SEMANTIC-GUIDED REPRESENTATION ENHANCEMENT FOR SELF-SUPERVISED MONOCULAR TRAINED DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised depth estimation has shown its great effectiveness in producing high quality depth maps given only image sequences as input. However, its performance usually drops when estimating on border areas or objects with thin structures due to the insufficient depth representation. In this paper, we address this problem by proposing a semantic-guided depth representation enhancement method, which promotes both local and global depth feature representations by leveraging rich contextual information. In stead of a single depth network as used in conventional paradigms, we propose an extra semantic segmentation branch to offer extra contextual features for depth estimation. Based on this framework, we enhance the local feature representation by sampling and feeding the point-based features that locate on the semantic edges to an individual *Semantic-guided Edge Enhancement module (SEEM)*, which is specifically designed for promoting depth estimation on the challenging semantic borders. Then, we improve the global feature representation by proposing a semantic-guided multi-level attention mechanism, which enhances the semantic and depth features by exploring pixel-wise correlations in the multi-level depth decoding scheme. Extensive experiments validate the distinct superiority of our method in capturing highly accurate depth on the challenging image areas such as semantic category borders and thin objects. Both quantitative and qualitative experiments on KITTI show that our method outperforms the state-of-the-art methods.

1 INTRODUCTION

Depth estimation is a long standing problem in computer vision community, which offers useful information to a wide range of tasks including robotic perception, augmented reality and autonomous driving, *etc.* Compared with depth estimation methods which rely on active vision or multi-view paradigms (Schonberger & Frahm, 2016; Li et al., 2019), estimating depth from only a single image is highly ill-posed, thus brings greater challenges for higher quality results.

In recent years, monocular depth estimation has witnessed new renaissances with the advent of deep learning (He et al., 2016; Jaderberg et al., 2015; Simonyan & Zisserman, 2014). By learning deep representations in a supervised manner, various of networks (Eigen & Fergus, 2015; Eigen et al., 2014; Laina et al., 2016) are capable of producing high quality depth maps thanks to the large corpus of training data. At the mean time, consider the lack of labeled training data for network training, recent advances (Zhou et al., 2017; Godard et al., 2019; 2017) show that monocular depth estimation can be accomplished in a self-supervised way. The network can be trained by unlabeled image sequences using two-view geometric constraints, while achieving comparable results with the supervised paradigm. The learning-based methods managed to handle the highly ill-posed monocular depth estimation problem by implicitly learning the mapping between visual appearance and its corresponding depth value.

However, despite the great effectiveness for learning-based depth estimation, these methods still struggle to conduct precise depth estimation on challenging image regions such as semantic category borders or thin object areas. For example, the estimated object depth usually fails to align with the real object borders, and the depth of foreground objects which have thin structures tends to be

submerged in the background. We attribute these phenomena to the insufficient depth representation that (1) the pixel-wise local depth information can not be well represented by current depth network, especially on highly ambiguous, semantic border areas, (2) current depth representations are not capable of well describing depth foreground/background relationships globally. These issues lead to the wrong depth estimation from the true scene structures, which hinders the further applications in the real-world tasks.

In this paper, we address this problem via enhancing both local and global depth feature representations for self-supervised monocular depth estimation via semantic guidance. As semantic segmentation conducts explicit category-level scene understandings and produces well-aligned object boundary detection, we propose an extra semantic estimation branch inside the self-supervised paradigm. The semantic branch offers rich contextual features which is fused with depth features during multi-scale learning. Under this framework, we propose to enhance the depth feature using semantic guidance in a *local-to-global* way. To improve the insufficient local depth representations on semantic category borders, inspired by the sampling and enhancing strategies used in semantic segmentation (Kirillov et al., 2020), **our first contribution** is to propose a *Semantic-guided Edge Enhancement Module* (SEEM) which is specially designed to enhance the local point-based depth representations that locate on the semantic edges. Different from the method in (Kirillov et al., 2020), we enhance the point features using multi-level of representations from different domains under a self-supervised framework. To be specific, we sample a set of point positions lying on the binary semantic category borders, and extract the point-wise features of the corresponding edge positions from the encoded feature, depth decoding feature as well as the semantic decoding feature. We then merge and enhance the point-wise features and feed them to the final depth decoding features to promote the edge-aware depth inference for self-supervision. For global depth representation enhancement, **our second contribution** is to propose a semantic-guided multi-level attention module to improve the global depth representation. Different from the conventional self-attentions (Fu et al., 2019; Vaswani et al., 2017) which are implemented as single modules on the bottleneck feature block, we propose to explore the self-attentions on different level of decoding layers. In this way, both semantic and depth representations can be further promoted by exploring and leveraging the pixel-wise correlations inside of their fused features.

We validate our method mainly on KITTI 2015 (Geiger et al., 2012), and the Cityscape (Cordts et al., 2016) benchmark is also used for evaluating the generalization ability. Experiments show that the proposed method significantly improves the depth estimation on category edges and thin scene structures. Extensive quantitative and qualitative results validate the superiority of our method that it outperforms the state-of-the-art methods for self-supervised monocular depth estimation.

2 Related Work

There exist extensive researches on monocular depth estimation, including geometry-based methods (Schonberger & Frahm, 2016; Enqvist et al., 2011) and learning-based methods (Eigen et al., 2014; Laina et al., 2016). In this paper, however, we concentrate only on the self-supervised depth training and semantic-guided depth estimation, which is highly related to the research focus of this paper.

