using PCA, in (Zhu et al., 2017; Manhardt et al., 2019), a 3D convolutional autoencoder is trained from voxelised training shapes to learn a shape parameter vector corresponding to the intermediate representation of the autoencoder in the low dimensional latent space. A network is trained to predict a small number of shape parameters together with the object's pose. While Zhu et al. (2017) use a reprojection loss similar to (Kundu et al., 2018) to train their network, Manhardt et al. (2019) use training instances with annotated reference shapes to calculate the loss by measuring the similarity between predicted shape and reference shape as the angle between the two parameter vectors.

A major drawback of data-driven approaches, especially when training a CNN, is the strong dependency on training data. On the one hand, training the huge amount of weight parameters that are typically contained in a CNN usually demands for large amounts of training data. Furthermore, the generalisability and consequently, the transferability of a CNN trained on specific data to other data sets exhibiting different statistics w.r.t. the data and content, have been found still to be an open problem. Purely data-driven approaches also suffer from what is called the *long-tail* problem (Wang et al., 2017), which refers to the problem of an imbalanced statistical distribution of the examples in the training data where only little data is available for classes in the tail of the distribution. This results in problems of these approaches w.r.t. their ability to generalise to barely or never seen examples or situations. To counteract the demand for large amounts of training examples as well as the *long-tail problem*, the usage of synthetic data is exploited (Manhardt et al., 2019). The generation of synthetic data for training allows the rendering of large numbers of training data, also including uncommon and statistically rare examples, compared to the usage of real-world data. Reference annotations are know for synthetic images, which makes manual annotation unnecessary. However, synthetic data usually lead to examples that differ from real ones w.r.t. the feature distributions caused by non-realistic illumination, reflectance and noise properties, etc.

Although research is performed to generate potentially photo-realistic renderings (Alhaija et al., 2018; Manhardt et al., 2019), training on synthetic data usually comes with a noticeable and significant drop in performance when applied to real data, which is caused by the domain gap. Research on domain adaptation is done to find solutions to overcome this effect. A survey on methods for domain adaptation is presented in (Wang and Deng, 2018). However, the requirement of handling the domain gap between synthetic and real data adds an additional complexity to data-driven approaches affected by this problem.

3.2 Model driven approaches

The line of work that is closest to the approach presented in this thesis initially detects and finally reconstructs the object of interest by fitting a 3D model to the image observations to reason about the object pose and shape. As the reconstruction of 3D objects from images is an ill-posed task, the usage of previously defined 3D shapes constrains and simplifies the problem. Another advantage of using 3D models as shape priors is their natural invariance w.r.t. the viewpoint. Whereas, for instance, a classifier for viewpoint estimation requires training data for each of the considered viewpoint classes, whose number will be large if more fine-grained the viewpoint estimation is required, resulting also in a large demand for training data, model driven approaches allow for model hypotheses from any viewpoint to be fitted to the image observations.

3.2.1 Shape priors

The way in which prior model knowledge is used varies in the literature. In (Ding et al., 2018; Murthy et al., 2017a), simple constraints on the vehicles' symmetry, about the centres of the wheels or the corners of the rooftop to be in a plane, and prior knowledge about the expected size are enforced during reconstruction. While these constraints give a rather basic and coarse representation of an object, CAD models allow a very detailed description of a 3D shape. A large variety of CAD models are available for all kind of different vehicle types and brands which can be used directly to guide the vehicle reconstruction (Güney and Geiger, 2015). However, without knowing the exact CAD model of interest a priori, it is intractable to run computations with each CAD model to find the most suitable one. Given a detected vehicle, Chabot et al. (2017) use a CNN to predict the closest 3D shape template from a set of rigid reference models and pick this model for further processing. However, potential errors in predicting the shape template may lead to erroneous reconstruction results. In this work, instead of enforcing a fixed shape for model fitting, softer constraints on the vehicle shape are implemented, which allow deformations of the shape according to the observations. For the task of vehicle re-identification Tang et al. (2019) predict the vehicle type, e.g. *compact car, sedan, limousine*, etc., using a CNN. However, the predicted type is not utilised in the context of vehicle reconstruction. In this work, a CNN which predicts the vehicle type is also used. A confidence-aware shape prior is presented which makes use of the type predictions by constraining shape deformations during the model fitting according to the predicted confidence scores for different vehicle types.

In contrast to rigid model instances, deformable shape representations learned from a set of reference shapes are more flexible and allow to cope with the large intra-class variability of vehicles. Therefore, they are used frequently for object reconstruction. As a consequence, the shape defining parameters are added to the list of target parameters for the estimation. In (Engelmann et al., 2016; Kundu et al., 2018; Manhardt et al., 2019), a Truncated Signed Distance Function (TSDF) is learned as a deformable shape prior for vehicles from a set of CAD models. Using a TSDF, the shape is represented in a voxel grid in which each voxel contains the truncated signed distance towards the object surface and thus, the shape manifold is implicitly represented as the zero-level of the TSDF. The common shape basis of the training set is learned by applying PCA to the TSDF representations of the training samples. Once the shape basis is learned, any deformed TSDF shape can be encoded by a low-dimensional shape vector. However, this representation only carries the information about the model surface while information such as semantic keypoint locations or wireframe edges, which can be significant cues for model fitting, are not explicitly contained. In contrast to that, an ASM is another deformable shape representation frequently used as shape prior for vehicles (Zia et al., 2013; Lin et al., 2014b; Murthy et al., 2017a; Ansari et al., 2018) that naturally contains keypoint information as it is learned by performing PCA on keypoints from training models (cf. Sec. 2.2). An ASM is also used in this work. However, we extend its keypoint based representation by defining a triangulated mesh and a wireframe topology based on the keypoints to achieve a joint representation for the vehicles surface, keypoints, and wireframe at the same time. These model entities are incorporated in the reconstruction process.

Such deformable shape priors are flexible but they can only be deformed in accordance with the variability contained in the training data. A common way of regularising the degree of admissible deformations during inference is to penalise deviations from the mean shape (Zia et al., 2013; Engelmann et al., 2016). However, this strategy is founded in the assumption that the object's shape variability follows an unimodal distribution which usually does not hold true. Especially in the case of vehicles, the shape variations result in several disjoint modes corresponding to different vehicle types, rather than following a single distribution (Lin et al., 2014b). Consequently, the applied regularisation of shape deformations is likely to enforce incorrect model shapes. In this thesis, the modes resulting from different vehicle types are learned together with the overall ASM representation. A CNN based prediction of the vehicle type from the image is used to guide the shape regularisation based on a newly proposed category-aware formulation of the shape prior.

3.2.2 Scene priors

To restrict the six dimensional and therefore large parameter space of the 6DoF pose, heuristics and prior knowledge about the scene can be employed. In the context of object reconstruction in street scenarios, often knowledge about the ground plane and the assumption that vehicles are located on that plane are incorporated. As a consequence, when the 3D ground plane is known, the 6DoF pose estimation problem reduces to an estimation of the 2D position on the plane and the estimation of the rotation about the plane's normal vector. In (Murthy et al., 2017b; Engelmann et al., 2016), who use images acquired from vehicle mounted platforms, the platform height and the assumption of a camera viewing direction to be parallel to the ground is used to reason about the location of the ground plane. However, the dynamic behaviour of the acquiring vehicle is likely to cause the assumption of the road to be parallel to the viewing direction to be incorrect, which, as a consequence, leads to errors in pose estimation, especially for objects in larger distance to the camera. To overcome the dependency on these strong heuristics, in this thesis, the ground plane is estimated from the 3D point cloud reconstructed from stereo images. By doing so, deviations from the co-planarity assumptions of camera viewing direction and the road surface are taken into account.

Ansari et al. (2018) go one step further and do not represent the ground by one common plane, but approximate the road surface by multiple planar patches. In particular, a road patch is estimated for every detected vehicle from 3D road points lying in a close neighbourhood of the respective vehicle. While in principle, this approach reduces the effect of deviations of the real surface from a strictly planar ground, it is prone to estimation errors of the local patches, which are likely to occur due to the small number of points used for the estimation and the resulting stronger effect of outliers and uncertainties on the estimation. Furthermore, it cannot be guaranteed that the local surface of the individual vehicles is always observable, e.g. due to occlusions.

Besides the utilisation of the ground plane as prior for potential vehicle poses, reasoning about free-space in the scene can be incorporated as regularisation of the pose parameters, but is rarely done in the literature. For the generation of 3D bounding box proposals, Chen et al. (2015) derive a free-space voxel grid from stereo-reconstructed 3D points by treating a voxel as occupied as soon as it contains a 3D point and penalise bounding box proposals that contain voxels of free-space. However, by deriving the binary state of being occupied or free-space from the occurrence of a single point within a voxel and by disregarding the information of local point densities, this approach is highly sensitive to outliers in the point cloud and the uncertainties of the 3D points. The position prior proposed in this thesis also uses the information about free-space in form of a 2D free-space grid map on the ground plane. However, instead of representing the occupancy of a cell by a binary state (occupied or not), a probability of being free-space is derived from the ratio of object and ground surface points that project to the respective cells. This approach is more robust against outliers and considers noise in the point cloud caused by reconstruction uncertainties.

3.2.3 Shape aware reconstruction

A common way to shape aware object reconstruction is to match entities such as keypoints, edges/contours, the surface or the silhouette of the model to its corresponding entities inferred from the image. The next few paragraphs describe the current state of the art of methodology following this strategy.

Edge/wireframe based reconstruction

Pioneer work on vehicle model fitting was based on model edge to image edge alignment (Tsin et al., 2009; Leotta and Mundy, 2009). The authors defined an appearance representation for a deformable vehicle model based on salient object edges that are likely to generate intensity edges such as occluding contours and part boundaries. Given sufficient 2D-3D correspondences between image and model edges, the pose and shape of the deformable model is estimated using iterative least squares adjustment. However, finding the 2D image edge corresponding to a 3D model edge, or more specifically finding corresponding point pairs on these edges, is non-trivial and a challenging task. On the one hand, finding sufficiently good correspondences for the alignment to converge to the correct solution highly depends on the model initialisation. Given initial pose and shape estimates as a starting point, candidates of point pairs on corresponding edges are searched along the normal vector of the rendered model edges in the image (Tsin et al., 2009; Leotta and Mundy, 2009). On the other hand, this procedure requires the desired intensity edges to be found in the first place. However, occlusions, illumination conditions, contrast, shadows or reflections are likely to cause outliers and consequently lead to incorrect correspondences.

An attempt to filter edge maps and extract semantically meaningful contours has been presented in (Isola et al., 2014) and is used in (Ortiz-Cayon et al., 2016) to reduce the number of outliers for the task of edge-to-edge alignment. Furthermore, a threshold for the angles between the edge normals of image and model edges is used as a criterion by Ortiz-Cayon et al. (2016) and Ramnath et al. (2014) to discard unlikely correspondence candidates. Still, the risk of incorrect matches and the need for good model initialisation remains.

In (Zhou et al., 2019), a CNN is used to infer edge maps for the task of reconstructing wireframes of buildings. While this approach shows promising results in only extracting desired edges, thus being robust to illumination, shadows or reflections, still no semantic information is extracted that would support the establishment of correspondences between model and image edges. Especially objects that exhibit symmetries in their wireframe representation, such as buildings and vehicles, suffer from missing semantic information behind the extracted edges. In this thesis, a CNN is trained to extract vehicle wireframes, too. However, in contrast to (Zhou et al., 2019), the CNN is trained to semantically distinguish between wireframe edges belonging to different sides of a vehicle, which allows an informed, more robust, and initialisation-invariant model-to-image-edge association.

Contour/silhouette based reconstruction

Another strategy for model based object reconstruction is to align the silhouette that results from the 3D model to a predicted segmentation mask of the target object. Prisacariu et al. (2012) train a Random Forest based on HoG and colour features for foreground-background classification and define an energy function considering pixelwise foreground and background matching scores to fit a deformable vehicle model to vehicle detections. A similar energy term is incorporated in (Dame et al., 2013), where foreground-background models are learned from reference segmentations for the parts of a DPM detector, which is used to infer pixel-wise foreground-background probabilities during test time. In (Kar et al., 2015), an instance segmentation method is applied to derive segmentation masks and a silhouette consistency term is used in the model fitting procedure. However, the mapping from a 3D shape model to a 2D image silhouette is highly multi-modal and ambiguous; the neglection of object details and structures within the silhouette is only one reason why this is the case. For example, a segmentation mask of a vehicle in a front and in a rear view is nearly the same, which makes the alignment highly sensitive to initialisation. Furthermore, it is also prone to segmentation errors and occlusions.

Surface based reconstruction

To reconstruct objects based on the 3D surface of an object's shape model requires depth or 3D information that can be derived from image observations (e.g. from stereo triangulation or Structure from Motion (SfM)) or can be obtained from laserscanning. Güney and Geiger (2015) use CAD models of vehicles to sample disparity patches from them and align these patches to disparity maps estimated from stereo images. Similarly to this, in (Menze et al., 2015) a 3D vehicle ASM is also estimated in accordance with a dense disparity map. In addition, images of a subsequent time epoch are used to also incorporate model constraints based on scene flow information. However, this thesis proposes a method for vehicle reconstruction using a single stereo image pair. In (Engelmann et al., 2016), a TSDF learned as a shape prior for vehicles is fitted to 3D point clouds obtained from dense stereo correspondences. Similarly, Xiao et al. (2016) fit a vehicle ASM to 3D points obtained from mobile laserscanning. 3D point reconstructions from multi-view images are used by Ortiz-Cayon et al. (2016) to fit a CAD model to detected vehicles. In these cases, model fitting is based on minimising the distances between the 3D points and the surface of the shape manifold. This procedure presents various difficulties. One of the major problems is the missing semantic information behind 3D points, required to relate them to corresponding parts of the vehicle model. Starting from an initial pose and shape, Engelmann et al. (2016) and Xiao et al. (2016) associate the 3D points to the closest point on the initial vehicle model hypothesis. The intersection points of the image rays with the model shape hypothesis are established as correspondences in (Ortiz-Cayon et al., 2016) to associate 3D points to the model surface. As a consequence, the minimisation procedure is similar to a point-to-surface iterative closest point (ICP) algorithm (Pomerleau et al., 2013) and highly depends on good pose initialisation. In case of imprecise initial poses, the model fitting is likely to fail. In this thesis, the requirement of good initial model registration is avoided by applying a Monte-Carlo based sampling strategy for optimisation. Our previous work has shown the operability of this strategy for model alignment based on stereo point clouds (Coenen et al., 2017).

Another problem arises from outliers in the 3D point cloud, resulting e.g. from detection, matching or segmentation errors, and points belonging to parts not represented by the generalised shape prior, like the vehicle's interior, antennae, mirrors, etc. To deal with this problem Xiao et al. (2016) propose to use a robust truncated distance function while Engelmann et al. (2016) apply the robust Huber norm for the alignment. Still, noisy point clouds and the increasing uncertainty of the stereo reconstructed 3D points with increasing distance from the camera remain challenging for reconstruction approaches that are only based on 3D points.

While 3D points provide explicit 3D information that is valuable especially for the position estimation of the object, they deliver a rather sparse and incomplete representation of the object, causing ambiguities w.r.t. shape and extent of the object. Further, when 3D points are used as the only data source, valuable image cues are disregarded completely for model fitting. To this end, Ortiz-Cayon et al. (2016) jointly align the model to 3D points and contours inferred from the image. Similarly to this, in (Dame et al., 2013) a depth consistency term between the model and depth estimations obtained from image sequences is combined with a segmentation consistency term between a segmentation mask inferred from the image and the segmentation mask resulting from the fitted model. These combinations allow for further improvement of model fitting. In this work, 3D information is also used jointly with 2D image information to exploit synergies of both domains. However, instead of using contours and segmentation masks, which do not carry semantic information and deliver rather weak constraints for model fitting as they preserve severe ambiguities w.r.t. pose and shape, semantic keypoints and wireframe edges are extracted from the images in this thesis, consequently allowing a more reliable model association.

Keypoint based reconstruction

Another strategy for shape-aware vehicle reconstruction is to match model keypoints with their corresponding keypoints detected in the image. One advantage of using semantic keypoints compared to the approaches described so far is the easier definition of correspondences between the image and model entities. Traditionally, handcrafted features, often based on HoG (Dalal and Triggs, 2005) features, are used for the model-to-keypoint alignment, e.g. in (Li et al., 2011; Zia et al., 2013; Bao et al., 2013), while recently CNNs are applied to detect keypoints, resulting in a better performance. For instance, Chabot et al. (2017) extend a RCNN for object detection to also regress keypoint coordinates in addition to the bounding box and the object class. A more frequently applied architecture for keypoint detection is a stacked-hourglass architecture (Pavlakos et al., 2017; Murthy et al., 2017b; Ding et al., 2018), in which multiple stacked U-Nets are used to infer keypoint probability maps from which the keypoint locations are derived, e.g., using non-maximum suppression. A U-Net-like architecture is also used in this thesis to predict keypoint heatmaps. However, in contrast to the listed previous works in which a mean-square error loss, which exhibits unfavourable properties w.r.t. the task of keypoint detection (cf. Sec. 4.3.5), is used to train the networks, we come up with a customised loss specifically tailored for keypoint prediction.

Given the 2D image keypoint locations and their corresponding vertices of the 3D model, one naive approach for model fitting is to minimize the reprojection error, a procedure which is also referred to as spatial resection in the literature, to solve for the 6DoF pose (Chabot et al., 2017). Considering a deformable model as shape prior in which the 3D model vertex coordinates are a function of a set of shape parameters, the 6DoF problem extends to a shape-aware spatial resection in which the shape parameters become part of the optimisation (Pavlakos et al., 2017). However, this naive procedure is problematic due to various reasons. On the one hand, it is intolerant to keypoint localisation errors, while at the same time the inferred 2D keypoint localisations are likely to be imprecise, leading to potentially large errors in 3D. On the other hand, it is not robust w.r.t to outliers such that potentially false keypoint detections influence the pose estimation directly and demand for robust strategies.

To overcome these problems, Pavlakos et al. (2017) use the respective confidence scores predicted by the CNN for each keypoint to derive a weight for the individual keypoint reprojection errors. This attempt relies on the assumption that imprecise or false keypoint detections are reflected by lower confidence scores. However, this assumption usually does not hold true. On the contrary, false detections often reveal very high confidence scores, which makes them hard to account for. The probabilistic approach presented in this thesis avoids the need for precise keypoint locations, because it is not built upon inferred 2D keypoint coordinates but on the raw keypoint probability maps instead. By incorporating the full keypoint probability distributions into the optimisation instead of only their inferred modes, the model fitting gains robustness w.r.t imprecise keypoint localisations caused e.g. by broad probability distributions. Performing the keypoint alignment on the raw heatmaps has also been done by Zia et al. (2013). The authors used a random forest (RF) classifier (Breiman, 2001), trained on gradient based handcrafted features for keypoint classification to derive the keypoint probability maps. The performance of a RF compared to deep learning based techniques, however, is qualitatively inferior as has been found by own previous work using a RF (Coenen et al., 2019) and a CNN (Coenen and Rottensteiner, 2019) for the prediction of vehicle keypoints.

To achieve a higher robustness w.r.t outliers, additionally to the keypoint weights based on the confidence scores mentioned above, Murthy et al. (2017a) introduce a weight that is derived from the prior probability of the keypoints being visible given the pose estimate. In this thesis self-occlusion is considered by only using potentially visible keypoints during inference, thus gaining robustness to false detections of self-occluded keypoints.

3.2.4 Optimisation

Usually, the problem of model fitting is formulated as an optimisation problem, in which, as comprehensively described above, e.g. the distance between image and backprojected model entities, such as keypoints or model edges, or the distance between observed 3D points and the 3D model surface is minimised. Formulating this goal as an objective function, the estimation of the target parameters corresponds to the minimisation of that function. Typically, the objective is non-linear w.r.t. the target parameters and typically non-convex, such that one common way for optimisation is to use local optimisation techniques by applying first-order approaches that are based on the Jacobian of the objective. Local optimisation techniques, however, require the availability of an already fairly good initialisation of the unknown parameters. In this context, Murthy et al. (2017a) apply an iterative least squares method for vehicle reconstruction to minimise the backprojection error of model keypoints and keypoints detected in the image. However, least squares optimisation in general is highly sensitive to outliers which are likely to occur in the context of keypoint detection due to imprecise or false positive keypoint detections. The goal of spatial resection based on vehicle keypoints is also defined in (Pavlakos et al., 2017), where gradient descent is applied for optimisation, though. Gradient descent is also used in (Engelmann et al., 2016) for the optimisation of an objective function which is based on the distance of 3D vehicle points to the surface of the 3D model that is utilised as shape prior. However, as non-linear least squares and gradient descent are methods for local optimisation, they do not guarantee finding the global optimum in case the objective function is non-convex. Instead, convergence to the correct minimum heavily relies on the initialisation of the parameters and consequently, already fairly accurate initial solutions are required. Often, coarse methods for object pose estimation are applied to generate initial estimates for the pose and as a consequence, the actual proposed approach for model fitting only serves as refinement of the pose and for the estimation of shape. Examples for this strategy can be found in (Engelmann et al., 2016), where the method of (Chen et al., 2015) for 3D bounding box prediction is applied for initialisation, or in (Zia et al., 2013), who make use of the detection and viewpoint prediction approach presented in (Pepik et al., 2015) to define start values for their optimisation approach.

Other than the approaches listed so far, Zia et al. (2013) do not use first order optimisation techniques for the pose and shape parameter estimation, but instead apply a stochastic hill-climbing approach for model fitting based on keypoint heatmaps predicted by a RF classifier. This approach is very similar to the Monte-Carlo based optimisation technique described in Sec. 2.3 and thus, closely related to the inference method that is applied in this thesis. In (Zia et al., 2013), multiple particles, each of which corresponds to an object hypothesis, are created as starting points with their corresponding scores computed from the objective function. Instead of computing the gradients, the particles are iteratively improved by randomly sampling new particles in their surrounding and preserving the best particles in each iteration. While this procedure in principle is able to account for the potential multi-modality of the posterior, the results in (Zia et al., 2013) nevertheless exhibit a high sensitivity to large errors in the initialisation. A reason for this can be the rather noisy keypoint probability maps that are produced by the RF, leading to ambiguities in the objective function. In this work, a Monte-Carlo based optimisation technique is also used due to the mentioned drawbacks of local solvers. In contrast to (Zia et al., 2013), we do not require any initialisation, proposing a better suited, more distinct, and unambiguous objective function for vehicle reconstruction.

3.3 Discussion

The previous sections have given an overview about existing and related work of image based pose estimation and reconstruction of vehicles. This section briefly summarises open questions and unsolved problems in the context of shape and pose recovery for vehicles. Based on these open points, our method will be proposed in the following chapter to tackle the identified research gap.

The analysis of related work for vehicle reconstruction shows that the majority of approaches that aim at a 3D pose and shape aware reconstruction of the object follow a model based path by leveraging a 3D shape prior and fitting it to the image. The ill-posed nature of 3D object reconstruction from 2D images makes purely data driven approaches difficult to solve due to the infinite number of combinations of object size, orientation and distance that would lead to the same appearance of the perspective projection. Constraints that restrict the ill-posed problem are hard to establish in data driven approaches, which makes the presented related work heavily dependent on the examples and conditions contained in the training data. Therefore, this thesis follows the line of model driven work, learning a 3D deformable shape prior to be fitted to the observations derived from stereo-images to compute the precise pose and shape of vehicles. Existing difficulties and unsolved problems of related methods that are tackled in the context of this thesis are discussed in the following paragraphs. The section starts with a discussion of observations used for the model fitting, followed by a discussion of prior knowledge that is employed to constrain the parameter space. The last paragraph discusses optimisation procedures applied for model fitting.

Observations

Aligning a 3D model to the image or to information derived from the image requires the establishment of correspondences between model entities and image observations. Most frequently, correspondences are derived from keypoint detections (Pavlakos et al., 2017), but also from reasoning about edges and contours (Ortiz-Cayon et al., 2016), or from reconstructed 3D points (Engelmann et al., 2016). As a consequence, the shape and pose of the model is determined by minimizing the distance between the assigned correspondences of the 3D model and the image observations of the respective entities.

The usage of semantic keypoints for model fitting has the advantage that the assignment of corresponding entity pairs is well defined. Most approaches make use of a point-wise, i.e. a single pixel representation of the keypoints, by deriving the image location for each keypoint from probability maps which are predicted by a CNN (Pavlakos et al., 2017; Chabot et al., 2017; Murthy et al., 2017b). The model fitting is usually done by minimising the reprojection error between detected image keypoints and backprojected model keypoints. However, this procedure is intolerant to false detections and is sensitive to imprecise 2D keypoint localisations leading to potentially large errors in 3D. Furthermore, the additional information contained in the probability maps is neglected. The incorporation of the raw keypoint probability maps into the model fitting procedure is rarely done in the literature. Yet, the probability distributions are likely to deliver valuable cues, e.g. regarding the uncertainties of the keypoint predictions. As an exception, in (Zia et al., 2013), a probabilistic approach for model fitting, which is based on probability maps for keypoints predicted by a Random Forest classifier, is presented. However, the probability maps obtained by the RF are very noisy, thus hampering the model fitting performance. In contrast to the RF applied in (Zia et al., 2013), this thesis presents a multi-task CNN for the generation of observations. The proposed approach for vehicle reconstruction incorporates keypoint probability maps predicted by one branch of that CNN, exploiting the more promising performance of deep learning for keypoint detection compared to a RF classifier. Furthermore, this thesis proposes the additional incorporation of other cues that were not used in (Zia et al., 2013).

In comparison to keypoints, edges representing the wireframe of the vehicles provide a more comprehensive description of the object and therefore deliver an advantageous representation for the task of model alignment. In existing work, inferring the object wireframe from the image is usually based on generic edges which are extracted based on image gradients (Ortiz-Cayon et al., 2016), expecting to actually find the desired edges which correspond to the projections of the object's wireframe. Starting from an initial pose, the assignment of image edges to backprojected model edges is essentially based on nearest neighbour association. This procedure reveals two severe problems which cause edge-wise representations to be rarely used for model fitting: firstly, shadows, reflections and lighting conditions are likely to cause undesired or missing edge detections, and secondly, the naive edge-to-edge association highly depends on the initialisation of the model. A learning based extraction of wireframe edges which, furthermore, allows the addition of a semantic component to the edges and therefore facilitates the establishment of edge-to-edge correspondences, has not been exploited in the literature for vehicle reconstruction so far. In this work, the extraction of model edges is built upon the hypothesis that, if it is possible to detect specific keypoints, it will also be possible to detect edges corresponding to the desired model wireframes without the explicit dependency on good image gradients. Furthermore, it should be possible to differentiate between edges belonging to different physical parts of the object during the detection, thus enriching the edge detections by semantic information which allows a more distinct and better defined association of extracted edges to model contours. To this end, this thesis proposes *semantic wireframe edges* as new feature for model fitting. The usage of a CNN is exploited for a contrast and illumination insensitive extraction of target vehicle wireframes, differentiated by their affiliation to one of the four vehicle sides (front, left, back, right). Due to the advantages regarding the usage of a probabilistic representation of the observations mentioned above, one branch of the multi-task CNN is trained to jointly produce probability maps for both, semantic keypoints as well as the semantic wireframe edges, in order to be incorporated into the model fitting procedure.

3D points reconstructed from stereo images are used in (Engelmann et al., 2016) for the task of model fitting by minimising the distance between the 3D points and the surface of the applied shape prior. While 3D points have the benefit of carrying valuable explicit 3D information, they do not contain semantic information and therefore, lead to similar point-to-model association problems as described above in the context of generic wireframe edges. Nearest neighbour relations are used to associate a 3D point to its corresponding point on the model surface (Engelmann et al., 2016), which leads to the requirement of fairly precise model initialisations or the need for additional observations which aid the procedure of model fitting. However, another aspect that becomes apparent from the literature review is that only few attempts are made to fuse different domains (e.g. 2D vs. 3D observations) or towards the fusion of different entities (e.g. keypoints vs. edges) for the purpose of object reconstruction. Mostly, methods are built on top of only one of the aforementioned domains or entities. The complementary effect or the benefit from redundancy that the combined incorporation of multiple entities from multiple domains can have on the reconstruction of objects is barely exploited. To fill this research gap, this thesis proposes a comprehensive probabilistic model as core of the reconstruction approach which simultaneously incorporates the keypoint and wireframe detections together with 3D point reconstructions, combining 2D image with 3D object space observations. Besides, the probabilistic formulation allows to derive a dynamic weighting between the different observations based on error propagation, thus reducing the number of hyperparameters.

Prior information

In exiting approaches, the incorporation of prior knowledge into the model fitting procedure is often limited to the usage of the shape priors and occasionally to an application of localisation constraints based on an estimate of the ground plane (Chen et al., 2015). In this work, the ground plane is not only estimated to simplify the 6DoF pose problem, but also to derive a representation of free space within the scene, which is incorporated as prior on admissible vehicle locations on the road surface. To this end, this thesis proposes a *probabilistic free-space grid map* which makes use of the 3D stereo point cloud to derive a probability distribution of free space in the scene.

To constrain shape deformations, restricting the shape prior to only result in geometrically valid shapes, the common approach is to penalise deviations from the mean shape determined from the training instances, e.g. (Zia et al., 2013). However, this procedure is founded on the assumption that the difference of vehicle shapes follow a unimodal distribution. Arguably, this assumption does not hold true and systematically impairs vehicle categories which differ from the average shape. Instead, in this thesis a multi-modal distribution is assumed for vehicle shapes, with different modes for different vehicle categories. In this context, a new shape prior is proposed which extends existing models by explicitly learning the modes of the shape variations corresponding to the different vehicle types. A prior term is introduced to the probabilistic formulation for model fitting which takes into account shape variations based on a predicted probability distribution for the respective vehicle's category. To this end, one branch of the multi-task CNN developed in this thesis is trained to predict the vehicle type and is used to derive a probability distribution for the vehicle type.

Prior information about the viewpoint is used for vehicle model fitting in (Engelmann et al., 2016; Zia et al., 2013). The authors make use of a coarse prediction of the viewpoint to initialise the vehicle model and to define start values for the optimisation procedure. If the viewpoint prediction is accurate enough, their presented approaches for model fitting will be able to deliver a refined estimate of the pose. However, in case of incorrect prior viewpoint predictions, model fitting is likely to fail due to its strong dependency on the initialisation, and erroneous viewpoint predictions cannot be corrected afterwards. A probabilistic prior representation of the viewpoint, which carries potential uncertainties of the viewpoint prediction in the form of a probability distribution and, consequently, reduces the sensitivity to erroneous predictions since it allows a correction of incorrect predictions when it is sufficiently supported by other observations, has not been used for vehicle reconstruction so far. Therefore, this thesis introduces a regulariser for the vehicle's orientation to the proposed approach for model fitting which makes use of probability distribution of the viewpoint, predicted by a dedicated branch of the multi-task CNN.

CNN

In order to detect object keypoints from images, frequently CNNs are applied in which U-Net-like architectures are used to predict keypoint heatmaps in a first step (Pavlakos et al., 2017; Murthy et al., 2017b; Ding et al., 2018) to finally derive the keypoint locations from them. For the training of the network, the authors make use of the mean-square error loss. However, due to the extremely small proportion of keypoint pixels w.r.t. non-keypoint pixels in the image, the MSE leads to an unfavourably broad shape of the loss function (a detailed explanation can be found in Sec. 4.3.5). To overcome this problem, a new custom loss is proposed in this thesis, introducing a different weighting for different groups of pixels for the loss computation. Furthermore, the CNN for the prediction of keypoint heatmaps is extended in this thesis to also produce heatmaps for *semantic wireframe edges* which introduced as a new feature for model fitting. This is in contrast to existing work, where generic edges are extracted for model fitting (Ortiz-Cayon et al., 2016) and the learning based extraction of semantic edges has not been exploited so far.

CNNs are also used for the prediction of the viewpoint of an object (Tulsiani and Malik, 2015; Su et al., 2015). Usually, the problem is formulated as a classification task, in which the network is asked to predict one of a discretised set of viewpoint classes. In this way, the degree of discretisation defines the granularity of the output. To achieve more fine-grained results more viewpoint classes need to be distinguished, leading to a higher complexity for the classification task (Su et al., 2015). Therefore, it is reasonable to expect the classification accuracy to decrease with an increasing number and therefore a finer definition of viewpoint bins. In order to profit from the higher classification accuracy for the coarse class definition and the finer level of detail of the fine class definition, a hierarchical solution is for the viewpoint classification is proposed in this thesis where multiple hierarchical layers of classes are combined, each containing different numbers of viewpoint classes.

Optimisation

Regarding the optimisation of pose and shape parameters, minimising the reprojection error between image and backprojected model entities is highly sensitive to gross errors and outliers, i.e. to imprecise and false detections. As in the literature usually local optimisers such as gradient descent or iterative least squares are used for optimisation (Engelmann et al., 2016; Pavlakos et al., 2017; Murthy et al., 2017b), the result also highly depends on a good initialisation of the parameters. An open research gap is to find methods that are insensitive to factors such as incorrect initialisation and detection errors. In this thesis, a Monte-Carlo based optimisation technique is developed, that does not require a good initialisation but which can handle the multi-modality of the objective function. In its principle, the proposed Monte-Carlo optimisation technique is similar to the stochastic hill-climbing method presented in (Zia et al., 2013). However, while Zia et al. (2013) require fairly accurate initialisations for their approach to find the optimal values, the optimisation procedure in this thesis is set up to be invariant w.r.t. the initialisation by defining a sampling strategy covering the whole range of possible solutions in the parameter space.

4 Methodology

This chapter presents the core of this thesis and proposes a new, fully automatic approach for the model based shape and pose recovery of initially detected vehicles from stereo imagery. Sec. 4.1 gives an overview of the method, describing the detailed problem statement and some required pre-processing operations. The subcategory-aware vehicle model which is learned as a shape prior for the model fitting approach is described in Sec. 4.2. Sec. 4.3 gives a detailed description of the new multi-task CNN which extracts the probabilistic information incorporated into the novel probabilistic model for vehicle reconstruction, which is finally presented in Sec. 4.4. A discussion about assumptions made in this thesis and about advantages and disadvantages of the proposed method is given in Sec. 4.5.

4.1 Overview

The essential goal of the proposed method is to recover the precise 3D pose (i.e. the 6DoF parameters of the position and orientation in 3D) as well as the type and shape of vehicles detected from street-level stereo images as input. Details and requirements on the input data are given in Sec. 4.1.1. The proposed processing pipeline is designed to deduce a 3D vehicle model which represents the detected vehicle best in terms of pose and shape. To this end, a 3D model is fitted to the observed data. The target parameters are implicitly contained in the fitted model. Sec. 4.1.2 gives a formal description of the problem statement and introduces the notion of the target parameters. Prior knowledge about the 3D layout of the observed scene is extracted and used to constrain the parameter space of the model fitting approach (cf. Sec. 4.1.3). A deformable vehicle model is learned as a shape prior (cf. Sec. 4.2) and is fitted to the detected vehicles. In the presented method, the detection of vehicles and their reconstruction are treated as decoupled tasks. To initially detect the vehicles visible in the stereo images, a state-of-the-art detection method is adapted and its output is tailored towards the requirements for the proposed vehicle reconstruction method (cf. Sec. 4.1.4).

Based on the initial vehicle detections, the core of the proposed method consists of **(1)** a novel subcategory-aware deformable vehicle model (Sec. 4.2), **(2)** a multi-task CNN (Sec. 4.3), trained to extract various pieces of semantic information from the vehicle detections to be incorporated into **(3)** an extensive probabilistic model that is designed to find the best fitting instance of the

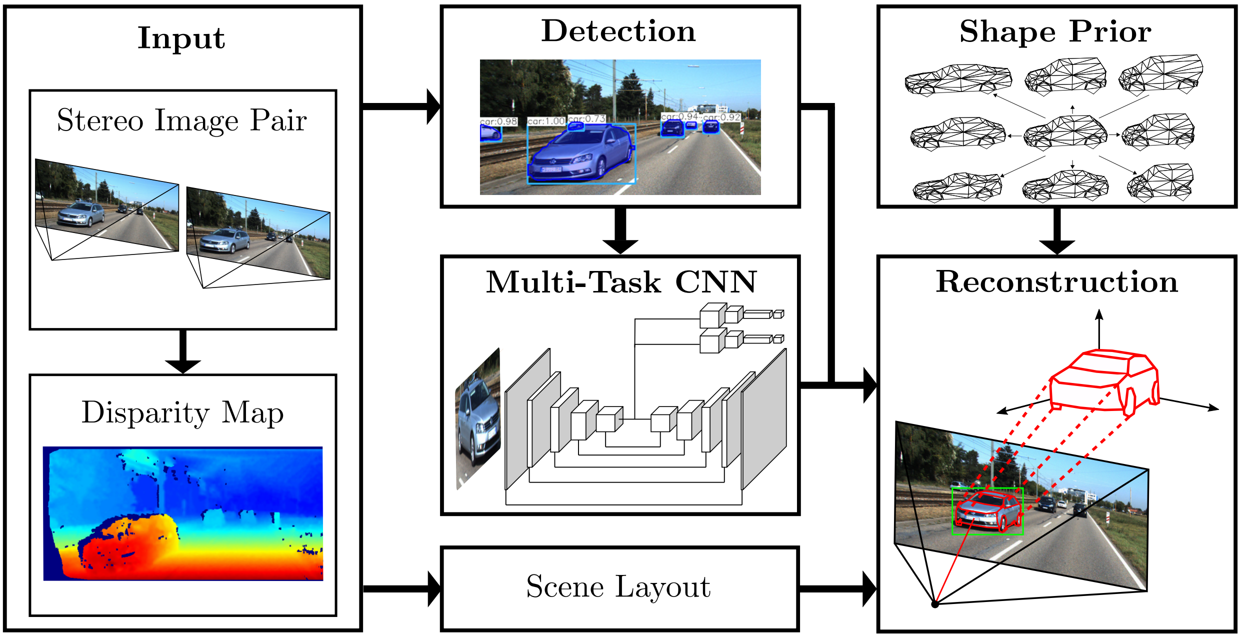


Figure 4.1: Overview of the proposed framework. The input of the method is a stereo-image pair and a disparity map to derive 3D information. In an intermediate step, vehicles and their segmentation masks are detected and global scene layout is derived from the 3D data. The vehicle detections are fed through the proposed CNN to derive probabilistic information about orientation, type and locations of vehicle keypoints as well as wire-frame edges. In the final reconstruction step, a parametrised instance of the deformable shape prior is fitted to each vehicle detection.

deformable model for each vehicle (Sec. 4.4). The general framework of the proposed approach is depicted in Fig. 4.1 and will be presented in the following sections.

