Exploiting single image depth prediction for mono-stixel estimation

Fabian Brickwedde[1,2][0000−0003−3527−9323], Steffen Abraham[1][0000−0003−1320−4275], and Rudolf Mester[2][0000−0002−6932−0606]

[1] Robert Bosch GmbH, Hildesheim, Germany
{Fabian.Brickwedde, Steffen.Abraham}@de.bosch.com
[2] VSI Lab, Goethe University, Frankfurt, Germany
mester@vsi.cs.uni-frankfurt.de

Abstract. The stixel-world is a compact and detailed environment representation specially designed for street scenes and automotive vision applications. A recent work proposes a monocamera based stixel estimation method based on the structure from motion principle and scene model to predict the depth and translational motion of the static and dynamic parts of the scene. In this paper, we propose to exploit the recent advantages in deep learning based single image depth prediction for mono-stixel estimation. In our approach, the mono-stixels are estimated based on the single image depth predictions, a dense optical flow field and semantic segmentation supported by the prior knowledge about the characteristic of typical street scenes. To provide a meaningful estimation, it is crucial to model the statistical distribution of all measurements, which is especially challenging for the single image depth predictions. Therefore, we present a semantic class dependent measurement model of the single image depth prediction derived from the empirical error distribution on the Kitti dataset.
Our experiments on the Kitti-Stereo'2015 dataset show that we can significantly improve the quality of mono-stixel estimation by exploiting the single image depth prediction. Furthermore, our proposed approach is able to handle partly occluded moving objects as well as scenarios without translational motion of the camera.

Keywords: mono-stixel, single image depth prediction, scene reconstruction, scene flow, monocamera, automotive

1 Introduction

Autonomous vehicles and driver assistance systems need to understand their environment including a geometric representation of the distances and motions as well as a semantic representation of the classification of each object. Additionally, to reduce the computational effort for higher-level vision applications, this representation should be compact.

Therefore, the stixel-world was introduced by Badino et al. [1] and extended to a multi-layer stixel-world by Pfeiffer et al. [21]. The stixel-world represents

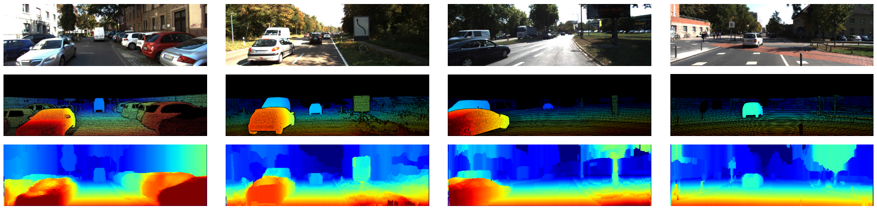

Fig. 1. Example depth prediction of our mono-stixel estimation. Top to bottom: image, ground truth depth, and our mono-stixel depth estimation. The color encodes the inverse depth from close (red) to far (dark blue). The ego vehicle is standing in the images of the third and fourth column. By exploiting the single image depth prediction our mono-stixel estimation approach provides reliable depth estimates even for partly occluded vehicles and scenarios without translational motion of the camera.

the scene as a composition of thin stick like elements, the stixels. Each stixel corresponds to a vertical rectangle in the image and stores the distance to the camera assuming an upright orientation of object stixel and lying orientation of ground stixel. Additionally to the type, segmentation and distance each stixel segment can consist of a label for the semantic class [23] and the motion of each stixel can be estimated using a Kalman-filter based tracking approach [20]. The mentioned works use stereo depth measurements to estimate the stixel-world. A recent work by Brickwedde et al. [2] presents the mono-stixel approach, a monocamera based stixel-model and estimation method. The mono-stixel model directly estimates the 2D-translational motion of each stixel as part of the segmentation problem and introduces a further mono-stixel type by distinguishing static and potentially moving object stixel. The mono-stixels are estimated based on a dense optical flow field and semantic segmentation leveraging the structure from motion principle for the static environment and the scene model assumption that moving objects stand on the ground plane. However, there are two limitations of the mono-stixel estimation approach in [2]. First, a translational motion of the camera is required. Second, the projection of a potentially moving object stixel on the ground plane only works as long as this part of the object really stands on the ground, the ground contact point is not occluded and the surface of the ground plane is estimated with high quality.

To overcome these limitations and improve the quality, we propose to exploit a deep learning based single image depth prediction for mono-stixel estimation as a further information. Thereby, the mono-stixel estimation serves as a fusion of an optical flow field, single image depth prediction and semantic segmentation supported by scene model assumptions. By exploiting the single image depth prediction the approach is able to handle partly occluded objects and a translational motion of the camera is not required anymore as shown in figure 1.