Self-supervised depth estimation. Self-supervised methods enable the networks to learn depth representation from merely unlabeled image sequences that they reformulate the depth supervision loss into the image reconstruction loss. Godard et al. (2017) and Garg et al. (2016) first propose the self-supervised method on stereo images, then Zhou et al. (2017) propose a monocular trained approach using a separate motion estimation network. Based on these frameworks, a large corpus of works seek to promote self-supervised learning from different aspects. For more robust self-supervision signals, Mahjourian et al. (2018) propose to use 3D information as extra supervisory signal, another kind of methods leverage additional information such as the optical flow (Ranjan et al., 2019; Wang et al., 2019b) to strengthen depth supervision via consistency constraints. In order to solve the loss deviation problems on non-rigid motion areas, selective masks are used to filter out the these areas when computing losses. Prior works generate the mask by the network itself (Zhou et al., 2017; Yang et al., 2017; Vijayanarasimhan et al., 2017), while the succeeding methods produce the mask by leveraging geometric clues (Bian et al., 2019; Wang et al., 2019a; Godard et al., 2019; Klingner et al., 2020), which is proved to be more effective. There also exist

other methods trying to enhance the network performance with traditional SfM (Schonberger & Frahm, 2016), which offer extra pseudo labels for depth estimation.

Depth estimation using semantics. Semantic information are shown to provide positive effects toward depth estimation framework in the previous works. The methods can be categorized into two groups by their way to use the semantic information. One group of the methods leverage the semantic labels directly to guide the depth learning. The works of Ramirez et al. (2018) and Chen et al. (2019a) propose depth constraints by leveraging the semantic labels of the scene. Klingner et al. (2020) and Casser et al. (2019) address the non-rigid motion issues by handling the moving object areas highlighted by the semantic map. Wang et al. (2020) use a divide-and-conquer strategy to conduct depth estimation with different semantic categories, and Zhu et al. (2020) offer a depth morphing paradigm with the help of semantic foreground edges. The other group of methods enhance the depth representation by feature manipulation. Chen et al. (2019a) generated depth and semantic maps by a single latent representation. Ochs et al. (2019) proposed a segmentation-like loss item for depth estimation. Guizilini et al. (2020b) leveraged a pre-trained semantic segmentation network and further conduct feature merging by PAC-based (Su et al., 2019) module.

In this paper, we propose an individual semantic network for depth estimation. But different with Guizilini et al. (2020b), our semantic branch shares the same encoder and is trained together with the depth network for better scene representation. We leverage the semantic information in both ways. The semantic edges are used to guide the edge-based point feature sampling for local depth depth representation enhancement. At the mean time, we utilize the semantic feature blocks to conduct multi-scale feature fusion for global depth representation enhancement. The proposed method constructs a holistic way for feature representation enhancing in self-supervised depth estimation.

3 Self-supervised Depth Estimation Framework

The proposed method builds upon the self-supervised depth estimation framework. An image triplet (I_{t-1}, I_t, I_{t+1}) is used as the input, where I_t acts as the target image and $I_{t'} \in \{I_{t-1}, I_{t+1}\}$ are the source images. During training, the target image I_t is fed into the depth network f_D to get the estimated depth $D_t = f_D(I_t)$, while the adjacent image pairs $(I_t, I_{t'})$ are put into the ego-motion network for the 6-DoF ego-motion estimations $T_{t'\to t}$. Then, the synthesized target images $I_{t'\to t}$ can be estimated using the source images $I_{t'}$, depth D_t and the ego-motions $T_{t'\to t}$, following the formulation from Zhou et al. (2017). The self-supervised loss can be calculated by

$$L(I_t, I_{t'\to t}) = L_p(I_t, I_{t'\to t}) + \lambda L_s(I_t, I_{t'\to t}), \tag{1}$$

where λ is the weighting factor between the photometric loss L_p and smoothness loss L_s. Following Godard et al. (2019), we implement the minimum photometric loss

$$L_p(I_t, I_{t'\to t}) = \min_{t'}\left(\frac{\alpha}{2}\left(1 - \text{SSIM}\left(I_t, I_{t'\to t}\right)\right) + (1-\alpha)\left\|I_t - I_{t'\to t}\right\|_1\right), \tag{2}$$

where SSIM is the Structural Similarity item (Wang et al., 2004), α is the weighting factor which is set to 0.85. The smoothness loss item (Godard et al., 2017) is implemented to smooth the depth map by the consistency between the image and depth gradient

$$L_s(I_t, I_{t'\to t}) = |\partial_x D_t|\, e^{-|\partial_x I_t|} + |\partial_y D_t|\, e^{-|\partial_y I_t|}, \tag{3}$$

where ∂_x, ∂_x denote gradient operation on x,y-axis. In our implementation, we also conduct auto-masking and multi-scale depth upsampling strategy as proposed in Godard et al. (2019) to further improve depth estimation.

4 The Proposed Method

In this paper, we promote the self-supervised depth estimation paradigm via proposing a semantic-guided representation enhancing depth network. The overview of the method is shown in Figure 1. We propose a multi-task framework which consists of the depth and semantic decoding branch with shared feature encoder. During the training stage, the semantic branch feeds contextual features to the depth branch to enhance depth comprehension. Under this multi-task framework, we propose a Semantic-guided Edge Enhancement Module (SEEM) to enhance the local depth feature representation on semantic category edges. Meanwhile, we enhance the global depth feature representations by