4.1.1 Input

The input to the proposed method consists of stereoscopic image pairs. The following assumptions are made w.r.t. the imagery to be processed by the present approach. The images are acquired at street level, e.g. by a stereo rig attached to a moving platform. The stereo head is calibrated so that the interior and relative orientation incl. the base-length are assumed to be known. The images are assumed to be rectified so that epipolar lines correspond to image rows. The synchronisation of the stereo pair is sufficiently accurate so that the influence of object and camera movements can be neglected.

The left stereo partner is defined to be the reference image. To derive 3D information, dense matching is applied to calculate a dense disparity map for every stereo pair. The state-of-the-art ELAS matcher (Geiger et al., 2011) is used for this purpose. The disparity images are used to reconstruct a 3D point cloud in the 3D model coordinate system ${}^{M}C$ from every pixel of the reference image via triangulation. The origin of the model coordinate system is defined to be the projection centre of the left camera. Its ${}^{M}X/{}^{M}Y$ plane is parallel to the image plane of the epipolar images and the ${}^{M}Z$-axis points into the viewing direction (cf. Fig. 4.2).

4.1.2 Problem statement

The ultimate goal of the proposed method is to detect the set of vehicles $\mathcal{V}$ visible in the given stereo pair and to recover their precise 6DoF poses in the coordinate system ${}^{M}C$, i.e. their poses relative to the reference camera of the stereo rig. For this purpose, we describe each stereo scene by a 3D ground plane $\Omega \in \mathbb{R}^3$. The extraction of the ground plane and the construction of a coordinate frame ${}^{\Omega}C$ attached to that plane are described in the subsequent Sec. 4.1.3. Further, each vehicle $v \in \mathcal{V}$ is associated with its state vector $\mathbf{s}$=($\mathbf{t}$, θ, γ). The state vector contains the pose and shape of the vehicle represented by its position $\mathbf{t} = [t_x, t_y, t_z = 0]$ on the ground plane in ${}^{\Omega}C$, its orientation θ, i.e. the rotation angle about the normal vector of the ground plane and a vector γ determining the shape of a deformable ASM representing the vehicle (cf. Sec. 4.2). It has to be noted that the 6DoF vehicle pose parameters w.r.t. the reference camera can easily be derived from the 2D pose $\mathbf{t}$ and θ on the ground plane, knowing the rigid transformation between the ground plane system ${}^{\Omega}C$ and the model system ${}^{M}C$ (cf. Sec. 4.1.3).

4.1.3 Scene layout

Prior to vehicle reconstruction, the stereo data is used to derive knowledge about the 3D layout of the scene, represented by the 3D ground plane and a probabilistic free-space grid map. Introducing the requirement for vehicles always to be located on the ground plane, estimating that plane reduces and constrains the parameter space of the model fitting approach. More specifically, the determination of the ground plane predetermines three of the 6DoF vehicle pose parameters (1 translational and 2 rotational parameters).

Ground plane

Given a street-level acquisition setup with an approximately horizontal viewing direction, the 3D points belonging to the ground plane correspond to the set of 3D points with the largest vertical coordinate (i.e. with the maximum value of ${}^{M}Y$; note that ${}^{M}Y$ points downwards). RANSAC (Fischler and Bolles, 1981) is applied to a user-defined percentage of the stereo point cloud exhibiting the largest vertical coordinate values to find the ground plane Ω as the plane of maximum support. All inliers of the final RANSAC consensus set are stored as ground points $\mathbf{X}_{\Omega}$. Additionally to the model coordinate system ${}^{M}C$, a local ground plane coordinate system ${}^{\Omega}C$ is defined. Its origin corresponds to the orthogonal projection of the model system origin to the ground plane. Its ${}^{\Omega}Z$-axis is defined to be identical to the direction of the plane normal vector and the ${}^{\Omega}X/{}^{\Omega}Y$ plane lies

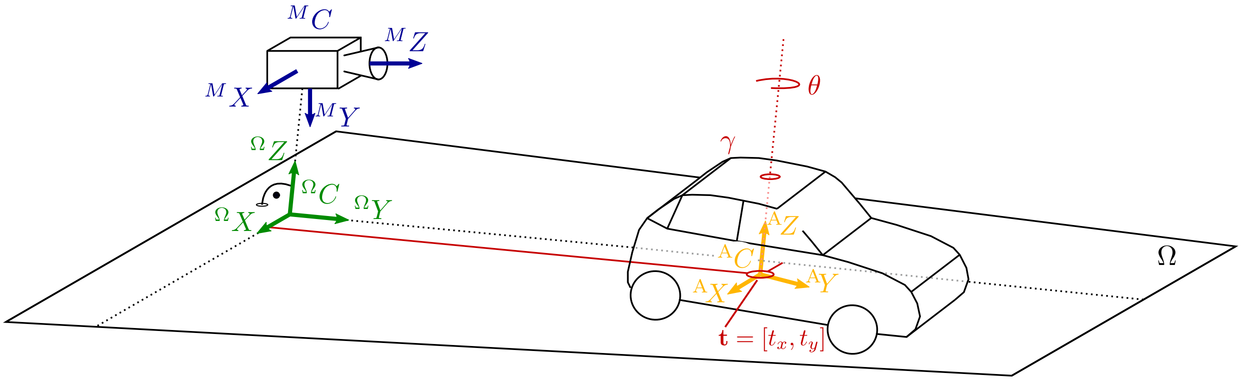


Figure 4.2: Illustration of the model coordinate system ^{M}C, the ground plane coordinate system $^{\Omega}C$, and the vehicle coordinate system ^{A}C, as well as the ground plane Ω and the vehicle model state parameters for position $\mathbf{t} = [t_x, t_y]$, orientation θ and shape γ.

in the ground plane (cf. Fig. 4.2). More specifically, the direction of $^{\Omega}Y$ is defined as the projection of the ^{M}Z axis to the ground plane and $^{\Omega}X$ completes $^{\Omega}C$ as an orthogonal right-handed system. By determining the rotation matrix Q and the translation vector T as the parameters of a rigid transformation between both systems, a point $^{M}\mathbf{X}$ in the model system can be transformed to the ground plane system by

$$^{\Omega}\mathbf{X} = Q \cdot {}^{M}\mathbf{X} + T \tag{4.1}$$

and vice versa.

Probabilistic free-space grid map Based on the ground plane points $\mathbf{X}_{\Omega}$ and all remaining points not belonging to the ground plane, the latter representing the set of arbitrary object points $\mathbf{X}_{Obj}$, it is possible to reason about free-space in the observed scene, i.e. areas on the ground plane not being occupied by any object. A probabilistic free-space grid map Φ is created to quantitatively represent the free-space areas. For this purpose, a grid is created in the ground plane consisting of square cells with a side length l_{Φ}. For each grid cell Φ_g with $g \in [1, G]$ the number of ground points n_{Ω}^{g} and the number of object points n_{Obj}^{g} whose orthogonal projection falls in the respective cell are counted. The probability ρ_g of each cell to be free-space is calculated according to

$$\rho_g = \frac{n_{\Omega}^{g}}{n_{\Omega}^{g} + n_{Obj}^{g}} \tag{4.2}$$

Grid cells without any projected points are marked as *unknown*. Fig. 4.3 visualises the free-space grid map for an exemplary scene.

In the probabilistic model proposed in this thesis for vehicle reconstruction, the free-space grid map is used to derive prior information about the vehicle's position (cf. Sec. 4.4).

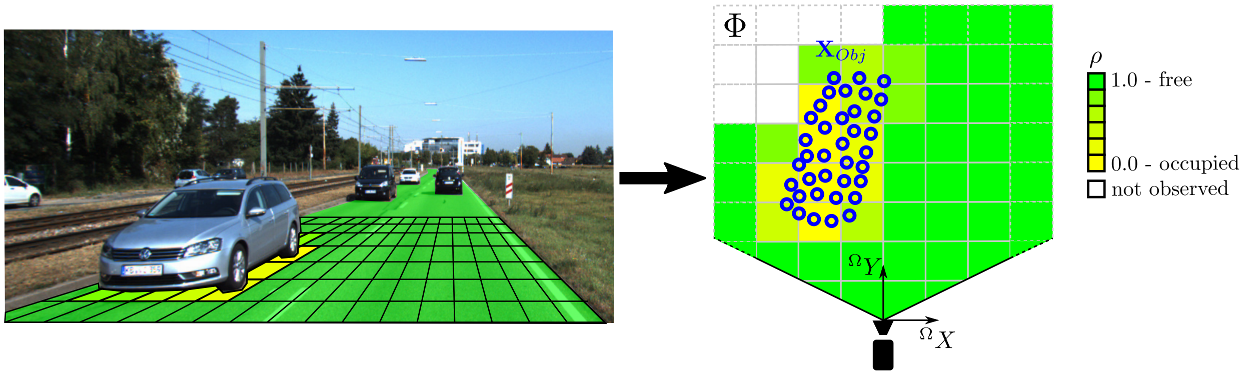


Figure 4.3: Visualisation of the probabilistic free-space grid map Φ. Left: Visualisation of the grid cells g in the image, coloured by the probability ρ_g of being free space. Right: visualisation of Φ from the same scene, but in top-view. The blue circles indicate the 3D triangulated object points $\mathbf{X}_{Obj}$.

4.1.4 Detection of vehicles

As an initial step of the method presented in this thesis, the presence and location of vehicles in the input images has to be discovered. This is a classical object detection and instance segmentation problem which has been extensively investigated for the last couple of years, with the result of very well performing object detection approaches being available today. One example is the Mask-RCNN (He et al., 2017) (cf. Sec. 2.1.2) which, besides its good performance in object detection, has the advantage of not only delivering bounding boxes containing instances of a specific object category: In addition, the Mask-RCNN infers a binary segmentation mask for each detected object instance. In this thesis, the Mask-RCNN, pre-trained on the COCO dataset (Lin et al., 2014a), is applied to the reference image to initially obtain the set of detected vehicles $\mathcal{V}$. Each vehicle $v \in \mathcal{V}$ is associated with an observation vector $\mathbf{o}^{det} = (\mathbf{X}_v, {}^{l}\mathcal{B}, {}^{r}\mathcal{B})$ containing a set of object points $\mathbf{X}_v$, as well as its bounding boxes ${}^{l}\mathcal{B}$ and ${}^{r}\mathcal{B}$ in the left and right image, respectively. To extract the vehicle points $\mathbf{X}_v$, the 3D points reconstructed from the disparities of the foreground pixels belonging to the respective segmentation mask are chosen initially. This initial set of 3D points is likely to contain outliers due to segmentation and/or matching errors. To isolate outliers, the number of neighbouring points within a user-defined radius is determined for each point in $\mathbf{X}_v$. Each point whose number of neighbours falls below a predefined threshold is treated as an outlier and rejected from $\mathbf{X}_v$. Naive nearest neighbour search for a large number of points is prohibitively slow. Kd-trees (Bentley, 1975) provide an efficient data structure for nearest neighbour search and hence are utilized for this purpose. More precisely, the approximate nearest neighbour search based on kd-trees proposed in (Muja and Lowe, 2009), which is much faster than the exact search with only a minor loss in accuracy, is exploited in this thesis. While the vehicle bounding boxes in the left image ${}^{l}\mathcal{B}$ are delivered as an output of the Mask-RCNN, the filtered point clouds $\mathbf{X}_v$ are used to deduce the corresponding bounding boxes in the right image. For this purpose, the 3D points $\mathbf{X}_v$ are projected to the right image, resulting in a set of image coordinates. The vehicle bounding box ${}^r\mathcal{B}$ is defined as the minimum axis parallel rectangle enclosing this set of image points. The approach for vehicle reconstruction presented in this thesis builds upon the vehicle detections by fitting a deformable vehicle model to each detected vehicle.

4.2 Subcategory-aware 3D shape prior

In this work, an Active Shape Model (ASM) is learned as a deformable shape representation for the object class *vehicle* according to the description in Sec. 2.2, to be incorporated as a shape prior for vehicle reconstruction. A set $\mathcal{K}$ of a total number of $C_\mathcal{K}$ of 3D keypoints were manually labelled for a variety of CAD models of vehicles belonging to one of a set of different vehicle types $\mathcal{T}$. (In the experiments of this thesis, seven vehicle types are distinguished with $\mathcal{T} = \{$*Compact Car, Sedan, SUV, Estate Car, Sports Car, Truck, Van*$\}$). Within $\mathcal{K}$, the keypoints represent a fixed structure, i.e. consistently for all CAD models, each keypoint describes the same semantic vehicle point, which, however, exhibits different coordinates in the vehicle's body frame ${}^AC$ for each CAD model. The body frame ${}^AC$ is a coordinate system attached to the vehicle with its origin at the centre of the vehicle's footprint, its ${}^AY$ axis pointing in forward direction of the vehicle and its ${}^AZ$ axis pointing upwards. Its ${}^AX$ axis completes ${}^AC$ as an orthogonal right-handed system (cf. Fig. 4.2). A synthesised model, deformed according to the shape parameters γ following Eq. 2.10, is denoted by $M(\gamma)$ and contains the deformed vertex positions of every keypoint in body frame coordinates. It has to be noted that in the practical realisation of the presented method, not all eigenvalues and eigenvectors of the ASM are considered. Instead, in order to reduce the dimensionality of the unknown shape parameter vector γ, the number n_s of considered shape parameters is restricted and is defined as a proper tradeoff between the complexity of the model and the quality of the model approximation (cf. Sec. 5.3.1). A fully parametrised instance of a 3D vehicle ASM in the ground plane coordinate system ${}^\Omega C$, denoted by $M(\mathbf{s})$, can be created by shifting and rotating the deformed model $M(\gamma)$ on the ground plane according to the translation vector $\mathbf{t}$ and a rotation about the ${}^AZ$ axis with the heading angle θ by

$$M_k(\mathbf{s}) = R_z(\theta) \cdot M_k(\gamma) + \mathbf{t} \tag{4.3}$$

where k is an index for the keypoints in $\mathcal{K}$ and $R_z(\theta)$ is a rotation matrix which performs the rotation according to the vehicle's orientation θ.

4.2.1 Geometrical representation

While the classical representation of an ASM only contains explicit information about the keypoints (Cootes et al., 1995), in this work, the ASM is enriched by an additional explicit definition of the model surface as well as a definition of model wireframe edges. In particular, a triangular mesh M_Δ is defined for the ASM vertices $\mathcal{K}$ to represent the model surface. Every entry of M_Δ contains a triplet of keypoints defining a single triangle. Automatic methods to generate the triangulated surface, such as *convex hull* (Barber et al., 1996) or *Delauney triangulation* (Cignoni et al., 1998) approaches were found to deliver partially incorrect surface representations for the vehicle ASM due to its non-convex shape or its unsuitable distribution of the keypoints. This is why the topology of the mesh representing the surface of the ASM is manually defined once and kept constant for all generated, deformed models. A visualisation of the defined triangular vehicle mesh can be found in Fig. 4.4. Note that keypoints exist (e.g. the center point of the wheels) which are not part of the triangulation in order to decrease the number of faces.

Moreover, edges connecting two keypoints are manually chosen to define a wireframe topology $M_\mathcal{W}$ of the vehicle model, consisting of both, *crease* edges that describe the outline of the vehicle, and *semantic* edges describing the boundaries between semantically different vehicle parts. Each entry of $M_\mathcal{W}$ contains a tuple of keypoints from $\mathcal{K}$ defining a single edge of the wireframe. The edges chosen for this purpose are depicted in Fig. 4.4. Furthermore, the wireframe $M_\mathcal{W}$ is subdivided into four groups of wireframe edges $M_\mathcal{W}^w$, each of which contains all edges in $M_\mathcal{W}$ belonging to the wireframe of one of the four vehicle sides $w \in \{\text{front, back, left, right}\}$. Note that the edges in the four wireframe subsets are not mutually exclusive, as a wireframe edge can belong to two of the distinguished vehicle sides. This distinctive wireframe representation adds further semantic information to the edges. The approach for vehicle reconstruction proposed in this thesis makes use of this semantic wireframe distinction by learning a CNN based detector for each of the wireframe representations $M_\mathcal{W}^w$ (cf. Sec. 4.3.4).

In addition to that, a subset of keypoints is chosen to contain the *appearance* keypoints $\mathcal{K}_A \subset \mathcal{K}$ for which an image based detector is learned as well (cf. Sec. 4.3.4). This set of keypoints contains a number of C_A keypoints with a potentially distinctive appearance, such as centre points of wheels, corner points of the wind shield and the rear window, front and back lights, etc. In comparison to the commonly used keypoint based ASM representation (Zia et al., 2013; Pavlakos et al., 2017), the ASM proposed in this thesis is extended by the explicit definition of the 3D surface, by the wireframe definition described earlier, as well as by the subcategory awareness which is described subsequently in Sec. 4.2.2. The triangulated surface as well as the wireframe and *appearance* keypoint definition can be seen in Fig. 4.4.

4.2.2 Mode Learning

A common practice in the literature is to constrain the shape parameters according to the deviations from the mean shape $\mathbf{m}$ (Zia et al., 2013; Engelmann et al., 2016), i.e. according to the deviations of γ from the zero vector (cf. Eq. 2.10). This strategy follows the assumption that the distribution

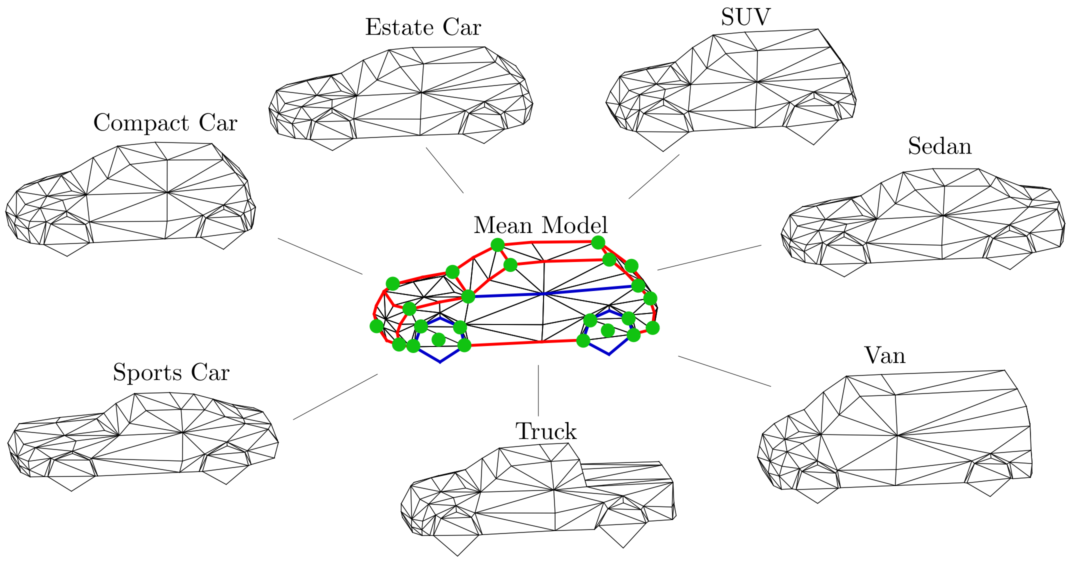


Figure 4.4: Visualisations of the ASM. In the centre, the mean model is depicted with *crease* edges in red and *semantic* edges in blue. The triangulated surface is shown in black, the *appearance* keypoints in green. Additionally to the mean model, deformed models corresponding to the modes γ^τ for seven vehicle categories are shown.

of vehicle shape deformations that are represented by the underlying principal components of the ASM corresponds to an unimodal distribution. However, rather than following a simple distribution having only one mode, there are several disjoint modes for different vehicle classes. As an example, let $\mathcal{T}$ only contain the types *Van* and *Sports Car*, then the mean model represents a shape which is neither a Van nor a Sports Car. Penalising the deviations from the mean shape is therefore not reasonable in this example. In contrast, in this work, the mode for each of the vehicle types in $\mathcal{T}$ is determined in addition to the general ASM formulation. During reconstruction, a prediction of the vehicle type (cf. Sec 4.3.2) is used to penalise deviations from the corresponding mode instead of the global mean shape. We denote the mode for every vehicle type $\tau \in \mathcal{T}$ by γ^τ, a vector representing the shape parameters that describe the mean shape $\mathbf{m}^\tau$ of all training exemplars associated to the respective type τ such that following Eq. 2.10 the mean category shape and its residuals $\mathbf{v}_\tau$ are expressed by

$$\mathbf{m}^\tau + \mathbf{v}_\tau = \mathbf{m} + \sum_{s=1}^{n_s} \gamma_s^\tau \sigma_s \mathbf{e}_s. \tag{4.4}$$

The modes are calculated according to a least-square fitting of the overall ASM to the mean shape of each type by

$$\gamma^\tau = (\mathbf{A}^\mathrm{T}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}(\mathbf{m}^\tau - \mathbf{m}), \tag{4.5}$$

where $\mathbf{m}$ is the mean model of all training exemplars. The mean shape of each type $\mathbf{m}^\tau$ is calculated in advance, given the CAD models and their annotated vehicle category. In this context, it has to be noted that the only additional annotation effort that is required for the proposed mode learning approach consists of associating a vehicle type label to each CAD model used for learning the ASM. The Jacobi matrix $\mathbf{A}$ is calculated from the partial derivatives of the linear Eq. 2.10 by the shape parameters γ_s. Given Eq. 2.10, the s^{th} column of $\mathbf{A}$ corresponds to the eigenvector $\mathbf{e}_s$ multiplied with the respective eigenvalue σ_s. With $C_{\mathcal{K}}$ keypoints defining the ASM and n_s parameters that are considered during vehicle reconstruction, the dimensions of the Jacobi matrix are equal to $[C_{\mathcal{K}} \cdot 3 \times n_s]$. As a consequence of the properties of $\mathbf{A}$ mentioned above and due to the orthonormality of the eigenvectors $\mathbf{e}_s$, $(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = diag(\sigma_1^2, ..., \sigma_{n_s}^2)$. Furthermore, the s'th row of $\mathbf{A}^{\mathrm{T}}(\mathbf{m}^\tau - \mathbf{m})$ can be expressed by the scalar product $\langle \mathbf{e}_s, \mathbf{m}^\tau - \mathbf{m} \rangle$ multiplied with the corresponding eigenvalue σ_s. Therefore, Eq. 2.10 can be simplified as the s^{th} component of the target shape vectors γ^τ can be calculated by

$$\gamma_s^\tau = \frac{1}{\sigma_s} \langle \mathbf{e}_s, \mathbf{m}^\tau - \mathbf{m} \rangle. \tag{4.6}$$

The mean ASM model as well as the deformed models according to the modes γ^τ are shown in Fig. 4.4.

4.3 Multi-Task CNN

As a first step of the model-based vehicle reconstruction, a mulit-task CNN is trained and used to infer semantic information to be used in the probabilistic model. The input to the network is an image showing a vehicle, cropped by the detection bounding box. The multi-task CNN consists of one common input branch and three individual output branches, each of them corresponding to one task, respectively. Unless specified differently, 3x3 filters are used in the convolutional layers and 2x2 filters with stride 2 are used for max pooling and upsamling operations. The overall architecture of the network can be seen in Fig. 4.5. The network output consists of a probability distribution $\Pi_{\mathcal{T}}$ for the vehicle's type, a probability distribution Π_{ϑ} for the vehicle's viewpoint, probability heatmaps $\mathcal{H}_{\mathcal{K}}$ for vehicle keypoints, and probability heatmaps $\mathcal{H}_{\mathcal{W}}$ for the vehicle wireframe edges. More details will be given in the following sections.

Notation: In this work, all vehicle detections in the reference (left) image, cropped by their respective bounding box, are fed into the network to infer the viewpoint probability distributions $\Pi_{\mathcal{T}}$ and Π_{ϑ}, the keypoint heatmaps ${}^l\mathcal{H}_{\mathcal{K}}$ and the wireframe heatmaps ${}^l\mathcal{H}_{\mathcal{W}}$. Additionally, the keypoint and wireframe heatmaps ${}^r\mathcal{H}_{\mathcal{K}}$ and ${}^r\mathcal{H}_{\mathcal{W}}$ are computed for the corresponding bounding box crops of the right image by feeding the detection window through the keypoint/wireframe branch. The results are gathered in an additional observation vector $\mathbf{o}^{cnn} = ({}^l\mathcal{H}_{\mathcal{K}}, {}^l\mathcal{H}_{\mathcal{W}}, {}^r\mathcal{H}_{\mathcal{K}}, {}^r\mathcal{H}_{\mathcal{W}})$.

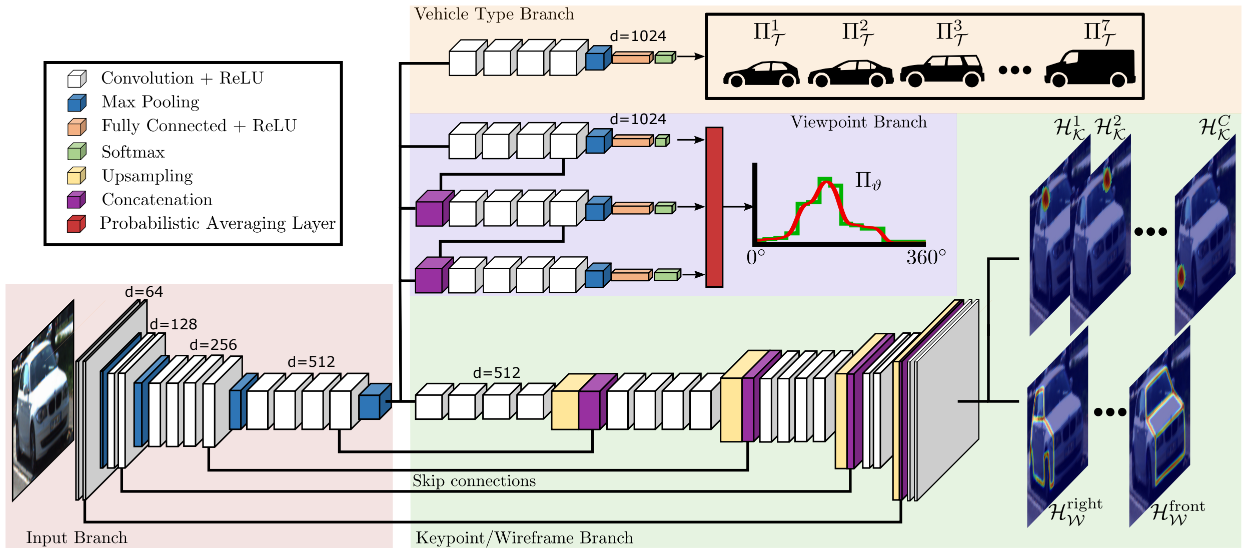


Figure 4.5: Architecture of the multi-task CNN. The input is a 3 channel image of size 224x224. The convolutional filters have size 3x3, max pooling and upsampling use filter size 2x2 and stride 2. The number of filters is denoted by d in the figure. Further explanations are given in the main text.

4.3.1 Input branch

The input to the network is a 3-channel image of size 224x224. Bilinear interpolation is used to resize image crops of the vehicle detections to the required size. The input branch acts as a shared backbone feature extractor of the CNN. The architecture of the VGG19 network (Simonyan and Zisserman, 2015) is adopted for this purpose (cf. Sec.2.1.2), defining the *input branch* to correspond to the first 16 layers of that network. Thus, this branch contains 12 convolutional layers and performs four max pooling operations. The number of filters d applied in each layer is given in Fig. 4.5. The feature map of size 14x14 produced by the *input branch* is forwarded to the task specific branches which are explained in the following sections.

4.3.2 Vehicle type branch

The *vehicle type branch* is used to obtain a prediction of the vehicle type for the target vehicle shown in the input image window. More specifically, this branch starts with a series of four convolutional layers followed by a max pooling layer. Together with the input branch, this branch complements the VGG19 architecture. Two fully connected layers are applied at the end of the branch and a softmax classification head delivers a class score for each of the vehicle type classes in $\mathcal{T}$ defined in Sec. 4.2. The scores are interpreted as probabilities for the target vehicle to belong to the respective classes. The scores are gathered in a vector $\Pi_{\mathcal{T}}$, thus representing the probability distribution for the vehicle type. This distribution is incorporated as prior information into the probabilistic model for vehicle reconstruction to introduce constraints on the deformations of the ASM used as shape prior.

4.3.3 Viewpoint branch

The viewpoint branch is added to the CNN in order to derive a probability distribution Π_{ϑ} for the vehicle's viewpoint ϑ, which can be incorporated as prior information about the vehicle orientation into the probabilistic approach for model fitting proposed in this thesis.

To estimate the vehicle orientation θ directly based on the detection window would require knowledge about the location of the window within the image. As can be seen in Fig. 4.6, the vehicle passing by the camera moves along a straight line and consequently possesses a constant orientation θ. Its viewpoint however, i.e. the angle w.r.t. the image ray through the center of the image patch showing the vehicle, changes, which leads to appearance changes apparently causing the vehicle to rotate. This is the reason why the *viewpoint branch* is designed to predict a probability distribution Π_{ϑ} for the vehicle viewpoint ϑ, which describes the aspect under which the vehicle is seen. As depicted in Fig. 4.7a, the viewpoint is defined as the angle between the reference image ray to the center of the vehicle and the vehicle's longitudinal axis ${}^{A}Y$ (red arrow in Fig 4.7a). Given the direction of the image ray ρ, the vehicle orientation θ can directly be computed from the viewpoint via

$$\theta = 180^{\circ} - \vartheta - \rho. \tag{4.7}$$

In order to derive a probability distribution for the viewpoint of the vehicle, which can later be used in our probabilistic formulation for model fitting, a classification network is set up rather than a regression network. The classes correspond to discretised orientation bins for the viewpoint estimation.

The design of the *viewpoint branch* follows two assumptions regarding the general behaviour of a viewpoint classifier: First, it is reasonable to expect the classification accuracy to decrease with an increasing number and therefore a finer definition of viewpoint bins. Secondly, classification errors are expected to primarily occur for vehicles with viewpoint angles close to the bin borders, i.e. being distributed close to the diagonal of the confusion matrix. Inspired by Yan et al. (2015), to consider the first assumption, three hierarchical layers of classes are combined for classification, containing different numbers (4, 8 and 16) of viewpoint classes. Thus, we strive to profit from the higher classification accuracy for the coarse class definition and the finer level of detail of the fine class definition. To consider the second assumption, the viewpoint bins of the individual layers overlap so that finer viewpoint classes do not share borders with coarser classes. The hierarchical division of viewpoint classes can be seen in Fig. 4.7b.

However, in contrast to Yan et al. (2015), no conditional hierarchical classification is applied, in which the output of the coarse classification layers decides about the execution of individual

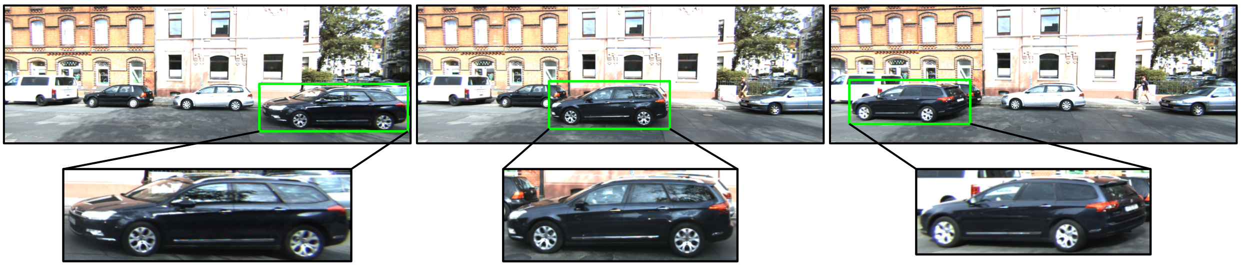
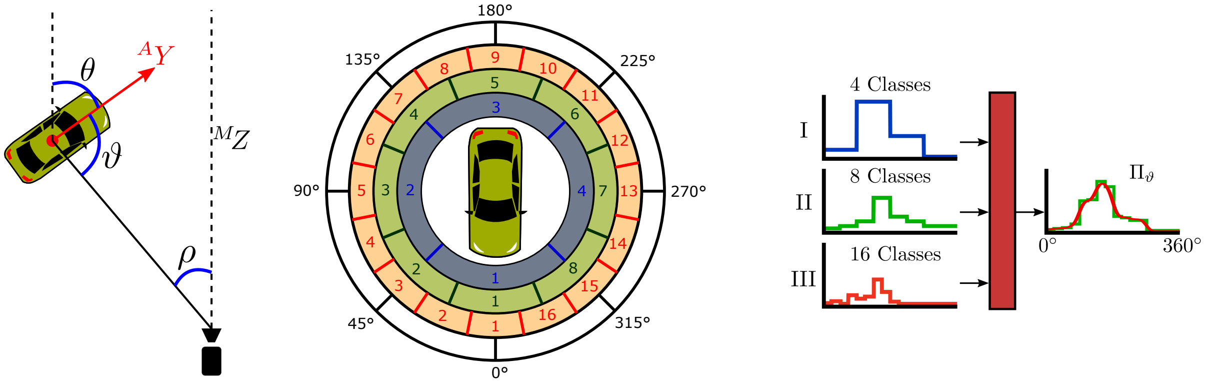


Figure 4.6: Viewpoint changes of a car passing by the camera. Top: Images of three subsequent time steps in which a car passes by the camera. Bottom: The cropped vehicle images. According to the viewpoint, the vehicle seems to rotate while the orientation in the camera model system stays constant.



(a) (b) (c)

Figure 4.7: Definition of the viewpoint angle (a) and of the hierarchical viewpoint classes (b). Details on the functionality of the *probabilistic averaging layer* are shown in (c). Further explanations are given in the main text.

classifiers for the finer layers, because in that case, the output for images that are passed on to an incorrect fine classifier cannot be corrected anymore. Instead, the *viewpoint branch* is split into three classification branches (Branch I-III), consisting of four convolutional layers followed by a max pooling layer and two fully connected layers (cf. Fig. 4.5). Each classification branch ends in a softmax classification head, one for each hierarchical class layer, and skip-connections are established from features extracted at coarser layers to the feature extraction pipeline of the next finer layers. In this way, the fine category classifiers can profit from the information extracted by the coarse classifiers but are independent from their predictions. The class-wise output scores of the classification heads are interpreted as probabilities. To derive a viewpoint probability distribution from each of the classification heads, a vector of length 720 is created, in which each entry is associated to the viewpoint discretised in 0.5° steps. Each value of the vector is set to the probability of the viewpoint class which the respective discretised viewpoint entry belongs to, according to the class definition shown in Fig. 4.7b. This results in the vectors to contain coarse-to-fine step functions representing the viewpoint probability distributions of each hierarchy layer (cf. Fig. 4.7c). These step functions are used as input to a *probabilistic averaging layer* producing the final probability distribution based on averaging the step functions, followed by a Gaussian kernel applied to the averaged function. The Gaussian kernel is used to smooth the result, reducing the effects of potential classification errors primarily expected at the bin borders. Fig. 4.7c visualises the functionality of the probabilistic averaging layer.

To summarise, the hierarchical class structure serves two purposes. On the one hand, we suppose to benefit from the more reliable output of the coarse layers while still leveraging the more detailed output of the finer layers. On the other hand, due to the overlapping viewpoint bins, the effect of misclassifications occurring between neighbouring viewpoint classes is mitigated. Compared to non-hierarchical approaches, e.g. (Tulsiani and Malik, 2015), we expect to achieve a more adequate probability distribution for the viewpoint by making use of the proposed hierarchical classifier.

4.3.4 Keypoint/Wireframe branch

This branch corresponds to a decoder network, upsampling the output of the input branch to the original input resolution. Following the vehicle keypoint and wireframe definitions made in Sec. 4.2, the goal of this branch is to predict the presence of the appearance keypoints and the wireframe edges in the image. Inspired by Newell et al. (2016) and Murthy et al. (2017a), it is trained to produce one heatmap $\mathcal{H}^c_{\mathcal{K}}$ for every *appearance* keypoint $c \in [1, C_A]$ in $\mathcal{K}_A$. Additionally, the network is adapted to also output heatmaps $\mathcal{H}^w_{\mathcal{W}}$ for the vehicle wireframe edges. More specifically, the network predicts one heatmap for each of the wireframe definitions associated to the four sides $w \in \{\text{front}, \text{back}, \text{left}, \text{right}\}$. A detailed definition of the *appearance* keypoints and wireframe edges the network is supposed to predict has been given in Sec. 4.2. The values at each pixel position of the resulting heatmaps correspond to a probability for the presence of the respective keypoint/wireframe edge at that position. Together with the input branch, the keypoint/wireframe network follows a symmetrical UNet-like (Ronneberger et al., 2015) architecture including skip connections between corresponding layers of the encoder and decoder blocks (cf. Sec. 2.1.2). Nearest neighbour upsampling is used to upsample the feature maps. The head of this branch consists of $C_A + 4$ convolution filters producing the C_A keypoint and the four wireframe heatmaps, using a sigmoid activation function (cf. Eq. 2.1) to produce pixel-wise outputs in the interval $[0, 1]$, denoting the probability for the occurrence of a keypoint or wireframe edge of a given type at the respective pixel position, respectively.

4.3.5 Training

The input branch is initialised from the corresponding layers of the VGG19 network (Simonyan and Zisserman, 2015), pre-trained on ImageNet (Russakovsky et al., 2015), and frozen during training. The remaining convolutional layers are initialised using the *He* initialiser described in (He et al., 2015). To train the proposed network, images of vehicles cropped by the tightly enclosing bounding box of the vehicles are used. A groundtruth viewpoint, keypoint annotations and the vehicle type (one out of $\mathcal{T}$) are available for each sample. The viewpoint classes needed for the viewpoint branch are derived from the groundtruth viewpoint angles according to our viewpoint class definition in Fig. 4.7b. The viewpoint branch is trained using a categorical cross-entropy loss (cf. Eq. 2.5) at each classification head given the groundtruth bin containing the groundtruth viewpoint of the training images. The categorical cross-entropy is also used to train the vehicle type branch.

For the keypoint/wireframe branch, training images are created using annotated image keypoints. Given a list of the annotated keypoint image coordinates for the keypoints $\mathcal{K}_A$, one reference image is created for every keypoint, respectively. When a keypoint is not visible, the corresponding reference image contains 0 everywhere. When a keypoint is visible, instead of setting a value of 1 at the corresponding annotated keypoint position, it is represented by a 2D Gaussian centred at the keypoint position in the reference image. This accounts for annotation uncertainties and for keypoints whose reference points on the vehicle are geometrically not precisely defined, like for instance the corner points of the roof or of the engine cover. Assuming the geometrical uncertainty of the keypoint definition on a vehicle to be r_K, the standard deviation of the 2D Gaussian s_G is computed according to

$$s_G = \frac{f_c}{\text{dist}_{veh}} \cdot r_K, \tag{4.8}$$

with the focal length f_c and the distance of the vehicle to the camera dist_{veh}. By doing so, the size of the Gaussian varies according to the distance of the vehicle from the camera and thus adapts to the size of the bounding box. This procedure leads to wider Gaussians for keypoints of vehicles which are close to the camera and narrow Gaussians for keypoints of vehicles that are further away.