2 Related work

Traditionally stixel estimation methods use stereo depth measurements like [1, 21]. Levi et al. [15] and Garnett et al. [9] propose a convolutional neural network, called the Stixel-Net for a monocamera based stixel estimation method. The convolutional neural network predicts the segment and depth of the closest stixel in each column, but does not provide a depth representation of the whole image and is more related to a freespace segmentation method. As discussed in the introduction, the mono-stixel approach by Brickwedde et al. [2] is highly related to our approach as a multi-layer mono-stixel estimation based on the structure from motion or multi-view geometry principle [12]. The 3D-position of static points in the scene can be reconstructed based on the image correspondences and the camera motion by triangulation. For example, SLAM methods like [18] and [5] jointly estimate the camera motion and image correspondences including their 3D-position in the scene. In general, this reconstruction is only known up to an unknown scale of the scene and camera motion in a monocamera setup. However, in autonomous applications the unknown scale can be estimated based on the known camera height above the ground [6, 19] or derived from an inertial measurement unit. But, there are still limitations of the structure from motion principle for moving objects or scenarios without translational motion of the camera. By exploiting the epipolar geometry or scene constraints some independent moving objects (IMO) are detectable based on the optical flow and camera motion [14, 22]. But some IMOs like oncoming vehicle are still not detectable. Therefore, the mono-stixel estimation approach in [2] proposes to distinguish static and potentially moving objects like vehicles based on a semantic segmentation. To reconstruct these objects some methods [2, 22] exploit the scene model assumption that these objects are connected with the surrounding static environment, for example, that a vehicle stands on the ground plane.

In the recent years, deep learning methods show impressive results for predicting the depth of the scene for a single image. Thus, these methods exploit totally different information than the multi-view approaches. These methods potentially learn the typical shape and size of objects and structures as well as the contextual information. One of the pioneering work is presented by Eigen et al. [4]. They propose a supervised learning approach for single image depth prediction, but it can also be trained in an unsupervised or self-supervised manner [30, 8, 11]. Providing additionally the statistical distribution of the predicted depth is still challenging. Kendall and Gal [13] distinguish two types of uncertainties: the aleatoric uncertainty, that refers to sensor noise and can not be reduced even with more training data and the epistemic or model uncertainty that could be explained away given enough training data. They propose to learn to predict the aleatoric uncertainty as part of a supervised learning approach and use Monte Carlo dropout to derive the epistemic uncertainty.

Single-view and multi-view depth predictions exploit totally different information with different benefits and drawbacks. This makes it powerful to fuse both depth prediction approaches [25, 7]. Alternatively, the methods [28, 27] propose to learn the multi-view and structure from motion principle directly. Thereby, the convolutional network can additionally exploit the single-view depth cues and seen as a fusion as well.

The contributions of our work are as follows: We present a mono-stixel estimation that fuses single image depth predictions with a dense optical flow field and semantic segmentation. Thereby, we significantly outperform previous mono-stixel estimation methods and overcome two main limitations. Our mono-stixel estimation method is able to provide reliable depth estimates even for scenarios without translational motion of the camera and is able to reconstruct moving objects even if the ground contact point is occluded. Furthermore, our approach can be seen as a fusion scheme that is supported by a semantic segmentation and scene model assumptions. For a statistical meaningful fusion, it is crucial to know the error distribution, which is especially challenging for the single image depth predictions. Therefore, we present a semantic class dependent measurement model for the single image depth prediction derived from the empirical error distribution on the Kitti dataset [17]. This analysis additionally gives some insights which parts of the scene are challenging for single image depth prediction.

3 Method

In this chapter, we present our mono-stixel estimation method. We mainly follow the mono-stixel model and segmentation algorithm proposed by Brickwedde et al. [2]. In the first section, we give a brief overview of that method and present how to adapt it to exploit the single image depth prediction for mono-stixel estimation. Thereby, our mono-stixel approach uses a pixel-wise semantic segmentation, dense optical flow field, and single image depth prediction as inputs. Furthermore, the camera motion is assumed to be known. In the second chapter, we derive a measurement model of the single image depth prediction based on the error statistic on the Kitti-Stereo'15 dataset [17]. Finally, the last chapter presents how to solve the mono-stixel segmentation problem.

3.1 Mono-stixel segmentation as energy minimization problem

We follow the mono-stixel model proposed in [2] that defines a mono-stixel as a thin stick-like planar and rigid moving element in the scene. To represent the whole scene the image of width w and height h is divided into columns of a fixed width w_s and each column k is segmented into N_k mono-stixels separately:

$$\begin{gathered}\mathbf{s}_k = \{s_i \mid 1 \leq i \leq N_k \leq h\} \\ \text{with } s_i = (v_i^b, v_i^t, c_i, p_i, \mathbf{t}_i, m_i)\end{gathered} \tag{1}$$

Each mono-stixel s_i consists of labels for the segmentation, represented by the bottom v_i^b and top v_i^t point in that column, a label for the semantic class c_i, labels to encode the inverse depth p_i and 2D-translational motion over the ground $\mathbf{t}_i$ and is of a given mono-stixel type m_i. The four mono-stixel types m_i are ground, static object, dynamic object, and sky stixel as defined in [2]. Each semantic class

c_i is directly associated with one mono-stixel type m_i. Thus, a ground stixel could be of the semantic class road, sidewalk or terrain, a static object stixel could be of the semantic class building, pole or vegetation, a dynamic object stixel could be of the type vehicle, two-wheeler or person and the sky stixels are of the semantic class sky. Furthermore, assumptions of the geometry and motion are defined for each stixel type. A ground stixel has a lying orientation, an object stixel has an upright orientation facing the camera center and a sky stixel is at infinite distance $p_i = 0$. The dynamic object stixel is the only type with independent motion, the other stixel types are assumed to be static $\mathbf{t}_i = 0$.