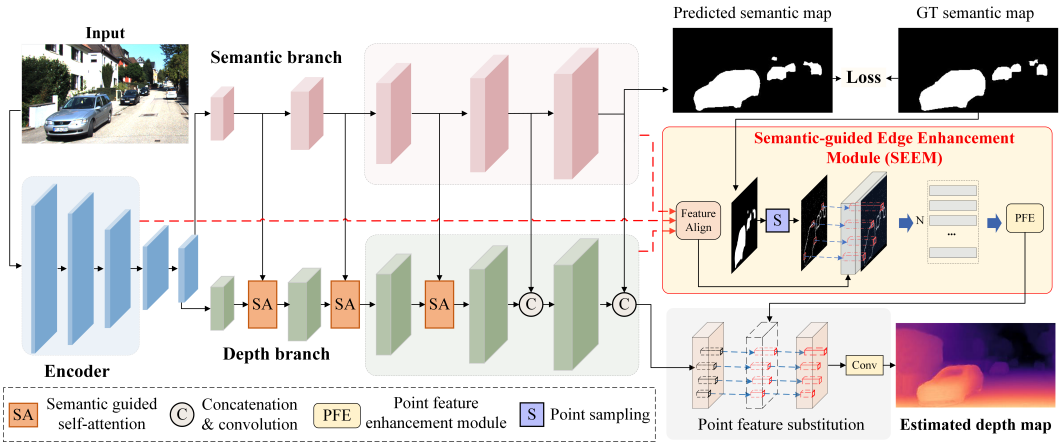


Figure 1: **The overview of the proposed architecture**. An extra semantic branch is proposed to offer contextual information for depth representations. Then, we enhance the local depth feature representation via sampling and enhancing semantic edge-based point features via the proposed *Semantic-guided Edge Enhancement module* (SEEM). The global feature representation is promoted by the proposed semantic-guided multi-layer self-attention (SA), which explores pixels feature correlations between fused depth and semantic features.

proposing the semantic-guided multi-level self-attention module, which improves both the semantic and depth feature representations by digging into the pixel-wise dependencies.

4.1 The Proposed Semantic Segmentation Branch

We extend the general depth network with an extra semantic segmentation branch to provide explicit contextual information for depth estimation. The semantic branch shares the same network structure and input feature blocks with the depth branch. During network training, it feeds the semantic features to the depth network in a multi-layer way, to facilitate better depth representation and used for further depth representation enhancement processing. Different with general segmentation tasks, it outputs the binary semantic mask, which implicitly encodes distance information by assigning foreground and background areas. This setting is empirically proved to offer abundant guidances for better depth estimation, while leading to less computational burden. Details can be found in Section C of the Appendix. Consider we build our method on the self-supervised scenarios, we use the semantic pseudo label generated by off-the-shelf method (Zhu et al., 2019) to supervise the semantic branch, which requires no groundtruth labels.

4.2 Semantic-guided Edge Enhancement Module (SEEM)

We improve depth estimation on semantic category edges via sampling and enhancing the local point features via the proposed *Semantic-guided Edge Enhancement Module* (SEEM). As shown in Figure 2, SEEM takes the encoding feature blocks F_E, depth decoding feature blocks F_D as well as the semantic decoding feature blocks F_S as input, and output a set of point features $F^{'} \in \mathbb{R}^{N \times C}$, where N is the number of sampled points, C denotes the output feature channel which is consistent with the the final output depth features. We first sample the local points according to the semantic edges. Then, we extract the point features from input feature blocks via grid-sampling, and enhance these feature representations via merging and feeding them into a point feature enhancement module. The resulting representative point features are used to replace the final output depth features on the corresponding positions, to further promote the depth representations on semantic category edges.

Semantic-guided edge point sampling strategy. A vanilla sampling strategy is to sample all points directly on semantic edges. However, it shows less robustness that (1) the semantic edge is produced by pseudo labels instead of GT labels, so it may not sample the real edges, and (2) the points nearby the semantic edges are also challenging areas for depth estimation, which should also be taken into consideration. To address these issues, we propose a sampling strategy which incorporates enough redundancies. We first compute the semantic edges response via convolving the binary semantic mask with the Sobel operator, and then extract the edges points set $\mathcal{S}_{\mathcal{E}} \in \mathbb{R}^{\mu N \times 2}$ which consists of μN points that exhibit largest responses from the edge response map. Besides the points lying

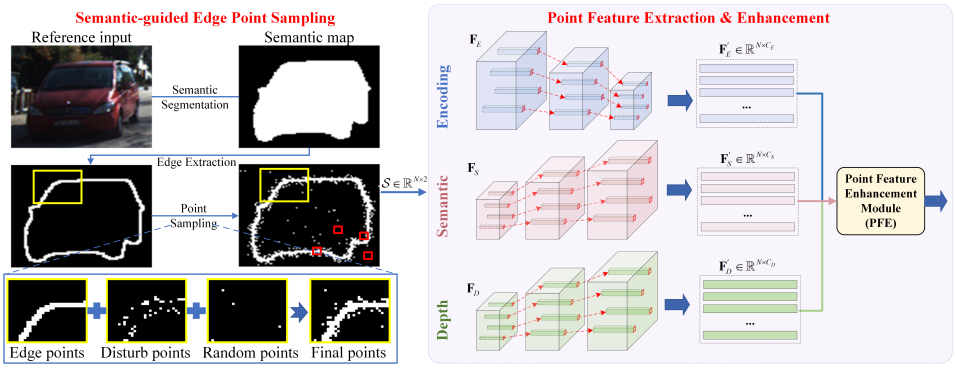


Figure 2: **The semantic-guided edge enhancement module (SEEM)**. The proposed method first samples and extracts the semantic edge-located point set which consists of the edge points, disturbed edge points and random points. Then, the points features are extracted via bilinear sampling and fed into the point feature enhancement module which includes a set of 1-D convolutions, to enhance the local point depth feature representations independently.

exactly on the edges, we choose another set of μN uniformly sampled random points in a small range $[-c, c]$, and use this random point set to disturb the point positions of $\mathcal{S}_{\mathcal{E}}$ to get a new disturbed edge point set $\mathcal{S}_{\mathcal{D}} \in \mathbb{R}^{\mu N \times 2}$, in order to sample the real ground truth edge points and nearby points. Finally, a point set $\mathcal{S}_{\mathcal{R}} \in \mathbb{R}^{(1-\mu N) \times 2}$ with $(1-2\mu)N$ points randomly sampled on the global image is selected to generalize the network's learning abilities. The final point set $\mathcal{S} \in \mathbb{R}^{N \times 2}$ is shown in Equation 4, where $\cup$ is the union operator.

