Proceedings of The 2021 IEEE International
Conference on Real-time Computing and Robotics
July 15-19, 2021, Xining, China

Bidirectional Weighted Loss with Feature Perception for Self-supervised Learning of Consistent Depth-pose

Fei Wang, Jun Cheng and Penglei Liu

***Abstract*—Photometric differences are widely used as supervision signals to train neural networks for estimating depth and camera pose from unlabeled monocular videos. However, these approaches are detrimental for model optimization because occlusions and moving objects in scene violate the underlying static scenario assumption. In addition, pixels in textureless regions or less discriminative pixels hinder model training. To solve these problems, we propose a bidirectional weighted photometric loss and constrain the loss using both a feature perception loss and a depth structure consistency loss. Extensive experiments and visual analysis demonstrate the effectiveness of the proposed method, in which we outperform existing state-of-the-art self-supervised methods.**

I. Method

Firstly, to enhance the perception ability of the model for weakly textured regions, we propose a bidirectional feature perception loss (L_{feat}^{bi}) shown in formula (1) based on the discovery that redundancies and noise in images can be removed by encoding larger-scale patterns, it can be inferred that the deep features extracted from raw images using an encoder network are more discriminative than the raw red-green-blue (RGB) image features in textureless regions.

$$L_{feat}^{bi} = \|f_{tgt} - \hat{f}_{ref}\| + \|f_{ref} - \hat{f}_{tgt}\| \tag{1}$$

where f_{tgt} and f_{ref} are the deep features extracted from the target and reference images using the encoder network, respectively. $\hat{f}_{ref}$ and $\hat{f}_{tgt}$ are the corresponding feature maps synthesized by warping reference/target feature maps to target/reference plane.

Secondly, we employ a bidirectional depth structure consistency loss (L_{dsc}^{bi}) shown in formula (2) to enforce consistency between the depth obtained from the multiview geometric transformation and the depth predicted from the corresponding frame.

$$\begin{aligned} L_{dsc}^{bi} &= L_{dsc}^{ref\to tgt} + L_{dsc}^{tgt\to ref} \\ &= \frac{\sum depth_{diff}(p_{ref})}{N_{ref}} + \frac{\sum depth_{diff}(p_{tgt})}{N_{tgt}} \end{aligned} \tag{2}$$

where $depth_{diff}(p_{\text{ref}})$ and $depth_{diff}(p_{tgt})$ stand for the errors between the depth obtained from the multiview geometric transformation and the depth predicted from the corresponding frame by DepthNet. N_{ref}, N_{tgt} denote the numbers of valid grid coordinates.

Finally, to handle moving objects and occlusions, we reweight the bidirectional photometric loss which can take full advantage of limited data, using both the proposed bidirectional camera flow occlusion masks, as shown in formula (5) and (6), and the proposed adaptive weights, as shown in formula (7) and (8).

$$\begin{aligned} L_p^{biw} &= M_{valid}^{ref\to tgt} * W_{aw}^{ref\to tgt} * L_{photo}^{ref\to tgt} \\ &\quad + M_{valid}^{tgt\to ref} * W_{aw}^{tgt\to ref} * L_{photo}^{tgt\to ref} \end{aligned} \tag{3}$$

$$\begin{aligned} M_{valid}^{ref\to tgt} &= 1 - \lambda_{occ}^{tgt} * M_{occ}^{ref\to tgt} \\ M_{valid}^{tgt\to ref} &= 1 - \lambda_{occ}^{ref} * M_{occ}^{tgt\to ref} \end{aligned} \tag{4}$$

$$\begin{aligned} M_{occ}^{ref\to tgt} &= \Gamma(\|u_{cam}^{ref\to tgt} + \hat{u}_{cam}^{tgt\to ref}\|^2, \\ &\alpha_1(\|u_{cam}^{ref\to tgt}\|^2 + \|\hat{u}_{cam}^{tgt\to ref}\|^2) + \alpha_2) \end{aligned} \tag{5}$$

$$\begin{aligned} M_{occ}^{tgt\to ref} &= \Gamma(\|u_{cam}^{tgt\to ref} + \hat{u}_{cam}^{ref\to tgt}\|^2, \\ &\alpha_1(\|u_{cam}^{tgt\to ref}\|^2 + \|\hat{u}_{cam}^{ref\to tgt}\|^2) + \alpha_2) \end{aligned} \tag{6}$$