The mono-stixel segmentation is defined as a 1D-energy minimization problem [2] for each column k:

$$\begin{aligned}\hat{\mathbf{s}} &= \arg\min_{\mathbf{s}} E(\mathbf{s}, \mathbf{f}, \mathbf{l}) \\ &= \arg\min_{\mathbf{s}} \left(\Psi(\mathbf{s}) + \Phi(\mathbf{s}, \mathbf{f}, \mathbf{l}, \mathbf{d}) \right), \end{aligned} \tag{2}$$

where $\Psi(\mathbf{s})$ represents the prior knowledge of the typical structure of street scenes. It consists of a gravity prior to encode that most of the objects typically stand on the ground plane, an ordering constraint to regard that one object might occlude another one and a flat ground plane prior which prefers small discontinuities of the height in the ground plane. Additionally, a constant value is added for each new stixel to prevent over-segmentation and regulates the model complexity. To model the scene prior we follow exactly the same equations as defined in [2] and refer the interested reader to that paper.

The data likelihood $\Phi(\mathbf{s}, \mathbf{f}, \mathbf{l}, \mathbf{d})$ rates the consistency of the stixel hypothesis based on the semantic segmentation $\mathbf{l}$, the optical flow $\mathbf{f}$ and single image depth prediction $\mathbf{d}$:

$$\Phi(\mathbf{s}, \mathbf{f}, \mathbf{l}, \mathbf{d}) = \sum_{i=1}^{N_k} \sum_{v=v_i^b}^{v_i^t} \left(\lambda_L \Phi_L(s_i, \mathbf{l}_v) + \lambda_F \Phi_F(s_i, \mathbf{f}_v, v) + \lambda_{SI} \Phi_{SI}(s_i, d_v) \right) \tag{3}$$

The probability is assumed to be independent across the rows v in that column and each data likelihood term is weighted by λ. The data likelihood terms $\Phi_L(s_i, \mathbf{l}_v)$ and $\Phi_F(s_i, \mathbf{f}_v, v)$ are defined as in [2]: $\Phi_L(s_i, \mathbf{l}_v)$ prefers stixels having a semantic class c_i with high class scores $\mathbf{l}_v(c_i)$ in the semantic segmentation and $\Phi_F(s_i, \mathbf{f}_v, v)$ rates the consistency of the stixel based on the optical flow $\mathbf{f}_v$. Therefore, the expected optical flow $\mathbf{f}_{exp,v}(s_i)$ at row v given a stixel hypothesis s_i is computed based on its inverse depth, relative motion to the camera and orientation defined by its mono-stixel type. This computation can be expressed by a stixel-homography [2]. The data likelihood $\Phi_F(s_i, \mathbf{f}_v, v)$ rates the difference between the expected $\mathbf{f}_{exp,v}(s_i)$ and measured optical flow $\mathbf{f}_v$ as the negative logarithm of the measurement model which is defined as a mixture model consisting of a normal distribution for inliers and a uniform distribution for outliers.

We propose to extend the data likelihood for the input of the single image depth prediction by $\Phi_{SI}(s_i, d_v)$. The output of our single image depth prediction is defined as a dense disparity map with $d_v = \frac{1}{Z_v}$, where Z_v is the z-coordinate of the 3D-position in camera coordinates. The data likelihood rates the difference between expected disparity $d_{exp,v}(s_i)$ of the corresponding stixel hypothesis s_i and the disparity measurement d_v of single image depth prediction for the pixel at row v. The expected disparity $d_{exp,v}(s_i)$ of the stixel hypothesis s_i is defined by the inverse depth p_i and stixel type m_i as:

$$d_{exp,v}(s_i) = \begin{cases} \frac{p_i(\mathbf{x} \cdot \mathbf{n}_i)}{\mathbf{x}_z} & \text{, if } m_i \in \ \{\text{static object, dynamic object, ground}\} \\ 0 & \text{, if } m_i = \text{sky} \end{cases} \tag{4}$$

where $\mathbf{x}$ is the ray of the pixel corresponding to row v and $\mathbf{n}_i$ is the normal vector defined by the orientation of the mono-stixel type m_i.

The measurement model of the single image depth prediction derived in the next section defines the statistical distribution of a disparity error dependent on the semantic class $p(d_{error}|c)$. Switching to the log-domain $\Phi_{SI}(s_i, d_v)$ is defined as the negative logarithm of this probability:

$$\begin{aligned} \Phi_{SI}(s_i, d_v) &= -\log\left(p(d_v - d_{exp,v}(s_i)|c_i)\right) \\ &= -\log\left(p(d_{error,v}|c_i)\right) \end{aligned} \tag{5}$$

3.2 Measurement model of single image depth prediction

To achieve a high performance of the mono-stixel estimation and meaningful fusion with the optical flow, it is crucial to model the statistical distribution of the disparity error of the single image depth prediction. Supervised learning methods are limited to the ground truth provided by the sensor which is typically limited by a certain range and view. For example, the Velodyne sensor in the Kitti dataset [17] only provides ground truth up to around 80 meters and only for the lower part of the image. Consequently, we propose to use a self-supervised learning method and follow the approach of Godard et al. [11] . However, this method does not provide uncertainties of the predicted depth and the aleatoric uncertainty estimation presented by Kendall and Gal [13] is not applicable for a self-supervised learning approach.