$$\mathcal{S} = \mathcal{S}_{\mathcal{E}} \cup \mathcal{S}_{\mathcal{D}} \cup \mathcal{S}_{\mathcal{R}}, \tag{4}$$

Point feature extraction and enhancement. For the input feature block among F_E, F_D, F_S, they are free to contain flexible layers of features. In this paper, we select the $0-2$ layers of features that we regard these features offer abundant information for image, semantics, and depth representations. We extract the point features via the sampled point set $\mathcal{S}$

$$F_j^{'} = F_j(S), j \in \{E, D, S\}. \tag{5}$$

In Equation 5, $F_j^{'} \in \mathbb{R}^{N \times C_j}, j \in \{E, D, S\}$ denote the extracted point features from the three input features blocks. In each input feature block, point feature extraction across different size of feature maps is conducted by bilinear interpolation (Jaderberg et al., 2015). We then feed these point-wise feature representations from different domains (*e.g.*, image, geometry and semantic domains) to a point feature enhancement module, which merges and enhances these point representations via a set of 1-D convolutions. The output is the point feature block $F \in \mathbb{R}^{N \times C}$

$$F^{'} = f_\theta(F_E^{'}, F_D^{'}, F_S^{'}), \tag{6}$$

where f_θ denotes the feature enhancement network, the output dimension C is consistent with the last output features of the depth branch. We then substitute the depth feature with enhanced point feature representations F on the corresponding position of $\mathcal{S}$. Since the module is specially designed for enhancing point-wise depth feature representations on semantic border areas, with the help of multi-domain feature representations, it's capable of conducting better depth estimation as is shown in the experiment section.

4.3 Semantic-guided Multi-level Self-Attention

In the scene perception task, the depth information is indeed, implicitly reflecting the contextual information in the scene. For instance, some semantic category areas such as cars or pedestrians often act as the foreground, thus they consistently exhibit smaller depth than their surrounding pixels. In our method, the semantic segmentation branch provides contextual information on each decoding layer of the depth branch. Consider the inherent correlation between depth and semantic representations as analysed above, we propose the multi-level self-attention which explores the pixel-wise correlation inside the fused depth and semantic feature, to further improve the global feature representations for better depth estimation.

Table 1: **Quantitative results on KITTI 2015**. The best results are in **bold** and the second best results are underlined. 'D' denotes depth supervision, 'S' and 'M' refer to self-supervision methods using stereo images and monocular images, respectively. 'Inst' and 'Sem' mean methods that leverage instance or semantic segmentation information. 'R50' refers to the method which uses Resnet50 as the backbone encoder. All monocular trained methods are reported without post-processing steps. The metrics marked in blue mean 'lower is better', while these in red refer to 'higher is better'. Our method outperforms the state-of-the-arts in most metrics by a large margin.

Method	Training	Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^2$
Godard et al. (2017)	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
Pillai et al. (2019)	S	0.112	0.875	4.958	0.207	0.852	0.947	0.977
Watson et al. (2019)	S	0.106	0.780	4.695	0.193	0.875	0.958	0.980
Godard et al. (2019)	MS	0.106	0.806	4.630	0.193	0.876	0.958	0.980
Zhou et al. (2017)	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Mahjourian et al. (2018)	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Yin & Shi (2018)	M	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Wang et al. (2018)	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Ranjan et al. (2019)	M	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Chen et al. (2019b)	M	0.135	1.070	5.230	0.210	0.841	0.948	0.980
Godard et al. (2019)	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Guizilini et al. (2020a)	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
Johnston & Carneiro (2020)	M	0.106	0.861	4.699	0.185	0.889	0.962	0.982
Casser et al. (2019)	M+Inst	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Chen et al. (2019a)	M+Sem	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Guizilini et al. (2020b)-(R50)	M+Sem	0.113	0.831	4.663	0.189	0.878	**0.971**	**0.983**
Ours	M+Sem	**0.105**	**0.764**	**4.550**	**0.181**	**0.891**	0.965	**0.983**

Conventional self-attention mechanisms are usually implemented on the bottleneck features as a single module. In this paper, we propose the semantic-guided multi-level attention which instead consists of a set of multi-level self-attentions that locate on the multi-level decoding layers. The attention is conducted on both depth and semantic features. In the depth decoding stage, suppose F_D^i and F_S^i the i-th level of feature maps from the depth and semantic branch, we concatenate the two features and fuse them via a set of convolutions to get the feature map F_A^i. For a given fused feature $F_A^i \in \mathbb{R}^{C_i \times H_i \times W_i}$, the self-attention module calculates the query feature, key feature and value feature $\{Q_i, Y_i, V_i\} \in \mathbb{R}^{C_i \times L_i}$ via convolution operation. H_i and W_i are height and width of features on the i-th level, and $L_i = H_i \times W_i$. The point-wise attention on the i-th level can be calculated by

$$A_i = softmax(Q_i \cdot Y_i^T). \tag{7}$$

Finally, the output feature can be formulate as the weighted matrix product of the value feature and the attention, added by the residual input feature

