Semantic and Optical Flow Guided Self-supervised Monocular Depth and Ego-Motion Estimation

Jiaojiao Fang and Guizhong Liu(✉)

Xi'an Jiaotong University, Xi'an 710049, China
liugz@xjtu.edu.cn

Abstract. The self-supervised depth and camera pose estimation methods are proposed to address the difficulty of acquiring the densely labeled ground-truth data and have achieved a great advance. As the stereo vision could constrain the predicted depth to a real-world scale, in this paper, we study the use of both left-right pairs and adjacent frames of stereo sequences for self-supervised semantic and optical flow guided monocular depth and camera pose estimation without real pose information. In particular, we explore (i) to construct a cascaded structure of the depth-pose and optical flow for well-initializing the optical flow, (ii) a cycle learning strategy to further constrain the depth-pose learning by the cross-task consistency, and (iii) a weighted semantic guided smoothness loss to match the real nature of a depth map. Our method produces favorable results against the state-of-the-art methods on several benchmarks. And we also demonstrate the generalization ability of our method on the cross dataset.

Keywords: Self-supervised learning · Monocular depth estimation · Camera pose estimation · Stereo vision

1 Introduction

Scene understanding is a crucial yet challenging problem in robotics and autonomous driving. One goal is to recognize and analyze the 3D scene structure and camera pose information from monocular 2D images. Traditional methods fail to model the ability of humans to infer the 3D geometric structure of a scene from a monocular image. Whereas with the rise of deep learning, several methods [5, 15] try to understand a scene in a supervised manner by large amounts of densely labeled data. But these methods are based primarily on the assumption that a plentiful of densely labeled ground-truth data is available which is costly and time-consuming [29].

Hence, some researchers attempt to address this problem by self-supervised learning from either stereo pairs [9] or video sequences [23] based on the scene structure reprojection error [10, 23] and multi-view geometric consistency [16]. Learning a mapping from a monocular image to a depth map in a self-supervised manner is challenging due to the high dimensions densely continuous-valued outputs, so more constraints are

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12890, pp. 465–477, 2021.
https://doi.org/10.1007/978-3-030-87361-5_38

needed to solve this problem. The self-supervised optical flow methods have achieved excellent access, as it is easier than the 3D scene understanding. Existing methods import the deep learning-based optical flow for depth-pose learning, and achieve competitive performance [2, 19, 22]. The deep learning-based optical flow extraction is expensive in space and time, while the iterative optimization-based TVNet [6] achieves competitive accuracy with the fastest feature extraction time. Furthermore, the depth values of the pixels within an object always be close and relative, and significant changes occurred at the boundary of an object. To improve the reliability of the depth-pose estimation and imitate the learning process of humans, semantic segmentations are introduced for multi-task learning to mutually positive transfer between semantic segmentation and depth estimation [1, 12, 31, 34]. While they always treat the foreground objects, i.e. cars, and the background equally, which is not squared with the factual scene structure. Thus the different depth distribution of each class should be treated differently. The scale-ambiguity issue is common in self-supervised monocular depth estimation, existing methods mainly incorporate the stereo information into the monocular videos-based methods by given the stereo relations of two image planes.

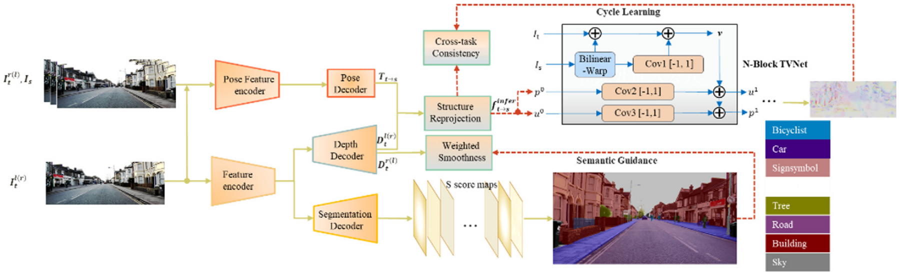

Fig. 1. An overview of the proposed pipeline for self-supervised depth and camera poses estimation. The depth network and pose network are learned collaboratively with the TVNet optical flow and guided by semantic segmentation.