$$W_{aw}^{ref\to tgt} = 1 - depth_{diff}^{ref\to tgt}(p_{ref}) \tag{7}$$

$$W_{aw}^{tgt\to ref} = 1 - depth_{diff}^{tgt\to ref}(p_{tgt}) \tag{8}$$

$$\Gamma(a,b) = \begin{cases} 1, & a<b \\ 0, & otherwise \end{cases} \tag{9}$$

where $L_{photo}^{ref\to tgt}$ and $L_{photo}^{tgt\to ref}$ are the corresponding photometric error functions that measure the differences between the target frame I_{tgt} and the corresponding synthesized frame $\hat{I}_{ref}^{pose}$ and between the reference frame I_{ref} and the corresponding synthesized frame $\hat{I}_{tgt}^{pose}$, respectively.

The total loss is shown in formula (10):

$$\begin{aligned} L_{total} &= \lambda_s^{bi} * L_s^{bi} + \lambda_{feat}^{bi} * L_{feat}^{bi} \\ &\quad + \lambda_p^{bi} * L_p^{biw} + \lambda_{dsc}^{bi} * L_{dsc}^{bi} \end{aligned} \tag{10}$$

where L_s^{bi} stands for the smoothness loss.

II. Results

In Tab. I, we compare the depth estimation results with those of current state-of-the-art self-supervised methods trained on the KITTI dataset and with the results of methods with parameters pretrained on the Cityscapes dataset and then fine-tuned on KITTI. To better understand the contribution of each element of the proposed objective function — the bidirectional weighted photometric function, which is composed of the bidirectional photometric function (L_p^{bi}) with bidirectional

Corresponding author: Jun Cheng.
Fei Wang, Jun Cheng, Penglei Liu are with the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China. They are also with The Chinese University of Hong Kong (email:{fei.wang2, jun.cheng, pl.liu}@siat.ac.cn).

Method	Data	Cap (m)	Resolutions	Error↓ AbsRel	SqRel	RMSE	RMSE log	Accuracy↑ $\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. [1]	K	80	128×416	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou et al. [1]	K+CS	80	128×416	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Yin et al. [2]	K	80	128×416	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Mahjourian et al. [3]	K+CS	80	128×416	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Ranjan et al. [5]	K	80	256×832	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Ranjan et al. [5]	K+CS	80	256×832	0.139	1.032	5.199	0.213	0.827	0.943	0.977
Bian et al. [6]	K	80	256×832	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Bian et al. [6]	K+CS	80	256×832	0.128	1.047	5.234	0.208	0.846	0.947	0.976
Ours	K	80	256×832	**0.1199**	**0.9474**	**4.9405**	**0.1965**	**0.8630**	**0.9569**	**0.9814**

TABLE I: Comparison of performance for monocular depth estimation on the KITTI dataset. K denotes that our models were trained only on KITTI, and CS+K means that the models were fine-tuned on KITTI after pretraining on the Cityscapes dataset. The best performance in each column is highlighted in bold.

Method	Cap (m)	Error↓ AbsRel	SqRel	RMSE	RMSE log	Accuracy↑ $\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline	80	0.1418	0.9628	5.2890	0.2222	0.8081	0.9406	0.9768
L_p^{bi}	80	0.1390	1.0420	5.2572	0.2198	0.8272	0.9417	0.9749
L_p^{bi} + M_{occ}^{bi}	80	0.1262	0.9592	4.8118	0.2026	0.8566	0.9535	0.9795
L_p^{bi} + M_{occ}^{bi} + L_{dsc}^{bi}	80	0.1234	0.9984	4.9396	0.1988	0.8585	0.9548	0.9806
L_p^{bi} + M_{occ}^{bi} + L_{dsc}^{bi} + W_{aw}^{bi}	80	0.1219	0.9833	4.9281	0.1980	0.8645	0.9558	0.9802
L_p^{bi} + M_{occ}^{bi} + L_{dsc}^{bi} + W_{aw}^{bi} + L_{feat}^{bi}	80	0.1199	0.9474	4.9405	0.1965	0.8630	0.9570	0.9814

TABLE II: Ablation studies on monocular depth estimation. The results were evaluated on the KITTI Eigen split with the depth capped at 80 m. δ represents the ratio between the estimated depth and ground truth depths.