$$F_{out}^i = \gamma(A_i \cdot V_i) + F_A^i, \tag{8}$$

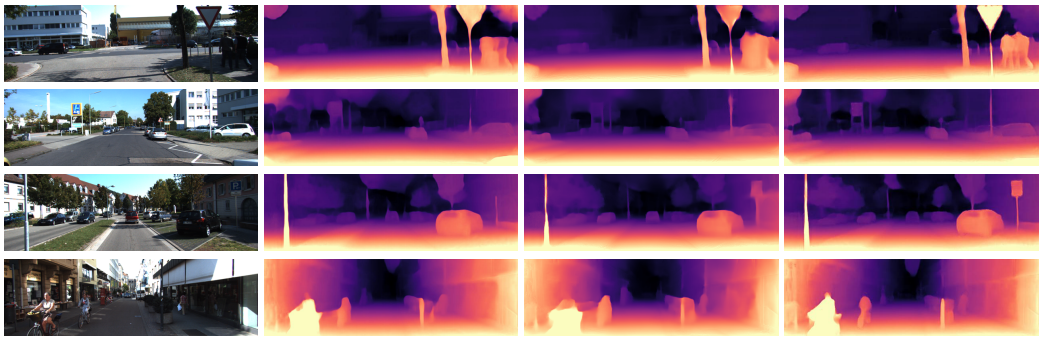
where γ is a learnable weight factor which is initially set to 0. In our paper, due to the limited computational capacity, we set $i \in [2, 4]$.

5 Experiments

In this section, we conduct extensive experiments to validate the superiority of the proposed method. We use standard KITTI benchmark (Geiger et al., 2012) to train and evaluate the proposed network. The Eigen's split (Eigen et al., 2014) along with Zhou's pre-processing scheme (Zhou et al., 2017) is selected which leads to 39810 images for training, 4424 images for validation and 697 images for testing. We also use the Cityscape dataset (Cordts et al., 2016) to evaluate the generalization ability of the proposed method. The dataset is directly tested on its test split which consists of 1525 images.

5.1 Implementation Details

Our method is build upon Monodepth2 (Godard et al., 2019), with ResNet-50 (He et al., 2016) initialized by the ImageNet (Deng et al., 2009) as the backbone. The code is implemented with Pytorch (Paszke et al., 2017) and trained with 1 NVIDIA Tesla V100 GPU. The input image size is 192×640. The motion network and depth branch are trained with self-supervision. For the semantic branch, the network is trained in the supervised manner by the pseudo label generated by off-the-shelf method, which requires no groundtruth labels. We use the pre-computed semantic labels from



Input | Godard et al. (2019) | Guizilini et al. (2020a) | **Ours**

Figure 3: **Qualitative comparison on KITTI 2015**. Our method shows its distinct advantages against other methods in generating sharp depth borders as well as thin structures that is well-aligned with the semantic contextual information.

Zhu et al. (2019) for training and testing on KITTI, and directly use the ground truth semantic labels when testing on Cityscape. After getting semantic labels, we conduct pre-processing to compute the binary semantic labels by assigning the pixels value of the traffic signs, different type of vehicles, and people to 1 as the foreground. During the point sampling process, we set $N = 3000$ that the number is enough to cover most of the edge areas on the images, μ is set to 0.4 by calculating the average number of edge points from randomly selected 10000 images on KITTI, and c is set to 3 empirically. The network is trained with batch size 12 for 19 epochs, with learning rate of 10^{-4} divided by 10 after 15 epochs. The final model is selected via the validation set.

5.2 KITTI Results

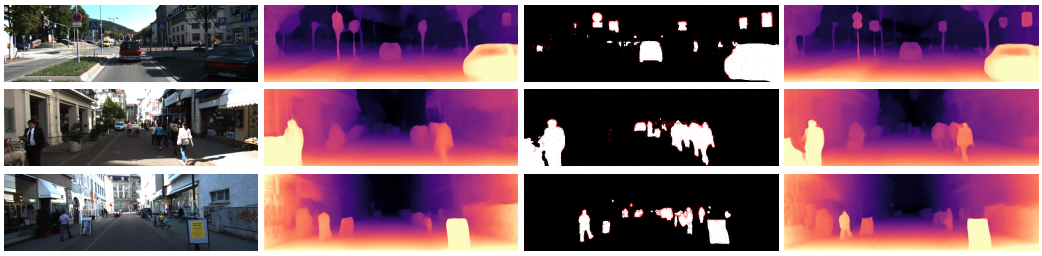
We compare our method with state-of-the-art self-supervised methods including stereo-trained methods, monocular-trained methods as well as other semantic-guided methods. For fair comparison, the image resolution of the latest methods is set to 192×640. We evaluate the latest semantic-guided method (Guizilini et al., 2020b) with its ResNet-50 backbone setting, which is the same with ours. Quantitative results are shown in Table 1. We observe that the proposed method outperforms all the current state-of-the-art self-supervised methods by a substantial margin, even including the stereo trained methods under the same input settings. The qualitative results are shown in Figure 3. We can see that compared with other recent state-of-the-art methods (Godard et al., 2019; Guizilini et al., 2020a), our method shows its remarkable advantages that our method (1) precisely recovers the depth contours of the semantic objects such as humans or traffic signs, even thin structures such as human legs or traffic sign poles can also be well estimated, (2) Our method precisely distinguishes between the depth foreground and background, which lead to better estimations of the global scene structures. More results can be found in the Appendix.

5.3 Ablation Studies

To further validate the effectiveness of the proposed contributions, We conduct ablation studies to analyze the performance improvements from individual contributions and their collaboration toward the baseline. The quantitative results are shown in Table 2. We observe that both SEEM module and the semantic-guided multi-level attention show distinct improvements, and their combination consistently improve the performance toward the baseline. We also show the visual improvements of our method in Figure 4. The baseline and the final results are shown in the second and final column. Semantic guidances are shown in the third column, where white/black denote the fore-

Table 2: **Ablation experiments**. Results of several versions of our proposed method on KITTI 2015 (Geiger et al., 2012) are reported. The best results are in **bold** and the second best are underlined.