In this paper, we propose a novel framework that jointly learns the semantic and optical flow-guided self-supervised monocular depth and camera pose from stereo video sequences. An overview of our proposed framework has been depicted in Fig. 1. Specifically, our main contributions are as follows:

1) We incorporate the structure of predicting the stereo views' disparity maps only based on a monocular image into the structure-from-motion (SfM) based self-supervised depth-pose learning framework, which takes full advantage of the constraints from spatial and temporal image pairs to improve upon prior art on monocular depth estimations. Our method can solve the scale ambiguity problem by stereo pairs without pose supervision.

2) We introduce the interpretable and simple optical flow, TVNet, into the self-supervised depth and pose learning framework to construct a cascaded structure with an end-to-end cycle learning strategy for well initializing the TVNet and better facilitating the depth-pose learning. Thus, the mismatching problem can be significantly alleviated, and the TVNet can be well-initialized.

3) As we always know that pixels labeled as 'sky' must accompany very large depth values and the depth values of the road are gradually increasing while the depth values of the pedestrian are almost the same. In this paper, we jointly train the depthpose and the semantic segmentation tasks and construct a weighted semantic guided region-aware smoothness loss to treat each class differently and make the predicted depth map closer to the real values.

2 Related Work

There are substantial studies on monocular depth estimation, including geometry-based methods [35] and learning-based methods [5]. In this paper, however, we concentrate only on the self-supervised depth-pose learning and semantic-guided depth estimation, which is highly related to the research topic of this paper.

Self-supervised depth estimation. Self-supervised methods enable learning the depth and pose networks from unlabeled images by substituting the direct ground-truth supervision with the new views synthesis loss. Godard et al. [9] and Garg et al. [7] proposed the stereo vision-based self-supervised methods, then Zhou et al. [23] proposed a self-supervised structure-from-motion-based learning framework that jointly learns the depth and camera pose from monocular video sequences. Based on this framework, a large corpus of works were proposed to promote self-supervised learning performance from different aspects. For more robust self-supervision signals, Mahjourian et al. [16] proposed a 3D geometric consistency loss which directly measured the whole structure of a scene, Shen et al. [32] introduced the epipolar geometry to measure the matching error, another kind of methods leveraged auxiliary tasks such as the optical flow [19, 22] to strengthen depth supervision via cross-task consistency. To deal with the dynamic objects, selective masks were used to filter out the unreliable information during training. Prior works generated the mask by the auto-learning network [23], while the recent methods produced the mask by geometric error guidance [10, 24, 26], which were proved to be more effective. Guizilini et al. [11] proposed a novel depth network architecture to improve the estimation performance. There also exist other methods trying to enhance the network performance with traditional SfM [3], which offer pseudo labels for depth estimation. Poggi et al. [18] focused on the uncertainty estimation for self-supervised monocular depth estimation. And recently some methods were proposed to solve the scale ambiguity problem [26, 32].

Semantic guided Depth estimation. Semantic information had been shown to provide positive effects on monocular depth estimation [1, 31]. The methods could be classified into two categories by the manner of using semantic information. One group of the methods used the outputs of semantic labels directly to guide the depth learning. Chen et al. [1] constrained the depth maps by leveraging the semantic labels of the scene. Klingner et al. [30] and Casser et al. [25] addressed the dynamic moving objects issues by the indication of the semantic map. The other group of methods enhanced the depth feature representation by semantic feature guidance [12, 27] proposed a segmentation-like loss term for depth estimation. Li et al. [28] fused these two manners and proposed an individual semantic network for better scene representation.

3 The Proposed Method

3.1 Cascaded Structure and Cycle Learning Strategy

Monocular depth, optical flow, and ego motion are coupled by the nature of 3D scene geometry. A reasonable combination of these tasks would improve the prediction performance. It contains two stages, the rigid 3D structure reasoning stage and the optical flow iterating stage. The first stage to infer scene structure is made up of two sub-networks, i.e. the DepthNet and the PoseNet. The depth maps and camera poses are regressed respectively and combined to produce the rigid scene flow. And the second stage is to feed the rigid scene flow into the TVNet based optical flow to form a cascaded structure for iterating the optical flow from these initial values. The TVNet [6] is obtained by imitating and unfolding the iterative optimization process of the TV-L1 method [34] and formulates the iterations as customized layers of a neural network. Thus it can be naturally connected with other related networks to form an end-to-end trainable architecture. It is not necessary to carefully design a complex network structure with unknown interpretability or to store the optical-flow features anymore. Since the Taylor expansion is applied in TVNet to linearize the brightness difference, the initial flow field should be close to the real field to ensure a smaller approximation error. Thus it is proper to construct a cascaded structure for depth-pose and optical flow learning. Furthermore, we constrain the cross-task geometric consistency check during training, which significantly enhances the coherence of the predictions and achieves impressive performance.