Therefore, to derive a measurement model we analyzed the empirical error distribution of the single image depth prediction approach by [11] on the Kitti dataset [17]. The error distribution shown in figure 2 mainly consists of two parts. First, a part with slowly decreasing tails that mainly models the distribution of large errors and has a triangular shape on a logarithmic scale. Second, one part that corresponds to a peak and high probabilities for small errors. To approximate this characteristic of the empirical density function we propose a mixture model that consists of a Laplacian distribution that mainly models the probability of large errors and a Gaussian distribution mainly for the low errors:

$$p(d_{error}) = \frac{1-\lambda}{\sqrt{2\pi}\sigma} e^{\frac{-d_{error}^2}{2\sigma^2}} + \frac{\lambda}{2b} e^{-\frac{|d_{error}|}{b}} \tag{6}$$

Figure 2 shows the approximated density function as an orange line for $\sigma = 0.0042$, $b = 0.02$ and $\lambda = 0.2$.

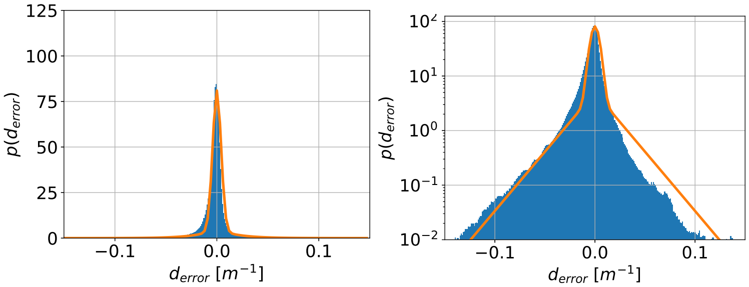

Fig. 2. Statistical distribution of the disparity error of the single image depth prediction [11]. The blue histograms show the empirical distribution of the error on Kitti-Stereo'15 [17] and the orange curve the approximated measurement model. The distribution is shown with a logarithmic scale of the frequency in the right diagram.

Furthermore, we identified that the error distribution highly depends on the semantic class as shown in figure 3. Especially roads, sidewalks, and vehicles work quite well. These classes follow strict model assumptions regarding surface, shape or size and are frequently represented in the training dataset. This observation motivates to model a class dependent measurement model. Therefore, we estimate the parameters σ_{c_i}, b_{c_i} and λ_{c_i} separately for each class as shown in table 1, which correspond to the density functions in figure 3 colored in orange.

Table 1. Class-dependent measurement model of single image depth prediction

Class c_i	**Road**	**Sidewalk**	**Terrain**	**Building**	**Pole**	**Vegetation**	**Vehicle**
σ_{c_i}	0.0032	0.006	0.007	0.0075	0.008	0.008	0.005
b_{c_i}	0.01	0.02	0.02	0.025	0.03	0.03	0.015
λ_{c_i}	0.15	0.1	0.1	0.2	0.3	0.3	0.2

For the semantic classes two-wheeler, person and sky there is not enough data for a reliable analysis of the statistical distribution. Therefore, we use the overall distribution in figure 2 for these classes. Based on the derived class dependent measurement model the term $\mathit{\Phi}_{SI}(s_i, d_v)$ in equation 5 is defined as:

$$\begin{aligned} \mathit{\Phi}_{SI}(s_i, d_v) &= -\log\left(p(d_{error}|c_i)\right) \\ &\approx min\left(-\log\left(\tfrac{1-\lambda_{c_i}}{\sqrt{2\pi}\sigma_{c_i}}\right) + \tfrac{d^2_{error}}{2\sigma^2_{c_i}},\ -\log\left(\tfrac{\lambda_{c_i}}{2b_{c_i}}\right) + \tfrac{|d_{error}|}{b_{c_i}}\right) \end{aligned} \tag{7}$$

3.3 Solving the mono-stixel segmentation problem

The mono-stixel segmentation problem is defined as the energy minimization problem in equation 2. To solve this segmentation problem we follow the proposed method in [2]. The optimization of the stixel types m_i and segmentation