Method	SEEM	SA	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	$\delta \leq 1.25$	$\delta \leq 1.25^2$	$\delta \leq 1.25^2$
Baseline	✗	✗	0.110	0.830	4.639	0.187	0.884	0.962	0.982
Baseline+SEEM	✓	✗	0.107	0.771	4.579	0.184	0.887	0.963	**0.983**
Baseline+SA	✗	✓	0.106	0.794	4.578	0.183	0.890	0.964	**0.983**
Baseline+SEEM+SA	✓	✓	**0.105**	**0.764**	**4.550**	**0.181**	**0.891**	**0.965**	**0.983**



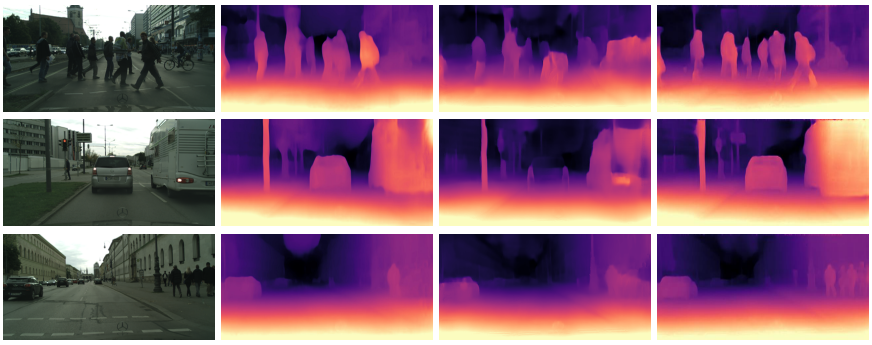
Input Baseline Sematic guidance **Ours**

Figure 4: **Qualitative results of the proposed method against the baseline**. The third column refers to the semantic guidance, in which the black and white area denote the foreground and background, and the red dots are the sampled points of the proposed SEEM module.

ground/background area, the red dots represents the sampled point positions. We can see from the figure that with the help of semantic guidances, the details of the final depth maps are significantly promoted, especially on the binary semantic edges, as well as the point-sampled areas. It indicates that our proposed contributions successfully leverage the semantic information to recover the detailed depth information on foreground semantic categories, as well as constraining the depth boundaries to be aligned with the semantic object borders.

5.4 CITYSCAPE RESULTS

We further validate the generalization ability of our proposed method by testing on Cityscape dataset (Cordts et al., 2016) with models trained only on KITTI (Geiger et al., 2012). The qualitative results are reported in Figure 5. When tested on the unseen datasets, state-of-the-art methods (Godard et al., 2019; Guizilini et al., 2020a) generate blurry object depth or even conduct wrong estimation. While our method conducts accurate and consistent depth estimation inside every object categories, which succeeds to distinguish the foreground from background. At the meantime, our method managed to recover detailed depth information on thin structures from even distant objects. We attribute these to the fact that our method benefits a lot from the semantic contextual information, which offers essential representation priors on the unseen datasets. Additional results can be found in the Appendix.



Input Godard et al. (2019) Guizilini et al. (2020a) **Ours**

Figure 5: Qualitative results on Cityscape (Cordts et al., 2016). All methods are trained on KITTI (Geiger et al., 2012) to evaluate the generalization abilities.

6 CONCLUSION

In this paper, we propose a semantic-guided feature representation enhancement paradigm for self-supervised monocular trained depth estimation. By proposing an extra semantic branch, our method learns semantic contextual information for depth estimation. We further enhance the depth feature representations in local and global manner, by the proposed semantic-guided edge enhancement module as well as the semantic-guided multi-layer self-attention mechanism. Extensive experiments validate the superiority of our method. In the future research, we plan to involve explicit constraints with semantic information for better depth estimation.

REFERENCES

Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems*, pp. 35–45, 2019.

Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8001–8008, 2019.

Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2624–2632, 2019a.

Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7063–7072, 2019b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pp. 2650–2658, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pp. 2366–2374, 2014.

Olof Enqvist, Fredrik Kahl, and Carl Olsson. Non-sequential structure from motion. In *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, 2011.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, 2019.

Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pp. 740–756. Springer, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. IEEE, 2012.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3828–3838, 2019.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020a.

Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pp. 2017–2025, 2015.

Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4756–4765, 2020.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. *arXiv preprint arXiv:2007.06936*, 2020.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248. IEEE, 2016.

Rui Li, Dong Gong, Jinqiu Sun, Yu Zhu, Ziwei Wei, and Yanning Zhang. Robust and accurate hybrid structure-from-moti. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 494–498. IEEE, 2019.

Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5667–5675, 2018.

Matthias Ochs, Adrian Kretz, and Rudolf Mester. Sdnet: Semantically guided depth estimation network. In *German Conference on Pattern Recognition*, pp. 288–302. Springer, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9250–9256. IEEE, 2019.

Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Asian Conference on Computer Vision*, pp. 298–313. Springer, 2018.

Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12240–12249, 2019.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11166–11175, 2019.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pp. 11–20. IEEE, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.

Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2022–2030, 2018.

Guangming Wang, Hesheng Wang, Yiling Liu, and Weidong Chen. Unsupervised learning of monocular depth and ego-motion using multiple masks. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4724–4730. IEEE, 2019a.

Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 541–550, 2020.

Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8071–8081, 2019b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2162–2171, 2019.

Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv preprint arXiv:1711.03665*, 2017.

Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992, 2018.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1858, 2017.