Our first stage aims to reconstruct the rigid scene structure with robustness towards non-rigidity and outliers. The DepthNet takes a single image $I_t^{l(r)} \in R^{H\times W\times 3}$ as input to regress a pair of pixel-wise stereo depth maps $(D_t^{l(r)}, D_t^{r(l)})$ and exploits accumulated scene priors for depth prediction. Similar to [9], our model does not require the relative pose between the stereo pair. The PoseNet takes the concatenated adjacent views $[I_t, I_s] \in R^{H\times W\times 6}$ as input to regress the relative 6DoF camera poses $T_{t\to s} = [R, t]$ from the target view I_t to the source viewsI_s, where$s \in \{t-1, t+1\}$. With the estimated depth and pose, the reprojected scene flow from the target image I_t to the source image I_s can be represented by

$$f_{t\to s}^{\inf er} = p - KT_{t\to s}D_t^{l(r)}(p)K^{-1}p \tag{1}$$

where $K \in R^{3\times 3}$ denotes the camera intrinsic and p denotes homogeneous coordinates of a pixel in the frame I_t. Here the estimated optical flow of the TVNet-15 [6] is represented as$f_{t\to s}^{pre}$ and can be optimized by the loss as follows:

$$L_{op} = \min_{u(p)} \sum (|u_1(p)| + |u_2(p)| + \lambda\rho(u(p))) \tag{2}$$

where $u(p) = (u_1(p), u_2(p))$ denotes the displacement of the position p from time t to the next frame $t+1$ of each iteration process and $\rho(u(p))$ is defined to penalize the brightness difference of adjacent time. The more implementation details are recommended to see the literature TVNet [6]. Instead of setting the initial value $u^0(p) = (0, 0)$, here we use the $u^0(p) = f_{t\to s}^{\inf er}(p)$ as the initial value of the TVNet.

As we all know that stereo view reconstruction can achieve more reliable results due to it is less affected by the illumination variation. In this paper, we adopt the minimum error among source views and a per-pixel binary auto-mask μ proposed by [10] to construct the photometric loss of the adjacent views as$L_{vs} = \mu \min_{s} pe(I_t, I_{s \to t})$, where $pe(,)$ is a mixture of L1-Norm and structural similarity (SSIM) difference.

We also use the left-right disparity consistency loss L_{lr} and the stereo image reconstruction losses L_{ap} introduced by [9] to constrain the depth maps. Thus the total reconstruction loss of our method can be expressed as:

$$L_{syth} = L_{vs} + \lambda_a L_{ap} + \lambda_c L_{lr} \tag{3}$$

where λ_a and λ_c are the weightings for the left-right image reconstruction loss and the left-right disparity consistency loss, respectively. The cross-task consistency on the depth-pose and optical flow estimation can further constrain the depth-pose learning procedure. Like other methods [26, 32], we compute a binary mask $M(p)$ based on the distribution of the consistency difference between$f_{t\to s}^{pre}(p)$ and$f_{t\to s}^{\inf er}(p)$ to filter out the outliers. The binary mask can be computed by:

$$M(p) = \{ \begin{matrix} 1, Percentile(|f_{t\to s}^{pre}(p) - f_{t\to s}^{\inf er}(p)|) < T_M \\ 0, \qquad\qquad\qquad Otherwise \end{matrix} . \tag{4}$$

Where pixel positions whose geometry consistency loss is above a percentile threshold T_M are filtered out. Thus the optical flow-guided depth and pose learning loss can be reformulated as:

$$L_{pf} = \sum_r \sum_s \sum_p M(p) |f_{t\to s}^{pre}(p) - f_{t\to s}^{\inf er}(p)|. \tag{5}$$

Where r indexes over different feature scales of the image, s indexes over source images, and p indexes over all pixels.