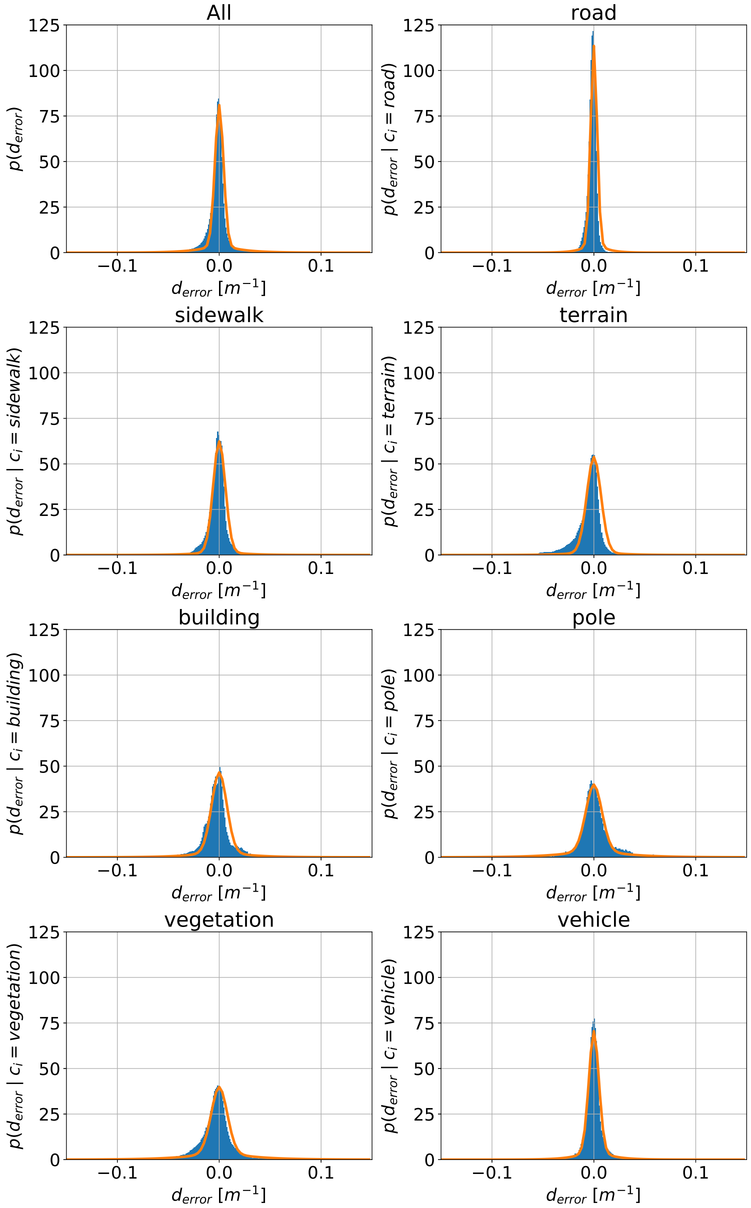

Fig. 3. Statistical distribution of the disparity error of the single image depth prediction [11] dependent on the semantic class. The blue histograms show the empirical distribution of the error on Kitti-Stereo'15 [17] and the orange curve the approximated measurement model.

labels v_i^b, v_i^t is formulated as a minimum path problem solved via dynamic programming. Each edge in the minimum path problem corresponds to a mono-stixel hypothesis of a given segmentation v_i^b, v_i^t and type m_i. To reduce the computational effort the semantic class c_i, inverse depth p_i and translational motion $\mathbf{t}_i$ are locally optimized for the corresponding image segment. We take that semantic class c_i that minimized $\Phi_L(s_i, \mathbf{l}_v)$ considering the association between semantic classes and the mono-stixel type m_i. Thereby, the semantic segmentation supports the segmentation as well as the distinction of the different mono-stixel types.

To estimate the inverse depth p_i and translational motion $\mathbf{t}_i$ we use a MLESAC-based approach [26] which, in contrast to the approach in [2], minimizes the optical flow as well as the single image depth prediction based data likelihood and serves as a statistical fusion:

$$\hat{p}_i, \hat{\mathbf{t}}_i = \underset{p_i, \mathbf{t}_i}{\arg\min} \sum_{v=v_i^b}^{v_i^t} \left(\lambda_F \Phi_F(s_i, \mathbf{f}_v, v) + \lambda_{SI} \Phi_{SI}(s_i, d_v)\right) \tag{8}$$

For static objects and ground stixels both data likelihood terms depend solely on the inverse depth p_i as the translational motion $\mathbf{t}_i$ is zero by definition. Therefore, the MLESAC-based estimation serves as a fusion and takes that inverse depth defined by one optical flow vector or one single image depth estimate in the corresponding image segment v_i^b, v_i^t that minimizes the cost term defined in equation 8. For dynamic object stixel the optical flow related-data likelihood term depends on two degrees of freedom, namely the linear combination of inverse depth and relative 2D-translation of that stixel to the camera $\tilde{\mathbf{t}}_i = p_i \mathbf{t}_{i,cam}$ as shown in [2]. But one degree of freedom, for example, the inverse depth p_i can be chosen freely. In contrast to that, the data likelihood term of the single image depth prediction only depends on the inverse depth p_i of the stixel, but is independent of the translational motion. Consequently, we separate the estimation in two parts. First, we take that inverse depth p_i defined by one single image depth estimate in the corresponding image segment v_i^b, v_i^t, that minimizes the following cost term:

$$\hat{p}_i = \underset{p_i}{\arg\min} \sum_{v=v_i^b}^{v_i^t} \left(\lambda_{SI} \Phi_{SI}(s_i, d_v)\right) \tag{9}$$

Second, we take that labels for the translational motion $\mathbf{t}_i$ defined by one optical flow vector in the corresponding image segment v_i^b, v_i^t that minimizes the optical flow based data likelihood for the given depth $\hat{p}_i$:

$$\hat{\mathbf{t}}_i = \underset{\mathbf{t}_i}{\arg\min} \sum_{v=v_i^b}^{v_i^t} \left(\lambda_F \Phi_F(s_i, \mathbf{f}_v, v)\right) \tag{10}$$

A hypothesis of the 2D-translational motion or inverse depth based on one optical flow vector can be estimated using the direct linear transform of the stixel-homography as explained in [2]. For each single image depth estimate, we can derive a hypothesis of the inverse depth of a stixel given its type by the inversion of equation 4. The scene model additionally rates the consistency of the estimated depth during optimization as a minimum path problem and prefers a mono-stixel segmentation consistent to the defined scene model.