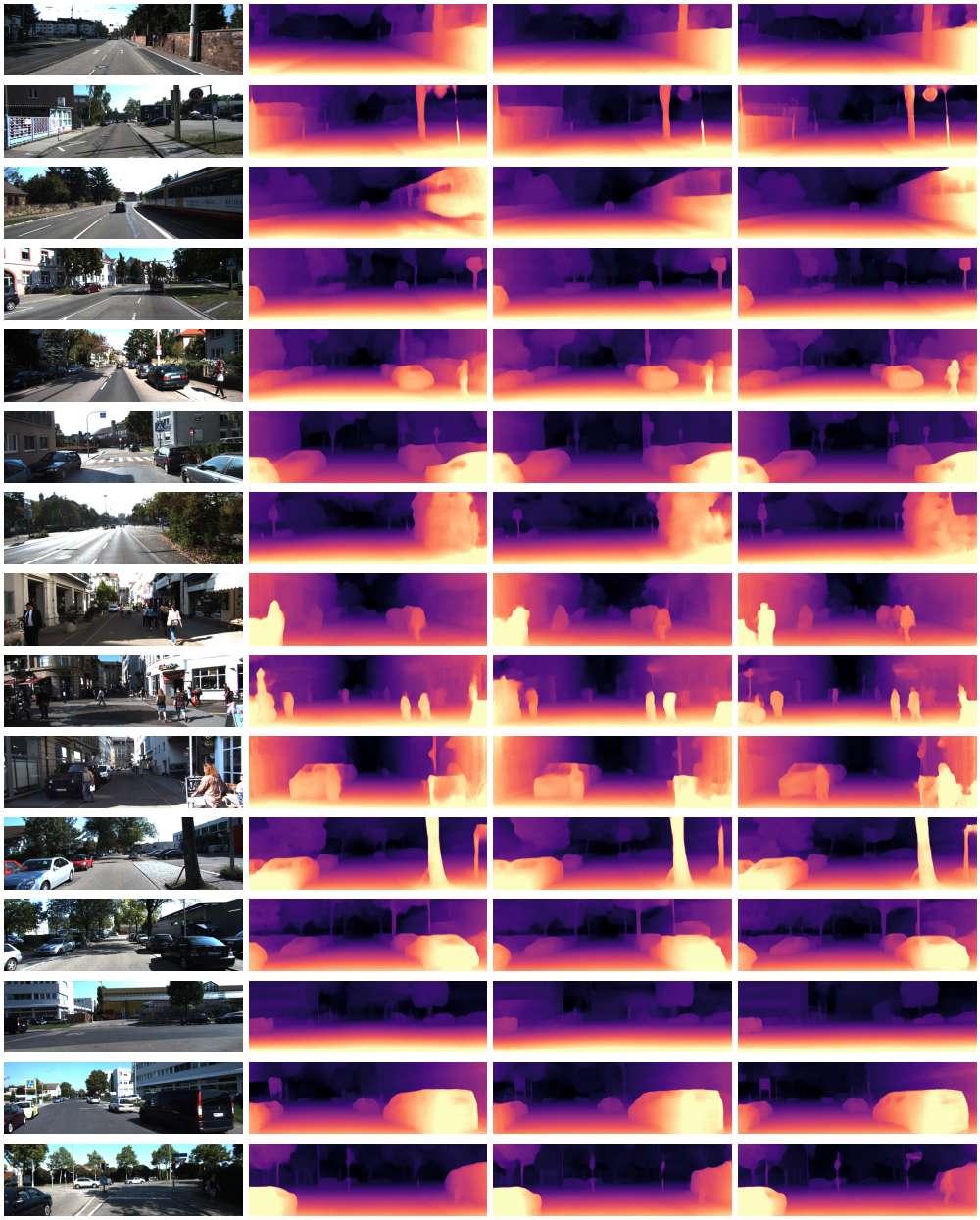
Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13116–13125, 2020.

Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8856–8865, 2019.