One training image for each of the defined wireframes belonging to one of the sides $w \in$ {front, back, left, right}, respectively, is automatically created from the given keypoint annotations. To this end, for every keypoint tuple representing an edge of the wireframe topology $M_{\mathcal{W}}^{w}$, all pixels being crossed by the line connecting the two keypoints are set to 1. A Gaussian kernel is used to blur the reference wireframes using the standard deviation s_G of the Gaussian as explained above. Note that by this procedure, no additional manual annotation of image wireframes is required.

Training of the keypoint/wireframe branch is done by comparing the predicted heatmaps to the reference heatmaps. In (Newell et al., 2016) and (Murthy et al., 2017a), a mean squared error (MSE) loss (cf. Eq. 2.6) is used for training, in which each pixel contributes equally to the loss.

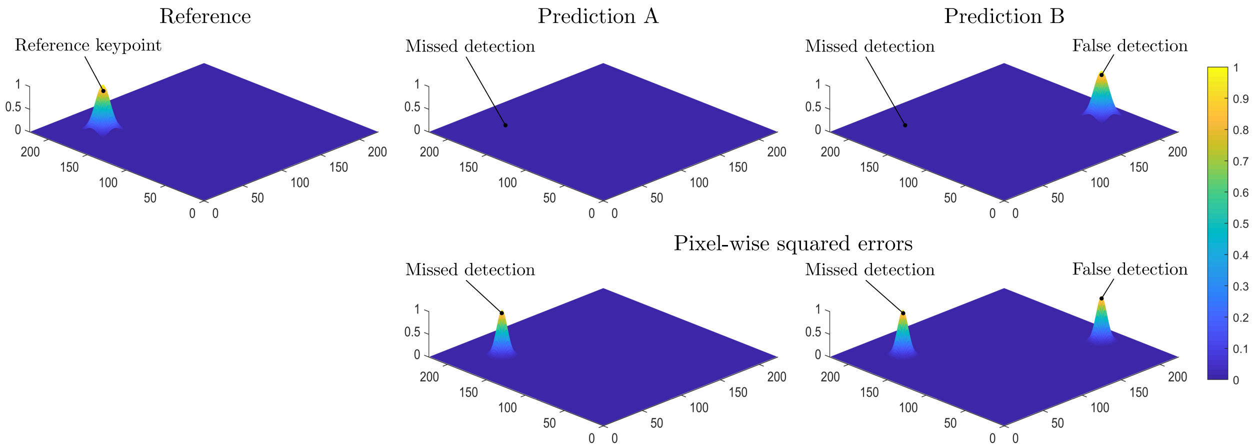


Figure 4.8: Toy example showing a reference keypoint heatmap and two potential predicted heatmaps for an image of size 224x224 in the top row. In *Prediction A*, no keypoint has been predicted at all, while in *Prediction B*, an erroneous keypoint has been predicted. The bottom row shows the pixel-wise squared error computed from comparing the two predicted heatmaps to the reference heatmap, respectively.

Table 4.1: Values of different loss metrics for the toy example shown in Fig. 4.8.

	Prediction A	Prediction B
$\mathcal{L}_{\text{MSE}}$	0.0030	0.0061
$\mathcal{L}_{\text{MSE}}^{\text{true}} + \mathcal{L}_{\text{MSE}}^{\text{false}}$	0.1818	0.1849
$\mathcal{L}_{\mathcal{KW}}$	0.1818	0.3667

In the use case of detecting keypoints and edges, the MSE leads to an unfavourably broad shape of the loss function due to the extremely small number of keypoint/wireframe pixels compared to non-keypoint/non-wireframe pixels in the groundtruth. A numerical toy example, depicted in Fig. 4.8 and Tab. 4.1, demonstrates this problem. Fig. 4.8 shows an exemplary reference keypoint heatmap as a 3D surface (left), as well as two variants of potential heatmap predictions, in both of which the reference keypoint has not been predicted correctly. While in *Prediction A*, no keypoint is detected at all, in *Prediction B* an erroneous keypoint is predicted. As can be seen from Tab. 4.1, which shows different loss metrics calculated for the described scenario, predicting a heatmap that only contains zero values already delivers a relatively small mean squared error loss, with $\mathcal{L}_{\text{MSE}}$ close to zero, due to the small proportion of non-zero pixels compared to the zero-pixels in the reference. Even predicting an erroneous keypoint (*Prediction B*), i.e. a false positive detection, does not increase $\mathcal{L}_{\text{MSE}}$ much. Due to this problem, very small gradients are obtained by the SGD using the standard MSE loss, leading to difficulties in training the network.

To overcome this problem, a new custom keypoint/wireframe loss $\mathcal{L}_{\mathcal{KW}}$ is applied with

$$\mathcal{L}_{\mathcal{KW}} = \mathcal{L}_{\text{MSE}}^{\text{true}} + \mathcal{L}_{\text{MSE}}^{\text{false}} + \mathcal{L}_{\text{MSE}}^{\text{pred}}. \tag{4.9}$$

Here, $\mathcal{L}_{\text{MSE}}^{\text{true}}$, $\mathcal{L}_{\text{MSE}}^{\text{false}}$ and $\mathcal{L}_{\text{MSE}}^{\text{pred}}$ correspond to individual MSE, each computed for a different subset of pixels. In $\mathcal{L}_{\text{MSE}}^{\text{true}}$ only pixels with groundtruth values larger than zero are considered, whereas $\mathcal{L}_{\text{MSE}}^{\text{false}}$ only considers pixels with groundtruth values equal to zero. In this way, the keypoint/wireframe and non-keypoint/non-wireframe pixels contribute to the loss with equal weight. As can be seen from Tab. 4.1, this leads to a more distinct loss compared to the standard MSE loss. However, comparing $\mathcal{L}_{\text{MSE}}^{\text{true}} + \mathcal{L}_{\text{MSE}}^{\text{false}}$ for *Predictions A* and *B*, only a small difference in the loss is visible. In other words, using the two MSE terms presented so far leads to better gradients for missed keypoint detections but still exhibits vague gradients in the case of erroneous predictions. Consequently, an additional term $\mathcal{L}_{\text{MSE}}^{\text{pred}}$ is added to the loss function. This term computes the MSE considering all pixels whose predicted probability exceeds a predefined threshold t_{pred}. Thus, this term puts emphasis on all (correct and incorrect) keypoint/wireframe detections and, therefore acts as an additional penalty of false positive detections (cf. Tab. 4.1).

The network is trained using the Adam optimizer (Kingma and Ba, 2015), a variant of stochastic mini-batch gradient descent. Learning rate decay is applied to improve training. Batch normalisation (Ioffe and Szegedy, 2015) is used and Dropout (Srivastava et al., 2014) is applied to the fully-connected layers. The definition of the required hyper parameters used in this thesis is given in Sec. 5.3.2. The network architecture and training was implemented in Keras (Chollet et al., 2015).

Data augmentation

To increase variation in the data during training, data augmentation is applied to the training data. The training images are horizontally flipped to double the amount of training data by adapting the viewpoint classes and keypoint/wireframe labels accordingly. The classes for the vehicle type remain unchanged. During training, random crops between 0 and 20% of the bounding box width are applied to the image width to simulate occlusions. We expect this to help the network to derive suitable features even if the vehicles are not fully visible. Further, a random value γ_c in the range of $[0.1, 2.0]$ is used to apply gamma correction to the training images. The gamma correction changes an image according to

$$O = I^{\gamma_c} \tag{4.10}$$

such that $\gamma_c < 1$ produces a brighter output image O than the input image I and vice versa. Applying a non-linear contrast change for data augmentation is mainly motivated by the intention to make the keypoint/wireframe branch insensitive to contrast variability in the data. The gamma correction is applied during training to enforce robustness against illumination conditions, background-foreground contrast, shadowing, etc.

4.4 Probabilistic vehicle reconstruction

Given the vehicle detections $v \in \mathcal{V}$ and the observation vector $\mathbf{o} = (\mathbf{o}^{det}, \mathbf{o}^{cnn})$ associated to each detection, a vehicle model $M(\mathbf{s})$ is fitted to each detection by finding the optimal state variables $\hat{\mathbf{s}} = (\hat{\mathbf{t}}, \hat{\theta}, \hat{\gamma})$ for position, orientation and shape. The observation vector $\mathbf{o}^{det}$ contains the 3D vehicle points $\mathbf{X}_v$ (cf. Sec. 4.1.4) and the observation vector $\mathbf{o}^{cnn}$ contains the heatmaps $\mathcal{H}_{\mathcal{K}}$ and $\mathcal{H}_{\mathcal{W}}$ for the keypoints and the wireframes (cf. Sec. 4.3), respectively. For the purpose of vehicle reconstruction, a probabilistic model is formulated that simultaneously fits the surface of the 3D ASM to the stereo reconstructed 3D points, matches model keypoints to the detected keypoints and aligns the model wireframe to the wireframes inferred by the CNN. The inferred scene knowledge in form of the free-space grid map as well as the probability distributions for orientation and the vehicle type are incorporated in the probabilistic model as state priors. Using a probabilistic formulation, the optimal state $\hat{\mathbf{s}}$ can be derived by maximising the posterior

$$p(\mathbf{s}|\mathbf{o}) \propto p(\mathbf{o}|\mathbf{s}) \cdot p(\mathbf{s}) \rightarrow \max . \tag{4.11}$$

In this work, we make the simplifying assumption that the observations $\mathbf{X}_v$, $\mathcal{H}_{\mathcal{K}}$, and $\mathcal{H}_{\mathcal{W}}$ are conditionally independent. Further, the assumption is made that the groups of state parameters, i.e. the position parameters $\mathbf{t}$, the orientation θ and the shape parameters γ, are also independent from each other. As a consequence, the likelihood $p(\mathbf{o}|\mathbf{s})$ can be factorised by individual likelihood terms, jointly incorporating both, 2D and 3D information derived from the stereo pairs and the CNN described in Sec. 4.3. The prior acts as a regularisation term on the state parameters and is factorised by one individual factor for each group of parameters, namely for position, orientation and shape, respectively. Neglecting the evidence $p(\mathbf{o})$, which does not depend on $\mathbf{s}$, the complete formulation of the posterior results in

$$p(\mathbf{s}|\mathbf{o}) \propto \underbrace{p(\mathbf{X}_v|\mathbf{s}) \cdot p(\mathcal{H}_{\mathcal{K}}|\mathbf{s}) \cdot p(\mathcal{H}_{\mathcal{W}}|\mathbf{s})}_{\text{Observation likelihood}} \cdot \underbrace{p(\mathbf{t}) \cdot p(\theta) \cdot p(\gamma)}_{\text{State prior}} . \tag{4.12}$$

This model is visualised in Fig. 4.9. The likelihood is composed of a *3D likelhihood* $p(\mathbf{X}_v|\mathbf{s})$ based on the 3D vehicle points $\mathbf{X}_v$ as well as the *keypoint* and *wireframe likelihoods* $p(\mathcal{H}_{\mathcal{K}}|\mathbf{s})$ and $p(\mathcal{H}_{\mathcal{W}}|\mathbf{s})$, which are based on the keypoint and wireframe heatmaps $\mathcal{H}_{\mathcal{K}}$ and $\mathcal{H}_{\mathcal{W}}$, respectively. The state priors for *position*, *orientation* and *shape* are derived based on the probabilistic free-space grid map Φ, the probability distribution for the viewpoint Π_{ϑ} and the prediction for the vehicle type $\Pi_{\mathcal{T}}$, respectively. An energy function E is defined, which corresponds to the negative logarithm of the posterior of Eq. 4.12 and which is minimised to find the optimal state parameters:

$$E(\mathbf{s}) = -\log p(\mathbf{X}_v|\mathbf{s}) - \log p(\mathcal{H}_{\mathcal{K}}|\mathbf{s}) - \log p(\mathcal{H}_{\mathcal{W}}|\mathbf{s}) - \log p(\mathbf{t}) - \log p(\theta) - \log p(\gamma). \tag{4.13}$$

The formulation of the individual likelihood and prior terms is explained in detail in the following sections.

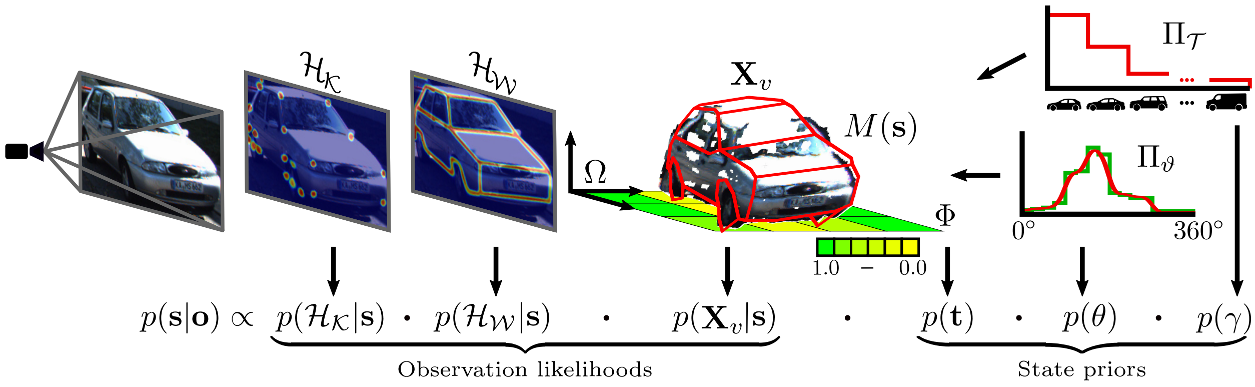

$\mathcal{H}_\mathcal{K}$
$\mathcal{H}_\mathcal{W}$
$\mathbf{X}_v$
$M(\mathbf{s})$
Ω
Φ
1.0 – 0.0
$\Pi_\mathcal{T}$
Π_ϑ
0°
360°
$p(\mathbf{s}|\mathbf{o}) \propto \underbrace{p(\mathcal{H}_\mathcal{K}|\mathbf{s}) \cdot p(\mathcal{H}_\mathcal{W}|\mathbf{s}) \cdot p(\mathbf{X}_v|\mathbf{s})}_{\text{Observation likelihoods}} \cdot \underbrace{p(\mathbf{t}) \cdot p(\theta) \cdot p(\gamma)}_{\text{State priors}}$

Figure 4.9: Visualisation of the probabilistic model. An ASM is jointly fitted to the reconstructed 3D points $\mathbf{X}_v$, the wireframe heatmap $\mathcal{H}_\mathcal{W}$ and the keypoint heatmaps $\mathcal{H}_\mathcal{K}$. The probabilistic free-space grid map Φ and the probability distributions for the viewpoint and the vehicle type Π_ϑ and $\Pi_\mathcal{T}$ act as regularisers on the state parameters $\mathbf{s}$.

4.4.1 3D likelihood

Given the 3D points $\mathbf{X}_v$, reconstructed from image points representing the surface of a vehicle (cf. Sec. 4.1.4), this likelihood aims at finding an instance of a deformed and transformed ASM representing the vehicle's surface in the best possible way. To achieve this goal, the *3D likelihood* is based on the distance of the 3D points $\mathbf{X}_v$ from the Model surface M_Δ of the model $M(\mathbf{s})$:

$$\log p(\mathbf{X}_v|\mathbf{s}) = -\frac{1}{P} \sum_{x \in \mathbf{X}_v} \frac{\mathcal{L}_\mathrm{H}(x, M_\Delta)}{2{\sigma_x}^2}. \tag{4.14}$$

Here, P is the overall number of 3D points in $\mathbf{X}_v$ and σ_x is the depth uncertainty of the individual 3D point x which is propagated from the stereo triangulation using an uncertainty for the disparity estimation σ_disp of 1 [px] using

$$\sigma_x = \frac{f_c \cdot B}{\mathrm{disp}^2} \cdot \sigma_\mathrm{disp}. \tag{4.15}$$

In this equation, f_c is the focal length, B is the stereo base-length, and disp is the disparity. To add robustness against possible outliers remaining in $\mathbf{X}_v$ we use the Huber loss (Huber, 1964) to calculate the distance $\mathcal{L}_\mathrm{H}(x, M_\Delta)$ using

$$\mathcal{L}_\mathrm{H}(x, M_\Delta) = \begin{cases} d(x, M_\Delta)^2 & \text{if } d(x, M_\Delta) \leq \sigma_x, \\ 2\sigma_x \cdot d(x, M_\Delta) - {\sigma_x}^2 & \text{otherwise.} \end{cases} \tag{4.16}$$

The Huber distance is more robust against outliers compared to the quadratic distance $d(x, M_\Delta)^2$ of a 3D point x to the model surface M_Δ. To determine $d(x, M_\Delta)$, the distance of the 3D point x to every triangle in M_Δ is calculated and the distance to the model surface is defined as the smallest distance found. This likelihood gives a measure of the correspondence of the 3D vehicle points with the surface of the vehicle model. Minimizing this term fits the 3D ASM to the 3D point cloud.

4.4.2 Keypoint likelihood

The *keypoint likelihood* is based on the intuition that, when backprojected to the image planes, the keypoints of the ASM representing the true vehicle should be supported by keypoint detections inferred from the image data at or close to the backprojected keypoint positions. Following this intuition, the keypoint heatmaps $\mathcal{H}_{\mathcal{K}}$ are used to derive this likelihood. For this purpose, the keypoints $\mathcal{K}_A$ of the model $M(\mathbf{s})$ are backprojected to the stereo images, resulting in a list of $c = [1, C_A]$ image points $\mathbf{u}_c^{l/r}$ for both, the left (l) and right (r) stereo images, respectively. Note that the CNN described in Sec. 4.3 is trained using only the keypoints being visible in the image and not being (self-)occluded. Consequently, the network is able to detect only visible keypoints, which is the reason why self-occlusion caused by the vehicle model itself is considered in the calculation of the keypoint likelihood. Ray tracing techniques can be used to reason about visible and self-occluded model keypoints. However, these techniques usually are computational expensive. Instead, the rigid topology of the ASM surface and the street-level acquisition setup allow the construction of a look-up table, storing the set of the visible model keypoints for every viewpoint $\vartheta \in [0, 360°]$. Using this look-up table, a boolean variable δ_c is associated to every image keypoint, with $\delta_c = 1$ if the keypoint c is visible and inside of the detection bounding box and $\delta_c = 0$ otherwise. The total number of visible keypoints is $U = \sum_{c=1}^{C} \delta_c$.

The keypoint likelihood is calculated by

$$\log p(\mathcal{H}_{\mathcal{K}}|\mathbf{s}) = -\frac{1}{2U} \sum_{i \in \{l,r\}} \sum_{c=1}^{C_A} \delta_c \cdot \log\left(1 - {}^{i}\mathcal{H}_{\mathcal{K}}^{c}(\mathbf{u}_c^{i})\right) \tag{4.17}$$

Here, $\mathcal{H}_{\mathcal{K}}^{c}(\mathbf{u}_c)$ denotes the output of the heatmap for the keypoint c at the location $\mathbf{u}_c$. This term thus models the likelihood of the model $M(\mathbf{s})$ based on the backprojected keypoint configuration and the keypoint probability distributions inferred by the CNN. Minimizing this likelihood fits the 3D ASM to the keypoints predicted in both images. The stochastical model of this likelihood is implicitly contained in the prediction of the keypoint probability maps since the geometric uncertainty of the 3D keypoints is considered in the process of creating the reference probability maps used for training. A detailed description was given in Sec. 4.3.5.

Related methods extract keypoint locations from the probability maps, e.g. using a greedy approach by selecting the maximum response in the heatmap as 2D keypoint location (Pavlakos et al., 2017), and perform model fitting based on the minimisation of the distance between backprojected and detected keypoints (Chabot et al., 2016; Pavlakos et al., 2017; Murthy et al., 2017b). As a consequence, these methods are sensitive to localisation errors and false detections. In contrast, the proposed likelihood is built on the raw keypoint probability maps. On the one hand, this avoids the need of precisely extracting keypoint locations from the probability maps, thus alleviating the effect of localisation errors, but allows the exploitation of the full probabilistic information instead. On the other hand, false keypoint detections and therefore outliers have less impact on the fitting procedure. This is because the presented keypoint likelihood is computed for a model hypothesis $M(\mathbf{s})$ and therefore, a falsely detected keypoint which does not coincide with its corresponding keypoint of the model hypothesis has no influence on the likelihood calculation, making the approach more robust compared to the related methods mentioned above.

4.4.3 Wireframe likelihood

The *wireframe likelihood* is based on a measure of similarity between the backprojected edges of the model wireframe $M_{\mathcal{W}}$ and the wireframe heatmaps $\mathcal{H}_{\mathcal{W}}$ inferred from the CNN. To this end, again considering self-occlusion following the same principle as for the model keypoints using a look-up table, we backproject the visible parts of the wireframe subsets $w \in \{\text{front}, \text{back}, \text{left}, \text{right}\}$ to the left and right images, resulting in binary wireframe images ${}^{l}I^{w}_{\mathcal{W}}$ and ${}^{r}I^{w}_{\mathcal{W}}$ with entries of 1 at pixels that are crossed by a wireframe edge and 0 everywhere else. To consider differences between the real image wireframe positions and the model wireframe caused by generalisation effects of the vehicle model, the wireframe images are blurred using a Gaussian filter

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right)}. \tag{4.18}$$

The generalisation error of the 3D model can be quantified by an uncertainty σ_M which is set to 10 cm in this work. We calculate a backprojection uncertainty for the model wireframe by performing error propagation on the backprojection of the model to the image. To this end, the image point $\mathbf{x}_M = [x_M, y_M]^T$ is defined as the backprojection of the centre point $\mathbf{X}_M = [X_M, Y_M, Z_M]^T$ of the model $M(\mathbf{s})$, such that $\mathbf{x}_M \sim f(\mathbf{X}_M)$. Using error propagation, the backprojection uncertainties σ_x and σ_y are computed according to

$$\sigma_{x/y} = \sqrt{\left(\frac{\partial f_{x/y}}{\partial X_M}\right)^2 \cdot \sigma_M^2 + \left(\frac{\partial f_{x/y}}{\partial Y_M}\right)^2 \cdot \sigma_M^2 + \left(\frac{\partial f_{x/y}}{\partial Z_M}\right)^2 \cdot \sigma_M^2}. \tag{4.19}$$

The applied Gaussian filter is defined according to the resulting backprojection uncertainties, leading to a stronger blurring effect when the vehicle is close to the camera and vice versa.

Given the heatmaps $\mathcal{H}_{\mathcal{W}}$ and the blurred images of the backprojected model wireframe $I_{\mathcal{W}}$, the wireframe likelihood is calculated according to

$$\log p(\mathcal{H}_{\mathcal{W}}|\mathbf{s}) = -\frac{1}{2} \sum_{i\in\{l,r\}} \sum_{w} \log\left(1 - BC({}^{i}I^{w}_{\mathcal{W}}, {}^{i}\mathcal{H}^{w}_{\mathcal{W}})\right) \tag{4.20}$$

where we use the Bhattacharyya coefficient $BC(\cdot,\cdot)$ (Bhattacharyya, 1943) as a similarity measure between the blurred wireframe images and the wireframe heatmaps:

$$BC(I_{\mathcal{W}}, \mathcal{H}_{\mathcal{W}}) = \sum_{b}^{B} \sqrt{I_{\mathcal{W}}(b) \cdot \mathcal{H}_{\mathcal{W}}(b)}. \tag{4.21}$$

In eq. 4.21, B is the overall number of pixels inside the respective detection bounding box ${}^{l,r}\mathcal{B}$ and $I_{\mathcal{W}}(b)$ and $\mathcal{H}_{\mathcal{W}}(b)$ are the values of the respective image at pixel b. It has to be noted that prior to computing the Bhattacharyya coefficient $BC(\cdot,\cdot)$, the images have to be normalised such that $\sum I_{\mathcal{W}} = \sum \mathcal{H}_{\mathcal{W}} = 1$, ensuring $0 \leq BC(\cdot,\cdot) \leq 1$. As a consequence, the negative logarithm of the wireframe likelihood term will become small if the backprojected wireframes correspond well to the wireframes predicted by the CNN.

Related methods for wireframe based model fitting extract generic edges based on gradients, establish correspondences between image and backprojected model edges, and perform model fitting by minimising the distance between the established correspondences (Leotta and Mundy, 2009; Ortiz-Cayon et al., 2016). This procedure comes with two major drawbacks. On the one hand, the generic gradient based edge extraction does not necessarily deliver the desired edges due to its sensitivity to contrast, shadows, reflectance, etc. On the other hand, establishing correct edge-to-edge correspondences highly depends on the assumption that the correct edges are extracted from the image in the first place, and also highly depends on the initialisation. The absence of semantic information behind the extracted edges further aggravates the problem of finding correct correspondences and leaves ambiguities during model fitting in the case of symmetric objects.

In this work, a CNN is trained to extract the model wireframe from the image. This allows the sole extraction of the target edges and reduces the sensitivity on contrast and the other disturbing factors mentioned above. Further, by distinguishing between the different vehicle sides during wireframe extraction, the edges carry semantically meaningful information. The semantic information allows a differentiated consideration of the distinguished wireframe edges and therefore evades potential ambiguities caused by symmetric object shapes. Last but not least, the proposed *wireframe likelihood* avoids the necessity of finding edge-to-edge correspondences. Instead, the likelihood is based in the similarity of the extracted and the backprojected wireframes of the model hypothesis $M(\mathbf{s})$, thus avoiding the error source of incorrect edge associations.

4.4.4 Position prior

The *position prior* is derived from the probabilistic free-space grid map Φ. It is calculated based on the amount of overlap between the minimum enclosing 2D bounding box $M_{\mathcal{B}}$ of the model $M(\mathbf{s})$ on the ground plane and the free-space grid map cells Φ_g with $g = [1, G]$ given their probability ρ_g of being free space (*unknown* cells are disregarded by setting $\rho_g = 0$):

$$\log p(\mathbf{t}) = \frac{\lambda_\Phi}{A_{\mathcal{B}}} \sum_{g=1}^{G} \log(1 - \rho_g) \cdot o(M_{\mathcal{B}}, \Phi_g). \tag{4.22}$$

$A_{\mathcal{B}}$ is the area of the model bounding box. The function $o(\cdot,\cdot)$ calculates the overlap between the model bounding box and a cell Φ_g. To this end, the intersecting polygon of $M_{\mathcal{B}}$ and Φ_g, represented by its n_g ordered vertices $\mathbf{v}_i = [x_i, y_i]$ with $i = [1, n_g]$, is extracted in a first step. Using the surveyor's area formula (Braden, 1986), the area of the intersecting polygon is calculated according to

$$o(M_{\mathcal{B}}, \Phi_g) = \frac{1}{2}\left|\sum_{i=1}^{n_g} x_i y_{i+1} - x_{i+1} y_i\right|, \tag{4.23}$$

where $x_{n_g+1} = x_1$ and $y_{n_g+1} = y_1$. If there is no intersection between $M_{\mathcal{B}}$ and Φ_g, $o(\cdot,\cdot)$ returns zero. As the free-space grid map is derived from the reconstructed 3D points, a factor λ_{Φ} is introduced to transfer the uncertainties of the 3D points to the calculation of the position prior. To this end, the factor $\lambda_{\Phi} = \min(1, \frac{l_\Phi}{\sigma_x})$ is used as a weight of this term based on the grid cell size l_Φ and the depth uncertainty σ_x of a reconstructed point x in the distance of the model $M(\mathbf{s})$. This prior penalises models that are partly or fully located in areas which are observed as not being occupied by 3D objects. Particularly, this prior establishes the constraint that free-space between the camera and the point cloud cannot be occupied by the model. It significantly helps to reduce ambiguities w.r.t. the position of the model and acts as substitute information where no observations are available.

While the 3D ground plane is used in related work to constrain the position of vehicles in the scene (Engelmann et al., 2016; Ansari et al., 2018), the exploration of free-space is rarely utilised as prior information. Chen et al. (2015) establish a free-space voxel grid from 3D points reconstructed from stereo images, in which a voxel is treated as occupied as soon as it contains a 3D point. They apply the voxel grid to the task of creating 3D bounding box proposals by penalising proposals that contain voxels of free-space. However, this approach is highly sensitive to noise in the 3D reconstruction and does not take into account the local density of 3D points, which can be an indicator for the certainty of occupancy. Besides, no distinction between occupied voxels that contain 3D points belonging to the ground surface or points belonging to any other object is made, which potentially leads to undesired effects of this prior. In the *position prior* proposed in this work, a probability of being free-space is computed from the ratio of ground vs. non-ground points for each grid cell rather than a binary state of either being occupied or not. By doing so, robustness against errors in the point cloud is gained on the one hand. On the other hand, the local density of the 3D point cloud is considered in the computation of the cell-wise probabilities, which allows to account for the uncertainties of the 3D points. By distinguishing between ground and non-ground points, the effect of cells being falsely classified as not free due to an occupation by a ground point is reduced.

4.4.5 Orientation prior

To calculate the *orientation prior* for the model $M(\mathbf{s})$, the probability distribution Π_ϑ for the vehicle viewpoint inferred by the multi-task CNN is used. The viewpoint $\vartheta_M(\theta)$ is computed for the model $M(\mathbf{s})$ from the model orientation θ using the relation in Eq. 4.7. The image ray direction

ρ is derived from the ray connecting the camera projection center and the centre of the vehicle model $M(\mathbf{s})$. The *orientation prior* is calculated according to

$$\log p(\theta) = \log \Pi_{\vartheta}\left(\vartheta_M(\theta)\right) + \log\left(\frac{1+\cos\left(\vartheta_{\mathrm{CNN}} - \vartheta_M(\theta)\right)}{2}\right). \tag{4.24}$$

$\Pi_{\vartheta}\left(\vartheta_M(\theta)\right)$ denotes the probability for the angle ϑ_M according to the output of the viewpoint classification branch of the CNN. As incorrect viewpoint classifications can be assumed to appear especially between neighbouring viewpoints, this term alone is prone to cause small orientation biases. This is why additionally, the cosine distance of the most likely viewpoint ϑ_{CNN} predicted by the vehicle CNN and the model viewpoint $\vartheta_M(\theta)$ is considered in this prior.

A potential symmetry of the object's shape is one example for a reason for ambiguities in the presented likelihood terms. For instance, experiments in Sec. 6.3 indicate that the *3D likelihood* exhibits ambiguities between the correct and the opposed viewing direction of the vehicles, caused by the symmetric shape of vehicles w.r.t. their lateral axes. Incorporating the prediction of the viewpoint in the *orientation prior* helps to constrain the model orientation and to resolve potential ambiguities present in the likelihood terms.

4.4.6 Shape prior

The *shape prior* formulation is based on the probability distribution $\Pi_{\mathcal{T}}$ for the vehicle types predicted by the CNN and acts as regularisation term for the shape of the model $M(\mathbf{s})$. Using

$$\log p(\gamma) = -\frac{1}{n_s}\sum_{\tau\in\mathcal{T}}\sum_{s=1}^{n_s}\Pi_{\mathcal{T}}^{\tau}\cdot\frac{(\gamma_s^{\tau}-\gamma_s)^2}{2\sigma_s^2}, \tag{4.25}$$

the *shape prior* term penalises deviations of the n_s considered model shape parameters γ_s from the predicted modes $\gamma_s^{\mathcal{T}}$ of the individual vehicle types. In the prior formulation, the deviations from each mode are weighted according to the probability scores $\Pi_{\mathcal{T}}$ predicted by the CNN with $\sum \Pi_{\mathcal{T}}^{\tau} = 1$. The parameter σ_s represents the ASM standard deviation of deformation, i.e. the square root of the eigenvalue, in the direction of the eigenvector associated to the s^{th} shape parameter as described in Sec. 2.2.

As it is reasonable to assume the ASM shape parameters of vehicles belonging to different categories not to follow a single normal distribution but rather a multi-modal distribution with each mode representing one vehicle category, penalising deviations from the overall mean shape, as is usually done in the literature (Zia et al., 2013; Engelmann et al., 2016), is an unfavourable procedure. The category aware *shape prior* formulation proposed here gives a more realistic and detailed constraint on the shape parameters. The confidence awareness in the prior term, achieved by considering the probability scores $\Pi_{\mathcal{T}}$ for all distinguished categories, reduces the sensitivity of the *shape prior* to potential uncertainties in the vehicle type predictions of the CNN.

4.4.7 Inference

To find the optimal pose and shape parameters for each detected vehicle, the energy function of Eq. 4.13 representing the probabilistic model is minimised. As this function is non-convex and discontinuous due to the changing visibility of keypoints/wireframes caused by self-occlusion, which leads to jumps in the objective function, a sequential Monte Carlo sampling procedure is applied as described in Sec. 2.3 to approximate the parameter set for which the energy function becomes minimal. To this end, the target parameters are sampled to generate model particles for the vehicle ASM. Starting from one or more initial parameter sets, a number of particles n_p are generated in each iteration $j = 1, ...n_{it}$ by jointly sampling the pose and shape parameters from a uniform distribution centred at the preceding parameter values. For the resampling step, the importance weight w_j^i that equals to the energy in Eq. 4.13 for every particle is calculated in each iteration j and the n_b highest weighted particles are introduced as initial seed particles for the next iteration. In each iteration, the size of the interval from which the parameters are sampled is reduced. In the following paragraphs, more details are given.

Initialisation: As described in Sec. 2.3, a common way for initialisation is to draw the initial particles $\mathbf{s}_0^i$ with $i = 1, ..., n_p$ from the prior distribution $p(\mathbf{s})$. With $p(\mathbf{s}) = p(\mathbf{t}) \cdot p(\theta) \cdot p(\gamma)$ the prior distribution heavily depends on the predictions of the CNN for the orientation and the shape. To account for potentially incorrect predictions, the initialisation procedure described in Sec. 2.3 is slightly adapted in this work. The most likely particle $\mathbf{s}_0^1 = (\mathbf{t}_0^1, \theta_0^1, \gamma_0^1)$ is determined from $p(\mathbf{s})$ by setting θ_0^1 to the most likely orientation given Π_ϑ and γ_0^1 to the most likely shape given $\Pi_\mathcal{T}$. The initial translation vector $\mathbf{t}_0^1$ is defined as the bounding box centre of the minimum 2D bounding box enclosing the 2D projections of the 3D vehicle points $\mathbf{X}_v$ to the ground plane. Instead of sampling all particles $\mathbf{s}_0^i$ from $p(\mathbf{s})$, the particles are sampled using an uniform distribution for position, orientation and shape, respectively, centred at the most likely particle $\mathbf{s}_0^1$. The initial interval boundaries are set to $\pm 180°$ and ± 1.5 [m] for the orientation and position parameters, respectively, and to ± 3 for the shape parameters. By choosing $\pm 180°$ as interval for the orientation angle, particles are allowed to take the whole range of possible orientations in the first iteration to be able to deal with incorrect initialisations, thus gaining robustness against initialisation errors.

Resampling: To resample the particles in each iteration $j = 1, ..., n_{it}$ the importance weights w_j^i are calculated for each preceding particle according to the energy function in Eq. 4.13. The best n_b particles, i.e. particles with the highest importance weights, are introduced as seed particles for the resampling step. In this work, $n_b = 10$ is chosen as number of seed particles. For each seed particle, an equal number of $N_j = \frac{n_p}{n_b}$ of offspring particles $\mathbf{s}_j$ are drawn from the importance distribution $\pi(\mathbf{s}_j | \mathbf{s}_{j-1}, N_{j-1})$. In this work, the importance distribution is defined as a uniform distribution centred at the respective seed particle. To encourage convergence of the applied MCM, the range for the respective parameters used to draw the particles from the uniform distributions is reduced by a factor 0.85^j in iteration j. As a consequence, the initial parameter range is reduced by the factor $0.85^{n_{it}}$ in the last iteration $j = n_{it}$. By forwarding multiple particles, the preservation of particles potentially belonging to different minima is enabled. As a consequence, the risk of getting stuck in a local minimum is reduced, which allows to deal with multi-modal distributions and local minima in the objective function.

Final result: The final values for the target parameters of pose and shape are defined after the last iteration and are set to the parameters of the particle achieving the lowest energy within the particle set of the final iteration.

In contrast to first order local optimisation approaches, e.g. used in (Engelmann et al., 2016; Pavlakos et al., 2017), whose convergence is highly sensitive to the initialisation and which consequently require good initialisation values for the target parameters, the Monte-Carlo optimisation procedure proposed in this work is insensitive to initialisation errors. While only a fairly coarse initialisation of the position is required, de facto no initial values for the orientation are needed as the initial range for drawing the particles is set to ±180° in the first iteration. However, as can be seen in the experiments (cf. Sec. 6.3), the initial orientation derived from the viewpoint prediction of the CNN is relatively accurate. If the initial orientations are expected to be very accurate, the robustness of the proposed optimisation method against initialisation errors can be traded for a potential speed up in convergence, by putting more trust into the initialisation and consequently reducing the orientation interval from which the particles are drawn. Moreover, the optimisation procedure presented here can be applied to arbitrary objective functions, without requirements concerning linearity, convexity or continuity. It also facilitates convergence and is able to cope with local minima.

4.5 Discussion

Based on the description of the methodology developed for vehicle reconstruction, this section discusses assumptions, advantages and limitations of the individual components of the proposed approach.

Setup and heuristics

The proposed method builds on stereo images as input, which allow the incorporation of valuable 3D constraints during object reconstruction, in contrast to approaches that use single images and thus, suffer from ambiguities left by the projection from 3D to 2D. However, the usage of stereoscopic images also adds requirements concerning data acquisition, causing the setup to be less flexible compared to a single camera setup. In the automotive industry as well as in academia, monocular acquisition setups and data sets are often presented, though frequently combined with one or more laserscanners. In view of this background, it should be noted that the presented approach conceptually can also be applied to single images by incorporating a monocular depth module (Godard et al., 2017) or by combining it with range observations of a laserscanner.

A strong constraint for a subset of the target pose parameters is derived from estimating the local ground plane (cf. Sec. 4.1.3) and by subsequently enforcing vehicles to be localised on that plane. This constraint follows the assumptions that the ground can be represented by a planar surface on the one hand, and that the ground plane can be accurately determined from the data on the other hand. While it is justifiable to suppose that these assumptions hold true in most real world street scenarios, the presented approach leaves opportunities to relax these assumptions and to become more generally valid. To overcome violations of this assumption, an additional height parameter can be added to the set of target parameters to allow distances to the ground plane larger than zero. On the other hand, Ansari et al. (2018) propose to represent the ground by locally planar patches instead of one common plane. Alternatively, more complex representations can be used, e.g. by representing the ground by polynomial surfaces.

Deformable shape prior

As shape regularisation on the ill-posed problem of 3D vehicle reconstruction, a new subcategory-aware shape prior is proposed in this thesis. As a basis, we make use of an Active Shape Model as e.g. used in (Zia et al., 2013) to learn a deformable representation for vehicles. However, significant extensions to the standard representation of the ASM are proposed in this thesis.