4 Experiments

In this chapter, the performance of our proposed mono-stixel estimation method is analyzed. First, we describe our experimental setup including the used metric. Second, we present the experimental results of our performance evaluation as well as some example results.

4.1 Setup and metric

The inputs of our method, the dense optical flow field, camera motion estimation, pixel-wise semantic labeling, and single image depth prediction are implemented as follows: For the dense optical flow we use the public available DeepFlow [29], the camera motion is provided by the method described in [10] and for the single image depth prediction we use the method proposed by Godard et al. [11] as discussed in section 3.2. For the semantic segmentation, we train our own fully convolutional network [16] based on the VGG-architecture [24]. We pretrain the network on the cityscape dataset [3] and fine-tune it on 470 annotated images of the Kitti dataset [17] as proposed in [23, 2].

Our experiments are performed on the Kitti-Stereo'15 dataset [17]. The dataset consist of 200 short sequences in street scenes with ground truth depth maps provided by a Velodyne laser scanner and 3D-CAD models for moving vehicles. In the first setup the optical flow and camera motion is computed on keyframes with a minimum driven distance of $0.5m$. These keyframes exist for 171 sequences. In the second setup, the optical flow and camera motion is computed on two consecutive images for all 200 sequences. This means that in the second setup also scenarios without or with a quite small translational motion of the camera are included in the dataset.

As a first baseline, we use the mono-stixel estimation approach described in [2] with the same optical flow and camera motion estimation as inputs. Comparing to that baseline shows if we can improve the performance of mono-stixel estimation by exploiting the single image depth predictions. The implemented baseline is exactly the approach described in [2]. However, due to some parameter tuning, we were able to achieve slightly better results than stated in the original paper. The baseline approach and our approach both use a stixel width of $w_s = 5$ and exactly the same parameters, for example, to define the scene model. Furthermore, we present the performance of both inputs to analyze if our approach serves as a suitable fusion of both information. Therefore, we implement a traditional structure from motion (SFM) approach [12] by triangulating each optical flow vector based on the camera motion. Again, the same optical flow

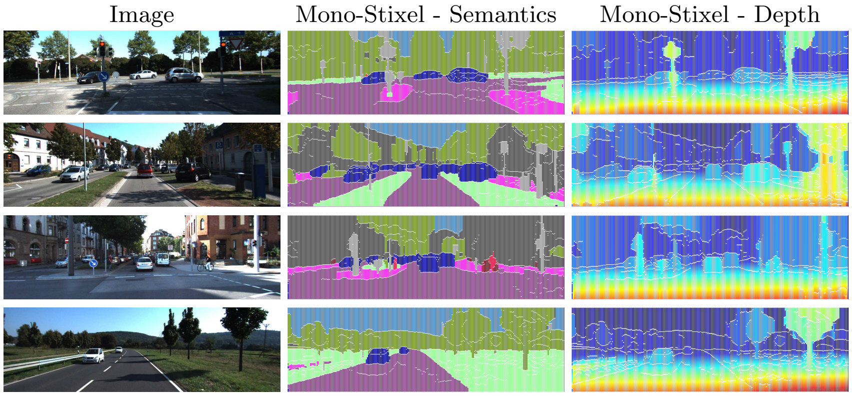

Fig. 4. Example depth and semantic scene representation of our proposed mono-stixel estimation method. The stixel color encodes the semantic class following [3] or the inverse depth from close (red) to far (dark blue), respectively.

and camera motion are used as for our approach. Additionally, the quality of the single image depth prediction in [11] is shown.

Moreover, we perform the experiments for three different measurement models for the single image depth prediction. First, our semantic class dependent mixture model proposed in section 3.2 (ours-SemMixture). Second, the same mixture model but independent of the semantic class (ours-Mixture). Third, a measurement model assuming a normal distribution of the inverse depth error (ours-Normal). The variance of the normal distribution is determined by the mean squared error of the inverse depth.

The results are compared using the depth metric by [4]. The metric measures the root mean squared error (RMSE) of the depth prediction, the mean absolute relative error (Rel. Error) and percentage of depths that fulfill some threshold δ. Note, that the metric is computed in the depth space and not in the inverse depth or disparity space.

4.2 Results

Figure 4 shows some example outputs of our proposed method. The segmentation of each mono-stixel is visualized by a white boundary and the color represents the semantic class or depth of that pixel encoded by the mono-stixel.

In figure 5 the performance of the depth representation is compared to the mentioned baselines. The first image shows that our approach is able to predict reliable depth estimates even for vehicles partly occluded by the image boundary or other objects. This is not the case for the mono-stixel approach in [2] that needs to observe the ground contact point of the mono-stixel for a reasonable depth estimate. The images in the second and third column show that the fusion supported by the scene model is able to correct errors of one of the inputs in many cases. For example, in the second column of figure 5 the depth of the bushes