3.2 Semantic Guided Depth Estimation

The existing depth estimation methods generally focus on pixel-wise disparity estimation and regard all pixels within an image as spatial homogeneity, which would lead to unfavorable disparity estimation along object boundaries. To overcome the limitation, we perform disparity estimation by leveraging semantic information to improve the quality of the depth maps. The semantic segmentation is derived by a neural network that implements a non-linear mapping between an input image I_t and the output scores $y_t \in R^{H \times W \times S}$ for all pixel indexes p and classes s. We thus define the semantic segmentation loss L_{seg} as:

$$L_{seg} = \omega_s CE(s_{GT}, s) \tag{6}$$

where $\omega_s CE()$ indicates the weighted cross-entropy loss, ω_s are the weights of each semantic class, $s \in S$ is the predicted semantic labels of each pixel p from a set of classes $S = \{1, 2, \cdots, S\}$ and s_{GT} denotes the ground truth labels from an additional disjoint dataset. To approximate the real distribution of the depth map, we proposed a region-aware smoothness loss function to constrain the depth values belonging to the same objects to be closer with their nearby pixels, while making a difference between the foreground objects and the background. $S_t = \varphi(y_t)$ is the operation that sets the maximum value along each channel as 1 and sets the remaining values as 0. Then the weighted region-aware disparity smoothness term is defined as:

$$L_{sm} = \Sigma_p(|\partial_x D_t^{l(r)}(p)| + |\partial_y D_t^{l(r)}(p)|) \odot \omega_{sf}(1 - S_t) \tag{7}$$

where $\partial_{\mathrm{x}}(\cdot)$ and $\partial_{\mathrm{y}}(\cdot)$ are the gradients of disparity in horizontal and vertical direction respectively $\odot$ denotes element-wise multiplication, S_t is the gradients of the segmentation map in each channel which mean the edges of all classes, and $\omega_{sf} \in R^S$ is the smoothness weights of all semantic classes. Thus the weighted semantic guided smooth factor is low (close to zero) on the boundary regions of the objects, high (close to one) on the foreground objects' central regions, and smaller on the background regions. The depth value of the nearby pixels within the edge of one semantic object should be almost the same for the foreground objects, while just be closer and gradually changed for the background of the nearby pixels. Thus the smoothness degree of each class should be different during training. The second term $1 - \nabla S_t$ is an edge detector operation to identify edges of the segmentation map.

Thus our final loss function becomes:

$$L_{final} = L_{syth} + \lambda_s L_{sm} + \lambda_f L_{pf} + \lambda_o L_{op}. \tag{8}$$

Where λ_s and λ_f are the weightings for the depth smoothness loss and the geometry consistency loss, respectively.

4 Experiments

In this section, we evaluated the quantitative and qualitative performance of our method, and compared it with the state-of-the-arts methods mainly on the KITTI dataset [8] for a fair comparison. We also evaluated the cross dataset generalization ability on the Make3D dataset [21].

4.1 Implementation Details

Parameters setting. The algorithm was implemented in the PyTorch [17] framework. For all the experiments, we set the weighting of the different loss components as $\lambda_s = 0.01$, $\lambda_f = 0.001$, $\lambda_a = 0.5$ and $\lambda_c = 0.5$. We trained our model for 40 epochs with the Adam [14] optimizer, Gaussian random initialization, and mini-batch size of 6. The learning rate was initially set to 0.0001 for the first 30 epochs and then dropped to 0.00001 for the reminder. The network took almost 28 h to train on a single Titan Xp for 20 epochs.

Network architecture. Our networks are based on SGDepth [30], where an encoder-decoder architecture with skip connections is employed. To ensure comparability to