Most importantly, instead of representing the class vehicle by a uni-modal shape distribution obtained by the PCA underlying the ASM, individual shape priors are learned for each of several different vehicle subcategories. In prior work, the ASM is exclusively used to constrain the shape parameters by penalising shape deviations from the mean model calculated from the training exemplars, e.g. (Zia et al., 2013). As a consequence of this procedure, vehicle types whose shape differs from the mean model, or types which are under-represented in the training set, are systematically disadvantaged. In contrast, this thesis proposes category-aware prior which considers the multi-modal distribution of vehicle shapes. Together with a prediction of the subcategory, the multi-modal shape representation allows for more detailed and more realistic constraints on the vehicle shape during the reconstruction. While the prediction of the vehicle type has been done in other work, as e.g. in (Tang et al., 2019), where it is used for the task of vehicle re-identification, it has not been incorporated in order to derive constraints on the shape parameters yet. The extension of the shape prior by the category-awareness adds a negligible manual labelling effort regarding the learning of the ASM, since only the information of the reference vehicle type is additionally required. However, a classifier has to be learned for the prediction of the vehicle type, which adds additional effort and the requirement of additional training data to the approach.

Moreover, this thesis proposes a richer representation of the ASM by defining a triangulated surface as well as a wireframe representation in addition to the standard keypoint representation (Zia et al., 2013), enriching the ASM by topological and semantic information. The richer representation of the shape model allows the joint consideration of different observation types such as keypoints, wireframe edges, and 3D points, in the reconstruction process. This is another essential contribution of this thesis, as it enables exploiting the synergistic effects anticipated from the complementary and/or redundant nature of the different observations, while in related work the reconstruction is nearly always based on a single observation type only.

Obviously, one limitation of using an ASM as shape prior is that a single ASM representation is not scalable to other object classes due to the requirement for consistent semantic keypoints in all object instances. While the ASM employed in this thesis covers a broad range of passenger cars, other object classes, as e.g. motorbikes or pedestrians, are excluded. To consider different object classes in the proposed reconstruction approach, dedicated object detectors and an individual ASM would have to be incorporated for each individual class.

Multi-task CNN

The new CNN presented in this thesis is trained to predict the vehicle type, the viewpoint, and heatmaps for keypoints and wireframes from an input image showing a single vehicle.

While other approaches also use CNNs to derive heatmaps for specific object keypoints, a new loss is proposed that improves the training for the task of keypoint prediction due to its creation of larger gradients for missing or false positive detections compared to the commonly used MSE loss. Mostly, prior work makes use of the keypoint heatmaps to localise the keypoints, often using a greedy approach by simply selecting the maximum response in the heatmap as keypoint location, e.g. (Pavlakos et al., 2017; Murthy et al., 2017b). This procedure is prone to localisation errors, cannot deal with multiple peaks in the keypoint distributions and neglects the valuable probabilistic information contained in the heatmaps. In contrast, the reconstruction approach proposed in this thesis refrains from localising the keypoints, thus reducing the error source of imprecise localisations, but builds upon the raw heatmaps instead, thus leveraging the full available probabilistic information. By doing this, we follow the procedure proposed in (Zia et al., 2013), but instead of using a RF classifier as proposed by the authors, resulting in very noisy keypoint heatmap predictions, the more promising performance of a CNN is exploited in this thesis.

A key component of the CNN and a central contribution of this thesis is the prediction of a new feature for model fitting, namely the prediction of *vehicle wireframe edges*. In contrast to existing work, which relies on generic image edges extracted by gradient based approaches (Ortiz-Cayon et al., 2016), the supervised prediction of the wireframe has no explicit dependency on proper gradient magnitudes, thus being less sensitive to contrast. Furthermore, it allows the direct prediction of target edges, thus being less sensitive to undesired gradients caused by e.g. reflections, shadows, etc. Moreover, learning to distinguish between wireframes belonging to different parts of the vehicle adds semantic information to the predicted edges, which allows to establish adequate associations between the image and model wireframes.

In real world scenarios, occasionally vehicles are partly occluded by other vehicles, which can lead to parts of two different vehicles being visible in the extracted bounding box crop. While this behaviour leads to ambiguities in predicting the viewpoint and vehicle type, it can be observed that keypoints and wireframes of both vehicles are predicted by the CNN and are consequently contained in the heatmaps. Both effects consequently lead to problems during the reconstruction step. A potential solution to reduce this problem, which has not been investigated in the scope of this thesis, can be to make use of the obtained segmentation masks to segment out the vehicle of interest and to find a proper way to remove the disturbing parts of the second vehicle (e.g. by blackening the respective image parts or by filling them with noise).

The proposed CNN is built on top of an object detector and thus, vehicle detection and the extraction of the proposed observations and a priori information by the CNN are performed as separate tasks. When applying a CNN also for the detection, a computationally more efficient opportunity would be do combine the detection and extraction of other variables in a common architecture where intermediate feature maps can be shared for both tasks, as e.g. done in (Chabot et al., 2017; Kundu et al., 2018). In particular, one possibility would be to incorporate the CNN presented in this thesis into the Mask RCNN (He et al., 2017) applied for the detection task. However, it can be argued that object detection is a highly vivid field of research, leading to the release of improved models with better performance and/or efficiency every year. The modular setup of this thesis retains the opportunity to replace the detection approach by a better one without the need of adapting and retraining the proposed multi-task CNN.

Although the architecture of the proposed CNN contains a shared backbone feature extractor (cf. Fig. 4.5) for the different tasks, pre-trained weights are used for feature extraction and the individual branches are trained individually on top of the feature extractor (cf. Sec. 4.3.5) due to lower requirements w.r.t. the amount of training data. Therefore, multi-task learning is only performed within the individual branches, e.g. for the joint prediction of the keypoint and wireframe heatmaps, or for the joint training of the different viewpoint classification heads. A joint training of the feature extractor using a multi-task loss could potentially enhance the performance of the CNN.

Probabilistic model

The factorisation of the likelihood and the prior used to formulate the probabilistic model which is proposed for vehicle reconstruction in this thesis is based on the following simplifications and assumptions. On the one hand, the observations incorporated in the likelihood, i.e. the 3D point cloud and the keypoint and wireframe heatmaps, are treated as conditionally independent observations, although they are derived from the same image data and thus, contain statistical dependencies. However, this simplification makes modelling of the likelihoods feasible, which would otherwise be very complex and impracticable. On the other hand, independence is also assumed for the state parameters, namely position, orientation and shape, which allows a factorisation of the state prior. Making the assumptions listed above, a comprehensive probabilistic model for vehicle reconstruction is formulated, in which information from both, the 2D image domain and 3D object space, is incorporated. Often, when different terms, e.g. corresponding to individual potentials, are incorporated into one energy formulation, weight factors are introduced for each individual potential to account for and control their different importance and impact on the overall objective, e.g. (Chen et al., 2015). Finding the optimal weights is usually done empirically or by learning them from data. However, the optimal weight values can vary for different data sets or even for different instances within the same data set, leading to a low generalisation capability of the formulation. The probabilistic model presented in this thesis does not require the definition of weights for the individual likelihood and prior terms. A stochastic model based on the uncertainties which are considered in the likelihoods is used to assign weights to the observations instead. The ability of applying error propagation to every piece of information derived from the data (e.g. the reconstructions of 3D points from the disparity estimates) enables a dynamic adaptation of the probabilistic model to the characteristics of the data and individual observations. Consequently, the model can also be readily generalised to different data sets, as it adapts to different sensor characteristics and observation uncertainties.

The probabilistic model is formulated to reconstruct a single detected vehicle based on a prior shape model that is fitted to the observations. Besides a common ground plane that is introduced to drastically reduce the search space for possible vehicle locations, no further global geometric constraints or context knowledge is incorporated to supervise the vehicle reconstruction. For instance, the presented approach does not prohibit the resulting vehicle reconstructions to erroneously overlap. Also, contextual geometric relations between vehicles are not explored in the probabilistic model. Such relations could be used to prevent vehicles from overlapping, to reason about mutual occlusions (Zia et al., 2015; Xiang et al., 2015) or to incorporate prior relational knowledge, e.g. that neighbouring vehicles are likely to be parallel to each other due to their typical arrangement in streets and parking lots.

Furthermore, being built on individual image pairs, the approach does not incorporate temporal constraints into the vehicle reconstruction. Considering the reconstruction of a vehicle from two temporally consecutive observations, differences in vehicle pose are restricted due to the set of possible movements the vehicle can perform within the time between the observations. A probabilistic motion model is a potential way to incorporate temporal constraints into vehicle reconstruction

(Engelmann et al., 2017). Besides, constraints on the vehicle shape can be established over time, since a vehicle visible in temporally different images has the same shape. However, this thesis targets an approach for the reconstruction of single vehicle detections based on individual image pairs. The consideration of global geometric and temporal consistency is considered as a potential extension of the presented approach in the future.

Inference

For optimisation of the established probabilistic model, an iterative Monte-Carlo sampling technique is applied. The key advantage of this technique is its applicability to arbitrary functions without requirements w.r.t. convexity or continuity and, thus, is suitable for the optimisation of the non-convex and potentially discontinuous objective function presented in this work. Furthermore, the setup of the Monte-Carlo approach developed in this work is insensitive to initialisation errors and local minima. However, particle based approaches can become computationally expensive and therefore time consuming, depending on the required amount of particles, as the objective function has to be evaluated for each particle. As far as real time capability is concerned it can be argued that each particle can be processed independently from the other particles, which allows parallelisation.

5 Experimental setup

In the preceding chapter, we proposed a probabilistic approach for the reconstruction of vehicles from stereo images. The goal of the experimental evaluation is to investigate the performance of the presented method on real world data and in realistic scenarios and to investigate its strengths and limitations. This chapter describes the experimental setup that is established for the evaluation. Sec. 5.1 starts with a description of the objectives and research questions leading the evaluation procedure. Sec. 5.2 introduces the data sets used for evaluation and Sec. 5.3 describes the parameter setting and training strategy employed for the investigations. The test setup defining the evaluation strategy and the evaluation criteria which serve as basis for the quantitative analysis are described in Sec. 5.4.

5.1 Objectives

The goal of this thesis is the precise 3D reconstruction of vehicles detected from stereo images. For this purpose, a model based approach is proposed which is based on an extensive probabilistic energy function and which strongly utilises observations and prior information inferred by a CNN. As a result of the vehicle reconstruction, 3D vehicle models are obtained which are parametrised by their 3D pose and shape. The information obtained by the CNN and the quality of the reconstructions can be quantitatively assessed based on available reference data. The strategy which is pursued for the evaluation is presented in Sec. 5.4. In order to evaluate if the initial research objectives introduced in Chapter 1 were achieved, the following questions are addressed in the evaluation to guide the empirical analysis:

> *(1) How good is the performance of the individual branches of the developed CNN? How well-suited are the probability distributions predicted by the CNN as observations and prior on the state parameters?*

To assess the suitability of the proposed CNN for the derivation of state priors, the CNN components which are employed for the derivation of the probability distributions for the vehicle viewpoint and the vehicle category are evaluated individually. The quality of the raw output of the individual CNN branches is analysed based on corresponding reference data for the vehicle viewpoint and the vehicle categories. Furthermore, the suitability of the predictions from the CNN for the task of vehicle reconstruction can be assessed by evaluating the effect of the individual likelihood and prior terms which consider the predicted information on the results of the reconstruction.

> *(2) Which contribution and potential benefit do the individual terms of the probabilistic model have on the quality of the results? How large is the synergistic potential achieved by the combination of the different observation likelihoods? Which effect on the results is obtained by incorporating the state priors to constrain the parameters?*

To assess the contribution of the individual constituents of the probabilistic model and to investigate their synergistic potential, different variants for the probabilistic model are defined and tested, each of them considering and neglecting a different set of likelihood and prior terms. The obtained results can be compared between the defined variants to investigate the effect of each term on the quality of the vehicle reconstructions.

> *(3) How good is the quality of the results that can be obtained by the presented approach? How sensitive is the approach w.r.t. the initialisation? What are the limitations and potential weaknesses of the proposed method? How good does the approach perform on different data sets? How does the performance compare to related state-of-the-art approaches?*

A detailed analysis of the results obtained by the proposed method on different data sets is conducted to expose its limitations and weaknesses. The generalisation ability of the method is validated by analysing its sensitivity to different data domains. Furthermore, the performance of the proposed method with respect to related state-of-the-art approaches is ascertained.

5.2 Test data

The empirical evaluation of the proposed method is conducted on two different real world data sets. Both data sets consist of stereo image pairs which were acquired by a synchronised and calibrated stereo camera rig, placed on a mobile platform, while driving in regular traffic on public roads in urban environments. One data set is taken from the publicly available KITTI benchmark suite (Geiger et al., 2012) and will therefore be referred to as *KITTI* data throughout this chapter. The second data set was created in the scope of this thesis and is based on the data presented in (Schön et al., 2018). This data set will be referred to as *ICSENS* data in the remainder of this thesis. Tab. 5.1 contains characteristics of the acquisition setups and of the image data, comparing both data sets. A detailed description of the data and the reference information is given in the following sections.

Table 5.1: Comparison of properties of the KITTI and the ICSENS data set. Characteristics of the acquisition setup, the images and the annotations are shown.

	KITTI	ICSENS
Acquisition setup		
Cameras	Point Grey Flea 2 (FL2-14S3C-C)	Allied Vision (AV MAKO G-234C)
Lenses	Edmund Optics (NT59-917)	Schneider Kreuznach (Cinegon 1.8/4.8-0902)
Focal length	4.0-8.0 mm (varifocal)	5.0 mm
Base length B	0.54 m	0.85 m
Camera height	1.64 m	1.93 m
Image properties (after rectification)		
Focal length f_c	721.5 [px]	793.6 [px]
Cropped image size [px]	1242x375	1934x860
Horizontal opening angle	80°	100°
Vertical opening angle	30°	55°
Annotation		
References	Oriented 3D bounding boxes	Fitted CAD models

5.2.1 KITTI benchmark

Regarding the KITTI data, the 3D object detection benchmark is used for the evaluation in this thesis. Reference labels are available for the training data set of that benchmark, which is therefore used to conduct experiments based on our own evaluation metrics, delivering more insights into properties and limitations of the proposed method compared to the official evaluation metrics of the KITTI test suite. Besides, the official KITTI evaluation metrics are designed to assess the joint performance of both, detection and pose estimation, while the focus of this work lies on the precise pose and shape reconstruction of vehicles, not on the detection. The benchmark consists of a training set with 7481 unordered stereo images, i.e. non-sequential images with manually annotated labels for objects like cars, pedestrians and bicycles. In this thesis, 260 of the training set images are used for training (cf. Sec. 5.3.2) and the remaining images are used for evaluating the proposed approach based on the dedicated metrics. The KITTI test set consists of 7518 stereo images for which no reference is provided. An official evaluation on the test set using the official KITTI metrics is used to compare the performance of the proposed method to the results of related methods which are reported in the KITTI leaderboard[1].

[1] http://www.cvlibs.net/datasets/kitti/eval_3dobject.php

In our own evaluation using the training set, all objects labelled as *car* are considered. For every object, the benchmark provides 2D image bounding boxes and oriented 3D object bounding boxes in the camera model system from which the 3D object location as well as the rotation angle about the vertical axis can be derived. The 3D bounding boxes were manually labelled by using the point clouds generated by a laserscanner which was part of the acquisition setup. The 2D image bounding boxes are derived from the backprojections of the 3D bounding boxes to the images and thus represent the entire object, independently from occlusions, instead of containing only the visible object parts. However, information about the level of object truncation and object occlusion is available. The values for truncation refer to the objects leaving image boundaries and are given as float values from 0 (non-truncated) to 1 (truncated) while the occlusion state indicates the vehicle occlusion due to other objects with 0 = fully visible, 1 = partly occluded, 2 = largely occluded, and 3 = unknown. According to the categorisation of the vehicles depending on their state of occlusion and truncation, three levels of difficulties (*easy, moderate*, and *hard*) are defined as shown in Tab. 5.2.

Table 5.2: Levels of difficulties considered in the KITTI data as evaluation criteria.

	easy	**moderate**	**hard**
max. occlusion level	0	1	2
max. truncation	0.15	0.30	0.50

5.2.2 ICSENS data set

Similar to the KITTI setup, a vehicle was equipped with a synchronised and calibrated stereo camera rig and stereoscopic image sequences were recorded in urban environments for the generation of the ICSENS data set (Schön et al., 2018). The acquisition took place during day time and dusk, leading to various and challenging lighting conditions in the images, contrary to the KITTI data, which exhibits almost always ideal lighting conditions. An overall number of 2289 vehicles were manually labelled in a subset of 1000 stereo images which were randomly extracted from the recorded sequences. All of the labelled stereo images are used for evaluation in this thesis. In contrast to the KITTI data set, which only delivers 2D and oriented 3D bounding boxes as references, a reference shape and the reference vehicle type are available for every labelled vehicle in addition to the 3D pose. A visual comparison of the provided references for the KITTI and the ICSENS data can be seen in Fig. 5.1.

In order to generate the reference data for the ICSENS data set, a set of 36 real-world dimensioned CAD models were collected via *Google's 3D Warehouse*[2]. Each CAD model was associated to one

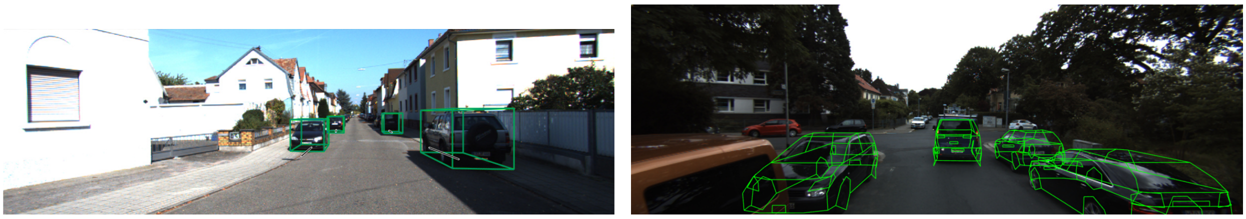
[2] https://3dwarehouse.sketchup.com

of the vehicle types in $\mathcal{T} = \{$*Compact Car, Sedan, SUV, Estate Car, Sports Car, Truck, Van*$\}$. When the models were collected, we made sure that the selected models cover all of the considered categories. To generate a reference label for a vehicle that is visible in the image, human annotators were asked to perform a two-step procedure.

In a first step, the CAD model which is most similar to the target vehicle in terms of its shape and category has to be manually selected from the set of reference CAD models. By implication, this selection defines the reference shape and category of the vehicle. It has to be noted that the categories of vehicle types do not have a clear definition and, therefore, the class boundaries are somewhat vague. The car body configuration, which is determined by the layout of the engine, passenger and luggage volumes, as well as the amount of pillars of a vehicle, which are the (near) vertical supports of a car's roof and its windows, are characteristics that can be used to distinguish different vehicle types. However, ambiguities in the distinction of vehicle types exist and are therefore likely to be contained in the reference labels for the types.

In a second step, the chosen CAD model is manually fitted to the vehicle visible in the image to generate the reference parameters for its pose. To this end, a 3D point cloud is generated from the disparity map which is derived from the stereo images using the method for dense stereo-matching described in (Geiger et al., 2011). For manual model fitting, a workflow based on *CloudCompare*[3], a software for 3D processing, has been developed. *CloudCompare* is used to visualise the stereo point cloud in 3D and to import the CAD model that has been chosen as reference model for the respective vehicle. Then, the CAD model is manually shifted and rotated to the desired pose. The correctness of the reference pose is checked by the annotator by alternately inspecting the quality of the fit of the CAD model to the 3D point cloud and the quality of the fit between the wireframes of the CAD model backprojected to the image and the actual appearance of the vehicle. An example of a fitting result is shown in Fig. 5.1b, where the backprojected wireframes of the selected and fitted CAD models are depicted on top of the left stereo image. The parameters of the final pose of the translated CAD model, i.e. its position in the model coordinate system as well as its orientation, are stored as reference pose parameters. It has to be noted that due to the restricted set of available CAD models and the discrepancy between the chosen model and the actual vehicle visible in the image, slight differences between the backprojected model shapes and the acquired vehicles are possible to occur (cf. Fig. 5.1). The minimum bounding box enclosing the backprojected wireframe in the image is defined as reference 2D image bounding box for each vehicle. Consequently, similar to the KITTI benchmark, the 2D reference bounding boxes comprise potentially occluded and therefore non-visible parts of the vehicles. Furthermore, in the reference data, two levels of difficulty are differentiated, and the difficulty level is stored as an additional piece of information for each reference vehicle. We differentiate between *easy* vehicles, which are fully visible in the images, and *difficult* vehicles, which are occluded or truncated, as additional

[3] http://www.cloudcompare.org



(a) KITTI

(b) ICSENS

Figure 5.1: Examples from the employed data sets. In (a), an example from the KITTI object detection benchmark is shown, which provides 3D bounding box references. In (b), an example of the ICSENS data set is shown, which provides fitted 3D CAD models as reference data as well as a reference vehicle type.

information for each vehicle. It has to be noted, that in contrast to the KITTI data set, where 3D laserscanner points were used to generate the reference data, the stereo point cloud generated from the estimated disparity map is used for the manual model fitting. Potential errors of the disparity estimations might therefore also be reflected in the references for the vehicle pose.

The images from both, the KITTI as well as the ICSENS data set, contain a subset of visible vehicles without reference labels. Mostly, those are vehicles that are heavily occluded or that are far away from the camera. For instance, in the KITTI data set, vehicles that have a bounding box height smaller than 40 [px] are not considered in the evaluation. Assuming an average height of about 1.5 m for passenger cars (cf. Sec. 5.3.1), a bounding box size of 40 [px] corresponds to a distance of about 25 m for the KITTI setup. Missing labels, however, prohibit the counting of false positive detections and therefore the computation of precision scores. The KITTI benchmark therefore provides *don't care* areas, in which potential detections are not counted as false positives. Introducing such areas in the ICSENS data set is a task to be accomplished in the future. As a consequence, the ICSENS data is not suited for a quantitative evaluation of the vehicle detection.

Images from both, the KITTI as well as the ICSENS data set, are provided after stereo rectification, i.e. as images corresponding to the normal stereo case. The cameras of both stereo setups were mounted with an approximately horizontal viewing direction, i.e. a viewing direction approximately parallel to the ground. Since the bottom part of the recorded images contains the engine hood of the acquisition vehicle, the KITTI and ICSENS images are cropped in order to remove the visible vehicle parts. Tab. 5.1 shows a comparison of the properties of the two data sets. As can be seen, the base length used for the acquisition of the ICSENS data (0.85 m) is significantly larger compared to the base length (0.54 m) used for the KITTI data set. Tab. 5.3 shows the geometric uncertainty of both data sets in form of the depth uncertainty σ_x of a stereo triangulated 3D point in different distances from the camera (5, 10, 15, 20, and 25 m), calculated according to Eq. 4.15. For the calculation, a standard deviation σ_{disp} of 1 [px] is assumed for the disparity estimation. As can be seen from Tab. 5.3, the larger base length of the ICSENS data set leads to better geometric properties represented by the distinctly smaller depth uncertainties after stereo triangulation. In particular the proposed *3D likelihood*, which makes use of the triangulated point clouds, but also the *keypoint* and *wireframe likelihoods*, in which both, the left and right stereo partner, are considered, are affected by these properties.

Table 5.3: Depth uncertainties σ_x [m] of a 3D point at different distances d calculated by applying error propagation to the stereo triangulation (cf. Eq. 4.15) using a standard deviation for the disparity of 1 [px].

d [m]	5 m	10 m	15 m	20 m	25 m
KITTI	0.06	0.26	0.58	1.03	1.61
ICSENS	0.04	0.15	0.33	0.59	0.93

5.3 Parameter settings and training

As described in Sec. 4, the proposed method requires a set of parameters which have to be defined by the user based on heuristics or empirical evidence. All free parameters of the proposed approach and their chosen values are presented in Tab. 5.4. The parameters for the free-space grid map and for inference, i.e. side length of the grid cells as well as the number of particles, iterations, and offspring were found empirically in this work. The parameters which have to be selected for the ASM and and the CNN are discussed in Sec. 5.3.1 and 5.3.2, respectively. It has to be noted that the selection of these values is made independently from the data sets, and thus, the same settings are applied for both data sets used to evaluate the presented methodology. Furthermore, the proposed method requires the learning of the category-aware ASM and the training of the proposed CNN. The learning and training procedures which are conducted for the experiments presented in this thesis are described in Sec. 5.3.1 and 5.3.2, as well.

5.3.1 Learning the ASM

The category-aware Active Shape Model, which is applied as a shape prior in this thesis and which is described in detail in Sec. 4.2, requires the definition of a set of type classes $\mathcal{T}$ and the availability of 3D keypoint annotations for a set of training exemplars. With $\mathcal{T} = \{$*Compact Car, Sedan, SUV, Estate Car, Sports Car, Truck, Van*$\}$, seven vehicle categories are distinguished in this thesis (cf. Fig. 4.4). To learn the ASM, $C_{\mathcal{K}} = 144$ 3D keypoints were manually labelled on a set of 36 different CAD vehicle models belonging to one of the considered vehicle types (cf. Tab. 5.4). This set of training exemplars is the same set that has been described in Sec. 5.2 and which has has been used to generate the reference data for the ICSENS data set. Fig. 5.2 shows the 36 CAD models and their associated vehicle types. Each model differs in shape and consequently also in its 3-dimensional

Table 5.4: Enumeration and definition of free parameters contained in the proposed methodology.

Parameter	Symbol	Value
Free-space grid map (Sec. 4.1.3)		
Side length of the grid cells	l_Φ	25 cm
Active Shape Model (Sec. 4.2)		
Overall number of ASM keypoints	$C_\mathcal{K}$	144
Number of *appearance* keypoints	C_A	36
Number of considered shape parameters	n_s	3
Multi-task CNN (Sec. 4.3)		
Mini-batch size	N	50
Initial learning rate	η	10^{-4}
Dropout rate		0.5
Keypoint uncertainty	$r_\mathcal{K}$	5 cm
Threshold for the custom loss $\mathcal{L}_{\mathcal{KW}}$	t_pred	0.05
Probabilistic model (Sec. 4.4)		
Uncertainty of a disparity estimate	σ_disp	1 [px]
Generalisation uncertainty of the ASM	σ_M	10 cm
Inference (Sec. 4.4.7)		
Number of particles	n_p	200
Number of iterations	n_it	10
Number of offspring particles	n_b	10

extent from the other models. Tab. 5.5 shows the statistics that are obtained from the variations in length, width, and height of the CAD models used in the training set in this thesis. It shows the mean value, standard deviation (std. dev.) as well as the minimum and maximum values for length, width, and height of the vehicles. Obviously, with a standard deviation of 0.40 m, the largest variations are present w.r.t. the length of the vehicles. With standard deviations of 0.20 m and 0.10 m, respectively, the variations in height and width are considerably smaller.

In total, 144 unique vehicle keypoints $\mathcal{K}$ are defined to represent the ASM. It can be noted that during the manual labelling process of the 3D keypoints, due to the symmetric shape of vehicles, only half of the keypoints need to be labelled. The set of *appearance keypoints* $\mathcal{K}_A \in \mathcal{K}$ considered in this work contains $C_A = 36$ individual keypoints and corresponds to the keypoints used in (Zia et al., 2013). A visualisation of the keypoints is shown in Fig. 4.4. However, in comparison to Zia et al. (2013), who used the 36 keypoints for the ASM representation, we use four times as many keypoints (144), which provides a more detailed and more flexible representation of the vehicle shapes. As a consequence of using 36 training exemplars, the PCA on which the ASM is based

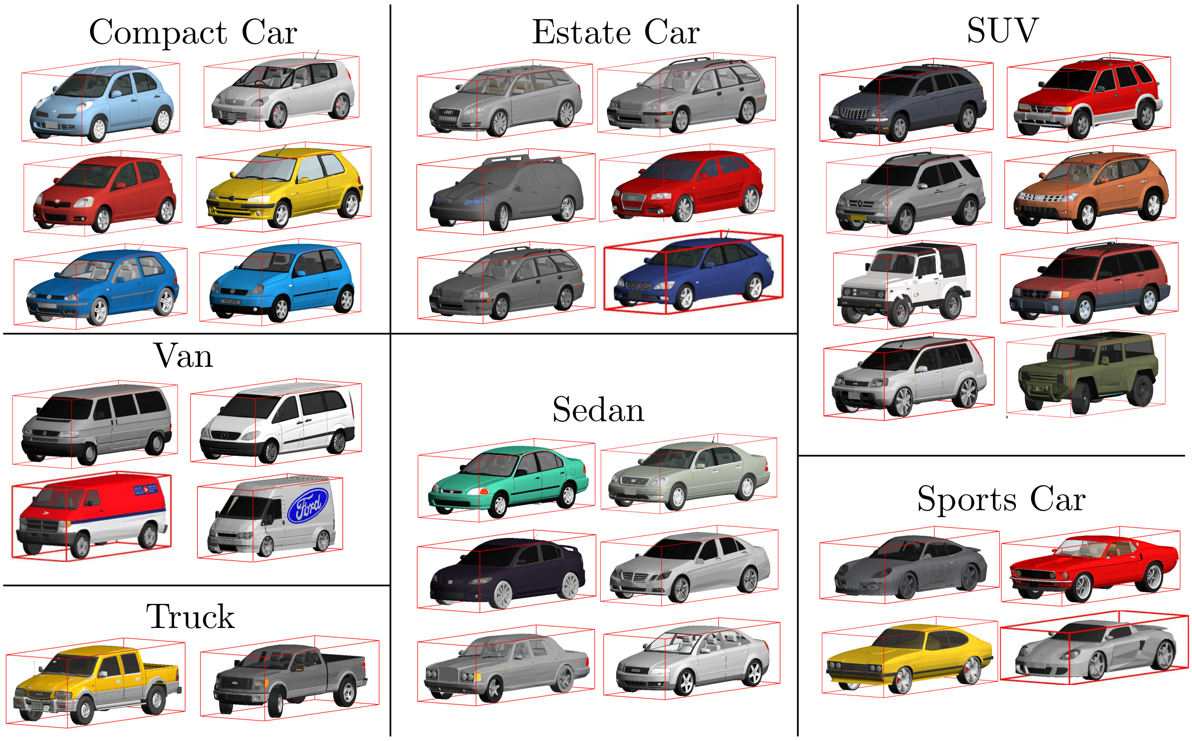
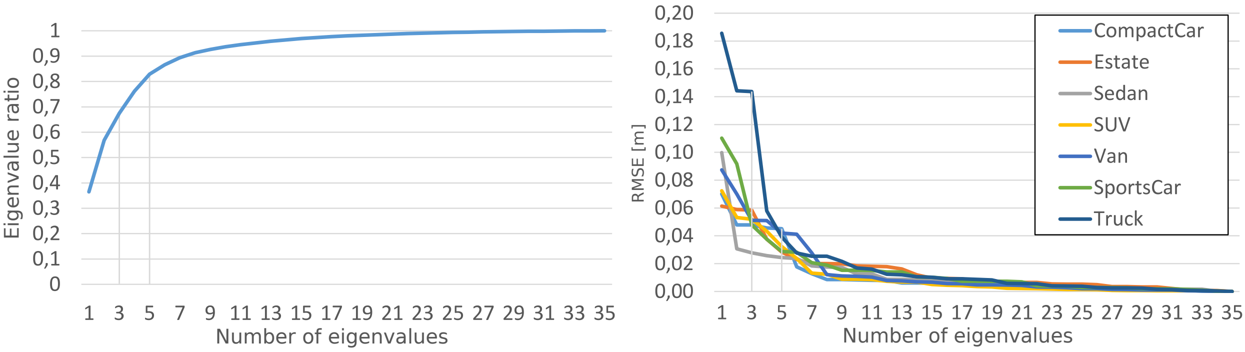


Figure 5.2: The CAD vehicle models which were used to learn the ASM, sorted by their associated vehicle types.

results in a maximum of 35 eigenvalues whose values are not zero. Fig. 5.3a shows the ratio of the sum of used eigenvalues and the overall sum of eigenvalues as a function of the number of used principal components. The resulting curve gives insights into the proportional distribution of the shape deformations that are covered by the used eigenvalues. As mentioned earlier, in order to reduce the dimensionality of the target shape parameter vector γ, the number n_s of considered principal components is restricted in the experiments. In this thesis, the first three principal components are selected for vehicle reconstruction, leading to three shape parameters that have to be estimated. According to Fig. 5.3a, the first three principal components contain about 70% of the variances of the training set. Arguably, while the main shape deformations are covered by the first three components, the remaining components represent the deformations of finer structures and details. In the proposed probabilistic model, however, only a subset of all keypoints and a rather coarse definition of the vehicle wireframe are considered in the image based likelihood terms, thus rendering the consideration of finer details in the ASM representation unnecessary. Likewise, the quality of the 3D point cloud considered within the *3D likelihood* term presumably does not allow the resolution of finer details either, due to its uncertainty properties. As a consequence, the consideration of only the first three principal components is considered as reasonable trade-off between the number of parameters that have to be estimated during the reconstruction and the level of detail of the model approximation.

Table 5.5: Statistical properties for the vehicle extent of the CAD training set which is used to learn the ASM.

	mean [m]	std. dev. [m]	max [m]	min [m]
length	4.35	0.40	5.70	3.55
width	1.80	0.10	2.34	1.65
height	1.49	0.20	2.12	1.11



(a) Ratio of the sum of used eigenvalues (in descending order) w.r.t the overall sum of eigenvalues.

(b) RMSE of the ASM fitting to the individual considered vehicle types.

Figure 5.3: Statistical properties of the learned ASM depending on the number of principal components used. Throughout the experiments of this thesis, the first three principal components are considered.

As explained int Sec. 4.2, a shape basis is learned for each of the considered vehicle types. In Fig. 5.3, the fitting error of the overall ASM to the mean shape $\mathbf{m}^{\tau}$ of the vehicle types τ resulting from Eq. 4.5 is shown as root mean square error (RMSE) of the keypoint coordinates, again as a function of the number of principal components used. The RMSE represents the generalisation error of the ASM caused by a restricted number of used eigenvalues. As can be seen, the errors resulting for the type *truck* using up to the first five components is larger by a factor 2 compared to the other vehicle types. This effect is presumably caused by the distinct dissimilarity of truck shapes compared to the other passenger vehicle types. While the RMSE of the type truck is 0.14 m when using the first three principal components, the RMSE of the remaining vehicle types is smaller than 0.06 m in all cases.

5.3.2 Training of the CNN

This section describes the strategy which was followed to train the CNN proposed in Sec. 4.3 for the experiments that are presented in this thesis. As mentioned earlier, the input branch (cf. Fig. 4.5) is initialised from its corresponding layers of the VGG19 network (Simonyan and

Zisserman, 2015), pre-trained on ImageNet (Russakovsky et al., 2015), and is frozen during the training procedure. The remaining convolutional layers are initialised using the *He* initialiser (He et al., 2015). For the presented experiments, the *type branch* is trained separately from the *viewpoint* and the *keypoint/wireframe* branches. To train the individual output branches, two different data sets are used. For the joint training of the *viewpoint* and the *keypoint/wireframe* branches, training images of vehicles are required including reference information of the vehicle's viewpoint as well as the image coordinates of the *appearance keypoints*. To this end, a subset of 260 images from the KITTI data set are used. The 2D image reference bounding boxes as well as the reference viewpoint angles are provided as annotations of the data set (Geiger et al., 2012). The authors of (Zia et al., 2015) labelled the 36 different *Appearance* keypoints in this subset of images and made the annotations publicly available. Together with the reference viewpoint angles, their keypoint annotations are used to train the *keypoint/wireframe branch* and the *viewpoint branch* for the experiments conducted in this thesis, while the *vehicle type branch* is frozen. Note that the KITTI images used for training are not used for evaluation in this thesis. To create the training data in the required format, the images are cropped by the provided reference bounding boxes. The viewpoint classes needed for the *viewpoint branch* are derived from the groundtruth viewpoint angles according to the viewpoint class definition shown in Fig. 4.7b. To train the *keypoint/wireframe branch*, the 36 labelled keypoints are used to create the 2D reference heatmaps for the keypoints and wireframe edges as described in Sec. 4.3.5. After training both branches, the *vehicle type branch* is trained independently from the remaining branches using the data set provided by (Yang et al., 2015), which contains images of vehicles including bounding box and vehicle type annotations. In (Yang et al., 2015), twelve different vehicle type classes are distinguished. To map the provided classes to the type definitions used in this work, some classes are merged. Tab. 5.6 shows the association of the type classes provided by (Yang et al., 2015) to the classes defined in this thesis.

The network is trained using the Adam optimizer (Kingma and Ba, 2015), a variant of stochastic mini-batch gradient descent with momentum, using the exponential decay rate for the 1st moment estimates $\beta_1 = 0.9$ and for the 2nd moment estimates $\beta_2 = 0.999$. A mini-batch size of $N = 50$ and an initial learning rate of $\nabla = 10^{-4}$ are applied. To improve training, the learning rate is dropped by a factor of 10^{-1} after 5 epochs with no improvement in the validation loss. Furthermore, batch normalisation (Ioffe and Szegedy, 2015) is used and Dropout (Srivastava et al., 2014) is applied

Table 5.6: Vehicle type class definition made in this thesis and the associated classes defined in (Yang et al., 2015).

Ours	Compact car	Sedan	SUV	Estate car	Sports car	Truck	Van
Yang et al. (2015)	Hatchback	Sedan Fastback	SUV Crossover	Estate car	Sports car Convertible Hardtop conv.	Pickup	MPV Minibus

to the fully-connected layers with a rate of 0.5. Data augmentation is applied according to the descriptions given in Sec. 4.3.5. The selected values for the hyperparameters are summarised in Tab. 5.4.

5.4 Evaluation strategy and evaluation criteria

This section describes the test setup and the strategy for the quantitative evaluation of the proposed approach for vehicle reconstruction and its components. Fundamental aspects which are thoroughly analysed in the evaluation concern the performance of the applied detection approach, the performance of the presented CNN, and the performance of the probabilistic model for vehicle reconstruction regarding the quality of the estimated vehicle poses and shapes as well as its performance in comparison to related methods. The test setup which is established for a detailed assessment of these aspects is defined in the following paragraphs. The quantitative evaluation is used to answer the research questions presented in Sec. 5.1.