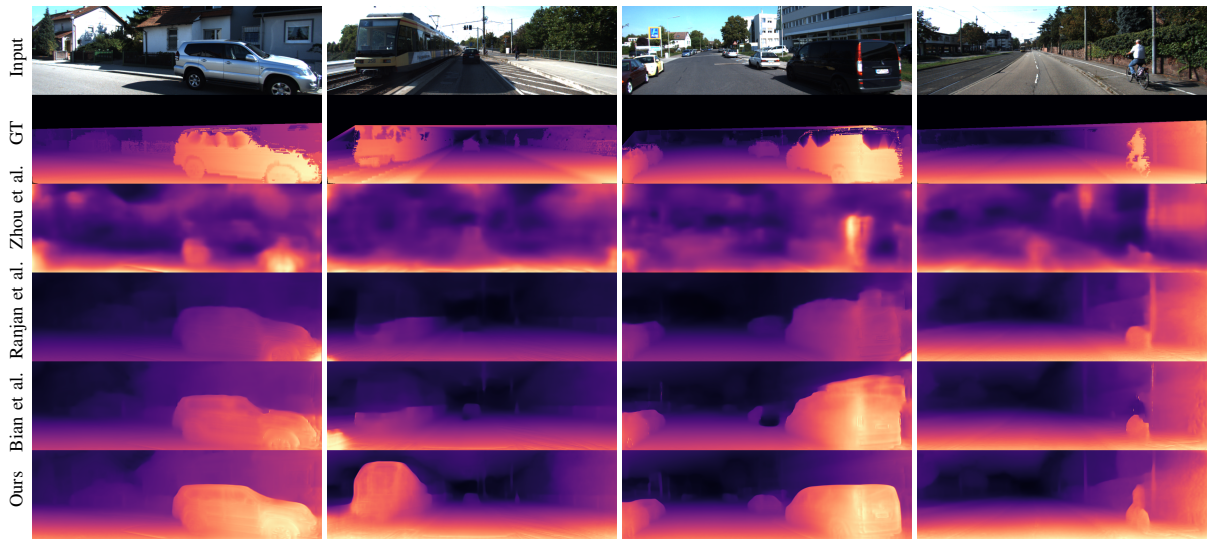

Fig. 1: Qualitative comparison of example results of our proposed self-supervised monocular depth estimation method with those of previous state-of-the-art methods as estimated on the KITTI dataset. The ground truth maps were obtained from sparse laser data for visualization only. The brighter an area in a depth map is, the closer it is to the camera.

camera flow occlusion masks (M_{occ}^{bi}) and adaptive weights (W_{aw}^{bi}), the bidirectional feature perception loss (L_{feat}^{bi}), and the bidirectional depth structure consistency loss (L_{dsc}^{bi}) — to the whole performance, we performed ablation studies, as shown in Tab. II. The qualitative comparison samples of our proposed self-supervised monocular depth estimation are shown in Fig. 1.

III. Conclusion

In this paper, we have presented an end-to-end selfsupervised learning pipeline that utilizes the task of view synthesis as the supervision signal for depth and camera pose estimation from unlabeled monocular video. Our experimental results indicate that the proposed method outperforms previous related works. The proposed bidirectional weighted photometric loss function can fully reveal information and handle the dynamic scene effectively. Secondly, the textureless regions can be paid more attention to by using our feature perception loss function in the scene. Moreover, we can enforce consistency between the depth maps, further improving the quality of estimated depth. Despite competitive performance on the benchmark evaluation,

the scale-drift issue that we need to align the estimation results using ground truth during evaluation still exists. Additionally, our method assumes the camera intrinsics are given, which forbids the use of random Internet videos with unknown camera types. We plan to address these problems in future work.

REFERENCES

[1] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1851–1858, 2017.

[2] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1983–1992, 2018.

[3] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5667–5675, 2018.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3828–3838, 2019.

[5] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12240–12249, 2019.

[6] Jia-Wang 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 *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, pages 35–45, 2019.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.