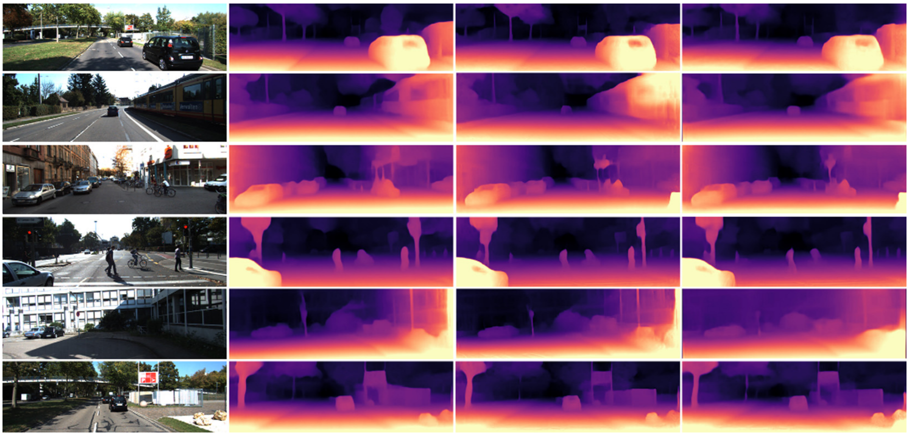

Fig. 2. The qualitative results of the proposed method on the KITTI dataset. The left, middle, and right columns show respectively input images, the state-of-the-art predicted depth maps Godard et al. [10], Klingner et al. [30] and the depths maps obtained by ours. Our method predicts sharper boundaries and fine-grained details on distant objects.

existing work [10, 12, 30], we choose the ResNet18 [13] pretrained on Imagenet [20] as encoder. The shared encoder was trained according to [30]. The depth head has two channels at each output layer and has 4 different spatial scales of the outputs. A sigmoid output σ is used to ensure the predicted depth to be within a reasonable range, which is converted to a depth map by $1/a\sigma + b$, where a and b are chosen to constrain the depth values within the range [0.1, 100]. For simplicity, the segmentation decoder uses the same architecture as the depth decoder, except for the last layer having S channels feature maps, whose elements are converted to class probabilities by a softmax function. The pose network's architecture is the same as in [30].

Datasets. We utilized one dataset to learn the semantic segmentation and another one for the self-supervised depth-pose and optical flow training. For training the semantic segmentation we utilized the Cityscapes dataset [4] while at the same time we use the KITTI dataset [8] for training the depth estimation. Similar to other state-of-the-art approaches we trained our model on the Eigen split dataset [5] which excluded 679 images from the KITTI dataset for testing and remove static frames following [23]. All the experiments were performed with image sequences captured by color cameras with fixed focal length. We resized images to 640 × 192 during training, but the network can be tested with arbitrary input image size, due to the depth and pose networks were all with the fully convolutional structure.

Augmentation and processing. We performed horizontal flips and the following data augmentations during training, with 50% chance: random brightness, contrast, saturation, and hue jitter with respective ranges of ±0.2, ±0.2, ±0.2, and ±0.1 as in [10]. All the images fed to the pose and depth networks are performed with the same augmentations. For results trained in stereo image pairs, we did not perform median scaling as the scale has to be learned by the stereo vision.

Table 1. Results on the KITTI dataset [8] using the Eigen split [5] compared with the state-of-the-art methods by the authors' report. 'S' and 'M' refer to stereo images and monocular images self-supervision, respectively. 'Inst' and 'Sem' indicate the instance or semantic information.

Method	Train	Error (lower is better)				Accuracy (higher is better)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Zhou et al. [23]	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Godard et al. [9]	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
Vid2Depth [16]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Yin et al. [22]	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
Chen et al. [1]	S	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Godard et al. [10]	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Godard et al. [10]	MS	0.106	0.818	4.750	0.196	0.874	0.957	0.979
Casser et al. [25]	M+Inst	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Bian et al. [26]	M	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Xue et al. [32]	M	0.113	0.864	4.812	0.191	0.877	0.960	0.981
Guizilini et al. [12]	M+Sem	0.117	0.854	4.714	0.191	0.873	0.963	0.981
ṇKlingner et al. [30]	M+Sem	0.113	0.835	4.693	0.191	0.879	0.961	0.981
Ours	MS+Sem	0.105	0.801	4.631	0.189	0.881	0.962	0.982

4.2 Main Results

In this section, we start by the comparison to multiple state-of-the-art methods, followed by an analysis of how the single components of our method improve the estimation results.