5.4.1 Detection

The proposed method for vehicle reconstruction builds upon initially detected vehicles. The procedure applied to derive the initial vehicle detections is described in Sec. 4.1.4 and makes use of the detection and instance segmentation method described in (He et al., 2017). In the evaluation of the results for pose and shape reconstruction, which is conducted in the following sections, all correctly detected vehicles are considered. To determine, whether a vehicle is detected correctly or not, the reference vehicle image bounding boxes are compared to the detected bounding boxes by computing the intersection-over-union (IoU) index

$$\mathrm{IoU} = \frac{A_{\text{overlap}}}{A_{\text{union}}}, \tag{5.1}$$

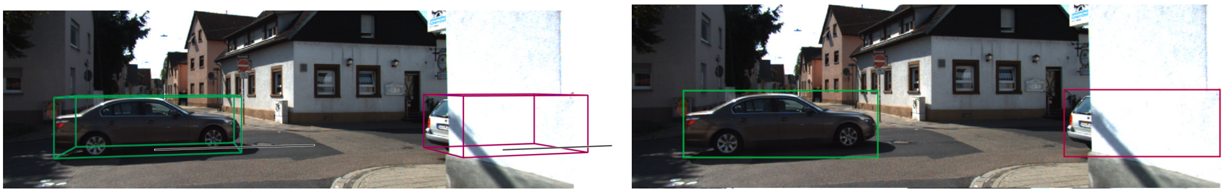
where A_{overlap} is the area of overlap between reference and detection bounding box, and A_{union} is their area of union. The official evaluation metric for 2D object detection proposed for the KITTI benchmark requires an IoU of 70% for a detection to be counted as correct [4]. However, as mentioned in Sec. 5.2, the reference image bounding boxes for both, the KITTI and the *ICSENS* data sets, are derived from backprojecting the reference 3D bounding box or reference CAD models to the image, respectively. As a consequence, the image bounding boxes contain the extent of the entire objects, independent from occluded and therefore non-visible object parts in the image. An example for reference bounding boxes of the KITTI images can be seen in Fig. 5.4, where the reference 3D bounding boxes are backprojected to the image to derive the 2D reference bounding boxes which consequently also include occluded objects. The detection approach applied in this thesis, however,

[4] http://www.cvlibs.net/datasets/kitti/

only detects the visible parts of the vehicles and the detection bounding boxes therefore only enclose the non-occluded object parts. Depending on the required IoU, this leads to the effect that actually correctly detected vehicles are considered as being not correctly detected, decreasing the number of true positive detections while simultaneously increasing the number of false positive detections. For this reason, in the evaluation of this thesis, the required IoU is relaxed. A car is considered to be a true positive detection if the IoU is larger than 50%. In the case of multiple detections for the same vehicle, one detection is counted as a true positive, whereas further detections are counted as false positives. To quantitatively assess the detection performance on both data sets, the values for

- Recall: percentage of all reference vehicles that are successfully detected,
- Precision: percentage of detections that actually are vehicles and
- F1 score: harmonic mean of recall and precision

are used. The evaluation of the detection results is presented in Sec. 6.1.



(a) Backprojected 3D bounding boxes.

(b) 2D bounding boxes.

Figure 5.4: Example of reference bounding boxes of the KITTI data set. Green boxes denote fully visible vehicles and red boxes denote occluded vehicles. The 2D bounding boxes are derived from the backprojected 3D bounding boxes and thus include the full extent of the object, independent from occlusions.

5.4.2 Multi-Task CNN

The CNN which is proposed in this thesis contains multiple branches that are used to generate additional observations in the form of probability maps for vehicle keypoints and the vehicle wireframe, and to derive prior distributions for the vehicle viewpoint and the vehicle type (cf. Sec. 4.3). A reference for the viewpoint is available for both test data sets, while a reference for the vehicle category is only available for the ICSENS data. The availability of these reference data allows an investigation of the performance of the *viewpoint* and *vehicle type* branches of the CNN, which is analysed in Sec. 6.2 based on the evaluation criteria described in this section. An evaluation of the individual branches allows to draw conclusions about the suitability of the proposed CNN for the derivation of prior information, which is subject of research question (1) in Sec. 5.1.

Evaluation of the *viewpoint branch*

As described in Sec. 4.3.3, the *viewpoint branch* of the CNN contains three classification heads (Branch I-III), one head for each hierarchical viewpoint class layer. As evaluation criterion, the overall accuracy (OA) is used, which is computed according to the ratio of the number of correctly classified images and the total number of images, and therefore represents the overall proportion of correct classifications in [%]. The OA that is achieved by the individual classification heads is analysed using the reference viewpoint information. To this end, the viewpoint class receiving the maximum confidence score is treated as the inferred class for the calculation of the OA. Furthermore, the proposed CNN uses a *probabilistic averaging layer* to produce the final probability distribution for the viewpoint Π_ϑ by combining the outputs of the individual heads. The viewpoint distribution as well as the most likely viewpoint ϑ_{CNN} which is derived from the distribution are considered in the presented *orientation prior* of the probabilistic model. To evaluate the beneficial effect of the hierarchical setup of the *viewpoint branch*, the overall classification accuracy that is achieved by deriving the viewpoint class of each hierarchical layer from ϑ_{CNN} can be compared to the OA of the single classification heads. In particular, this comparison is able to reveal if potential classification errors of individual classification heads can be mitigated due to the hierarchical setup. The evaluation is conducted based on the results that are achieved by the CNN when applied to the vehicles of the KITTI and of the ICSENS data sets.

Evaluation of the *vehicle type branch*

The *vehicle type branch* of the CNN delivers classification scores for each of the distinguished vehicle type classes. The overall accuracy is used as metric to evaluate the quality of the classification results. As explained in Sec. 5.2.2, the class boundaries between some the considered vehicle type classes are not clearly defined. The vague definition of the class boundaries is expected to be reflected by a broad distribution of the confidence scores predicted by the *vehicle type branch* over the concerned classes. To account for this effect in the evaluation, different values of OA are reported: The *Top-1* OA, which reports the percentage of correct classifications that are obtained by using the class exhibiting the highest confidence score as the predicted one. In addition, the *Top-2* and *Top-3* OA values are analysed. To obtain the *Top-2* and *Top-3* accuracies, a sample is considered to be a true positive if the reference class is among the two or three classes having the highest confidence scores, respectively. For the evaluation, the ICSENS data set is used, because it contains a reference for the vehicle type.

5.4.3 Probabilistic model for vehicle reconstruction

The core of the experimental evaluation contains the analysis of the results that are obtained by the proposed method for the 3D vehicle reconstruction. The main goals of the analysis are to evaluate the contribution of the individual components of the probabilistic model (question 2 in Sec. 5.1), to evaluate the quality of the results which can be obtained by the full model, and to investigate its limitations (both question 3 in Sec. 5.1). To be able to obtain detailed insights into the behaviour and distribution of the errors in the estimation of pose and shape, suitable evaluation criteria to asses the quality of the reconstructed vehicles are proposed. The following paragraphs give a description of the evaluation strategy and evaluation criteria applied for a thorough analysis of the probabilistic model.

Ablation studies of the model components

To asses the impact of the components of the probabilistic model developed in this thesis, the effects of the individual likelihood and prior terms are analysed in Sec. 6.3. In order to asses the contributions of the individual constituents, different variants for the probabilistic model are defined, each of them considering a different set of likelihood and prior terms. The ablation studies of the majority of model components are performed on the KITTI training set and the results obtained by the different variants are compared and analysed w.r.t. the research question (2) asked in Sec. 5.1. A set of selected variants are also evaluated on the ICSENS data.

In a first set of experiments the effects of the individual likelihood terms and their potential benefits are evaluated (Sec. 6.3.1). For that purpose, several variants of the probabilistic model are defined, each consisting of a different set of likelihood terms. In this set of experiments, the state prior terms are not used for the reconstruction, so that they correspond to variants of a maximum likelihood estimation of the vehicle shape and pose. Note that in this estimation, the regularisation of shape deformations by the category-aware shape prior is omitted. However, constraints on the shape prior still have to be introduced to avoid unconstrained shape variations of the ASM, which may result in geometrically invalid vehicle shapes. One option would be to restrict the shape parameters to a predefined interval, e.g. $[-3, 3]$, which would allow shape variations within $3\sigma_s$ of the variations contained in the training set (cf. Sec. 2.2). To achieve a probabilistic formulation, in this setting the shape parameters are chosen to be regularised by penalising deviations from the mean ASM shape, as it is also done in comparable related methods, e.g. (Zia et al., 2013; Engelmann et al., 2016). By doing so, the category-aware shape prior formulation proposed in this thesis (Eq. 4.25) is replaced by

$$\log p(\gamma) = -\frac{1}{n_s} \sum_{s=1}^{n_s} \left(\frac{\gamma_s}{2\sigma_s} \right)^2 . \tag{5.2}$$

The potential benefit of the category-aware shape prior in comparison to the regularisation setting defined here is analysed in later sections. The variants and the likelihood terms used in the individual variants are shown in Tab. 5.7. More details are given in the subsequent paragraphs.

Table 5.7: Settings and variants for the likelihood formulations. The symbols indicate whether a term is considered in the respective variant (✓) or not (✗). The likelihood terms are those defined in Sec. 4.4.

	Init(-)	$p(\mathbf{X}_v \mid \mathbf{s})$	$p(\mathcal{H}_\mathcal{K} \mid \mathbf{s})$	$p(\mathcal{H}_\mathcal{W} \mid \mathbf{s})$
Init(-)	✓	✗	✗	✗
Base	✓	✓	✗	✗
Base+K	✓	✓	✓	✗
Base+W	✓	✓	✗	✓
Base+K+W	✓	✓	✓	✓

Init(-): As described in Sec. 4.4.7, the proposed parameter initialisation is based on the probability distributions of the state priors. To be consistent with the assumption of neglecting the prior information in this set of experiments, an *uninformed* initialisation is used for the maximum likelihood estimation, denoted by *Init(-)*. In the absence of prior information for the target state variables the parameter initialisation is purely based on the stereo reconstructed 3D points of the vehicle $\mathbf{X}_v$. The position of the initial particle $\mathbf{s}_0^1 = (\mathbf{t}_0^1, \theta_0^1, \gamma_0^1)$ is determined as explained in Sec. 4.4.7: The minimum 2D bounding box, enclosing the 2D projections of the 3D vehicle points on the ground plane, is created and the initial position $\mathbf{t}_0^1$ is set to its centre. The particle orientation θ_0^1 is set to the orientation of the first semi-major axis of the box. The shape parameter vector γ_0^1 is initialised as zero vector, thus representing the mean vehicle shape model.

Base: In this variant, the baseline setting is defined in which only the *3D likelihood* term is considered for model fitting. The *3D likelihood* $p(\mathbf{X}_v|\mathbf{s})$ is chosen to define the baseline model because it is exclusively based on the ASM and the reconstructed 3D points and, therefore, this variant does not require any supervised learning. The observations generated by the CNN (keypoint and wireframe heatmaps) are not considered here. Successively, the additional likelihood terms are added in the next steps to evaluate their impact on the results.

Base+K: This variant performs model fitting based on the *3D likelihood* $p(\mathbf{X}_v|\mathbf{s})$ in combination with the *keypoint likelihood* $p(\mathcal{H}_\mathcal{K}|\mathbf{s})$ to evaluate the impact of the keypoint predictions on the reconstruction.

Base+W: Similar to the previous setting, in this setting the model fitting is based on the combination of the *3D likelihood* $p(\mathbf{X}_v|\mathbf{s})$ and the *wireframe likelihood* $p(\mathcal{H}_\mathcal{W}|\mathbf{s})$. The effects of the wireframe predictions can be compared to the effect of the keypoint predictions when considered for the vehicle reconstruction.

Base+K+W: This setting corresponds to the complete maximum likelihood estimation of the presented probabilistic model. Compared to the previous settings, the potential synergy between all likelihood terms can be assessed.

In a second block of experiments, the effect of the state prior terms on the vehicle reconstruction is analysed (Sec. 6.3.2). To this end, we perform model fitting based on the *3D likelihood* $p(\mathbf{X}_v|\mathbf{s})$ as baseline setting and evaluate the effect of considering the individual state priors during the reconstruction. Tab. 5.8 contains the chosen combinations of prior terms within the probabilistic model that are evaluated in this context. Again, as long as not all prior terms are considered, the *uninformed* initialisation *Init(-)* is used to instantiate the vehicle model. As soon as all prior terms are incorporated for vehicle reconstruction, the *informed* procedure for model initialisation *Init(+)* proposed in Sec. 4.4.7 is applied.

Table 5.8: Settings and variants for the evaluation of state prior terms. The symbols indicate whether a term is considered in the respective variant (✓) or not (✗). The prior terms are those defined in Sec. 4.4.

	Init(-)	Init(+)	$p(\gamma)$	$p(\mathbf{t})$	$p(\theta)$
Base+S	✓	✗	✓	✗	✗
Base+S+P	✓	✗	✓	✓	✗
Init(+)	✗	✓	✗	✗	✗
Base+S+P+O	✗	✓	✓	✓	✓

Base+S: In the first setting, the *shape prior* term $p(\gamma)$ according to Eq. 4.25 is added to the model alignment based on the *3D likelihood*.

Base+S+P: In this setting, the *position prior* $p(\mathbf{t})$ is added to the probabilistic model by considering the probabilistic free-space grid map during the model alignment.

Init(+): As the initialisation strategy proposed in this thesis (cf. Sec. 4.4.7) for instantiating the model parameters is based on the prior probabilities for the viewpoint and the shape, which are inferred by the CNN, the variants of model alignment which have been defined so far and which did not exploit all state priors were built upon the uninformed initialisation *Init(-)* to be consistent with assumptions on the availability of information. When using the full prior formulation during the model alignment, i.e. if all the information required for the proposed parameter initialisation can be assumend to be available, the *informed* initialisation *Init(+)* can be applied. In the *Init(+)* setting, the quality of pose parameters that can already be achieved by this *informed* initialisation is evaluated. Note that consequently, for the *Init(+)* variant, no likelihood term is used and therefore, model fitting is not performed in this variant.

Base+S+P+O: Finally, to assess the potential of the complete state priors, the *shape, position*, and *orientation priors* are jointly considered as regularisers on the state parameters during alignment.

Analysis of the full model for vehicle reconstruction

In this analysis, the evaluation of the vehicle reconstructions using the complete probabilistic formulation of the energy function of Eq. 4.12, referred to as the **Full** model, is conducted (Sec. 6.4) to assess the quality of the results that can be obtained by the presented approach (research question (3) in Sec. 5.1). The *Full* model makes use of all proposed likelihood and prior terms (cf. Tab. 5.9), fusing observations and priors from the 2D image and the 3D object space domain. Until now, every setting was built on top of the *base* variant, thus using the *3D likelihood* for model fitting. To evaluate the importance of the 3D observations for the vehicle reconstruction, a variant form of the *Full* model is proposed, referred to as **Full_Img**, which neglects all terms derived from the explicit 3D information (the *3D likelihood* and the *position prior*) and which consequently uses only the image information (cf. Tab. 5.9). 3D information is only used implicitly in this variant due to the usage of keypoint and wireframe heatmaps of both, the left and right, stereo images. The results for pose (Sec. 6.4.1) and shape (Sec. 6.4.2) obtained by the *Full* model are evaluated on both, the KITTI and the ICSENS data sets. To get further insights into the limitations of the presented method, an in-depth analysis of the results obtained by the *Full* model is conducted using the KITTI data set (Sec. 6.4.3). In this context, reconstruction errors are analysed as functions of the vehicle viewpoint and of the vehicle distance, and as potential systematic errors are analysed as well.

Table 5.9: Settings of the **Full** and the **Full_Img** variant. The symbols indicate whether a term is considered in the respective variant (✓) or not (✗). The likelihood and prior terms are those described in Sec. 4.4.

	Init(+)	$p(\mathbf{X}_v\|\mathbf{s})$	$p(\mathcal{H}_\mathcal{K}\|\mathbf{s})$	$p(\mathcal{H}_\mathcal{W}\|\mathbf{s})$	$p(\gamma)$	$p(\mathbf{t})$	$p(\theta)$
Full	✓	✓	✓	✓	✓	✓	✓
Full_Img	✓	✗	✓	✓	✓	✗	✓

Criteria for pose and shape evaluation

To evaluate the vehicle reconstruction, the resulting pose and shape of each fitted 3D vehicle model are compared to the reference data for location, orientation, and shape of the vehicles.

Pose metrics: To achieve detailed insights into the quality of pose reconstructions, position and orientation estimates are reported in three stages. The values for $\mathbf{t}_{25}$, $\mathbf{t}_{50}$, and $\mathbf{t}_{75}$ report the percentage of determined vehicle positions whose euclidean distance $d_{\mathbf{t}}$ to the reference position is smaller than 0.25 m, 0.50 m, and 0.75 m, respectively. Similarly, θ_5, θ_{10}, and $\theta_{22.5}$ show the percentage of estimated vehicle orientations whose distance to the reference orientation d_θ is less than 5°, 10°, and 22.5°, respectively. For a joint evaluation of position and orientation, the number of pose estimates that are correct in both, position and orientation, are reported in $\mathbf{t}_{75} + \theta_5$ considering the 0.75 m and 5° thresholds. Furthermore, to distinguish between estimates of the position in lateral (across the camera's line of sight) and longitudinal (along the camera's line of sight) direction, $\mathbf{t}^{\mathrm{lat}}_{25/50/75}$ and $\mathbf{t}^{\mathrm{lon}}_{25/50/75}$ report the percentage of lateral and longitudinal position estimates whose error is smaller than 0.25 m, 0.50 m, and 0.75 m, respectively. The root mean square (RMS) error is used to report the quadratic mean of the errors for position $\varepsilon^{\mathbf{t}}_{\mathrm{rms}}$ and orientation $\varepsilon^{\theta}_{\mathrm{rms}}$ with

$$\varepsilon^{\mathbf{t}/\theta}_{\mathrm{rms}} = \sqrt{\frac{1}{n_{\mathrm{veh}}} \sum_{i=1}^{n_{\mathrm{veh}}} \left(d^i_{\mathbf{t}/\theta} \right)^2}, \tag{5.3}$$

with n_{veh} representing the number of vehicles that are considered to calculate the RMS error. The results for the orientation estimates of some of the tested variants reveal systematic errors with the majority of incorrect orientation estimates being wrong by about 180°, rendering the RMS unsuited as an overall error metric. Due to this, the RMS is only used to assess the errors for the set of vehicles whose distance or angle difference to the reference, respectively, is smaller than the defined thresholds. As a consequence, an RMS error is calculated for each of the position and orientation metrics defined above. In order to derive a global error metric for all vehicles, a more robust measure of the average error is used by reporting the median of the position errors $\varepsilon^{\mathbf{t}}_{\mathrm{Med}}$ and the median of the orientation errors $\varepsilon^{\theta}_{\mathrm{Med}}$. In addition, the median absolute deviation $\sigma^{\mathbf{t}/\theta}_{\mathrm{MAD}}$ is reported to assess the variability of the errors w.r.t. the median with

$$\sigma^{\mathbf{t}/\theta}_{\mathrm{MAD}} = k \cdot \mathrm{median}\left(\left| d^i_{\mathbf{t}/\theta} - \varepsilon^{\mathbf{t}/\theta}_{\mathrm{Med}} \right| \right), \tag{5.4}$$

and with $k = 1.4826$ as a constant factor (Hampel et al., 1986).

Shape metrics: Given the 3D vehicle reconstructions, an analysis of the quality of the reconstructed vehicle shapes is conducted. In case of the KITTI data, which only provides 3D bounding boxes as reference for the vehicles, the vehicle dimensions are the only criteria based on which the inferred shape can be evaluated. To this end, average errors and average absolute errors are computed from the differences of the reference and the inferred length, width, and height of the vehicles for the KITTI data set. In contrast to the reported average of absolute errors, the signed average errors give insights into potential biases. While a normal distribution of errors would result in average errors close to zero, a larger deviation from zero indicates an estimation of the vehicle dimensions being systematically to large (in case of negative average errors) or to small (in case of positive average errors), respectively. Instead of 3D bounding boxes, the ICSENS data provides CAD models as reference. In order to compute the same evaluation metrics for the ICSENS data, the minimum 3D bounding box enclosing the CAD models is derived from the reference.

In addition, the reference CAD models are used to compute an error metric based on the distances between corresponding keypoints of the reference model and the estimated model (in the body coordinate system ${}^{A}C$ of the vehicles). To this end, the RMS error d_{rms} of the euclidean distances $d_{\mathcal{K}}^{c}$ between corresponding keypoints $c = [1, C_{\mathcal{K}}]$ is computed for every vehicle:

$$d_{\mathrm{rms}} = \sqrt{\frac{1}{C_{\mathcal{K}}} \sum_{c=1}^{C_{\mathcal{K}}} \left(d_{\mathcal{K}}^{c}\right)^{2}} \tag{5.5}$$

This error represents the quality of the estimated shape w.r.t. the reference shape for a single vehicle. To assess the average quality of shape estimation that is achieved on the whole data set, the geometric mean $\varepsilon_{\mathrm{rms}}^{\mathcal{K}}$ of d_{rms} is computed with

$$\varepsilon_{\mathrm{rms}}^{\mathcal{K}} = \sqrt{\frac{1}{n_{\mathrm{veh}}} \sum_{i=1}^{n_{\mathrm{veh}}} (d_{\mathrm{rms}}^{i})^{2}}. \tag{5.6}$$

Analysis of further aspects

In Sec. 6.4.3, the influence of the distance and the viewpoint on the quality of the pose estimates is analysed in order to get further insights into the strengths and limitations of the proposed approach. To this end, we analyse the performance of vehicle reconstruction as a function of distance and viewpoint. As far as the dependency of the results from the distance is concerned, we distinguish vehicles at four distance classes (distances of 5-10 m, 10-15 m, 15-20 m, and >20 m, respectively). The angles under which the vehicles are observed are grouped into five viewpoints, namely *front, front-side, side, back-side*, and *back*. The exact definition of the five viewpoint classes is depicted in Fig. 5.5. For the analysis of the limitations the KITTI data set is used, because it contains more vehicles, which is important to obtain representative values for the viewpoints and distance classes.

5.4.4 Comparison to related methods

In order to compare the performance of the proposed method to the performance of related methods, the test data set and evaluation metrics of the official KITTI benchmark can be used. However, the focus of that benchmark lies on object detection. As a consequence, metrics to assess the quality of pose estimation are coupled with the performance of detection (cf. (Geiger et al., 2012) for a detailed description of the error metrics). To assess the orientation accuracy, the official metric of the KITTI benchmark is the *Average Orientation Similarity* (AOS), which multiplies the *Average Precision* (AP) of the detector with the average cosine distance similarity for the orientation. However, as explained in Sec. 5.4.1, evaluating the performance of vehicle detection of the method proposed here delivers rather pessimistic results due to the discrepancy in the definition of the bounding boxes provided as reference and those derived from the detection. To still be able to compare the the

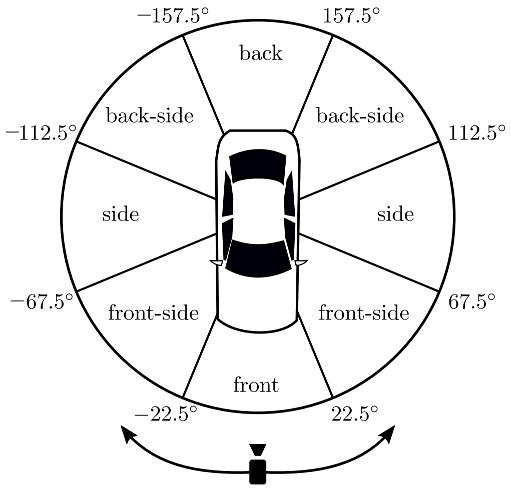


Figure 5.5: Viewpoint categories that are distinguished in the analysis on the influence of the viewpoint on the results of the pose estimation.

orientation estimates to the results achieved by related methods, we apply the *Orientation Score* (OS) metric that was proposed in (Mousavian et al., 2017), and which factors out the 2D detector performance by computing the ratio between AOS over AP. Thus, OS is a metric that is unaffected by the detector's performance and can be used to assess the quality of the estimated orientation values.

6 Results and discussion

In this thesis, a probabilistic approach for the reconstruction of vehicles detected from stereo images is proposed. This chapter presents and discusses the conducted experiments and the results that are achieved by the presented method. First, the performance of the object detector which is applied for the detection of vehicles (Sec. 6.1) and the performance of the developed CNN are analysed (Sec. 6.2). Afterwards, an ablation study of the individual components of the probabilistic approach for vehicle reconstruction is presented (Sec. 6.3). The performance of the complete method is thoroughly evaluated in Sec. 6.4. A comparison of the results that are achieved by the proposed method and results achieved by related methods is given in Sec. 6.5. The chapter concludes with a discussion of the presented results in Sec. 6.6.

6.1 Detection

The initial step of the proposed method consists of detecting the vehicles that are visible in the stereo images. For each detected vehicle instance, the proposed procedure is conducted with the goal to derive a parametrised 3D reconstruction of the vehicle. To this end, it is important that the applied approach for vehicle detection (cf. Sec. 4.1.4) achieves a high detection rate, as undetected vehicles cannot be considered in the reconstruction approach. On the other hand, mainly from a computational point of view, it is also important that the rate of false detection is kept low as every false detection leads to an unnecessary execution of the reconstruction procedure.

Tab. 6.1 shows the percentage of correctly detected vehicles (Recall) for both, the KITTI and the ICSENS data sets. As described in Sec. 5.2, meaningful precision and F1 scores can only be delivered for the KITTI data set, because in contrast to the ICSENS data, vehicles which are visible in the images but which are not contained in the reference are annotated as *don't care* areas, so that it is possible not to count them as false positive detections if they are detected correctly. Furthermore, because false positive detections cannot be associated to one of the *easy, moderate* or *hard* categories, the precision and F1 score can only be computed for the overall data set. Using the KITTI data set, 99.1% and 97.0% of the vehicles associated to the easy and moderate categories, respectively, are detected. For the hard category, the recall is considerably lower. One possible reason for this effect is the discrepancy between the definition of reference bounding boxes and these bounding boxes obtained from the detection, a problem which has already been addressed

Table 6.1: Performance metrics resulting from the applied Mask RCNN on the KITTI and ICSENS data sets.

KITTI	easy	moderate	hard	overall
Recall [%]	99.1	97.0	87.2	92.9
Precision [%]	———	———	———	98.1
F1 score	———	———	———	95.9

ICSENS	easy	difficult	———
Recall [%]	91.9	89.2	———

in Sec. 5.4.1. Another reason is that an increasing degree of object occlusion negatively effects the performance of the detector due to the missing observations for the occluded object parts. The precision achieved for the entire KITTI data set is 98.1% which means that 1.9% of all detections are actually not vehicles. However, it has to be noted that a part of the declared false positive detections are caused by the mentioned discrepancy of reference and detection bounding boxes, where the required value for the IoU is not achieved.

The recall values achieved on the ICSENS data set are considerably lower. Here, the recall is 91.9% and 89.2% for the easy and difficult categories, respectively. It has to be noted that the category *difficult* comprises all vehicles that are associated to the *moderate* and *hard* difficulty levels in the KITTI benchmark. In order to find a potential explanation for the discrepancy of the recall values obtained on the KITTI and the ICSENS data sets, missing detections in the ICSENS data are analysed. The ICSENS data were acquired in the context of research on collaborative positioning of vehicles in dynamic networks (Schön et al., 2018). To this end, three vans serving as sensor nodes in the network were equipped with multi-sensor platforms, and different driving scenarios were designed in order to generate sensor data with mutual observations. A detailed description of the sensor platforms and the scenarios is given in (Schön et al., 2018). As a consequence of the scenario design towards the generation of mutual observations, a considerable proportion of the vehicles visible in the ICSENS data consists of the other measurement vans. However, for some reason, the Mask RCNN (He et al., 2017) applied for the detection in this work, frequently fails to detect the other measurement vehicles. Fig. 6.1 shows selected examples of images from the ICSENS data set, in which the measurement vehicle in the centre of the image could not be detected by the Mask RCNN (first three rows), while in the image shown in the last row, the same van has successfully been detected. The missing detections of the described vans is responsible for a considerable proportion of the lower recall values for the ICSENS data in comparison to the values achieved for the KITTI data. The right column of Fig. 6.1 shows the backprojected wireframes of the ASMs which are estimated by the method proposed in this thesis. As the reconstruction procedure is conducted for every detected vehicle, obviously no ASM is fitted to undetected vehicles. In the subsequent sections, which focus on the evaluation of the quality of the reconstruction, only the true positives are considered, i.e., only vehicle detections that are associated to a reference object.

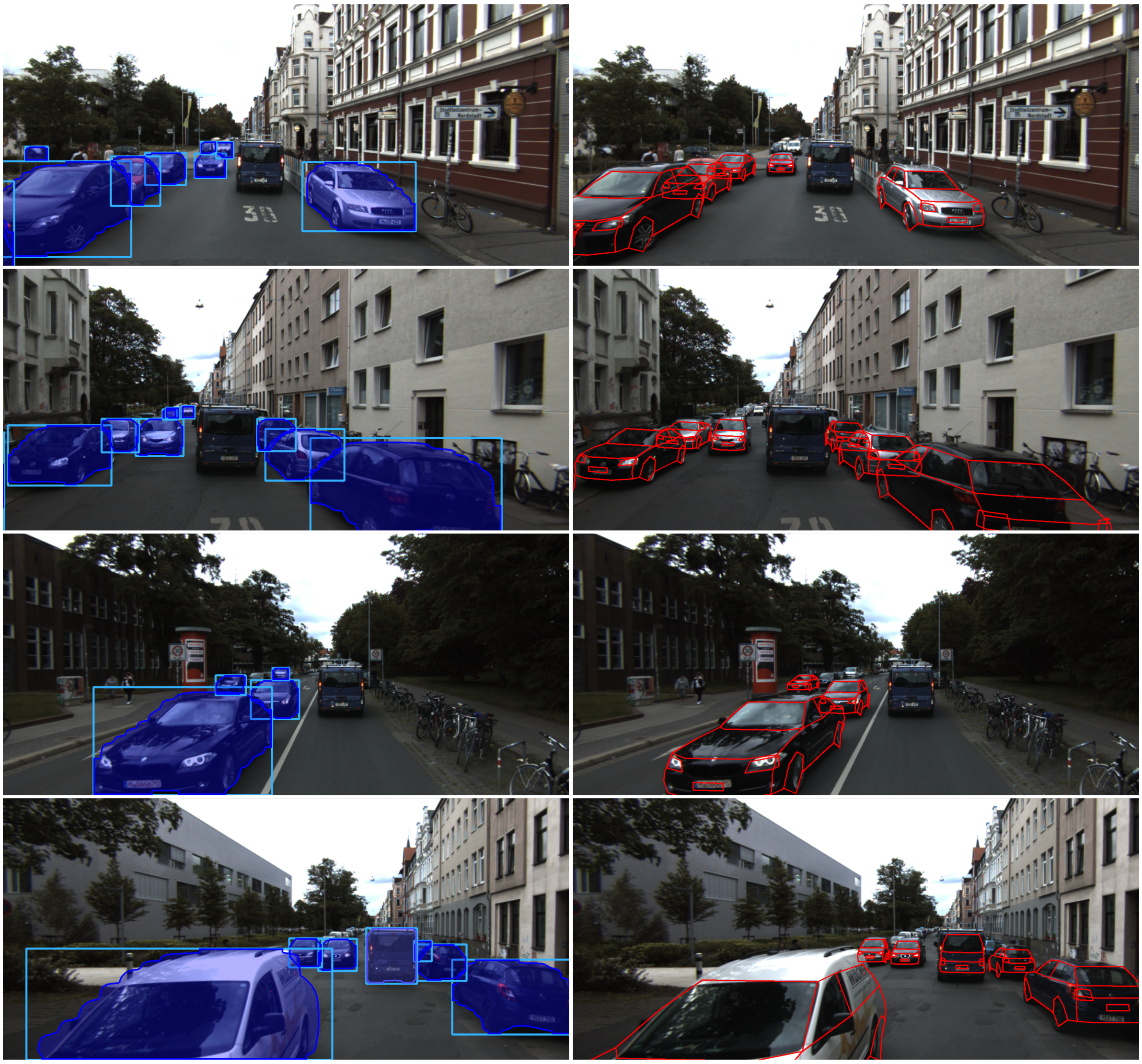


Figure 6.1: Left column: Detection results of the Mask RCNN applied on the ICSENS data. Bounding boxes are shown in light blue and the instance segmentation masks in transparent blue. Right column: The wireframes of the estimated ASM are backprojected to the images. An ASM can only be fitted to the vehicles that are detected. The first three rows show the frequently occurring missed detection of the van in the centre of the images. However, the very same van has been successfully detected and reconstructed in the image shown in the last row.

6.2 Evaluation of the CNN components

To derive prior information about the vehicle orientation and the vehicle type, the *viewpoint* and the *vehicle type branch* of the CNN are used. Both branches are classification branches, designed to output a confidence score for each of the set of considered viewpoint and type classes, respectively. The confidence scores are interpreted as probabilities for the occurrence of the respective classes in order to derive a probability distribution which is incorporated in the probabilistic model as a state prior . This section presents a study of the performance of the individual branches by evaluating the quality of their classification in order to answer the research question (1) in Sec. 5.1.

6.2.1 Evaluation of the *viewpoint branch*

In Sec. 4.3.3, the hierarchical class structure of the viewpoint classes is presented, which consists of three hierarchical class layers, distinguishing between 4, 8, and 16 different classes for the viewpoint angle, respectively. The *viewpoint branch* consists of three classification heads (Branch I - III), where each head is used for the prediction of the different classes that are contained in the hierarchical class definitions. In a probabilistic averaging layer, the outputs of the individual heads are merged to derive a probability distribution. The setup of the *viewpoint branch* followed the intuition that the classification accuracy decreases with an increasing number of viewpoint classes. The hierarchical setup was chosen to be able to profit from the expectedly more reliable output of the coarse class layers while still leveraging the more detailed output of the finer layers. One way to evaluate if the initial assumption and the promised effect of the hierarchical setup holds true is to analyse the performance of the individual classification heads and to compare the results to the result that is obtained by merging the outputs those single branches. To this end, Tab. 6.2 shows the overall accuracies (OAs) that are obtained by the individual classification heads on the KITTI and the ICSENS data sets. Furthermore, the class of each hierarchical layer can be derived from the viewpoint ϑ_{CNN}, which is the viewpoint having the maximum value in the probability distribution computed by the *probabilistic averaging layer* (cf. Sec 4.3.3). The overall accuracy achieved by using these classes is also presented in the table.

Regarding the results that are obtained on the KITTI data, the assumption that a finer definition of viewpoint classes leads to a lower classification accuracy by using the Branch I - III classification heads is verified. While in the *easy* category, an OA of 96.6% is achieved for the classification of four viewpoint classes, i.e. the class definition, where each class corresponds to a viewpoint interval of a width of 90°, the OA of 91.1% for the classification of 16 classes is significantly lower. This effect is even more distinctive for the *moderate* and *hard* categories. Furthermore, the classification based on ϑ_{CNN}, i.e. on the combined usage of the classification branches, improves the results obtained by only using the single classification branches in all cases. This observable pattern supports the rationale on which the hierarchical setup of the *viewpoint branch* is founded and indicates that the

Table 6.2: Overall accuracies achieved by the *viewpoint branch* and its components for the classification of viewpoint classes on the KITTI and the ICSENS data sets.

		4 classes		8 classes		16 classes	
	OA [%]	Branch I	ϑ_{CNN}	Branch II	ϑ_{CNN}	Branch III	ϑ_{CNN}
KITTI	easy	96.6	98.5	93.3	94.2	91.1	91.4
	moderate	95.4	97.1	90.1	91.4	87.6	88.2
	hard	82.8	85.2	76.3	78.1	72.7	73.6
ICSENS	easy	78.6	83.6	71.6	77.0	75.1	70.6
	difficult	78.5	83.6	70.4	75.3	72.3	69.6

probability distribution that is obtained by the combined use of the classification branches improves over the distributions that are achieved by the individual branches.

The results for the OA obtained on the ICSENS data set for 4 and 8 classes support the conclusion that were drawn based on the results for the KITTI data. Here, the classification accuracy derived from ϑ_{CNN} surpasses the results that are achieved by the individual branches by a significant margin. However, counterintuitively, the OAof 75.1% and 72.3% for the *easy* and *difficult* categories obtained by Branch III for 16 viewpoint classes are larger than the OAs obtained by Branch II for only 8 classes. Due to this fact, the OA using the class derived from ϑ_{CNN} for the 16 class layer is smaller compared to the OA achieved by the usage of Branch III alone.

Apart from determining the overall classification accuracy from the predicted viewpoint ϑ_{CNN}, its quality w.r.t. the continuous reference viewpoint can be assessed. To this end, the reader is referred to the analysis of the *Init(+)* variant, which is presented in Secs. 6.3 and 6.4, as this variant reports the quality of the initial vehicle models which are instantiated by using ϑ_{CNN} in order to define the initial orientation.

6.2.2 Evaluation of the *vehicle type branch*

In this section, the performance of the *vehicle type branch* is analysed based on the overall accuracy of the vehicle type classifications. As mentioned in Sec. 5.2, the association of a vehicle to one of a set of defined vehicle types can be an ambiguous task in some cases, even for human annotators. Depending on the vehicle types that are to be distinguished, the transitions between the appearance of different type classes can be fluent. As a consequence, for some vehicles, the decision of assigning them to one class or another can be difficult due to a vague and not stringent definition of the type classes themselves. Transferring these properties to the expected results of a vehicle type classifier leads to the expectation that the partially vague class definitions are reflected by a larger amount of class confusion on the one hand, and by the prediction of confidence scores that lack of a distinct maximum on the other hand. In Tab. 6.3 the OAs computed from the classification results of the *vehicle type branch* on the ICSENS data set are shown. To analyse the effect of potential confusions due to vague class boundaries between specific vehicle types, the Top-1 - Top-3 overall accuracies are reported in the table.