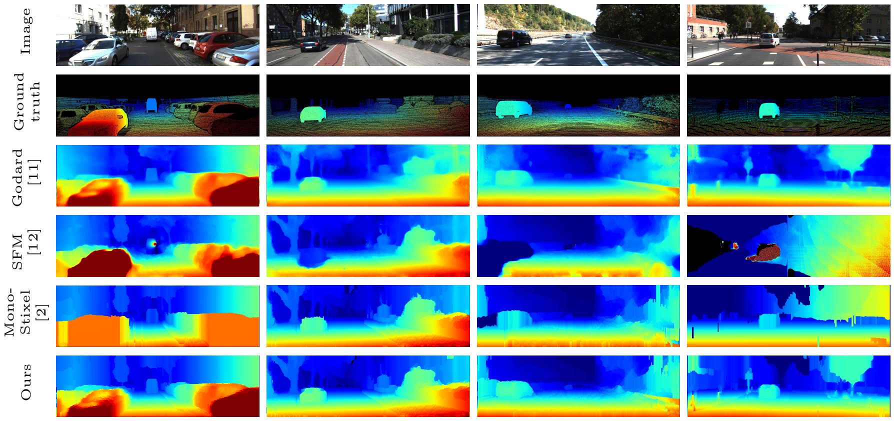

Fig. 5. Example performance of the depth reconstruction of our proposed mono-stixel estimation method compared to the mono-stixel estimation method of Brickwedde et al. [2], structure from motion (SFM) [12] and single image depth prediction by Godard et al. [11]. The color encodes the inverse depth from close (red) to far (dark blue). Invalid negative depth values are colored black. The ego vehicle is standing in the scenario of the last column.

and building in the right part of the image mainly follows the depth defined by the optical flow. But, our approach is also able to correct errors in the optical flow as shown in the third column behind the vehicle or for the guideline on the right side, even though these are parts of the scene with high parallax. The last column additionally shows a scenario with standing ego vehicle. The structure from motion baseline completely fails in that situation, the mono-stixel approach in [2] reconstructs a flat ground plane, projects the dynamic objects on that plane, but fails for the static objects, whereas our approach provides a reasonable depth reconstruction of the whole scene.

Including the single image depth prediction does not have a significant effect on the number of mono-stixels and thus the compactness of the representation. The mean number of mono-stixels per image of the approach in [2] is 1853 which corresponds to 7.4 mono-stixels per column. Compared to that our approach gives out 1944 mono-stixels per image or 7.8 mono-stixels per column in average. Note, that the parametrization is more focused on quality than on compactness. By changing the parameter the number of mono-stixel could be reduced significantly but at the expense of the quality due to higher discretization effects.

Table 2 shows the performance of our method compared to the mentioned baselines for the keyframe-based subset evaluated for all parts of the scene. The results show that our proposed semantic class dependent measurement model outperforms the class independent counterpart and especially the measurement model assuming a normal distribution. Furthermore, our approach significantly improves the quality of the mono-stixel estimation by exploiting the single image

Table 2. Results on Kitti-Stereo'15 for the 171 keyframes evaluated on all parts of the scene

Method	RMSE	Rel. Error	$\delta < 1.1$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SFM [12]	9.36 m	29.11 %	67.02 %	76.81 %	82.21 %	86.02 %
Godard [11]	5.14 m	9.84 %	71.00 %	87.52 %	96.17 %	98.56 %
Mono-Stixel [2]	6.05 m	11.99 %	74.08 %	89.71 %	95.85 %	97.64 %
ours-Normal	4.71 m	8.37 %	79.73 %	92.58 %	97.31 %	98.76 %
ours-Mixture	**4.57 m**	8.01 %	80.76 %	92.70 %	97.36 %	**98.79 %**
ours-SemMixture	**4.57 m**	**7.97 %**	**81.36 %**	**93.04 %**	**97.45 %**	**98.79 %**

depth prediction and serves as a suitable fusion, which is shown by the fact, that the performance is better than each input solely. In table 3 and 4 the same experiment is evaluated for the static and moving parts of the scene separately. For moving objects the structure from motion baseline fails completely and therefore our depth estimation mainly follows the single image depth prediction. Due to the discretization effect and errors in the semantic segmentation, our performance is slightly lower than the single image depth prediction for moving objects. However, for the whole scene, our approach is significantly better as shown in Table 2, the representation is more compact and the translational motion of the moving object stixels is additionally provided. Furthermore, compared to the mono-stixel estimation approach in [2], we show that our approach significantly outperforms the scene model-based reconstruction of moving objects.

Table 3. Results on Kitti-Stereo'15 for the 171 keyframes evaluated on static parts of the scene

Method	RMSE	Rel. Error	$\delta < 1.1$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SFM [12]	7.02 m	16.11 %	76.67 %	87.49 %	92.77 %	94.98 %
Godard [11]	5.38 m	10.26 %	69.32 %	86.23 %	95.78 %	98.43 %
Mono-Stixel [2]	6.01 m	11.20 %	77.54 %	90.81 %	96.03 %	97.71 %
ours-SemMixture	**4.51 m**	**7.79 %**	**81.29 %**	**92.63 %**	**97.37 %**	**98.81 %**

Table 4. Results on Kitti-Stereo'15 for the 171 keyframes evaluated on moving objects

Method	RMSE	Rel. Error	$\delta < 1.1$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SFM [12]	18.51 m	115.17 %	3.12 %	6.15 %	12.36 %	26.71 %
Godard [11]	**3.2 m**	**7.1 %**	**82.18 %**	**96.07 %**	**98.81 %**	**99.48 %**
Mono-Stixel [2]	6.32 m	17.21 %	51.14 %	82.43 %	94.65 %	97.21 %
ours-SemMixture	4.98 m	9.17 %	81.83 %	95.73 %	97.98 %	98.68 %

In table 5 the same experiment is shown, but with the optical flow and camera motion computed on consecutive frames. Thus, there are still many scenarios with a moving camera, but also some cases without any or a quite small translational motion. Consequently, the performance of the SFM and mono-stixel [2] baselines drop significantly. For example by around 7% and 5% for the accuracy threshold of $\delta < 1.1$. Our approach is still able to handle situations without translational motion and thereby the quality only decreases slightly by around 1% for the same accuracy threshold. The small deterioration is explainable by the fact, that there are lower parallax configurations for the consecutive frames. However, the optical flow is still useful even in standstill situations to support the distinction between static and moving objects and to estimate the translational motion of the moving objects.