Depth Estimation Results on KITTI. We first provided a comparison with the state-of-the-arts self-supervised methods on the monocular depth evaluation by the Eigen split in Table 1. To be fair for all methods, we used the same crop manner as [10] and evaluated the prediction with the same resolution as the input image. The evaluation metrics conformed to the one used in [10], and the depth value was capped to 80 m during evaluation. Our method outperforms all comparable baselines, where we compare to methods that use only video sequences as supervision on the KITIT dataset. As the resolution of the input image highly dependent on the estimation performance, we reported the results at the middle resolutions 640 × 192 for a fair comparison. Due to fairness, we do not compare against results with online refinement [1] or employing a more efficient network architecture [11], as such techniques can improve any methods. Qualitative results were shown in Fig. 2. We could observe that our method was able to reconstruct small objects such as traffic signs and could achieve the state-of-the-art performance.

Depth Estimation Results on Make3D. To illustrate the generalization ability of our method, we evaluated our model trained only on the KITTI dataset on the Make3D test set of [21]. Make3D consists of only RGB/Depth pairs and without stereo image. Qualitative results were shown in Fig. 3, note that our model is only trained on the KITTI dataset, and directly tested on Make3D. These results would be further improved with more relevant training data.

Table 2. Odometry results on the KITTI [8] odometry dataset. Results show the average absolute trajectory error, and standard deviation, in meters.

Methods	Sequence 09	Sequence 10	# frames
Garg et al. [7]	0.013 ± 0.010	0.012 ± 0.011	3
Zhou et al. [23]	0.021± 0.017	0.020 ± 0.015	5
Mahjourian et al. [16]	0.013± 0.010	0.012 ± 0.011	3
GeoNet [22]	0.012± 0.007	0.012 ± 0.009	5
Ranjan et al. [19]	0.012± 0.007	0.012 ± 0.008	5
Monodepth2 [10]	0.017± 0.008	0.015 ± 0.010	2
SGDepth [30]	0.019± 0.010	0.016 ± 0.010	2
Ours	0.011 ± 0.007	0.012 ± 0.008	2

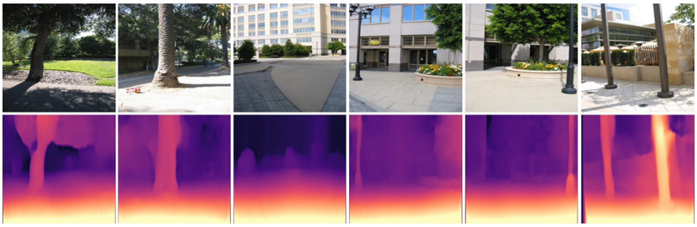

Fig. 3. Illustration of examples of depth predictions on the unseen Make3D dataset [21].

Pose Estimation Results on KITTI Odometry. While we mainly concentrated on better depth estimation, we also compared our pose networks with competing methods on the KITTI odometry dataset since the two tasks are inter-dependent. The KITTI odometry dataset contains 11 driving sequences with ground-truth poses available (and 11 sequences without ground-truth). We evaluated the pose error on sequences 09 and 10. Competing methods typically feed more frames to their pose network for improving their performance. We had observed that with the joint optical flow learning, the result of visual odometry would be improved. We measured the Absolute Trajectory Error (ATE) over N-frame snippets (N = 3 or 5), as measured in [10]. As showed in Table 2, our method outperformed other state-of-the-art approaches.

Table 3. Ablation studies. Results for different variants of our model trained on KITTI 2015 [8] using the Eigen split.

Method	Error (lower is better)				Accuracy (higher is better)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\Delta < 1.25^2$	$\Delta < 1.25^3$
Baseline	0.115	0.903	4.863	0.193	0.877	0.959	0.977
+Stereo	0.113	0.863	4.767	0.192	0.875	0.957	0.980
+Cascaded-cycle learning	0.106	0.842	4.723	0.191	0.877	0.959	0.979
+ Weighted Semantic guidance	0.105	0.801	4.631	0.189	0.881	0.962	0.982

4.3 Ablation Studies

To verify how each component of our model contributed to the overall performance, we performed ablation studies by changing various components of our model based on and listed the results in Table 3. The study was performed on the Eigen split for better comparison. We see that each component of our model could promote the estimation performance of the baseline model [10], and all our components combined lead to a significant improvement. As expected, integrating stereo information into a monocular model could increase accuracy.

Benefits of the cascaded structure by cycle learning strategy for depth estimation. We also compared our only depth estimation model with the jointly depth-pose and TVNet learning framework with zero initialization. Our method performed better than the model that simply jointed training three networks together on the KITTI Eigen split dataset. We further verified that the cycle learning could address the mismatching problem which could be reflected by the end-to-end point error.