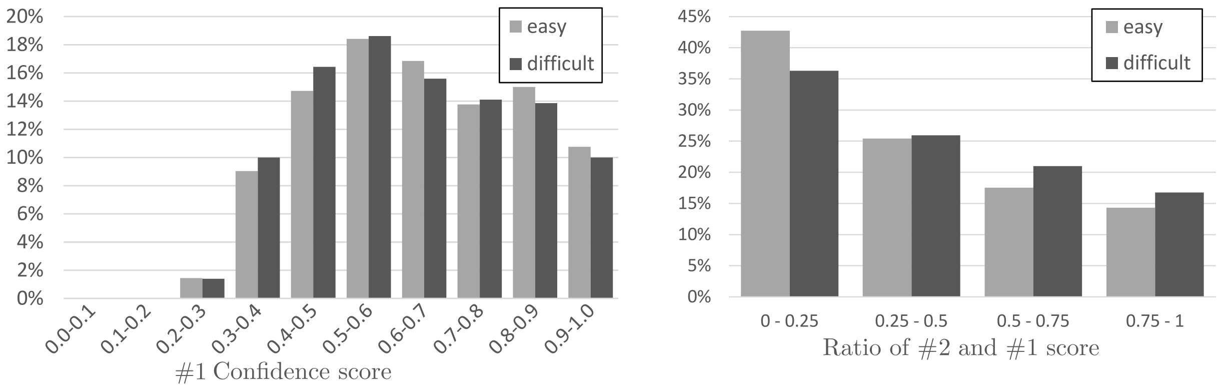
Tab. 6.3 shows that the OA of vehicle types is relatively low. If only the class having the highest score is considered (Top-1 OA), it is at 58.0% and 57.5% for the *easy* and *difficult* categories, respectively. On the one hand, the different data domains used for training and testing can be a potential factor limiting the performance of the classifier on the ICSENS data. On the other hand, confusions in the classification results that are caused by the problems described above can potentially cause a significant contingent of incorrect classifications. To get deeper insights into the behavior of the *vehicle type branch*, Fig. 6.2 shows selected properties of the confidence scores predicted by the classifier for the type classes. One indicator for the vagueness of the class definitions to actually be a problem for the classifier is the magnitude of the #1 confidence score, i.e. the score with the largest value among all classes. A histogram of the #1 scores obtained by the *Vehicle type branch* on the ICSENS data is shown in Fig. 6.2a. Another indicator for the same phenomenon is the ratio between the #2 and #1 scores, i.e. the ratio between the second largest and largest confidence scores, lying in the range [0,1]. A larger value for that ratio indicates a higher uncertainty of the classifier about the distinction between the two classes associated to the scores. A histogram of the ratio between the #2 and #1 confidence scores obtained on the ICSENS data is shown in Fig. 6.2b.

As can be seen in Fig. 6.2a, only about 10% of all vehicles obtain a #1 confidence score of 0.9 or higher for one of the classes. Instead, the largest amount of more than 18% of the data achieve #1 scores between 0.5 and 0.6. Interpreting the distribution allows the conclusion that in a considerable amount of cases, the certainty of the classifier to predict the correct class is relatively small. As Fig. 6.2b reveals, this conclusion is further stressed by the relatively large proportion of data (14 and 16% for the *easy* and *difficult* levels), for which the ratio of #2 and #1 confidence scores is larger than 0.75, i.e. the confidence score for the second most probable class is almost as large as the one for the most probable class. Another 18-21% of the data exhibit a confidence score ratio between 0.5 and 0.75. Taking into account the OA achieved for the Top-2 and Top-3 evaluation shown in Tab. 6.3, which permit confusions between the first two and first three most confidently predicted classes, increases the accuracy by about 13.5 and 21.8%, respectively.

To sum up, the prediction of the vehicle type by the *vehicle type branch* results in comparably low overall Top-1 accuracies. However, the uncertainties of the prediction are reflected by likewise low confidence scores for the predictions. One reason for this behaviour may be the vagueness of the class definition mentioned earlier. However, the confidence-aware *shape prior* term (Eq. 4.25) which makes use of the output of the *vehicle type branch* considers the prediction uncertainties and is therefore able to handle classification errors caused by a potentially indistinct definition of vehicle

Table 6.3: Overall classification results of the *vehicle type branch* on the KITTI data. Top-1 - Top-3 OAs [%] are shown.

	Top-1	Top-2	Top-3
ICSENS			
easy	58.0	71.6	79.8
difficult	57.5	70.8	79.3



(a) Histogram of the #1 confidence scores obtained by the *vehicle type branch* on the ICSENS data.

(b) Histogram of the ratio of #2 and #1 score of the *vehicle type branch* on the ICSENS data..

Figure 6.2: Distribution and properties of the classification confidence scores obtained by the *vehicle type branch* on the ICSENS data.

type assignments. The effect of the *shape prior* term leveraging the predictions of the vehicle type is investigated in Sec. 6.3.2.

While the analysis of the *viewpoint* and the *vehicle type branches* presented in this section answers the first research question of Sec. 5.1 w.r.t. the performance of the individual branches of the CNN, the suitability of their output for the derivation of state priors is investigated in Sec. 6.3.2.

6.3 Ablation studies of the model components

This section presents an ablation study of the individual components of the proposed probabilistic model for vehicle reconstruction in order to answer research question (2) of Sec. 5.1 by investigating the desired benefit from the individual likelihoods and priors. To this end, different variants of combining different constituents of the model were presented in Sec. 5.4.3. In Sec. 6.3.1, the likelihood terms and their contribution on the quality of the results are investigated. The effect and benefit of the state prior terms is evaluated in Sec. 6.3.2. The ablation study is performed on the KITTI data.

6.3.1 Analysis of the observation likelihoods

To evaluate the overall performance of the model variants using the likelihood terms, Tab. 6.4 shows the numerical results for the different position and orientation metrics achieved by each individual setting defined in Tab. 5.7. Tab. 6.5 contains the error metrics achieved by the different likelihood variants. To enable detailed insights into the orientation estimates, Fig. 6.3 shows a cumulative histogram (Fig. 6.3a) and a histogram (Fig. 6.3b) of orientation errors for the *easy* level. The behavioural pattern of the results for the *easy* level is representative for the other difficulty levels, although the absolute values of the results differ.

Table 6.4: Quantitative pose estimation results for the likelihood variants on the KITTI training set. The best achieved values for the respective metrics are printed in bold font.

	in [%]	Init(-)	Base	Base+K	Base+W	Base+K+W
easy	$\mathbf{t}_{25}$	15.6	39.9	41.6	**42.3**	40.7
	$\mathbf{t}_{50}$	40.6	71.8	71.6	**73.6**	73.0
	$\mathbf{t}_{75}$	66.1	87.3	87.1	**89.0**	88.5
	θ_{5}	9.2	57.0	73.3	84.8	**84.9**
	θ_{10}	13.5	68.8	81.9	94.3	**95.2**
	$\theta_{22.5}$	18.8	74.5	84.6	96.8	**97.9**
	$\mathbf{t}_{75} + \theta_{5}$	6.9	54.1	68.3	**78.0**	77.6
moderate	$\mathbf{t}_{25}$	13.7	36.4	38.8	**39.2**	38.6
	$\mathbf{t}_{50}$	36.2	67.5	68.7	70.2	**70.6**
	$\mathbf{t}_{75}$	58.9	82.6	83.9	85.3	**85.9**
	θ_{5}	8.7	52.9	68.4	77.3	**78.7**
	θ_{10}	13.0	64.5	76.5	86.2	**88.4**
	$\theta_{22.5}$	18.2	70.2	80.1	89.4	**91.8**
	$\mathbf{t}_{75} + \theta_{5}$	5.6	49.7	63.8	71.6	**72.5**
hard	$\mathbf{t}_{25}$	11.5	32.5	34.9	**35.4**	**35.4**
	$\mathbf{t}_{50}$	30.7	60.8	62.7	64.1	**65.0**
	$\mathbf{t}_{75}$	51.2	75.2	77.1	78.8	**79.8**
	θ_{5}	8.4	47.7	61.8	68.6	**70.3**
	θ_{10}	12.9	58.0	69.2	76.9	**79.6**
	$\theta_{22.5}$	18.0	63.4	72.8	80.2	**83.2**
	$\mathbf{t}_{75} + \theta_{5}$	5.0	44.2	57.2	63.5	**64.6**

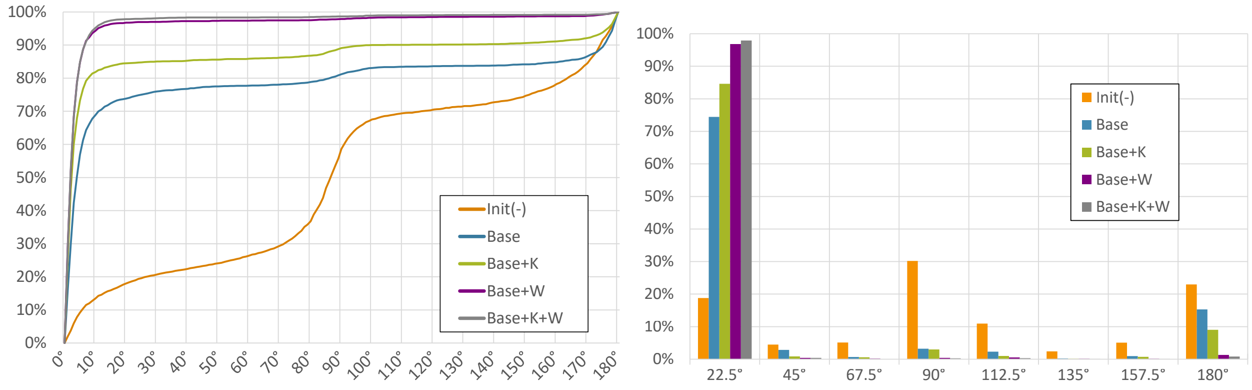
Init(-): Tab. 6.4 shows that the initialised position is estimated within a radius of 0.75 m from the true position in 66.1% of the cases and in a radius of 0.25 m in only 15.6% of the cases for fully visible vehicles. The numbers decrease significantly for occluded or truncated objects. As only 3D points reconstructed from the parts of the vehicle that are visible to the camera are used for the initialisation, partial occlusion and self-occlusion lead to bounding boxes that do not cover the

Table 6.5: Error metrics for the likelihood variants on the KITTI training set. The smallest values achieved for the different metrics among all settings are printed in bold font.

		Init(-)	Base	Base+K	Base+W	Base+K+W
easy	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.59	0.31	**0.30**	**0.30**	0.31
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.38	0.25	0.25	**0.24**	**0.24**
	$\varepsilon^{\theta}_{\text{Med}}$[°]	86.5	3.9	2.2	**1.9**	**1.9**
	$\sigma^{\theta}_{\text{MAD}}$[°]	72.7	4.5	2.3	1.9	**1.8**
moderate	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.65	0.34	**0.32**	**0.32**	**0.32**
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.43	0.27	0.27	**0.26**	**0.26**
	$\varepsilon^{\theta}_{\text{Med}}$[°]	88.1	4.5	2.5	2.2	**2.1**
	$\sigma^{\theta}_{\text{MAD}}$[°]	83.9	5.4	2.7	2.2	**2.1**
hard	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.73	0.39	**0.36**	**0.36**	**0.36**
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.49	0.33	0.32	0.31	**0.30**
	$\varepsilon^{\theta}_{\text{Med}}$[°]	87.6	5.6	2.9	2.5	**2.4**
	$\sigma^{\theta}_{\text{MAD}}$[°]	75.5	7.4	3.5	2.8	**2.7**

whole vehicle but only the visible parts. Consequently, deriving the initial position from the centres of these underestimated bounding boxes results in shifts w.r.t. the true centre. The rationale behind the initialisation of the vehicle orientation from the semi-major bounding box axis is derived from the heuristic assumption about the proportion of the vehicle dimensions (usually the side-length is longer than the front/back-width). However, even if this assumption holds true, it introduces ambiguities w.r.t. the two possible opposed directions of the semi-major axis. Furthermore, the heuristic can be violated in the case of bounding boxes that are too small due to occlusions. In these cases, the initial orientation may suffer from an offset to the true orientation of abount 90°. The described phenomena are visible in Fig. 6.3 where, apart from the peak at orientation errors smaller than 22.5°, the peaks in the histogram correspond to orientation errors of about 90° and 180°, respectively. Consequently, the number of correct orientation estimations is quite poor, with a maximum of 18.8% for the easy category and the coarse error threshold of 22.5°. Accordingly, the median errors for position and orientation for this variant are large with up to 0.73 m for the position and up to 88.1° for the orientation (cf. Tab. 6.5). However, as the subsequent sections will show, the proposed approach is robust against these initialisation errors.

Base: In this setting, the ASM is fitted to the triangulated point cloud. This procedure already delivers significantly better results for the pose compared to the initial parameters of the *Init(-)* setting. For the $\mathbf{t}_{75}$ and $\theta_{22.5}$ metrics of the *easy* level, Tab. 6.4 shows that the number of correct position estimates improves from 66.1 to 87.3% , while the number of correct orientation estimates almost triples and achieves 74.5% of correct fits. As a consequence, the median errors achieved by this setting are drastically reduced (cf. Tab. 6.5). It is interesting to see the change of the orientation error distribution shown in the Figs. 6.3. The increase of correct orientation estimates



(a) Cumulative histogram of orientation errors.

(b) Histogram of orientation errors.

Figure 6.3: Histograms of absolute differences between estimated and reference orientations for the maximum likelihood variants on the KITTI training data (*easy* level).

primarily reduces the incorrect orientation initialisations lying in the intermediate orientation bins ($45 - 157.5°$), especially in the 90° bin. Compared to this, the reduction of errors in the last bin is relatively small. As a consequence, the large majority of wrong orientation estimates resulting from the *base* variant are incorrect by about 180° and, thus, correspond to an opposite orientation compared to the true one. This phenomenon is suspected to be caused by the symmetries of the 3D shape of vehicles w.r.t. their lateral axis, leading to ambiguities between the two opposite viewing directions that cannot be resolved by the *3D likelihood* term.

One goal of extending and composing the probabilistic formulation by multiple likelihoods derived from both, the 3D and 2D domains, is to exploit the redundancy obtained by using additional observations in order to resolve ambiguities that are potentially inherited by the individual likelihoods. The effects of adding the *keypoint* and a *wireframe likelihood* are analysed in the following paragraphs.

Base+K: This variant performs model fitting based on the *3D likelihood* in combination with the *keypoint likelihood*. Tab. 6.4 shows that considering the *keypoint likelihood* leads to similar or slightly improved results regarding the position estimation compared to the *base* variant. The improvement is larger for occluded vehicles in the *moderate* and *hard* levels than for fully visible (*easy*) vehicles. In contrast to the small benefits regarding the position, for the orientation estimation there are larger improvements of up to 10.1% for the 22.5° threshold and of even up to 16.3% for the 5° threshold due to considering the *keypoint likelihood*. According to this observation, the median error for the position improves only slightly compared to the *base* setting, while the median orientation error improves by about 45% in all difficulty levels. As Fig. 6.3b reveals, the improvements on the orientation results lead to a reduction of erroneous estimations mainly in the last (180°) orientation bin. It is reasonable to assume that the introduction of semantic knowledge in the *keypoint likelihood*, especially the differentiation between keypoints belonging to front and back parts and to left and right parts of the vehicles, helps to reduce the ambiguities between vehicle models with opposite viewing directions.

Base+W: Similarly to the previous setting, in this setting the model fitting is based on the combination of the *3D likelihood* and the *wireframe likelihood*. The effects of the wireframe predictions can be compared to the effect of the keypoint predictions when considered for the vehicle reconstruction. Comparing the results to the *Base+K* variant, the improvement is even larger, especially regarding the number of correct orientation estimations, which increase drastically as can be seen in Tab. 6.4. The large improvement of the orientation estimates achieved by the *wireframe likelihood* compared to the *keypoint likelihood* seems surprising. It can possibly be explained by the differences between the point-like representation of the keypoints and the linear representation of the wireframe. An alignment of 3D model keypoints to its detected 2D counterparts in the image inherits a certain degree of geometrical instability, because small localisation errors on the image plane can result in large uncertainties in 3D space. In contrast to the point-like representation by keypoints, a representation by edges leads to geometrically more stable alignments. It can be assumed that these properties lead to a better performance of the *wireframe likelihood* for model alignment compared to the *keypoint likelihood*.

Base+K+W: This setting uses all likelihood terms introduced in Sec. 4.4. Compared to the previous settings, the potential synergy of all likelihood terms can be assessed. According to Tab. 6.4, in the *easy* category, i.e. for fully visible vehicles, the number of correct position estimates decreases slightly compared to the *Base+W* variant, while at the same time subtle improvements in the number of correct orientation estimates are achieved. For occluded vehicles (*moderate* and *hard* levels), the improvement w.r.t. the orientation estimation is even more distinct and in these cases also the position estimation receives benefits from considering both, the *keypoint* and *wireframe likelihoods*, during model fitting. Except for the *easy* category, where the median error for the position is slightly larger compared to the median error achieved by the *Base+K* and *Base+W* variant, the *Base+K+W* setting achieves the smallest median errors and median absolute deviations for both, the position and orientation estimates. Regarding the number of reconstructed vehicles which are correct in both, position and orientation, according to the $\mathbf{t}_{75} + \theta_5$ metric, the consideration of the complete maximum likelihood formulation achieves best results for the *moderate* and *hard* levels of difficulty. These findings answer the question (2) in Sec. 5.1 w.r.t. the contribution of the individual likelihood terms to the vehicle reconstruction by demonstrating the benefit from the joint consideration of all likelihood terms for the quality of the reconstructions.

Qualitative results obtained by the *Base+K+W* variant are shown in Fig. 6.4, which contains visualisations of the probability maps for the keypoints and wireframes as well as the wireframe of the reconstructed ASM backprojected to the left stereo image. To be able to show the probability maps for the individual keypoints and wireframe definitions in one image, the heatmaps of all keypoints and wireframes are superimposed and the maximum value among all heatmaps for each pixel is shown.

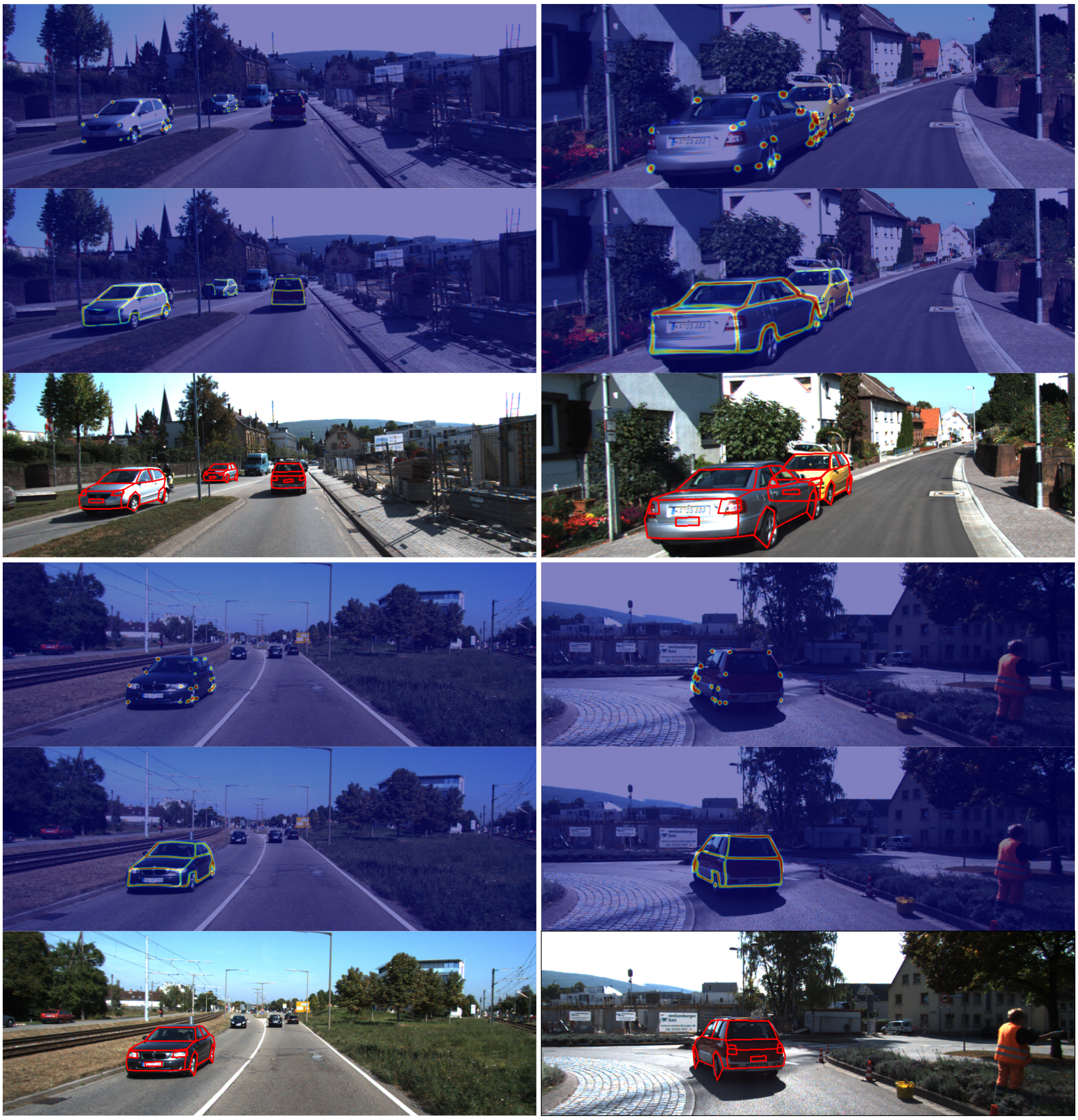


Figure 6.4: Qualitative results obtained by the *Base+K+W* variant on four images from the KITTI data set. For every image, a triplet consisting of the probability maps for vehicle keypoints (top) and vehicle wireframes (middle) superimposed to the left input image are shown (the cold to warm colour coding represents low to high probabilities). Furthermore, the backprojected wireframes of the reconstructed vehicles are depicted (bottom).

6.3.2 Analysis of the state priors

In this section, the effect of the state priors on the vehicle reconstruction is analysed. To this end, we perform model fitting based on the *3D likelihood* as baseline setting and evaluate the effect of considering the individual state priors during the reconstruction. The results for the *Init(-)* and *base* settings are recalled for comparison. The resulting values for the proposed evaluation metrics are shown in Tab. 6.6 and for the error metrics in Tab. 6.7. The detailed histograms revealing the error distribution of the orientation estimates are depicted in Fig. 6.5.

Table 6.6: Quantitative pose estimation results for the state prior variants on the KITTI training set. The best achieved values for the respective metrics are printed in bold font.

	in [%]	Init(-)	Base	Base+S	Base+S+P	Init(+)	Base+S+P+O
easy	$\mathbf{t}_{25}$	15.6	39.9	45.0	45.9	15.3	**46.6**
	$\mathbf{t}_{50}$	40.6	71.8	74.9	76.7	40.5	**77.1**
	$\mathbf{t}_{75}$	66.1	87.3	88.8	91.3	66.2	**92.1**
	θ_5	9.2	57.0	60.4	66.9	42.0	**77.3**
	θ_{10}	13.5	68.8	69.5	76.1	71.3	**95.1**
	$\theta_{22.5}$	18.8	74.5	73.1	79.6	97.8	**98.6**
	$\mathbf{t}_{75} + \theta_5$	6.9	54.1	58.3	64.1	28.0	**73.0**
moderate	$\mathbf{t}_{25}$	13.7	36.4	40.9	41.2	13.5	**43.0**
	$\mathbf{t}_{50}$	36.2	67.5	71.0	72.1	36.0	**74.2**
	$\mathbf{t}_{75}$	58.9	82.6	84.3	85.9	59.0	**89.2**
	θ_5	8.7	52.9	56.0	60.3	37.4	**72.5**
	θ_{10}	13.0	64.5	64.9	69.0	66.6	**89.8**
	$\theta_{22.5}$	18.2	70.2	69.0	73.3	93.6	**95.3**
	$\mathbf{t}_{75} + \theta_5$	5.6	49.7	53.6	57.2	23.5	**68.0**
hard	$\mathbf{t}_{25}$	11.5	32.5	36.3	36.6	11.3	**38.9**
	$\mathbf{t}_{50}$	30.7	60.8	64.0	65.3	30.6	**68.9**
	$\mathbf{t}_{75}$	51.2	75.2	76.9	78.7	51.2	**83.8**
	θ_5	8.4	47.7	50.4	53.7	32.8	**66.9**
	θ_{10}	12.9	58.0	58.4	61.7	60.4	**82.9**
	$\theta_{22.5}$	18.0	63.4	62.3	65.9	87.2	**89.4**
	$\mathbf{t}_{75} + \theta_5$	5.0	44.2	47.6	50.6	19.7	**61.5**

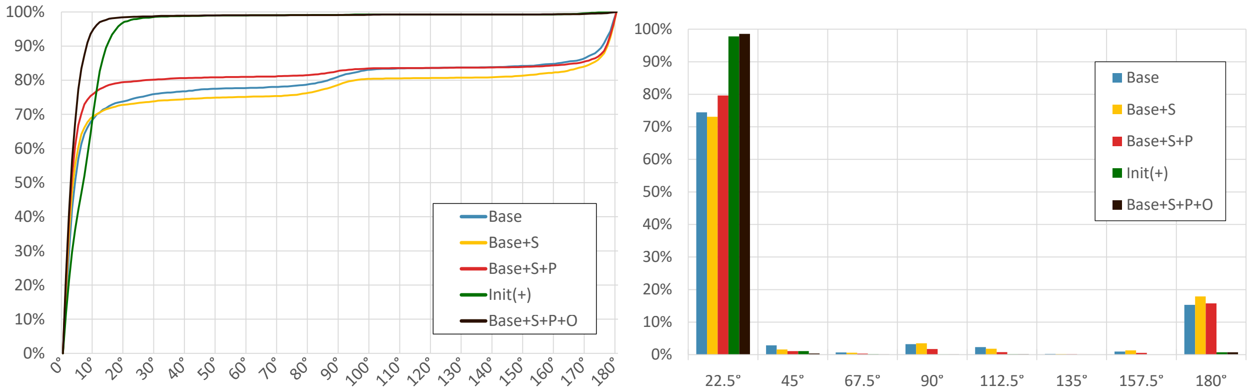
Base+S: In this setting the *shape prior* term is added to the model alignment based on the *3D likelihood*. As can be noticed in Tab. 6.6, the number of correct position estimates, especially w.r.t. to the finer-grained metrics $\mathbf{t}_{25}$ and $\mathbf{t}_{50}$, is significantly improved by the incorporation of the category-aware shape prior, with improvements of up to 5.1%. It is reasonable to conclude that the category-aware regularisation of the model shapes results in a better representations of the observed vehicles w.r.t. to their dimensions. When only considering the *3D likelihood* for model alignment, there are errors of shape reconstructions and therefore errors in the estimated vehicle dimensions,

Table 6.7: Error metrics for the state prior variants on the KITTI training set. The smallest values achieved for the different metrics among all settings are printed in bold font.

	in [%]	Init(-)	Base	Base+S	Base+S+P	Init(+)	Base+S+P+O
easy	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.59	0.31	0.28	**0.27**	0.59	**0.27**
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.38	0.25	0.22	0.22	0.38	**0.21**
	$\varepsilon^{\theta}_{\text{Med}}[^\circ]$	86.5	3.9	3.3	2.8	6.6	**2.5**
	$\sigma^{\theta}_{\text{MAD}}[^\circ]$	72.7	4.5	3.8	3.0	6.1	**2.5**
moderate	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.65	0.34	0.31	0.30	0.65	**0.29**
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.43	0.27	0.25	0.23	0.43	**0.22**
	$\varepsilon^{\theta}_{\text{Med}}[^\circ]$	88.1	4.5	3.8	3.4	7.4	**2.7**
	$\sigma^{\theta}_{\text{MAD}}[^\circ]$	83.9	5.4	4.7	3.9	6.4	**2.7**
hard	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.73	0.39	0.35	0.34	0.73	**0.32**
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.49	0.33	0.30	0.29	0.49	**0.26**
	$\varepsilon^{\theta}_{\text{Med}}[^\circ]$	87.6	5.6	4.9	4.2	8.3	**3.1**
	$\sigma^{\theta}_{\text{MAD}}[^\circ]$	75.5	7.4	6.5	5.3	7.0	**3.2**

caused by the fact that the observed 3D points are only available for a part of the vehicle because no points are observed on the vehicle side facing away from the camera. Introducing a shape regularisation that is aware of the vehicle category, which in many/most cases is directly related to the vehicle dimensions, leads to better regularisation constraints on the shape and consequently to enhanced results for the position estimates. At the same time, the consideration of the *shape prior* also significantly improves the results for orientation estimation w.r.t. the θ_5 metric while the number of correct orientation estimates in the $\theta_{22.5}$ metric is slightly decreased. As can be seen in Fig. 6.5, the additional erroneous orientation estimates are to be found almost exclusively in the last orientation error bin (180°). Both, the median errors for position and orientation, are notably improved by the consideration of the *shape prior* term (cf. Tab. 6.7). To conclude, the category-aware shape prior proposed in this thesis leads to an increase in correct position estimations, probably due to a better representation of shape and dimension by the *shape prior*, and to a better quality of the orientation estimations due to an increased number of correct orientation estimations within the 5° threshold. Consequently, the suitability of the type predictions as prior information for the vehicle shape (research question (1), Sec. 5.1) could be demonstrated. However, the shape prior obviously reintroduces ambiguities between vehicles having opposite viewing directions.

Base+S+P: In this setting, the *position prior* is added to the probabilistic model by considering the probabilistic free-space grid map during the model alignment. The intention behind this prior term is to prevent vehicle models from being located in areas that are likely to be unoccupied according to the observations. The *position prior* thus enforces shifts and/or rotation on models that are violating the consistency between the footprint of the model and the free-space grid map. Tab. 6.6 shows that the estimation of the orientation as well as the one of the position benefits from



(a) Cumulative histogram of absolute orientation errors.

(b) Histogram of absolute orientation errors.

Figure 6.5: Histograms of absolute differences between estimated and reference orientations for the prior term variants on the KITTI training data (*easy* level).

this extension for fully visible as well as occluded vehicles. The number of joint correct position and orientation estimates ($\mathbf{t}_{75} + \theta_5$ criterion) increases by a significant margin of up 5.7% in the *easy* level. An interesting observation can be made from Fig. 6.5: Considering free-space leads to a reduction of incorrect orientation estimations mainly in the intermediate bins while the number of orientation results in the first and last bin increases compared to the *Base* model. This is an expected and natural effect of the *position prior* because, due to the symmetry properties of vehicles, it is not able to distinguish between two vehicles with opposite headings.

Init(+): In the *Init(+)* setting, the model parameters for shape and orientation are defined as the most likely parameter set derived from the prediction of the CNN for the vehicle type and the orientation. The prediction of the viewpoint and the vehicle type has already been analysed in Sec. 6.2. In this section, the quality of pose parameters that are determined by this informed initialisation is evaluated.

As can be seen from Tab. 6.6, using the *Init(+)* and the *Init(-)* initialisations leads to almost the same results for the position parameters. Consequently, the median errors for the position achieved by the two different initialisation procedures are identical (cf. Tab. 6.7). However, significant differences occur in the initialisation of the orientation. While only 18.8% of the initial models of the *Init(-)* setting are within 22.5° of the correct values in the *easy* level, this number is at 97.8% for the *Init(+)* variant. For the *moderate* and *hard* categories, the numbers are lower, but the performance pattern comparing the two initialisation variants is similar. However, with 42.0% of model initialisations that differ less than 5° from the reference values, it has to be noted that a large gap exists between coarse and fine quality of the orientation instantiations. Nevertheless, the median errors for the orientation decrease from more than 85° for the *Init(-)* variant to less than 9° for the *Init(+)* variant. Regarding the research question (1) in Sec. 5.1, the large amount of correct orientation results for the $\theta_{22.5}$ metric is considered as highly suitable for instantiation of the vehicle models for the subsequent reconstruction.

Base+S+P+O: Finally, to assess the full potential of the state priors, the *shape, position*, and *orientation priors* are jointly considered as regularisers for the state parameters during alignment. As can be seen from Tab. 6.6, the best results among all prior settings are achieved by this variant combining all prior terms. This observation underlines the value of each individual prior formulation proposed in this thesis for the model alignment. Comparing the full prior formulation to the *Base+S+P* variant shows that including the *orientation prior* in model alignment leads to huge improvements regarding the quality of the orientation estimates, demonstrating the beneficial effect of the viewpoint predictions (research question (1), Sec. 5.1). In the *easy* level, the number of correct orientations improves by 10.4 for the θ_5 metric and 19.0% for θ_{10} and $\theta_{22.5}$ metrics, respectively. For the more challenging difficulty levels, the amount of improvement is even higher. At the same time, the consideration of the *orientation prior* also leads to a slight benefit on the number of correct position estimates, which slightly improve throughout all levels of difficulties by up to 5.1% for the $\mathbf{t}_{75}$ metric and up to 2.3% for the $\mathbf{t}_{25}$ metric. Tab. 6.7 shows that the joint consideration of all priors leads to the smallest median errors for both, position and orientation throughout all levels of difficulty. The analysis of this section demonstrates the beneficial effect of incorporating the proposed state priors into the approach for model fitting, thus answering the second part of the research question (2) in Sec. 5.1.

While the values for orientation and position estimates for the coarse metrics $\mathbf{t}_{75}$ and $\theta_{22.5}$ are already fairly high with up to 98.6 and 92.1% by only considering the *Base* model and the state priors, the achieved results for the fine metrics are significantly lower. In the following section the combination of both, the full likelihood as well as the complete prior of the probabilistic model is analysed.

6.4 Analysis of the full model for vehicle reconstruction

This section contains the analysis of the complete probabilistic formulation of the energy function that is presented in this thesis (cf. Sec. 4.4), referred to as the **Full** model. The *Full* model is tested on both, the KITTI and the ICSENS data sets. The results will be investigated w.r.t. the quality of the estimated pose parameters (Sec. 6.4.1) and w.r.t. the quality of the estimated shape parameters (Sec. 6.4.2). An analysis of the limitations of the full proposed probabilistic model is conducted in Sec. 6.4.3. The investigations conducted in this section strive for answers to the questions asked in the context of the research question (3) in Sec. 5.1.

6.4.1 Evaluation of the pose

The quality of the pose parameters that is achieved on the KITTI and ICSENS data sets is analysed in this section and compared the *Base+K+W* and *Base+S+P+O* variant to asses the potential of the full formulation of the energy function.

Results on the KITTI data

Tab. 6.8 contains the pose estimation results achieved by the *Full* model on the KITTI data set including the RMS errors for position and orientation that are computed from all instances that are correct according to the respective pose metrics (the results for the *Base+K+W* and the *Base+S+P+O* are recalled for comparison). Besides, the results achieved by the *Full_Img* variant, neglecting all terms which are based on explicit 3D information, are shown to assess the importance of 3D data for the reconstruction. In addition, Tab. 6.9 shows the error median error metrics resulting from the tested variants on the KITTI data.

According to Tab. 6.8, when describing the results of the *Full* probabilistic model compared to the *Base+K+W* and *Base+S+P+O* variants, a distinction has to be made between the results for position and the results for the orientation. Counterintuitively, the amount of correct position estimates decreases in the *easy* category when combining the likelihood terms and prior terms in the *Full* model, compared to both, the *Base+K+W* and the *Base+S+P+O* settings. This decrease is less distinct for the *moderate* and *hard* categories, where the decrease of correct position estimates of the *Full* model w.r.t. the *Base+S+P+O* model is smaller, and in fact an increase of correct position estimates using the *Full* model can be observed compared to the *Base+K+W* model. The *Base+S+P+O* variant delivers the best results for the position estimates throughout all difficulty levels, with 5.5 and 8.0% more correct estimates in the *easy* category for the $\mathbf{t}_{75}$ and $\mathbf{t}_{25}$ metrics, respectively. In the *moderate* and *hard* level, the differences are smaller, with 3.0/5.4%, and 2.0/3.9% for the $\mathbf{t}_{75}/\mathbf{t}_{25}$ metrics, respectively.

In contrast, regarding the results for the estimated orientations, the *Full* model achieves the highest number of correct estimates throughout all levels of difficulty. While the numbers of correct orientation estimates for the $\theta_{22.5}$ metric achieved by the *Base+S+P+O* settings were already fairly high, with up to 98.6% in the *easy* category and up to 89.4% in the *hard* category, only small improvements of up to 1.4% are achieved by the *Full* model. The improvements created by the *Full* model compared to the *Base+K+W* and the *Base+S+P+O* models within the finer-grained θ_5 metric are up to 12.8% in the *easy* category and up to 11.7% in the *hard* category, which is remarkably large. The number of vehicle reconstructions that are correct w.r.t. both, the $\mathbf{t}_{75}$ and θ_5 metrics, are largest for the *Full* model throughout all difficulty levels. In accordance with the observations described so far, the median errors resulting from the *Full* model are slightly larger compared to the other variants regarding the position, but slightly smaller w.r.t. the orientation.

Table 6.8: Quantitative pose estimation results for the full probabilistic model and selected variants on the KITTI training set. The best achieved values for the respective metrics are printed in bold font.

		Base+K+W		Base+S+P+O		Full		Full_Img	
		[%]	$\varepsilon_{\text{rms}}^{\mathbf{t}/\theta}$	[%]	$\varepsilon_{\text{rms}}^{\mathbf{t}/\theta}$	[%]	$\varepsilon_{\text{rms}}^{\mathbf{t}/\theta}$	[%]	$\varepsilon_{\text{rms}}^{\mathbf{t}/\theta}$
easy	$\mathbf{t}_{25}$	40.7	0.15 m	**46.6**	0.16 m	38.6	0.16 m	26.7	0.16 m
	$\mathbf{t}_{50}$	73.0	0.27 m	**77.1**	0.26 m	70.3	0.27 m	52.9	0.29 m
	$\mathbf{t}_{75}$	88.5	0.36 m	**92.1**	0.34 m	86.6	0.36 m	70.8	0.40 m
	θ_{5}	84.9	2.2 °	77.3	2.5 °	**90.1**	2.1 °	81.0	2.2 °
	θ_{10}	95.2	3.1 °	95.1	3.8 °	**97.8**	2.8 °	92.5	3.3 °
	$\theta_{22.5}$	97.9	3.8 °	98.6	4.4 °	**98.9**	3.1 °	98.2	4.5 °
	$\mathbf{t}_{75} + \theta_{5}$	77.6	-	73.0	-	**79.8**	-	59.0	-
moderate	$\mathbf{t}_{25}$	38.6	0.15 m	**43.0**	0.16 m	37.6	0.16 m	25.0	0.16 m
	$\mathbf{t}_{50}$	70.6	0.26 m	**74.2**	0.26 m	69.9	0.28 m	49.8	0.29 m
	$\mathbf{t}_{75}$	85.9	0.36 m	**89.2**	0.38 m	86.2	0.36 m	67.3	0.40 m
	θ_{5}	78.7	2.2 °	72.5	2.5 °	**85.4**	2.1 °	76.6	2.2 °
	θ_{10}	88.4	3.1 °	89.8	3.8 °	**93.5**	2.9 °	88.5	3.3 °
	$\theta_{22.5}$	91.8	4.0 °	95.3	5.1 °	**96.2**	3.8 °	95.0	4.9 °
	$\mathbf{t}_{75} + \theta_{5}$	72.5	-	68.0	-	**76.5**	-	55.5	-
hard	$\mathbf{t}_{25}$	35.4	0.16 m	**38.9**	0.16 m	35.0	0.16 m	22.6	0.16 m
	$\mathbf{t}_{50}$	65.0	0.27 m	**68.9**	0.27 m	65.7	0.28 m	45.7	0.29 m
	$\mathbf{t}_{75}$	79.8	0.36 m	**83.8**	0.35 m	81.8	0.37 m	62.5	0.41 m
	θ_{5}	70.3	2.2 °	66.9	2.5 °	**78.6**	2.1 °	69.3	2.2 °
	θ_{10}	79.6	3.2 °	82.9	3.8 °	**87.0**	3.0 °	81.9	3.5 °
	$\theta_{22.5}$	83.2	4.3 °	89.4	5.4 °	**90.8**	4.2 °	89.1	5.3 °
	$\mathbf{t}_{75} + \theta_{5}$	64.7	-	61.5	-	**70.1**	-	50.0	-

Comparing the performance of the *Full_Img* variant, which does not consider the terms based on explicit 3D information, to the performance of the *Full* model, a significant decrease in the numbers of both, position and orientation metrics, is visible (cf. Tab. 6.8). The decrease of performance is also reflected by the median errors for position and orientation, shown in Tab. 6.9. While the median errors for the orientation increase by 0.3° up to 0.5° for the *Full_Img* variant, and are still smaller compared to the errors achieved by the *Base+S+P+O* setting, the median errors for the position increase by a factor up to 1.5 and are clearly the highest among all the settings shown in Tab. 6.9. This behaviour demonstrates the importance of the 3D information for the estimation of the orientation but especially for the estimation of the position.