Table 5. Results on Kitti-Stereo'15 for the 200 consecutive frames evaluated on all parts of the scene

Method	RMSE	Rel. Error	$\delta < 1.1$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SFM [12]	16.06 m	59.20 %	59.55 %	69.71 %	76.27 %	81.31 %
Godard [11]	5.20 m	9.68 %	71.63 %	87.91 %	96.25 %	98.58 %
Mono-Stixel [2]	7.26 m	13.90 %	69.68 %	87.09 %	94.53 %	96.80 %
ours-SemMixture	**4.88 m**	**8.24 %**	**80.27 %**	**92.60 %**	**97.28 %**	**98.73 %**

5 Conclusions

We have presented an extension of the mono-stixel estimation by exploiting the recent advantages in single image depth prediction. The mono-stixel estimation serves as a statistical fusion of the single image depth prediction and optical flow supported by scene model assumptions and semantic segmentation. For a statistically reasonable fusion, we tackle the challenging problem of providing a statistical error distribution for deep learning based single image depth estimates in a self-supervised learning approach and proposed a semantic class dependent measurement model derived by the empirical error distribution.

Our approach is able to significantly improve the quality of mono-stixel estimation and handle partly occluded moving objects as well as scenarios without translational motion of the camera. Both cases might be highly relevant for a driver assistance system or autonomous vehicles.

References

1. Badino, H., Franke, U., Pfeiffer, D.: The Stixel World - A compact medium level representation of the 3D-world. In: Joint Pattern Recognition Symposium. pp. 51–60. Springer (2009)
2. Brickwedde, F., Abraham, S., Mester, R.: Mono-Stixels: Monocular Depth Reconstruction of Dynamic Street Scenes. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–7. IEEE (2018)

3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
5. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: European Conference on Computer Vision. pp. 834–849. Springer (2014)
6. Fanani, N., Stürck, A., Ochs, M., Bradler, H., Mester, R.: Predictive monocular odometry (PMO): What is possible without RANSAC and multiframe bundle adjustment? Image and Vision Computing (2017)
7. Fcil, J.M., Concha, A., Montesano, L., Civera, J.: Single-view and multi-view depth fusion. IEEE Robotics and Automation Letters **2**(4), 1994–2001 (Oct 2017). https://doi.org/10.1109/LRA.2017.2715400
8. Garg, R., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision. pp. 740–756. Springer (2016)
9. Garnett, N., Silberstein, S., Oron, S., Fetaya, E., Verner, U., Ayash, A., Goldner, V., Cohen, R., Horn, K., Levi, D.: Real-time category-based and general obstacle detection for autonomous driving. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
10. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3D reconstruction in real-time. In: Intelligent Vehicles Symposium (IV), 2011 IEEE. pp. 963–968. Ieee (2011)
11. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
12. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
13. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in neural information processing systems. pp. 5574–5584 (2017)
14. Klappstein, J.: Optical-flow based detection of moving objects in traffic scenes. Ph.D. thesis (2008)
15. Levi, D., Garnett, N., Fetaya, E., Herzlyia, I.: Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In: BMVC. pp. 109–1 (2015)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
17. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
18. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
19. Pereira, F.I., Ilha, G., Luft, J., Negreiros, M., Susin, A.: Monocular visual odometry with cyclic estimation. In: Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on. pp. 1–6. IEEE (2017)
20. Pfeiffer, D., Franke, U.: Modeling dynamic 3D environments by means of the Stixel World. IEEE Intelligent Transportation Systems Magazine **3**(3), 24–36 (2011)
21. Pfeiffer, D., Franke, U.: Towards a global optimal multi-layer Stixel representation of dense 3D data. In: BMVC. vol. 11, pp. 51–1 (2011)

22. Ranftl, R., Vineet, V., Chen, Q., Koltun, V.: Dense monocular depth estimation in complex dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4058–4066 (2016)
23. Schneider, L., Cordts, M., Rehfeld, T., Pfeiffer, D., Enzweiler, M., Franke, U., Pollefeys, M., Roth, S.: Semantic Stixels: Depth is not enough. In: Intelligent Vehicles Symposium (IV), 2016 IEEE. pp. 110–117. IEEE (2016)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Tateno, K., Tombari, F., Laina, I., Navab, N.: CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2017)
26. Torr, P.H., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding **78**(1), 138–156 (2000)
27. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: DeMoN: Depth and motion network for learning monocular stereo. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
28. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: Sfm-net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804 (2017)
29. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: Large displacement optical flow with deep matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1385–1392 (2013)
30. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)