Effect of the weighted semantic guided smoothness loss. We saw that the weighted semantic guided smoothness loss could not only achieve improvement of the quantitative results than the edge-aware first order smooth loss and the second-order smooth loss, the qualified results could also be better improved. Semantic information could improve depth estimation in all cases.

More specifically, we applied the exact network architecture of Godard et al. [9] that predicted the disparity maps for stereo views from a monocular image by randomly using the left or right view as input instead of only using the left image as input, which would solve the problem of unawareness of structural information from the other view, and promoted the depth and pose estimation.

5 Conclusions

In this paper, we have presented a self-supervised semantic and optical flow-guided depth-pose learning pipeline from stereo sequences with an unknown stereo pose. This framework takes full advantage of the constraints on the unlabeled stereo and temporal image pairs by predicting stereo disparity maps from a monocular image. Furthermore, we import the explainable TVNet into the self-supervised depth-pose learning and construct a cascaded structure with a cycle learning strategy for better depth-pose and optical flow learning. Finally, we propose a weighted semantic guided smoothness loss to treat the foreground objects and background region differently for predicting more natural and reasonable depth maps. Experiments show that our method can exceed the existing self-supervised method, and can generalize well to the unseen dataset.

References

1. Chen, P., Liu, A.H., Liu, Y., Wang, Y.F.: 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 (2019)
2. Chen, Y., Schmid, C., Sminchisescu, C.: 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 (2019)
3. Zhan, H., Weerasekera, C.S., Bian, J., Reid, I.: Visual odometry revisited: What should be learnt?. In: IEEE International Conference on Robotics and Automation, pp. 4203–4210 (2020)
4. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
5. 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)
6. Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., Huang, J.: End-to-end learning of motion representation for video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6016–6025 (2018)

7. Garg, R., Kumar, B,G.V., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: European Conference on Computer Vision, pp. 740–756 (2016)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361 (2012)
9. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
10. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3828–3838 (2019)
11. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3D packing for self-supervised monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
12. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. In: International Conference on Learning Representations (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: International Conference on 3D Vision, pp. 239–248 (2016)
16. Mahjourian, R., Wicke, M., Angelova, A.: 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)
17. Paszke, A., et al.: PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of NeurIPS, Vancouver, BC, Canada, pp. 8024–8035 (2019)
18. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. Ranjan, A., et al.: 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)
20. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
21. Saxena, A., Sun, M., Ng, A.Y.: Make3D: learning 3D scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 824–840 (2009)
22. Yin, Z., Shi, J.: 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)
23. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: 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)
24. Wang, G., Wang, H., Liu, Y., Chen, W.: Unsupervised learning of monocular depth and ego-motion using multiple masks. In: IEEE 2019 International Conference on Robotics and Automation (ICRA), pp. 4724–4730 (2019)

25. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos. Proc. AAAI Conf. Artif. Intell. **33**, 8001–8008 (2019)
26. Bian, J., et al.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. Adv. Neural Inf. Process. Syst. 35–45 (2019)
27. Choi, J., Jung, D., Lee, D., Kim, C.: Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. In: Workshops at the 34th Conference on Neural Information Processing Systems (2020)
28. Li, R., He, X., Zhu, Y., Li, X., Sun, J., Zhang, Y.: Enhancing self-supervised monocular depth estimation via incorporating robust constraints. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3108–3117 (2020)
29. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: sparsity invariant CNNs. In: International Conference on 3D Vision (2017)
30. Klingner, M., Termohlen, J., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: European Conference on Computer Vision, (2020)
31. Meng, Y., et al.: SIGNet: semantic instance aided unsupervised 3D geometry perception. In: Proceedings of CVPR, Long Beach, CA, USA, pp. 9810–9820, June 2019
32. Shen, T., Luo, Z., Zhou, L., et al.: Beyond photometric loss for self-supervised ego-motion estimation. In: International Conference on Robotics and Automation, pp. 6359–6365 (2019)
33. Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z., Ang, Jr.: Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020)
34. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. Pattern Recogn. 214–223 (2007)
35. Schonberger, J.-L., Frahm, J.-M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)