Appendix

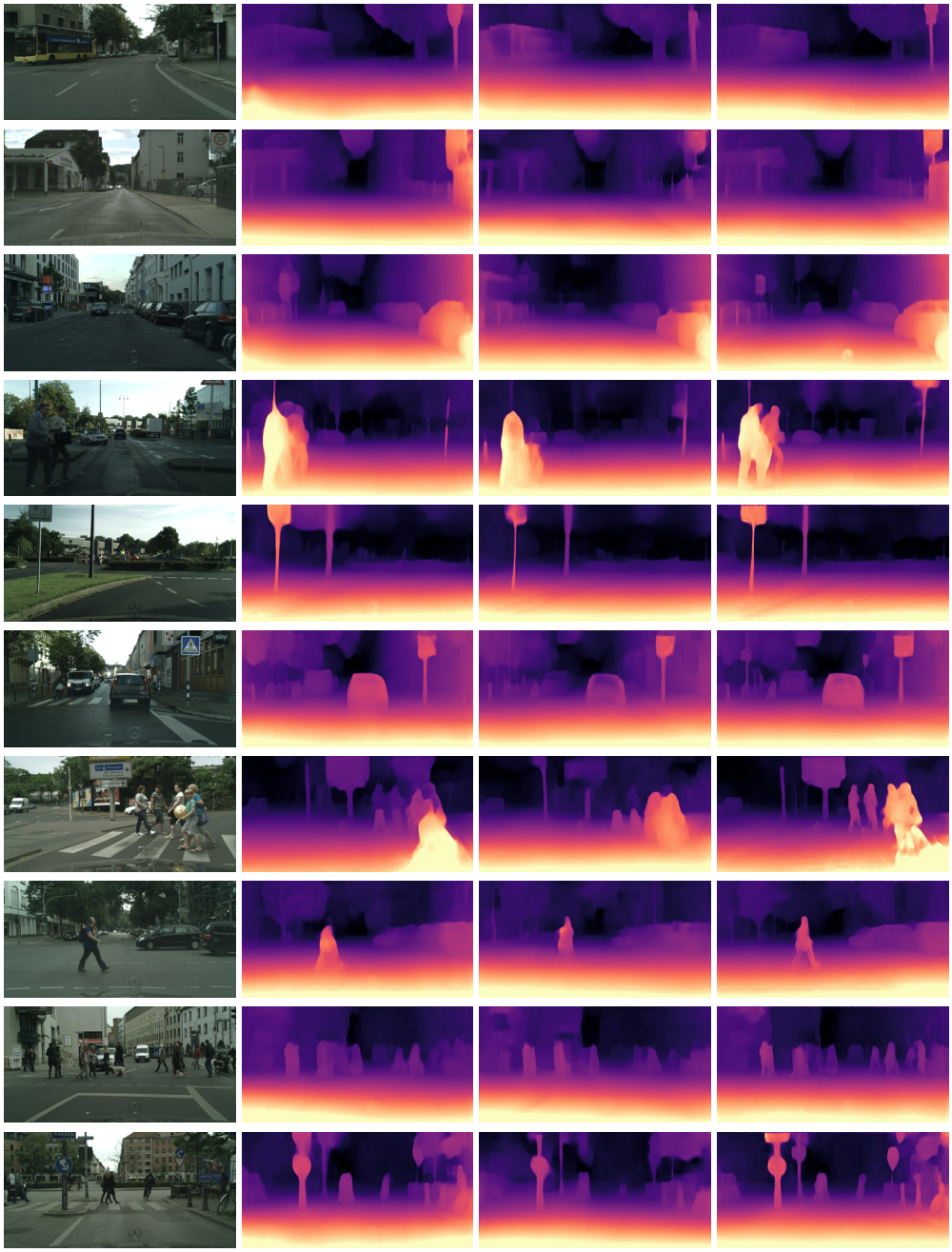
A ADDITIONAL EXPERIMENTAL RESULTS

We offer additional qualitative comparisons for depth on benchmark KITTI (Geiger et al., 2012) and Cityscape (Cordts et al., 2016). All models are trained using only KITTI. As shown in Figure 6 (KITTI) and Figure 7 (Cityscape), the proposed method performs the best that it produces sharp depth edges as well as detailed depth information from thin structures.



input | Godard et al. (2019) | Guizilini et al. (2020a) | **ours**

Figure 6: **Additional qualitative comparisons on KITTI dataset**. All models are trained on KITTI (Geiger et al., 2012) Eigen split (Eigen et al., 2014).



input | Godard et al. (2019) | Guizilini et al. (2020a) | **ours**

Figure 7: **Additional qualitative comparisons on Cityscape**. All models are trained on KITTI (Geiger et al., 2012) Eigen split (Eigen et al., 2014).

B KITTI IMPROVED GROUND TRUTH RESULTS

We use KITTI (Geiger et al., 2012) with improved groundtruth (Uhrig et al., 2017) to further validate the effectiveness of our method. The test set is part of the Eigen split (Eigen et al., 2014), which accounts for 93% (652/691) of the original test frames. Quantitative results are shown in Table 3. Our method performs the best against other methods.

Method	Training	Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^2$
Godard et al. (2017)	S	0.109	0.811	4.568	0.166	0.877	0.967	0.988
Pillai et al. (2019) +pp	S	0.090	0.542	3.967	0.144	0.901	0.976	0.993
Godard et al. (2019)	S	0.085	0.537	3.868	0.139	0.912	0.979	0.993
EPC++	MS	0.123	0.754	4.453	0.172	0.863	0.964	0.989
Godard et al. (2019)	MS	**0.080**	0.466	3.681	0.127	0.926	0.985	0.995
Zhou et al. (2017)	M	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Mahjourian et al. (2018)	M	0.134	0.983	5.501	0.203	0.827	0.944	0.981
Yin & Shi (2018)	M	0.132	0.994	5.240	0.193	0.833	0.953	0.985
Wang et al. (2018)	M	0.126	0.866	4.932	0.185	0.851	0.958	0.986
Ranjan et al. (2019)	M	0.123	0.881	4.834	0.181	0.860	0.959	0.985
EPC++	M	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Godard et al. (2019)	M	0.090	0.545	3.942	0.137	0.914	0.983	0.995
Ours	M	0.081	**0.437**	**3.635**	**0.124**	**0.927**	**0.987**	**0.996**

Table 3: **Quantitative results on KITTI improved ground truth**. The best results are in **bold** and the second best results are underlined. 'S' and 'M' refer to self-supervision methods trained using stereo images and monocular images, respectively. Results are presented without any post-processing. The metrics marked in blue mean 'lower is better', while these in red refer to 'higher is better'. Our method outperforms the state-of-the-arts in most metrics by a large margin.

C COMPARISON BETWEEN BINARY/FULL SEMANTIC SEGMENTATION BRANCH

To further validate the effectiveness of the proposed binary semantic branch. We conduct quantitative comparison between the semantic segmentation branches supervised by full/binary labels respectively. For the semantic branch trained by full semantic groundtruth, we change the network to conduct pixel-wise predictions of 20 categories. And the full semantic labels are pre-processed into one-hot vectors for supervision. Different from the network settings illustrated in the main text, the semantic information are directly concatenated with the depth feature without additional contributions. The results shown in Table 4 validate that there are no distinct performance difference between the two networks. Thus, we choose the binary semantic branch in our final settings as it well encodes the semantic contextual information with the preferable simplicity.

Table 4: **Quantitative comparison between different semantic segmentation branches**. 'Full semantic label' refers to the semantic branch supervised by full category labels (20 categories), while 'Binary semantic label' represents the model trained by the binary semantic label.

Groundtruth type	Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta \leq 1.25$	$\delta \leq 1.25^2$	$\delta \leq 1.25^2$
Full semantic label	0.107	0.835	4.675	0.185	0.888	0.963	0.982
Binary semantic label	0.108	0.784	4.588	0.184	0.887	0.963	0.983