For 90.1% of the detected and fully visible vehicles (*easy* level), the *Full* probabilistic model proposed in this thesis delivers an orientation estimate that is correct within 5° compared to the reference, while even 97.8% of the reconstructed vehicles have an orientation estimate with an

Table 6.9: Error metrics for the *Full* model and selected variants on the KITTI training set. The best achieved values for the respective metrics are printed in bold font.

		Base+K+W	Base+S+P+O	Full	Full_Img
easy	$\varepsilon^{\mathbf{t}}_{\mathrm{Med}}$ [m]	0.31	**0.27**	0.33	0.47
	$\sigma^{\mathbf{t}}_{\mathrm{MAD}}$ [m]	0.24	**0.21**	0.26	0.40
	$\varepsilon^{\theta}_{\mathrm{Med}}[^\circ]$	1.9	2.5	**1.7**	2.0
	$\sigma^{\theta}_{\mathrm{MAD}}[^\circ]$	1.8	2.5	**1.6**	2.0
moderate	$\varepsilon^{\mathbf{t}}_{\mathrm{Med}}$ [m]	0.32	**0.29**	0.33	0.50
	$\sigma^{\mathbf{t}}_{\mathrm{MAD}}$ [m]	0.26	**0.22**	0.26	0.44
	$\varepsilon^{\theta}_{\mathrm{Med}}[^\circ]$	2.1	2.7	**1.8**	2.2
	$\sigma^{\theta}_{\mathrm{MAD}}[^\circ]$	2.1	2.7	**1.7**	2.2
hard	$\varepsilon^{\mathbf{t}}_{\mathrm{Med}}$ [m]	0.36	**0.32**	0.36	0.55
	$\sigma^{\mathbf{t}}_{\mathrm{MAD}}$ [m]	0.30	**0.26**	0.29	0.50
	$\varepsilon^{\theta}_{\mathrm{Med}}[^\circ]$	2.4	3.1	**2.0**	2.5
	$\sigma^{\theta}_{\mathrm{MAD}}[^\circ]$	2.7	3.2	**2.1**	2.7

error smaller than 10°. The numbers are somewhat smaller for the occluded or truncated vehicles in the more challenging categories. In terms of correctly estimated vehicle positions, 86.7% of the detected vehicles exhibit position errors of less than 75 cm, in the *easy* level. Considering an error threshold of 50 cm, 70.3% are still correct, while this number reduces to 38.6% for the error threshold of 25 cm. These numbers slightly decrease for the *moderate* and *hard* categories. A reason for the position errors can be the fact that the position of the vehicles, being represented by the vehicle's centre points which are lying in the inside of the vehicles, are an entity which is never directly observed in the images, but instead is derived from the reconstructed 3D vehicle shape. As a matter of fact, a vehicle is never entirely visible in the image but instead, only one or two of the four vehicle sides are observed, while the remaining parts of the vehicle are averted from the camera and therefore are invisible. As a consequence, the extent of the vehicle in the viewing direction is always unobserved and ambiguous. Model based approaches as proposed in this thesis constrain and derive the full extent of the object using statistically learned shape priors. As an example, the length of a vehicle which is observed from the front is hard to determine. As, for instance, the shape of the frontal part of a compact car can be very similar to an estate car, determining the extent of the vehicle is highly ambiguous and can cause differences of 1-2 m. As the vehicle position is derived from the reconstructed model, the remaining ambiguities of object extent in the viewing direction cause errors in the position estimates, expecting errors to especially occur in the viewing direction of the camera. To verify this hypothesis, an analysis of the position estimates, distinguished by errors in lateral (across the viewing direction of the camera) and longitudinal (along the viewing direction of the camera) directions is conducted. Tab. 6.10 contains the numbers of correctly estimated lateral and longitudinal vehicle coordinates, denoted by $\mathbf{t}^{\mathrm{lat}}$ and

Table 6.10: Number of correct longitudinal and lateral position estimates of the *Full* model on the KITTI training set.

	Lateral position			Longitudinal position		
[%]	$\mathbf{t}_{25}^{\mathrm{lat}}$	$\mathbf{t}_{50}^{\mathrm{lat}}$	$\mathbf{t}_{75}^{\mathrm{lat}}$	$\mathbf{t}_{25}^{\mathrm{lon}}$	$\mathbf{t}_{50}^{\mathrm{lon}}$	$\mathbf{t}_{75}^{\mathrm{lon}}$
easy	88.0	97.3	98.8	43.3	73.4	88.7
moderate	84.1	95.5	97.9	43.4	74.3	88.8
hard	79.3	91.4	95.6	42.0	71.4	85.5

$\mathbf{t}^{\mathrm{lon}}$ where a coordinate is considered to be correct if its absolute difference from the reference is smaller than 25, 50 and 75 cm, respectively. This table shows the obvious differences between the results for the estimated lateral and longitudinal positions. While the number of estimated lateral coordinates that are within 75 cm of the reference lies between 95.6 and 98.8% for all difficulty categories, these numbers are approximately 10% lower for the longitudinal position estimates. This discrepancy becomes even larger considering 25 cm as threshold for an estimate to be counted as correct. In the lateral direction, up to 88.0% of the reconstructed vehicles exhibit a position which is correct within 25 cm. In the longitudinal direction, only half of this number is achieved. One reason for this observation can be the uncertainty of determining the depth, which increases with an increasing distance of the vehicles to the camera. A detailed investigation on the influence of the vehicle distance on the reconstruction results is presented in Sec. 6.4.3. Furthermore, the observation of obtaining worse results for the longitudinal compared to the lateral position supports the hypothesis that the errors in estimating the overall vehicle position are likely caused by the ambiguities in determining the spatial extent of the vehicles in the longitudinal direction.

To get deeper insights into the distribution of longitudinal position errors, Fig. 6.6 contains a histogram of the signed longitudinal position errors. The figure shows the histogram of errors for position estimates being closer to the camera compared to the reference in orange, and the histogram of erroneous estimates being further away from the camera in blue. For all three categories, *easy, moderate* and *hard*, an identical pattern can be observed, indicating a clear systematic effect in the longitudinal error distribution. The majority of incorrect position estimates, namely 78.8 - 83.6%, result from estimated positions that are further away from the camera compared to the true position. This effect can be caused by two possible reasons or a combination of both. Either, the vehicle is reconstructed with a shift in the viewing direction, or, as explained above, the extent of the vehicle is estimated to be too large in the direction of the vehicle side averted to the camera. Sec. 6.4.2, which provides an analysis of the results for the reconstructed vehicle shapes, supports the latter assumption.

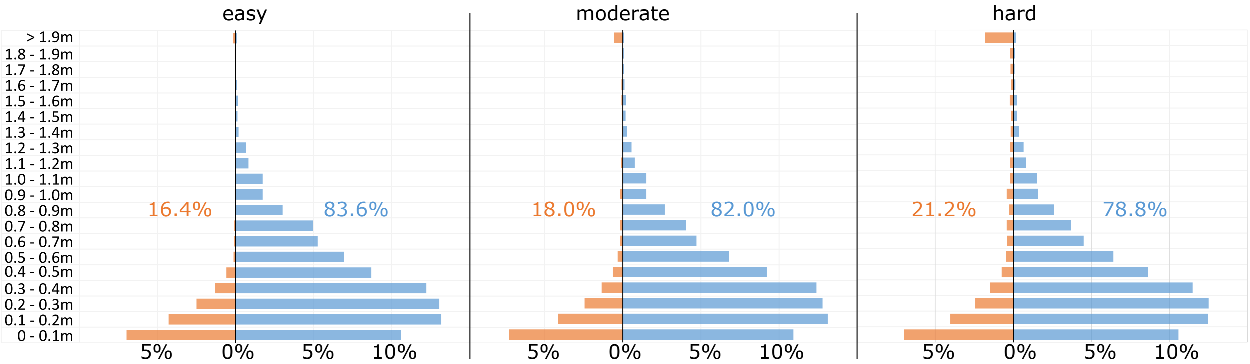


Figure 6.6: Histograms of signed longitudinal position errors resulting from the KITTI data. The orange bars demonstrate errors in the direction of the camera, the blue bars demonstrate errors in the opposed camera direction.

Results on the ICSENS data

In addition to the evaluation of the performance on the KITTI data set, the *Base+K+W*, the *Base-+S+P+O*, as well as the *Full* settings of the proposed method were also applied to the ICSENS data. Tab 6.11 contains the corresponding quantitative results that are obtained by the mentioned settings. In the context of the research objective (3) presented in Sec. 5.1, the main intention of analysing the results of the proposed method on a second and independent data set is to answer the following questions:

- Do the different components which are incorporated in the probabilistic model lead to the same pattern and the same effect as on the first data set and thus, support the previous conclusions about their potential benefits or drawbacks?
- Are the obtained quantitative results comparable to the results obtained on the first data set, potentially indicating the transferability of the proposed method to the input data and its generalisation ability?

The quality of the initialisation, which is evaluated in the *Init(+)* variant in Tab. 6.8, is comparable to the initialisation quality achieved on the KITTI data (cf. Tab. 6.6). In fact, the number of initial positions of the different metrics are slightly higher (e.g. +5.6%) for the $\mathbf{t}_{25}$ metric in the *easy* category), which is probably caused by the smaller depth uncertainties of the 3D points of the ICSENS data (cf. Tab. 5.3). The large numbers of 96.8% (*easy*) and 92.7% (*difficult*) which are already achieved by the *Init(+)* setting for the $\theta_{22.5}$ metric once again demonstrate the good performance of the *viewpoint branch* (research question (1) in Sec. 5.1), even on the ICSENS data, i.e. on data from a different domain as the one the *viewpoint branch* has been trained on.

The benefit of combining all likelihood terms with the state priors in the *Full* formulation on the results for the orientation estimation, which could be observed on the KITTI data (cf. Tab. 6.8), is

Table 6.11: Quantitative pose estimation results for the *Full* probabilistic model and selected variants on the ICSENS data set. The best values for the respective metrics are printed in bold font.

		Init(+)		Base+K+W		Base+S+P+O		Full	
		[%]	$\varepsilon_{\mathrm{rms}}^{\mathbf{t}/\theta}$	[%]	$\varepsilon_{\mathrm{rms}}^{\mathbf{t}/\theta}$	[%]	$\varepsilon_{\mathrm{rms}}^{\mathbf{t}/\theta}$	[%]	$\varepsilon_{\mathrm{rms}}^{\mathbf{t}/\theta}$
easy	$\mathbf{t}_{25}$	20.9	0.16 m	38.3	0.16 m	**48.3**	0.16 m	44.7	0.16 m
	$\mathbf{t}_{50}$	48.7	0.31 m	76.9	0.28 m	81.1	0.26 m	**82.5**	0.27 m
	$\mathbf{t}_{75}$	71.4	0.43 m	90.7	0.35 m	93.5	0.33 m	**93.9**	0.33 m
	θ_{5}	52.4	2.9 °	69.0	2.6 °	68.5	2.7 °	**73.8**	2.5 °
	θ_{10}	84.6	5.0 °	88.3	4.0 °	91.0	4.2 °	**92.2**	3.9 °
	$\theta_{22.5}$	96.8	6.6 °	94.0	5.2 °	96.7	5.1 °	**97.0**	4.7 °
	$\mathbf{t}_{75} + \theta_{5}$	37.2	-	65.5	-	65.3		**70.5**	
difficult	$\mathbf{t}_{25}$	15.8	0.16 m	37.3	0.16 m	41.4	0.16 m	**43.3**	0.16 m
	$\mathbf{t}_{50}$	38.3	0.31 m	73.7	0.28 m	75.2	0.27 m	**78.1**	0.27 m
	$\mathbf{t}_{75}$	58.7	0.45 m	87.6	0.35 m	88.7	0.35 m	**90.5**	0.34 m
	θ_{5}	47.9	2.9 °	66.2	2.6 °	64.9	2.7 °	**71.6**	2.5 °
	θ_{10}	80.2	5.1 °	84.7	4.0 °	87.6	4.2 °	**89.3**	3.9 °
	$\theta_{22.5}$	92.7	6.8 °	90.1	5.2 °	93.6	5.3 °	**94.1**	4.9 °
	$\mathbf{t}_{75} + \theta_{5}$	29.5	-	62.5	-	60.5	-	**67.4**	-

also visible from the results on the ICSENS data. The *Full* model delivers the best results, although the improvement over the other variants is less distinct compared to the results on the KITTI data. However, in contrast to the observations made from Tab. 6.8, in which the combination of likelihood and prior terms lead to a decrease in the amount of correct position estimates, the *Full* model also improves the position results on the ICSENS data. An exception is the $\mathbf{t}_{25}$ metric, where the *Base+S+P+O* setting achieves the best results for the *easy* category. Nevertheless, the overall tendency appearing from Tab. 6.11 attests the beneficial effect of the joint consideration of the proposed likelihoods and state priors.

Comparing the results achieved on the two different data sets by the *Full* model, different phenomena are observable. Regarding the results for determining the position, it is apparent that the performance on the ICSENS data reveals significant better numbers compared to the performance on the KITTI data. In the *easy* category, the difference in the amount of correct position estimates ranges from 4.5 - 11.3% for the different evaluation metrics. In the previous sections, the depth uncertainty has been identified as a strong impact factor on the performance of the position reconstruction. An explanation for the comparably better results for the position on the ICSENS data thus can be the larger baseline of the ICSENS stereo rig, and consequently, a comparably lower depth uncertainty. Besides, as described in Sec. 5.2, the references for the ICSENS data are generated based on the same disparity maps that are used in the proposed method. Potential

Table 6.12: Error metrics for the *Full* model and selected variants on the KITTI training set. The best achieved values for the respective metrics are printed in bold font.

		Init(+)	Base+K+W	Base+S+P+O	Full
easy	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.51	0.31	**0.26**	0.28
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.37	0.21	**0.19**	**0.19**
	$\varepsilon^{\theta}_{\text{Med}}[°]$	4.7	3.2	3.4	**2.8**
	$\sigma^{\theta}_{\text{MAD}}[°]$	3.8	2.9	3.0	**2.6**
difficult	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.64	0.32	0.30	**0.29**
	$\sigma^{\mathbf{t}}_{\text{MAD}}$ [m]	0.48	0.23	0.23	**0.21**
	$\varepsilon^{\theta}_{\text{Med}}[°]$	5.2	3.3	3.5	**2.9**
	$\sigma^{\theta}_{\text{MAD}}[°]$	4.2	3.2	3.2	**2.8**

errors in the disparity estimations therefore also contributed to the generation of the references and consequently may cause slightly better results on this data.

Regarding the results obtained for the orientation, the performance of the coarse viewpoint estimation is comparable for both data sets. With up to 97.0 and 98.9% of correct orientation estimates, respectively, for the $\theta_{22.5}$ criterion, the proportion of vehicles whose estimated orientation is closer to the reference than 22.5° is similar. However, the achieved numbers for the finer-grained orientation evaluation criteria are distinctly lower for the ICSENS data set, especially for the θ_5 metric, which is 18.4% and 17.7% lower in the *easy* and *difficult* levels, respectively, compared to the θ_{10} criterion. As a consequence of the worse orientation estimates, the RMS orientation errors as well as the median errors for the orientation obtained on the ICSENS data are significantly larger compared to the KITTI data. As the majority of the likelihood and prior terms of the probabilistic model are based on the predictions of the CNN, a potential reason of the decreased performance on the ICSENS data set may be given by the domain gap impacting the performance of the CNN, which is mainly trained on KITTI data and therefore might perform better on data from the same domain.

6.4.2 Evaluation of the shape

So far, only the quality of the poses that result from the vehicle reconstruction was evaluated. This section provides an analysis of the reconstructed vehicle shapes. Tab. 6.13 shows the evaluation metrics for the shape which are explained in Sec. 5.4.3 for the KITTI data and Tab. 6.14 shows the shape metrics for the ICSENS data. As explained in Sec. 5.4.3, in contrast to the reported average of absolute errors, the signed average errors give insights into potential systematics. In this context, a larger deviation of the signed errors from zero indicates an estimation of the vehicle dimensions being systematically to large (in case of negative average errors) or too small (in case of positive average errors), respectively.

Table 6.13: Shape evaluation results on the KITTI data. The dimensions of the reference bounding boxes are compared to the vehicle dimensions resulting from the reconstruction.

		Average absolute errors [m]			Average errors [m]		
		length	width	height	length	width	height
easy	Base	0.36	0.10	0.23	-0.08	-0.05	0.22
	Base+S	0.32	0.08	0.19	-0.03	-0.05	0.18
	Full	0.37	0.11	0.12	-0.23	-0.09	0.11
moderate	Base	0.39	0.10	0.23	-0.12	-0.06	0.22
	Base+S	0.32	0.09	0.19	-0.06	-0.06	0.18
	Full	0.37	0.11	0.13	-0.23	-0.10	0.11
hard	Base	0.40	0.10	0.22	-0.13	-0.06	0.22
	Base+S	0.33	0.09	0.18	-0.07	-0.06	0.18
	Full	0.38	0.11	0.13	-0.23	-0.10	0.11

In Tab. 6.13, the results of the *Base*, *Base+S* and the *Full* models are shown, to assess the influence of the *shape prior* term on the shape estimation, as well as the final results of the entire probabilistic model. Comparing the average absolute errors of the *Base* variant and the *Base+S* variant, the consideration of the proposed *shape prior* leads to significantly better results for the vehicle length and height, which are improved by up to 7 and 4 cm, respectively. This observation underlines the benefit of the category-aware ASM and *shape prior* presented in this thesis compared to using the commonly applied regularisation of the shape by penalising deviations from the mean ASM shape, as it is done in the *Base* variant. The consideration of the *Full* model increases the errors in length but gives the distinctly best results for the object height. As a general observation, it can be noted that the errors resulting for the different difficulty levels are of almost the same size, probably caused by the applied ASM as shape prior, which acts as a regulariser on the shape independently from the observability of the vehicles. Furthermore, according to the results, the largest errors in the estimated vehicle dimensions occur for the vehicle length, followed by the height. The smallest errors are achieved for the vehicle width. One potential reason for the smaller errors for the estimated vehicle width can be that the variations of the width of different vehicles are lower compared to variations in the length and height and the maximum possible error is therefore already smaller.

Inspecting the signed average errors for the full model to see potential systematic errors, it can be noted that the magnitude of the average error for the height is almost as large as the average absolute error, indicating that the height is systematically estimated too small. A potential explanation for this is that no keypoint of the ASM is explicitly defined for the point of the vehicle with the maximum height. The curvature of the roof is likely to cause the actual height of the vehicle to be larger than the maximum height inferred from the keypoints. As a consequence, the height derived from the ASM reconstructions will be smaller compared to the 3D bounding box height, which covers the full vehicle extent. The second observation that can be made from the signed average errors is that the error of length and width estimates also have a high magnitude and a negative sign, indicating that mostly, the dimensions of the vehicles are estimated too large. In combination with the observation made earlier in Sec. 6.4.1 in Fig. 6.6, which revealed that the position estimates are systematically too far away from the camera, two possible conclusions can be drawn. Either the ambiguities between object size and object distance cause the reconstructed vehicles to be scaled and shifted along the image ray as a whole, or the distance of the vehicle to the camera is estimated correctly and the errors in the position estimates result from estimating the vehicles to be too long in the viewing direction of the camera. To check these hypotheses, Fig. 6.7 shows the relation between average absolute errors of the vehicle size and the absolute position error. In this context, the average absolute errors of the vehicle size is computed from the absolute differences between the diagonals of the footprint of the reference and the estimated vehicle bounding boxes, because the diagonal enables a combined consideration of both aspects, the length and the width of a vehicle. The relation between the errors in vehicle size and position is clearly visible, as an increasing error in the position comes along with an increased error of the estimated vehicle size. This relation might support the conclusion that errors in estimating the vehicle's size are potentially responsible for a part of the position errors. As could be seen from Fig. 6.6, the number of vehicles with position errors larger than 1.2 m is very low. As a consequence, the small number causes that the average values for these categories are not representative and that already few outliers distort the average dimension errors, leading to the irregular behavior of the graph in Fig. 6.7 appearing for position errors larger than 1.2 m.

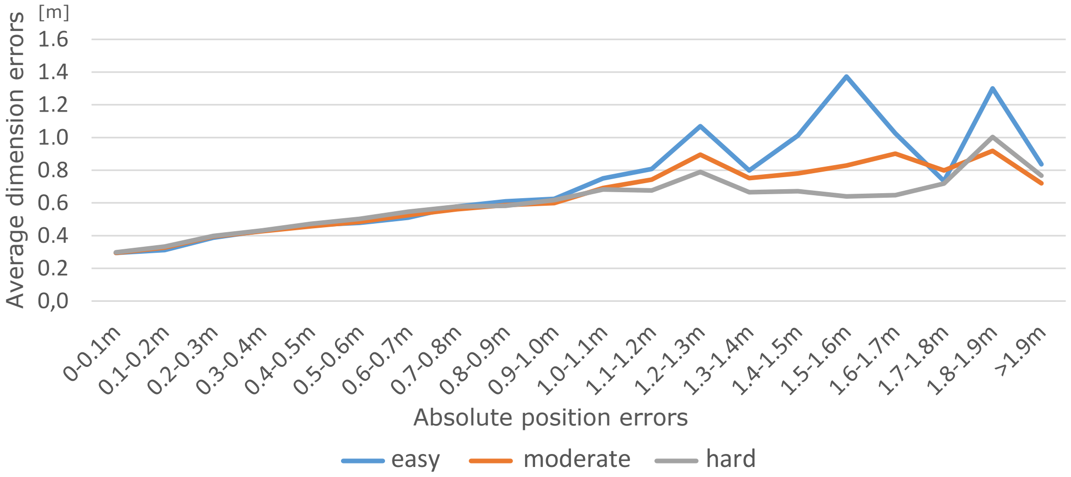


Figure 6.7: Relation between errors in the estimated vehicle footprint dimension and position errors. The diagonals of the bounding boxes' footprints are used to compute average dimension errors from the absolute differences between the diagonals.

Regarding the dimension errors obtained on the ICSENS data (cf. Tab. 6.14), the pattern and magnitude of average absolute errors is comparable to the one obtained on the KITTI data set. The

Table 6.14: Shape evaluation results on the ICSENS data. The dimensions of the reference CAD models and the reconstructed ASM are compared. Besides, the keypoint based RMS error is computed from the euclidean distances of corresponding keypoints.

		Average absolute errors [m]			Average errors [m]			Keypt RMS error
		length	width	height	length	width	height	$\varepsilon_{\text{rms}}^{\mathcal{K}}$ [m]
easy	Base+K+W	0.44	0.13	0.21	-0.10	-0.04	-0.13	0.27
	Base+S+P+O	0.38	0.09	0.14	0.01	0.03	-0.01	0.24
	Full	0.45	0.13	0.23	-0.11	-0.05	-0.14	0.27
difficult	Base+K+W	0.43	0.12	0.20	-0.07	-0.02	-0.09	0.26
	Base+S+P+O	0.38	0.10	0.14	0.04	0.04	0.00	0.24
	Full	0.42	0.12	0.20	-0.08	-0.03	-0.10	0.26

largest discrepancies between reference and estimation occur w.r.t. the length of the vehicles, while the vehicle width is estimated with the smallest errors. Comparing the different tested variants, the smallest errors are obtained by the *Base+S+P+O* setting, in which only the *3D likelihood* under consideration of the state priors is used, which again is an indicator for the benefit of the proposed *shape prior* term. Remarkably, the signed average errors of this variant are close to zero, indicating a symmetric distribution of the dimension estimates around the reference. Comparing the results of the *Full* model to the *Base+S+P+O*, average absolute errors are slightly increased. This observation is consistent with those made for the KITTI dataset and allows to draw similar conclusions. The average keypoint error is consistently smallest (24 cm) for the *Base+S+P+O* variant and slightly increases for the *Full* model. As the results for the dimension errors imply, the largest impact on the keypoint error is expected to result from errors in the longitudinal vehicle direction.

6.4.3 Analysis of further aspects

While its overall performance has been investigated and analysed in the previous sections, a more detailed analysis will be conducted to highlight the strength and limitations of the proposed reconstruction approach (research question (3) in Sec. 5.1) in the following paragraphs. The influence of the distance and viewpoint in and under which the vehicle is observed on the quality of the reconstruction is analysed in this section. Tab. 6.15 shows the performance results of the *Full* model differentiated by the vehicle distances and viewpoints. The depth uncertainties σ_x of a stereo-reconstructed 3D point in the considered distances, assuming an uncertainty of the disparity σ_{disp} of 1 [px] are also shown in Tab. 6.15. The angles under which the vehicles are observed are grouped into five viewpoints, namely *front, front-side, side, back-side*, and *back*. The definition of the viewpoint groups was shown in Fig. 5.5.

Table 6.15: Results for pose estimation using the *Full* model on the KITTI data. This table reports results of the pose evaluation metrics depending on distance and viewpoint of the vehicles by distinguishing between different distance (left) and viewpoint categories (right), respectively.

		Vehicle distance				Vehicle viewpoint				
		5 - 10 m	10 - 15 m	15 - 20 m	>20 m	front	front-side	side	back-side	back
	σ_x [cm]	6 - 26	26 - 58	58 - 103	> 103					
easy	$\mathbf{t}_{25}$ [%]	60.4	54.7	28.0	17.3	52.8	34.6	15.7	35.5	34.3
	$\mathbf{t}_{50}$ [%]	93.9	89.7	63.2	40.0	86.5	62.2	59.7	70.7	64.3
	$\mathbf{t}_{75}$ [%]	99.0	97.6	84.6	67.6	97.2	82.6	81.8	84.8	82.6
	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.21	0.23	0.40	0.58	0.24	0.37	0.43	0.32	0.36
	θ_5 [%]	94.2	94.1	89.2	83.7	96.8	88.4	70.4	88.0	88.9
	θ_{10} [%]	99.3	99.1	97.6	95.4	99.5	97.9	89.3	95.7	97.9
	$\theta_{22.5}$ [%]	99.7	99.6	98.8	97.6	99.6	99.3	95.6	96.8	99.1
	$\varepsilon^{\theta}_{\text{Med}}$ [°]	1.5	1.5	1.7	2.1	1.2	1.7	3.2	1.8	1.8
moderate	$\mathbf{t}_{25}$ [%]	61.0	50.8	26.7	17.6	50.1	34.3	11.1	32.9	36.0
	$\mathbf{t}_{50}$ [%]	93.6	86.6	61.2	42.5	83.6	63.5	48.3	65.5	67.7
	$\mathbf{t}_{75}$ [%]	98.9	95.9	83.7	67.8	95.4	81.6	73.1	82.9	84.7
	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.20	0.25	0.42	0.57	0.25	0.37	0.52	0.37	0.35
	θ_5 [%]	91.3	89.2	83.7	78.2	94.8	82.2	55.4	74.1	88.0
	θ_{10} [%]	96.9	95.6	92.5	89.5	97.6	91.2	74.4	84.7	96.9
	$\theta_{22.5}$ [%]	98.2	97.7	95.6	93.7	98.0	94.9	85.2	91.2	98.5
	$\varepsilon^{\theta}_{\text{Med}}$ [°]	1.5	1.6	1.9	2.3	1.3	1.9	4.0	2.5	1.8
hard	$\mathbf{t}_{25}$ [%]	59.2	46.1	24.9	16.3	49.4	33.8	9.3	28.6	35.2
	$\mathbf{t}_{50}$ [%]	92.1	80.3	57.2	39.2	82.2	62.4	38.5	58.0	66.4
	$\mathbf{t}_{75}$ [%]	97.5	90.9	78.4	63.3	94.3	79.2	58.1	75.4	83.3
	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.21	0.28	0.44	0.60	0.26	0.37	0.63	0.42	0.35
	θ_5 [%]	87.8	81.7	76.5	70.6	92.9	75.2	41.9	60.8	86.5
	θ_{10} [%]	93.7	89.1	85.5	81.5	95.9	84.7	56.6	71.4	95.5
	$\theta_{22.5}$ [%]	95.9	92.6	89.4	86.4	96.7	89.9	66.3	79.0	97.6
	$\varepsilon^{\theta}_{\text{Med}}$ [°]	1.6	1.8	2.1	2.6	1.3	2.2	7.1	3.4	1.8

Influence of the vehicle distance

Tab. 6.15 shows that the distance of the vehicle from the camera strongly affects the ability of correctly estimating the vehicle's position and also affects, although less strikingly, the quality of orientation estimates. In the *easy* level, it can be noticed that while the percentage of correct position estimates lies at 99.7% for the $\mathbf{t}_{75}$ criterion and at 60.4% for $\mathbf{t}_{25}$ for vehicles in a distance between 5 and 10 m, the percentages decrease to 67.6% and 17.3% for vehicles further away from the camera than 20 m. Accordingly, the median error of the position estimates also increases drastically almost by a factor of three from 21 cm for vehicles in a distance between 5 and 10 m to 58 cm for vehicles being more distant than 20 m. The numbers for the *moderate* and *hard* categories are comparable and show the same pattern. It can be assumed that the increasing depth uncertainty of distant 3D points is responsible for this effect. While an error of 1 [px] in the disparity leads to uncertainties of 6-25 cm for triangulated 3D points in a distance of 5-10 m, the uncertainty is already 103 cm for points in a distance of 20 m. As a consequence, the ability to precisely estimate the position decreases with an increasing distance of the vehicle from the camera.

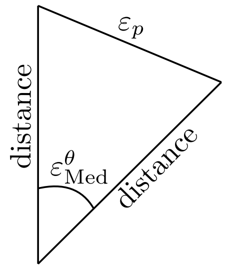
While an increasing distance of the vehicle also negatively effects the estimation of the orientation, the influence is less distinct compared to the effect on the position estimates. In all categories (*easy, moderate*, and *hard*), the ability of estimating the precise orientation within the θ_5 criteria is affected most by the vehicle distance, while the number of correct orientation estimates considering the $\theta_{22.5}$ metric is still relatively high (86.4%) for vehicles at a distance > 20 m even in the *hard* difficulty level. However, the average orientation error of reconstructions from vehicles further away from the camera than 20 m is more than twice as large as for vehicles at a distance of 5-10 m.

In order to draw a comparison between the quality of the results for the orientation and for the position of the vehicles, the *perpendicular error* ε_p resulting from the median orientation errors $\varepsilon^{\theta}_{\mathrm{Med}}$ in dependency on the distances can be computed (cf. the figure in Tab. 6.16) and can be compared to the median errors of the position $\varepsilon^{\mathbf{t}}_{\mathrm{Med}}$ achieved in the respective distances. Tab. 6.16 uses the median errors for the orientation achieved on the KITTI data set in the distinguished intervals for the distances from Tab. 6.15 to compute the perpendiculars ε_p and in order to compare them to the median errors for the position achieved on the KITTI data. As can be seen from Tab. 6.16, in a distance larger than 10 m, the effect of the obtained median errors for the orientation on the perpendicular is larger than the median errors of the position for all difficulty levels. In a distance of 20 m, ε_d is almost twice as large as $\varepsilon^{\mathbf{t}}_{\mathrm{Med}}$. This behaviour can for instance be relevant in the context of applications related to collaborative autonomous driving, in which the determined pose of the vehicles is introduced as *vehicle to vehicle* (V2V) observations for the task of collaborative positioning (Knuth and Barooah, 2009).

Influence of the vehicle viewpoint

The right-hand side of Tab. 6.15 shows the evaluation of the vehicle positions and orientations depending on the viewpoint under which the vehicle is observed. Regarding the ability of determining the orientation in dependency on the viewpoint, a clear pattern is observable. The largest numbers of correct orientation estimates are achieved for vehicles observed from the front. The second best performance is achieved for vehicles having the opposite viewpoint (i.e. back). A strikingly lower amount of correct orientation estimates for both, the coarse as well as the fine evaluation metrics, is obtained for vehicles observed under the side view. This effect is also reflected by the viewpoint-

Table 6.16: In order to compare the results achieved for the orientation with the results achieved for the position, the perpendicular ε_p is computed using the median orientation errors of Tab. 6.15.



		Vehicle distance			
		5 - 10 m	10 - 15 m	15 - 20 m	>20 m
easy	$\varepsilon^{\theta}_{\text{Med}}$[°]	1.5	1.5	1.7	2.1
	ε_p [m]	0.13 - 0.26	0.26 - 0.39	0.45 - 0.59	> 0.92
	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.21	0.23	0.40	0.58
moderate	$\varepsilon^{\theta}_{\text{Med}}$[°]	1.5	1.6	1.9	2.3
	ε_p [m]	0.13 - 0.26	0.28 - 0.42	0.50 - 0.66	> 1.00
	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.20	0.25	0.42	0.57
hard	$\varepsilon^{\theta}_{\text{Med}}$[°]	1.6	1.8	2.1	2.6
	ε_p [m]	0.14 - 0.28	0.31 - 0.47	0.55 - 0.73	> 1.13
	$\varepsilon^{\mathbf{t}}_{\text{Med}}$ [m]	0.21	0.28	0.44	0.60

wise median orientation errors, which are up to almost five times higher for vehicles belonging to the side category compared to vehicles being observed from the front.

A very similar pattern is observable in the results achieved for the position estimates. According to Tab. 6.15, the most favourable viewpoint to determine the vehicle's position is the front view. The results show that the biggest difficulties in estimating the position occur when the vehicle is observed from a side view.

An attempt to explain the observed behaviour could follow the argumentation that the front face of a vehicle contains the largest geometrical variations and, consequently, delivers a more suitable geometry for the model fitting based reconstruction. While the front face of a vehicle contains parts like the grill, the engine cover, and the wind shield, and therefore exhibits textural and geometrical variations due to the diversity in depth of the different parts, the side face of a vehicle resembles a planar surface instead. These textural and geometric properties potentially affect the quality of object reconstruction.

6.5 Comparison to related methods

This section conducts a comparison of the results obtained by the proposed method to the results obtained by related methods from the literature in order to answer the last question asked in the research objective (3) in Sec. 5.1. Tab. 6.17 shows the results of related methods on the KITTI test data obtained from evaluating the results of the proposed method on the official KITTI benchmark[1], compared to the results achieved on the same test data by the method proposed

[1] http://www.cvlibs.net/datasets/kitti

Table 6.17: Comparison to related methods based on the performance of the KITTI test set. The *orientation score* (OS) allows a mere comparison of the orientation estimates without the impact of the detection performance.

	easy			moderate			hard		
Method	AOS	AP	OS	AOS	AP	OS	AOS	AP	OS
Pepik et al. (2015)	82.15	79.09	96.28	66.72	63.58	95.29	49.01	46.59	95.06
Chen et al. (2015)	93.04	91.44	98.28	88.64	86.10	97.13	79.10	76.52	97.73
Chen et al. (2016)	92.33	91.01	98.57	88.66	86.62	97.69	78.96	76.84	97.31
Xiang et al. (2017)	90.81	90.67	99.84	89.04	88.62	99.52	79.27	78.68	99.25
Ku et al. (2019)	93.63	93.29	99.63	88.50	87.45	98.81	73.36	72.26	98.50
Manhardt et al. (2019)	76.56	75.32	98.38	70.16	68.14	97.12	61.15	58.98	96.45
Ours (Full)	78.17	77.21	98.77	47.70	46.58	97.65	36.90	35.70	96.75

in this thesis. As discussed in Sec. 5.4.4, the metrics applied by the KITTI benchmark for pose evaluation are coupled with the performance of detection. In order to evaluate how the method performs solely on orientation estimation, Mousavian et al. (2017) proposed the *Orientation Score* (OS) metric which factors out the 2D detector performance by computing the ratio between AOS over AP.

As can be seen from Tab. 6.17, our results for AP, which measure the quality of the vehicle detections, are significantly lower compared to most of the other methods. The reason for this has already been discussed in this thesis (cf. Sec. 6.1) and is caused by the fact that only the visible vehicle parts are detected by the approach applied in this thesis, while the reference bounding boxes of the KITTI benchmark comprise the entire objects. However, regarding the OS, our method is en par with or outperforms most of the shown related methods in the *easy* and *moderate* categories. Compared to the related works, regarding the *hard* category, our methods reveals a larger decrease of the OS. Throughout all difficulty levels, the best results for OS are achieved by the methods presented by Xiang et al. (2017); Ku et al. (2019). However, it has to be noted that the different CNN architectures which are presented in their respective work for pose estimation were trained using the entire KITTI training data set of almost 8000 images from the same domain, while the approach presented in this thesis was trained on only 260 images from the KITTI training set.

In order to compare the evaluation metrics developed in the context of this thesis to results achieved by other state-of-the-art approaches, we make use of the detection and pose estimation results of (Mousavian et al., 2017), which are provided by the authors for a subset of 3799 images from the KITTI training data. Tab. 6.18 contains the comparison of results of the proposed *Full* model and the results from Mousavian et al. (2017) on the same set of images. The images that were used for training in this work were excluded from the evaluation of both, the results of Mousavian et al. (2017) and ours. Our method significantly outperforms Mousavian et al. (2017) in terms of the

Table 6.18: Comparison of the proposed method to the results achieved by Mousavian et al. (2017) based on the evaluation metrics proposed in this thesis.

	in [%]	$\mathbf{t}_{25}$	$\mathbf{t}_{50}$	$\mathbf{t}_{75}$	θ_5	θ_{10}	$\theta_{22.5}$	$\mathbf{t}_{75} + \theta_5$
easy	Mousavian et al. (2017)	19.5	38.4	55.1	87.1	94.9	**98.9**	54.8
	Ours (Full)	**42.7**	**75.2**	**89.5**	**88.7**	**97.0**	98.5	**88.5**
mod.	Mousavian et al. (2017)	16.8	34.0	49.2	78.7	90.1	**96.1**	48.1
	Ours (Full)	**39.3**	**72.4**	**87.8**	**83.5**	**92.3**	95.6	**85.8**
hard	Mousavian et al. (2017)	13.9	29.1	42.9	68.6	80.2	88.1	40.4
	Ours (Full)	**35.5**	**66.9**	**81.9**	**75.5**	**84.7**	**89.0**	**78.6**

number of correct position estimates, especially for the fine evaluation metric $\mathbf{t}_{25}$, by a factor of up to 2.5. However, for a fair comparison it has to be noted, that in (Mousavian et al., 2017) no stereo information is used. Regarding the orientation estimates, while the results for the coarse level of $\theta_{22.5}$ are approximately at par, the probabilistic model of this thesis leads to significantly better results for the finer evaluation metrics, particularly for the *moderate* and *hard* categories. As a consequence, the number of vehicle reconstructions that are considered as correct in both, position and orienation, obtained by our approach is considerably larger.

6.6 Discussion

This section provides a discussion of the results described earlier in this chapter in order to evaluate the accomplishment of the research objectives introduced in Chapter 1. This thesis presented an extensive probabilistic model for the ill-posed task of reconstructing vehicles in 3D from stereo imagery and from information derived from the images by the proposed CNN. The Bayesian model comprises different likelihood formulations and prior terms, jointly incorporating 3D and 2D information in the overall model. The core of the experimental evaluation aimed at answering the research questions raised in Sec. 5.1 w.r.t.

- the performance of the developed CNN and the suitability of its predicted probability distributions for the derivation of likelihoods and state priors (research question (1)),
- the synergistic effect of fusing different likelihoods and the benefit of the established prior terms as regularisers for the ill-posed problem (research question (2)),
- the performance, sensitivity, limitations and generalisation capability of the proposed method (research question (3)).

For the sake of clarity, this section discusses these questions by reviewing the likelihood terms (Sec. 6.6.2) and the state priors (Sec. 6.6.1), as well as the full model (Sec. 6.6.3) and the applied inference strategy (Sec. 6.6.4).

6.6.1 Likelihood terms

3D likelihood: The *3D likelihood* is formulated based on the target to minimise the distance between the 3D stereo-reconstructed vehicle point cloud and the 3D ASM. It is used as a baseline for the evaluation against which the benefits of including additional terms are assessed. The missing information about the correspondences of the 3D points to their exact point-wise counterpart on the vehicle surface required a heuristic association of point-to-model-surface by establishing correspondences between the 3D points and the closest point on the model. This strategy, coupled with the rather noisy stereo-reconstructed point cloud and the symmetry properties of vehicle shapes, leads to the expectation for the *3D likelihood* to be highly sensitive to the initialisation and ambiguous in estimating the orientation due to the shape symmetries. The results reported in Tab. 6.4 and Fig. 6.3 verify the latter expectation, because the majority of erroneously estimated orientations exhibit an opposite viewing direction compared to the true one. However, as can also be seen from the results, the applied Monte-Carlo based technique for optimisation achieves a high level of robustness w.r.t. initialisation, because the majority of even strongly incorrect initial orientations are corrected, omitting the incorrect estimates caused by the ambiguities of the *3D likelihood*. Yet, the results show that the predominate strength of this likelihood is the estimation of the position which already achieves more than 87% of correct estimates for the coarsest metric and a median error of 31 cm.

Keypoint likelihood: This likelihood performs the model fitting based on the probability of the configuration of backprojected model keypoints derived from the keypoint probability maps. In contrast to the previously described *3D likelihood*, the keypoints describe physically defined and unique points of the vehicle and therefore allow a defined association between the model and the image entities. As a consequence, the expected effect of this term is mainly the reduction of the orientation ambiguities that were exposed in the *3D likelihood*. As shown by the experimental results, the consideration of the *keypoint likelihood* leads to a significant improvement of the orientation estimates, while the effect on the position is only small.

Wireframe likelihood: Similar to the *keypoint likelihood*, the *wireframe likelihood* is based on the probability of the model backprojection, but based on its wireframe and the probability for the occurrence of wireframe edges inferred from the image. Arguably, line-wise features deliver a more stable and more comprehensive representation for 3D object reconstruction compared to point-wise features. This property possibly explains the significantly superior results, particularly w.r.t. the orientation, achieved by the *wireframe likelihood* compared to the *keypoint likelihood*. While early approaches for vehicle reconstruction made use of generic gradient based edge detections, entailing problems w.r.t. the dependency on suitable gradients and the establishment of correct edge-to-model associations, recent work mostly utilise point-like features for model fitting. This thesis revives the idea of using edge-wise features and proposes a deep learning based approach for the extraction of specific wireframe edges and shows superior results using the edge-wise representation compared to using keypoints for the object reconstruction.

The likelihoods using the observations generated by the CNN provide insights into research question (1) by demonstrating the good suitability of the predicted keypoint and wireframe heatmaps for model fitting. Furthermore, fusing all proposed likelihoods within the probabilistic model is able to slightly improve the pose estimation, as results reported in Tab. 6.4 have shown. Regarding research question (2), this observation verifies the benefit of combining different terms in the likelihood, as proposed in this thesis, in order to fuse different 2D and 3D features for vehicle reconstruction.

6.6.2 State priors

Shape prior: Estimating the entire shape of objects from images is a highly ambiguous task as the complete extent of the object is never observed due to the self-occlusion of the object parts averted from the camera. For this reason, for the reconstruction of vehicles, shape priors are often learned from a representative set of vehicle examples to constrain the shape estimation. Instead of applying a regulariser as a shape prior that penalises variations from the statistical average shape, which is common practice in the literature, a finer-grained prior was proposed in this thesis, learning and considering different modes in the statistical distribution of the present shape deformations for vehicles. A prediction of the vehicle type by a dedicated branch of the proposed CNN delivered the prior information to constrain the parameter space for the shape. As can be deduced from the empirical results that have been presented in Tab. 6.6 and Sec. 6.4.2, the developed category-aware prior delivers improved results in terms of the estimated vehicle shape as well as for the estimated vehicle positions. Both parameters are closely related, as the position, represented by the centre point of the vehicle model, is directly correlated with the vehicle dimensions and thus, to the vehicle shape.

Position prior: To apply constraints to the vehicle position, a probabilistic free-space grid map has been proposed in this thesis to prevent vehicle reconstructions from being located in free areas in the scene. Although this term introduces prior information about the vehicle positions, the quantitative benefit obtained in the empirical evaluation of the position is relatively small (cf. Tab. 6.6). Instead, considering the *position prior* leads to significant improvements of the orientation estimates. Errors in the orientations of vehicles reconstructed solely on the basis of the *3D likelihood* can be compensated by the *position prior*, i.e. by considering the violation of the free-space constraints in the probabilistic model.

Orientation prior: As mentioned above, the reconstruction based on the *3D likelihood* is affected by the symmetry of vehicle shapes. Consequently, for a relatively large number of vehicles, the orientation is estimated to be opposed to the correct one by 180°. As can be seen from Fig. 6.5, the *orientation prior* helps to drastically reduce this number of reverse vehicle reconstructions. The incorporation of the prior distribution for the orientation leads to impressive results for vehicle reconstructions, that are correct within the coarse range of 22.5°. However, a significant drop of precisely estimated orientations that are correct within 5° is noticeable, caused by the limited granularity of the discretisation of the viewpoint classes which are distinguished to derive the probability distribution for the orientation.

To sum up, introducing the prior terms into the model fitting procedure leads to significant improvements of shape, position and orientation estimates, underlining the suitability of the type and viewpoint predictions by the CNN and demonstrating the value of the proposed state priors as regularisation of the state parameters (research question (1) and (2)). Yet, the discretisation of the viewpoint constitutes a natural limit of the achievable precision for the orientation estimates.

6.6.3 Full model

Combining both, the likelihood terms as well as the state priors in the *full* probabilistic model delivers the best results in estimating the vehicles' orientations. Especially the precision of the orientation estimates benefits from combining likelihoods and state priors, which answers the research question (2) by revealing the desired synergy between the terms proposed in this thesis (cf. Tabs.6.8 and 6.11). Also for the position estimates the benefit of the joint consideration of likelihood and prior terms in the *full* probabilistic model was visible for the ICSENS data, indicating the generalisation capability of the proposed method (research question (3)). However, a slight decrease of correct position estimates is noticeable for the KITTI data when the image based likelihoods are considered additionally to the *base* and prior formulation.

On the basis of the results that have been achieved in the context of the empirical evaluation, a set of general conclusions and limitations of the proposed method can be derived in order to provide answers to research question (3). While more than 90% of the vehicle orientations can be estimated in the *easy* category with an error smaller than 5° and an average error of less than 2°, the number of correct position estimates decreases drastically depending on the desired precision. While almost 87% of vehicles are correct within 75 cm, the percentage decreases to 70.3% for a 50 cm and to less than 39% for a 25 cm error tolerance. The analysis of Tab. 6.15 shows that the impact of the vehicle distance to the camera is much higher on the results for the position, compared to the results of the orientation. The increasing depth uncertainty with increasing object distance clearly constitutes a limiting factor for the pose estimation, especially for the determination of the position. A second limiting factor for a successful model fitting is the degree of truncation and occlusion of the vehicles. The quantitative analysis has shown the impact of different occlusion levels on the quality of pose estimates (cf. Tabs. 6.8 and 6.11). Particularly the orientation estimates suffer from increasing object occlusions, while the impact on the position estimates is also noticeable, though less distinct.

While most investigations were conducted on the KITTI data set, the proposed method has also been tested on the ICSENS data set. One reason for doing so was to analyse the generalisation capability of the approach (research question (3)). Although the proposed CNN producing the intermediate information which is incorporated in the probabilistic model has mainly been trained on data from the KITTI benchmark, the results for the pose estimation on the ICSENS data are of a similar quality as the results on the KITTI data. In fact, the position estimates even exceed the results achieved on the KITTI data. Having identified the depth uncertainties as one limiting factor on the vehicle reconstructions, it has to be noted for a fair comparison that the stereo setup used for the acquisition of the ICSENS data exhibits a significantly larger baseline and therefore better properties of the depth estimations. The results for the orientation estimates obtained for the coarse evaluation metrics are also comparable to the numbers achieved on the KITT data, indicating a certain degree of insensitivity to the domain gap and the generalisation capability of the probabilistic model. In this context, it has to be pointed out that the lighting conditions and consequently, the illumination properties in the images partly exhibit strong differences between the KITTI and ICSENS data. Yet, the results obtained for the coarse evaluation metrics appear to be unaffected by these differences, indicating the value of learning an illumination insensitive extraction of keypoints and especially of the wireframe edges. However, the accuracy of the obtained orientation estimates is distinctly lower comparing the average error of 3.9° obtained on the ICSENS data for the orientation estimates that are correct within 22.5° to the 2.2° average error achieved on the KITTI data (for the *easy* level). The lower accuracy is also reflected by the distinctly smaller number of orientation estimates which are within 5° of the correct values, which amounts to 70.3 in case of the ICSENS data compared to 90.1% achieved on the KITTI data (cf. Tabs. 6.8 and 6.11). The comparison of the results achieved by the proposed method to the results achieved by related work demonstrated that the developed approach is competitive to state-of-the-art methods.

6.6.4 Inference

In order to optimise the objective function derived from the probabilistic model, due to its non-convex and non-continuous nature, an iterative Monte-Carlo based optimisation has been established in this thesis, desired to be insensitive to the parameter initialisation and local optima. As the empirical results have shown (cf. in Fig. 6.3), the inference procedure is able to derive correct pose estimates even if the initialisation is completely wrong, which answers the research question asked in (3) w.r.t. the sensitivity of the proposed method to the initialisation. In fact, while at least a coarse initialisation for the position is required, no initial information regarding the viewpoint is needed at all in order to achieve a successful inference. In contrast to related work, where often mature approaches are employed to derive initial pose estimates which are refined afterwards (Zia et al., 2013; Engelmann et al., 2016), the method proposed in this thesis has no demands on good pose estimates as start values.

7 Conclusion and outlook

This thesis proposed a method for the 3D reconstruction of vehicles from street-level stereo images. This chapter draws conclusions from the presented approach and its experimental evaluation and provides an outlook to promising future directions of work.

Inspired by approaches based on shape priors, a deformable object model is learned in this thesis in order to fit it into the observations for the derivation of the state parameters for each vehicle. In this context, this thesis addressed the proposal of an improved shape model for the representation of vehicles. Furthermore, the development of a CNN has been addressed for the derivation of valuable observations and prior information for the task of model fitting. The formulation of a new Bayesian model for the vehicle reconstruction has been proposed, in which the likelihood and the prior are factorised by individual terms, fusing the different observations and prior information.

The commonly used ASM as deformable shape prior represented by a mean object shape and a set of principal components to encode the deformations has been extended in this thesis so that a shape model is learned for each of a set of vehicle types. In combination with a prediction of the vehicle type using the proposed CNN, the new shape model allows a finer-grained shape representation which leads to improved results for the estimated vehicle shape and the vehicle position compared to the standard representation of the shape prior. In addition to the shape model, prior terms for the vehicle position and its orientation have been introduced to constrain the parameter space. To establish the position prior, the probabilistic derivation of free-space from the stereo observations has been introduced in this work. In order to derive prior information on the vehicle's orientation, the presented CNN predicts a probability distribution for the vehicle's viewpoint using a newly proposed hierarchical classifier structure. Both the priors for orientation and position have shown to be valuable contributions to the reduction of pose ambiguities for the vehicle reconstruction.

The likelihood of the proposed Bayesian model is factorised by three individual terms, fusing the observations of different entities, such as keypoints, wireframes and 3D points. To this end, the standard representation of the ASM based on keypoints has been enriched by defining an edge topology to represent a vehicle wireframe, and by defining a triangulated mesh to represent the surface of the model. Both representation deliver valuable cues for the derivation of observation likelihoods. This thesis proposed to use the developed CNN in order to derive probability maps for keypoints and for wireframe edges, the latter being introduced as new feature in this thesis, to be used as observations carrying semantic information. A keypoint and wireframe likelihood is formulated based on the predicted probability maps and the backprojected model keypoints and wireframes, respectively. In addition, reconstructed 3D points are used to derive a likelihood based on the consistency of the point cloud with the model surface. While the explicit 3D information incorporated in the 3D likelihood has proven to be particularly beneficial for the position estimation, the 3D points, however, lack semantic information, which leads to ambiguities in estimating the orientation, an effect which is also enforced by the potentially symmetric shape of vehicles. By additionally incorporating a keypoint- and wireframe based likelihood, both of them incorporating the observations inferred by the proposed CNN, helps to reduce the aforementioned ambiguities w.r.t. the orientation. The benefit of the individual likelihood and prior terms and the synergistic effect achieved from their joint consideration in the developed probabilistic model for vehicle reconstruction could be demonstrated in the empirical evaluation.

For optimisation, an iterative Monte-Carlo based approach has been established, which is able to deal with non-convexities and non-continuities of the objective function. Besides, the proposed inference procedure turns out to be insensitive to erroneous initialisations. The good quantitative results achieved by the proposed method on two different data sets with average errors better than 3° for the orientation and about 30 cm for the position, demonstrate the success of the approach for the 3D vehicle reconstruction from stereo images w.r.t. the research objectives raised in Chapter 1. A comparison to related work shows that the performance of the proposed approach is on par with current state-of-the-art methods.

However, limitations of the model exist. They are related to the occurrence of larger occlusions or vehicles observed under unfavourable viewing conditions such as larger distances, which impairs the ability of precise reconstruction. Occlusions of a vehicle caused by other objects cause a reduction of available observations and consequently lead to incomplete keypoint/wireframe predictions and point clouds as well as potentially incorrect predictions of the vehicle type and its viewpoint. Recent research has shown promising results on recovering the appearance of invisible parts of partially occluded objects (Yan et al., 2019) or on point cloud completion from incomplete observations (Stutz and Geiger, 2018). Such approaches deliver potential solutions to overcome reconstruction problems related to occlusions. Furthermore, the proposed method provides an approach for vehicle reconstruction based on one individual stereo pair, i.e. observations from one epoch in time. An extension of the probabilistic model to a sequence of images acquired at subsequent epochs and the tracking of vehicles would enable the incorporation of multiple observations of the same object. Suitable constraints, e.g. on plausible changes in the position and orientation between two time steps, can for instance be enforced by proper motion models (Engelmann et al., 2017). Besides, the reconstruction of each vehicle is treated individually in this thesis. Global constraints, that e.g. prevent vehicle reconstructions to coincide, or contextual relations between vehicles are not exploited so far but provide opportunities for improvement.

Moreover, the probabilistic model as well as the proposed shape prior, being composed of individual components, provide the flexibility to introduce further terms in future developments. Any entity that can be represented by the ASM and which can be inferred from the image can be used to formulate a likelihood to be included in the Bayesian model. In this context, the presented multi-task CNN delivers the flexibility to be extended by additional branches for the prediction of further observations. While our results indicate that wireframe edges are better suited for model fitting compared to point-wise features such as keypoints, segment-wise entities, such as semantically segmented vehicle parts, could even provide more stable features. A first attempt of matching object models to images based on segmented parts is presented in (Barowski et al., 2019). However, it has only been tested on synthetic data so far. Yet, such an approach could enrich the probabilistic model by further likelihoods.

Furthermore, the performance of the proposed CNN can potentially be enhanced by different adaptations. In this thesis, the VGG19 network is used as backbone feature extractor of the proposed CNN (cf. Sec. 4.3.1). Pre-trained weights are used for the shared weight layers of the backbone network and frozen during training. As a consequence, the individual branches of the proposed CNN are trained separately and independent from each other. Investigations could be conducted to analyse whether the utilisation of a different backbone architectures (e.g. the usage of the ResNet architecture (He et al., 2016)) can improve the performance of the CNN. Besides, it has been shown in the literature that deep learning applications can benefit from multi-task learning (Kendall et al., 2018). Consequently, a joint learning of the shared weights could further enhance the predictions of the proposed CNN.

The learning of additional weights and extending the CNN to predict further features, as mentioned above, is likely to lead to the requirement of additional and more training data. However, the generation of reference data for training is labour-intensive. In order to reduce the dependency on expensively labelled training data, the suitability of synthetic data for training the CNN can be investigated. In this context, Alhaija et al. (2018) and Manhardt et al. (2019) presented approaches for the generation of photo-realistic renderings of vehicle models as data for the training of CNNs. However, due to the domain gap, training on synthetic data usually comes with a noticeable drop in performance when applied to real data. Potential solutions to overcome this effect can be found in the context of research on domain adaptation (Wang and Deng, 2018).

The listed possibilities leave space for promising future directions in which the concept for vehicle reconstruction developed in this thesis can serve as a powerful and flexible basis.

Bibliography

Alhaija, H. A., Mustikovela, S. K., Mescheder, L., Geiger, A. and Rother, C., 2018. Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes. *International Journal of Computer Vision (IJCV)* 126(9), pp. 961–972.

Angelopoulou, A., Psarrou, A., Garcia-Rodriguez, J., Orts-Escolano, S., Azorin-Lopez, J. and Revett, K., 2015. 3D Reconstruction of medical Images from Slices automatically landmarked with growing Neural Models. *Neurocomputing* 150, pp. 16–25.

Ansari, J. A., Sharma, S., Majumdar, A., Murthy, J. K. and Krishna, K. M., 2018. The Earth ain't Flat: Monocular Reconstruction of Vehicles on Steep and Graded Roads from a Moving Camera. In: *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 8404–8410.

Bao, S. Y., Chandraker, M., Lin, Y. and Savarese, S., 2013. Dense Object Reconstruction with Semantic Priors. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1264–1271.

Barber, C. B., Dobkin, D. P. and Huhdanpaa, H., 1996. The Quickhull Algorithm for Convex Hulls. *ACM Transactions on Mathematical Software (TOMS)* 22(4), pp. 469–483.

Barowski, T., Szczot, M. and Houben, S., 2019. 6DoF Vehicle Pose Estimation using Segmentation-Based Part Correspondences. In: *IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 573–580.

Bentley, J. L., 1975. Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the Association for Computing Machinery (CACM)* 18(9), pp. 509–517.

Bhattacharyya, A., 1943. On a Measure of Divergence between two Statistical Populations defined by their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* 35, pp. 99–109.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning.* Springer, New York.

Boyd, S. and Vandenberghe, L., 2004. *Convex Optimization.* Cambridge University Press, New York, NY, USA.

Braden, B., 1986. The Surveyor's Area Formula. *The College Mathematics Journal* 17(4), pp. 326–337.

Breiman, L., 2001. Random Forests. *Machine Learning* 45(1), pp. 5–32.

Chabot, F., Chaouch, M., Rabarisoa, J., Teuliere, C. and Chateau, T., 2016. Accurate 3D Car Pose Estimation. In: *IEEE International Conference on Image Processing (ICIP)*, pp. 3807–3811.

Chabot, F., Chaouch, M., Rabarisoa, J., Teuliere, C. and Chateau, T., 2017. Deep MANTA: A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1827–1836.

Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S. and Urtasun, R., 2016. Monocular 3D Object Detection for Autonomous Driving. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2147–2156.

Chen, X., Kundu, K., Zhu, Y., Berneshawi, A. G., Ma, H., Fidler, S. and Urtasun, R., 2015. 3D Object Proposals for accurate Object Class Detection. In: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 28, pp. 424–432.

Chollet, F. et al., 2015. Keras. https://keras.io.

Cignoni, P., Montani, C. and Scopigno, R., 1998. DeWall: A fast divide and conquer Delaunay triangulation algorithm in E^d. *Computer-Aided Design* 30(5), pp. 333–341.

Coenen, M. and Rottensteiner, F., 2019. Probabilistic Vehicle Reconstruction Using a Multi-Task CNN. In: *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 822–831.

Coenen, M., Rottensteiner, F. and Heipke, C., 2017. Detection and 3D Modelling of Vehicles from terrestrial stereo Image Pairs. In: *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-1/W1, pp. 505–512.

Coenen, M., Rottensteiner, F. and Heipke, C., 2019. Precise Vehicle Reconstruction for autonomous Driving Applications. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. IV-2/W5, pp. 21–28.

Cootes, T. F., Taylor, C. J. and Cooper, D. H., 1995. Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding (CVIU)* 61(1), pp. 38–59.

Cortes, C. and Vapnik, V., 1995. Support-Vector Networks. *Machine Learning* 20(3), pp. 273–297.

Dalal, N. and Triggs, B., 2005. Histograms of oriented Gradients for Human Detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1, pp. 886–893.

Dame, A., Prisacariu, V. A., Ren, C. Y. and Reid, I., 2013. Dense Reconstruction using 3D Object Shape Priors. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1288–1295.

Ding, W., Li, S., Zhang, G., Lei, X. and Qian, H., 2018. Vehicle Pose and Shape Estimation through Multiple Monocular Vision. In: *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 709–715.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T., 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In: *Proceedings of the 31st International Conference on Machine Learning (PMLR)*, Vol. 32(1), pp. 647–655.

Doucet, A., de Freitas, N. and Gordon, N., 2001. *Sequential Monte Carlo Methods in Practice*. Springer New York, chapter An Introduction to Sequential Monte Carlo Methods, pp. 3–14.

Du, G., Wang, K. and Lian, S., 2019. Vision-based Robotic Grasping from Object Localization, Pose Estimation, Grasp Detection to Motion Planning: A Review. In: *arxiv 1905.06658*.

Engelmann, F., Stückler, J. and Leibe, B., 2016. Joint Object Pose Estimation and Shape Reconstruction in urban Street Scenes using 3D Shape Priors. In: *German Conference on Pattern Recognition (GCPR)*, Lecture Notes in Computer Science, Vol. 9796, Springer, Cham, pp. 219–230.

Engelmann, F., Stueckler, J. and Leibe, B., 2017. SAMP: Shape and Motion Priors for 4D Vehicle Reconstruction. In: *IEEE Winter Conference on Applications of Computer Vision*, pp. 400–408.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D., 2010. Object Detection with discriminatively trained part-based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 32(9), pp. 1627–1645.

Fischler, M. A. and Bolles, R. C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the Association for Computing Machinery (CACM)* 24(6), pp. 381–395.

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for autonomous Driving? The KITTI Vision Benchmark Suite. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361.

Geiger, A., Roser, M. and Urtasun, R., 2011. Efficient Large-Scale Stereo Matching. In: *Asian Conference on Computer Vision (ACCV)*, Lecture Notes in Computer Science, Vol. 6492, Springer, Berlin, pp. 25–38.

Glorot, X. and Bengio, Y., 2010. Understanding the Difficulty of Training deep feedforward Neural Networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Vol. 9, pp. 249–256.

Godard, C., Mac Aodha, O. and Brostow, G. J., 2017. Unsupervised Monocular Depth Estimation With Left-Right Consistency. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 270–279.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning.* MIT Press, Cambridge, Massachusetts, USA.

Grabner, A., Roth, P. M. and Lepetit, V., 2018. 3D Pose Estimation and 3D Model Retrieval for Objects in the Wild. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3022–3031.

Güney, F. and Geiger, A., 2015. Displets: Resolving stereo Ambiguities using Object Knowledge. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4165–4175.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J. and Seung, H. S., 2000. Digital Selection and analogue Amplification coexist in a Cortex-inspired Silicon Circuit. *Nature* 405(6789), pp. 947–951.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. and Stahel, W. A., 1986. *Robust Statistics: The Approach based on Influence Functions.* Wiley-Interscience, New York, New York, USA.

Han, J. and Moraga, C., 1995. The Influence of the Sigmoid Function Parameters on the Speed of Backpropagation Learning. In: *International Workshop on Artificial Neural Networks*, pp. 195–201.

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask R-CNN. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Helmholtz, H. v., 1867. *Handbuch der physiologischen Optik.* Vol. 3, Leopold Voss, Leipzig.

Hödlmoser, M., Micusik, B., Pollefeys, M., Liu, M. and Kampel, M., 2013. Model-Based Vehicle Pose Estimation and Tracking in Videos Using Random Forests. In: *International Conference on 3D Vision (3DV)*, pp. 430–437.

Huber, P. J., 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35(1), pp. 73–101.

Ioffe, S. and Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *International Conference on Machine Learning (ICML)*, pp. 448–456.

Isola, P., Zoran, D., Krishnan, D. and Adelson, E. H., 2014. Crisp Boundary Detection Using Pointwise Mutual Information. In: *European Conference on Computer Vision (ECCV)*, pp. 799–814.

Janai, J., Güney, F., Behl, A. and Geiger, A., 2020. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. *Foundations and Trends in Computer Graphics and Vision* 12(1-3), pp. 1–308.

Kar, A., Tulsiani, S., Carreira, J. and Malik, J., 2015. Category-Specific Object Reconstruction From a Single Image. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1966–1974.

Kendall, A., Gal, Y. and Cipolla, R., 2018. Multi-Task Learning using Uncertainty to weigh Losses for Scene Geometry and Semantics. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491.

Kingma, D. and Ba, L., 2015. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations (ICLR)*.

Knuth, J. and Barooah, P., 2009. Distributed collaborative Localization of multiple Vehicles from relative Pose Measurements. In: *Conference on Communication, Control, and Computing*, pp. 314–321.

Kroese, D., Brereton, T., Taimre, T. and Botev, Z., 2014. Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics* 6(6), pp. 386–392.

Ku, J., Pon, A. D. and Waslander, S. L., 2019. Monocular 3D Object Detection Leveraging Accurate Proposals and Shape Reconstruction. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11867–11876.

Kundu, A., Li, Y. and Rehg, J. M., 2018. 3D-RCNN: Instance-level 3D Object Reconstruction via Render-and-Compare. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3559–3568.

LeCun, Y., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D. and Henderson, D., 1990. Handwritten Digit Recognition with a Back-Propagation Network. In: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 2, pp. 396–404.

Leibe, B., Leonardis, A. and Schiele, B., 2006. *An Implicit Shape Model for combined Object Categorization and Segmentation.* Lecture Notes in Computer Science, Vol. 4170, Springer Berlin Heidelberg, pp. 508–524.

Leotta, M. J. and Mundy, J. L., 2009. Predicting high Resolution Image Edges with a generic, adaptive, 3-D Vehicle Model. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1311–1318.

Li, Y., Gu, L. and Kanade, T., 2011. Robustly Aligning a Shape Model and Its Application to Car Alignment of Unknown Pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33(9), pp. 1860–1876.

Liebelt, J. and Schmid, C., 2010. Multi-view Object Class Detection with a 3D geometric Model. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1688–1695.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature Pyramid Networks for Object Detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L., 2014a. Microsoft COCO: Common Objects in Context. In: *European Conference on Computer Vision (ECCV)*, pp. 740–755.

Lin, Y.-L., Morariu, V. I., Hsu, W. and Davis, L. S., 2014b. Jointly Optimizing 3D Model Fitting and Fine-Grained Classification. In: *European Conference on Computer Vision (ECCV)*, pp. 466–480.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.

Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W. and Fan, X., 2019. Accurate Monocular 3D Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 6851–6860.

Manhardt, F., Kehl, W. and Gaidon, A., 2019. ROI-10D: Monocular Lifting of 2D Detections to 6D Pose and Metric Shape. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2069–2078.

Menze, M., Heipke, C. and Geiger, A., 2015. Joint 3D Estimation of Vehicles and Scene Flow. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. II-3, pp. 427–434.

Mousavian, A., Anguelov, D., Flynn, J. and Kosecka, J., 2017. 3D Bounding Box Estimation Using Deep Learning and Geometry. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7074–7082.

Muja, M. and Lowe, D. G., 2009. Fast approximate nearest Neighbors with automatic Algorithm Configuration. In: *International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 331–340.

Murthy, J. K., Krishna, G. V. S., Chhaya, F. and Krishna, K. M., 2017a. Reconstructing Vehicles from a single Image: Shape Priors for Road Scene Understanding. In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 724–731.

Murthy, J. K., Sharma, S. and Krishna, K. M., 2017b. Shape Priors for real-time monocular Object Localization in dynamic Environments. In: *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 1768–1774.

Newell, A., Yang, K. and Deng, J., 2016. Stacked Hourglass Networks for Human Pose Estimation. In: *European Conference on Computer Vision (ECCV)*, pp. 483–499.

Ortiz-Cayon, R., Djelouah, A., Massa, F., Aubry, M. and Drettakis, G., 2016. Automatic 3D Car Model Alignment for Mixed Image-Based Rendering. In: *International Conference on 3D Vision (3DV)*, pp. 286–295.

Ozuysal, M., Lepetit, V. and Fua, P., 2009. Pose Estimation for Category specific multiview Object Localization. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pp. 778–785.

Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. G. and Daniilidis, K., 2017. 6-DoF Object Pose from Semantic Keypoints. In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2011–2018.

Pepik, B., Stark, M., Gehler, P. and Schiele, B., 2012. Teaching 3D Geometry to Deformable Part Models. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3362–3369.

Pepik, B., Stark, M., Gehler, P. and Schiele, B., 2015. Multi-View and 3D Deformable Part Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37(11), pp. 2232–2245.

Poggio, T., Torre, V. and Koch, C., 1985. Computational Vision and Regularization Theory. *Nature* 317(26), pp. 314–319.

Pomerleau, F., Colas, F., Siegwart, R. and Magnenat, S., 2013. Comparing ICP Variants on Real-World Data Sets. *Autonomous Robots* 34(3), pp. 133–148.

Prisacariu, V. A., Segal, A. V. and Reid, I., 2012. Simultaneous Monocular 2D Segmentation, 3D Pose Recovery and 3D Reconstruction. In: *Asian Conference on Computer Vision (ACCV)*, pp. 593–606.

Ramnath, K., Sinha, S. N., Szeliski, R. and Hsiao, E., 2014. Car Make and Model Recognition using 3D Curve Alignment. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 285–292.

Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 28, pp. 91–99.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241.

Rumelhart, D., Hinton, G. and Williams, R., 1986. Learning Representations by back-propagating Errors. *Nature* 323(6088), pp. 533–536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3), pp. 211–252.

Schön, S., Brenner, C., Alkhatib, H., Coenen, M., Dbouk, H., Garcia-Fernandez, N., Fischer, C., Heipke, C., Lohmann, K., Neumann, I., Nguyen, U., Paffenholz, J.-A., Peters, T., Rottensteiner, F., Schachtschneider, J., Sester, M., Sun, L., Vogel, S., Voges, R. and Wagner, B., 2018. Integrity and Collaboration in Dynamic Sensor Networks. *Sensors* 18(7), pp. 2400–2421.

Simonyan, K. and Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: *International Conference on Learning Representations (ICLR).*

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research* 15(1), pp. 1929–1958.

Stutz, D. and Geiger, A., 2018. Learning 3D Shape Completion from Laser Scan Data with weak Supervision. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1955–1964.

Su, H., Qi, C. R., Li, Y. and Guibas, L. J., 2015. Render for CNN: Viewpoint Estimation in Images Using CNNs Trained With Rendered 3D Model Views. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2686–2694.

Tang, Z., Naphade, M., Birchfield, S., Tremblay, J., Hodge, W., Kumar, R., Wang, S. and Yang, X., 2019. PAMTRI: Pose-Aware Multi-Task Learning for Vehicle Re-Identification Using Highly Randomized Synthetic Data. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 211–220.

Tekin, B., Sinha, S. N. and Fua, P., 2018. Real-Time Seamless Single Shot 6D Object Pose Prediction. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 292–301.

Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T. and van Gool, L., 2007. Depth-From-Recognition: Inferring Meta-data by Cognitive Feedback. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8.

Tsin, Y., Genc, Y. and Ramesh, V., 2009. Explicit 3D Modeling for Vehicle Monitoring in Non-overlapping Cameras. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 110–115.

Tulsiani, S. and Malik, J., 2015. Viewpoints and Keypoints. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1510–1519.

Wang, M. and Deng, W., 2018. Deep visual Domain Adaptation: A Survey. *Neurocomputing* 312, pp. 135–153.

Wang, Y., Tan, X., Yang, Y., Liu, X., Ding, E., Zhou, F. and Davis, L. S., 2018. 3D Pose Estimation for Fine-Grained Object Categories. In: *European Conference on Computer Vision Workshops (ECCV Workshops)*.

Wang, Y.-X., Ramanan, D. and Hebert, M., 2017. Learning to Model the Tail. In: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 30, pp. 7029–7039.

Xiang, Y., Choi, W., Lin, Y. and Savarese, S., 2015. Data-Driven 3D Voxel Patterns for Object Category Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1903–1911.

Xiang, Y., Choi, W., Lin, Y. and Savarese, S., 2017. Subcategory-Aware Convolutional Neural Networks for Object Proposals and Detection. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 924–933.

Xiang, Y., Schmidt, T., Narayanan, V. and Fox, D., 2018. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In: *Robotics: Science and Systems*, Vol. XIV.

Xiao, W., Vallet, B., Schindler, K. and Paparoditis, N., 2016. Street-Side Vehicle Detection, Classification and Change Detection using mobile Laser Scanning Data. *ISPRS Journal of Photogrammetry and Remote Sensing* 114, pp. 166–178.

Xu, B. and Chen, Z., 2018. Multi-Level Fusion Based 3D Object Detection From Monocular Images. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2345–2353.

Yan, X., Wang, F., Liu, W., Yu, Y., He, S. and Pan, J., 2019. Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 7618–7627.

Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W. and Yu, Y., 2015. HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2740–2748.

Yang, L., Luo, P., Change Loy, C. and Tang, X., 2015. A Large-Scale Car Dataset for Fine-Grained Categorization and Verification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3973–3981.

Yang, M.-D., Chao, C.-F., Huang, K.-S., Lu, L.-Y. and Chen, Y.-P., 2013. Image-based 3D Scene Reconstruction and Exploration in Augmented Reality. *Automation in Construction* 33, pp. 48–60.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H., 2014. How transferable are Features in Deep Neural Networks? In: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 27, pp. 3320–3328.

Zeiler, M. D., Krishnan, D., Taylor, G. W. and Fergus, R., 2010. Deconvolutional Networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2528–2535.

Zhou, Y., Qi, H., Zhai, Y., Sun, Q., Chen, Z., Wei, L.-Y. and Ma, Y., 2019. Learning to Reconstruct 3D Manhattan Wireframes From a Single Image. In: *IEEE International Conference on Computer Vision (ICCV)*.

Zhu, R., Kiani Galoogahi, H., Wang, C. and Lucey, S., 2017. Rethinking Reprojection: Closing the Loop for Pose-Aware Shape Reconstruction From a Single Image. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 57–65.

Zia, M. Z., Stark, M. and Schindler, K., 2015. Towards Scene Understanding with detailed 3D Object Representations. *International Journal of Computer Vision (IJCV)* 112(2), pp. 188–203.

Zia, M. Z., Stark, M., Schiele, B. and Schindler, K., 2013. Detailed 3D Representations for Object Recognition and Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35(11), pp. 2608–2623.

Curriculum Vitae

Personal Information

Name	Maximilian Alexander Coenen
Date of birth	23.04.1987 in Cologne, Germany

Work Experience

Dec. 2016 - May 2020	Leibniz University Hannover Research Training Group 2159: Integrity and Collaboration in Dynamic Sensor Networks (i.c.sens) *Research Member*
Nov. 2015 - May 2020	Leibniz University Hannover Institute of Photogrammetry and GeoInformation *Research Associate*
May 2015 - Oct. 2015	Leibniz University Hannover Institute of Photogrammetry and GeoInformation *Student Assistant*

Education

2012 - 2015	Leibniz University Hannover Course Navigation and Field Robotics *Master of Science*
2007 - 2012	University of Bonn Course Geodesy and Geoinformation *Bachelor of Science*
1997 - 2006	Gymnasium Haus Overbach, Jülich-Barmen *A level*

Acknowledgements

I want to express grateful thanks to all persons who contributed to the success of this dissertation which was funded by the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens [GRK2159].

First and foremost, my deep gratitude goes to my supervisor Prof. Dr. techn. Franz Rottensteiner for giving me the trust and the freedom to develop my research ideas and for his guidance, advise, and expertise throughout my studies and my doctorate.

As chair of the Institute of Photogrammetry and GeoInformation (IPI), I thank Prof. Dr.-Ing. habil. Christian Heipke for giving me the opportunity to conduct my research and to do my PhD at the IPI. Besides, I thank him and all other professors involved in the RTG i.c.sens for fruitful discussions during the PI meetings, the retreats, and in private. Besides, I thank Prof. Dr.-Ing. Claus Brenner and Prof. Dr.-Ing. Olaf Hellwich for acting as referees of this dissertation and for providing valuable remarks.

I also want to express my gratitude to my friends and colleagues from the IPI and from the i.c.sens for the nice time and the warm and friendly working atmosphere. In particular, I would like to thank Max M., Nico, Uyen, and Hani for making the office a place where I always loved and enjoyed to be, for the inspiring discussions, both professional and private, and for sharing the smaller and larger challenges of a PhD student.

Finally, I cordially thank my parents for their loving support and faith throughout my life. I also thank my brothers, sisters, and my friends for their encouragement and for being there in every situation. My special gratitude goes to my wife Stephanie for her unconditional love, her understanding, and her strength during restless and demanding times.