SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

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

Fei Wang

fei.wang2@siat.ac.cn

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China

2021 IEEE International Conference on Real-time Computing and Robotics

Proposed Method

1.1 Bidirectional Photometric Loss

$$L_p^{bi} = L_p^{ref\to tgt}(I_{tgt}, \hat{I}_{ref}) + L_p^{tgt\to ref}(\hat{I}_{tgt}, I_{ref}) \tag{1}$$

where $L_p^{ref\to tgt}(I_{tgt}, \hat{I}_{ref})$ and $L_p^{tgt\to ref}(\hat{I}_{tgt}, I_{ref})$ are the corresponding photometric error functions,

$I_{i\in\{tgt,ref\}}$ denotes image sequences, $\hat{I}_{i\in\{tgt,ref\}}$ denotes the corresponding synthesized image sequences.

Proposed Method

1.2 Bidirectional Weighted Photometric Loss

$$L_p^{biw} = (1 - M_{occ}^{ref\to tgt}) * W_{aw}^{ref\to tgt} * L_p^{ref\to tgt}(I_{tgt}, \hat{I}_{ref}) + (1 - M_{occ}^{tgt\to ref}) * W_{aw}^{tgt\to ref} * L_p^{tgt\to ref}(\hat{I}_{tgt}, I_{ref}) \quad (2)$$

$$M_{occ}^{ref\to tgt} = \Gamma(\| u_{cam}^{ref\to tgt} + \hat{u}_{cam}^{tgt\to ref} \|^2, \alpha_1(\| u_{cam}^{ref\to tgt} \|^2 + \| \hat{u}_{cam}^{tgt\to ref} \|^2) + \alpha_2) \quad (3)$$

$$M_{occ}^{tgt\to ref} = \Gamma(\| u_{cam}^{tgt\to ref} + \hat{u}_{cam}^{ref\to tgt} \|^2, \alpha_1(\| u_{cam}^{tgt\to ref} \|^2 + \| \hat{u}_{cam}^{ref\to tgt} \|^2) + \alpha_2) \quad (4)$$

where M_{occ} denotes camera flow occlusion mask, W_{aw} denotes adaptive weights obtained from difference between depths, u_{cam} denotes camera flow obtained by the transformed image coordinates, $\hat{u}_{cam}$ denotes the synthesized camera flow, $\Gamma(\cdot)$ stands for an indicator function.

Proposed Method

1.3 Bidirectional Feature Perception Loss

$$L_{feat}^{bi} = \| f_{tgt} - \hat{f}_{ref} \| + \| f_{ref} - \hat{f}_{tgt} \| \quad (5)$$

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

Proposed Method

1.4 Bidirectional Depth Structure Consistency Loss

$$L_{dsc}^{bi} = L_{dsc}^{ref \to tgt} + L_{dsc}^{tgt \to ref} = \frac{\sum depth_{diff}(p_{ref})}{N_{ref}} + \frac{\sum depth_{diff}(p_{tgt})}{N_{tgt}} \tag{6}$$

where $depth_{diff}(\cdot)$ stands for the errors between the depth obtained from the multiview geometric transformation and the depth predicted from the corresponding frame by DepthNet, $N_{i \in \{ref, tgt\}}$ denotes the numbers of valid grid coordinates.

SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Proposed Method

Finally, the total loss function, as shown formula (7), is employed as the supervision signal to train neural networks for estimating depth and camera pose from unlabeled monocular videos in a self-supervised fashion:

$$L_{total} = \lambda_s^{bi} * L_s^{bi} + \lambda_{feat}^{bi} * L_{feat}^{bi} + \lambda_p^{bi} * L_p^{biw} + \lambda_{dsc}^{bi} * L_{dsc}^{bi} \quad (7)$$

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

SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Experiment Results

2.1 Comparison of performance

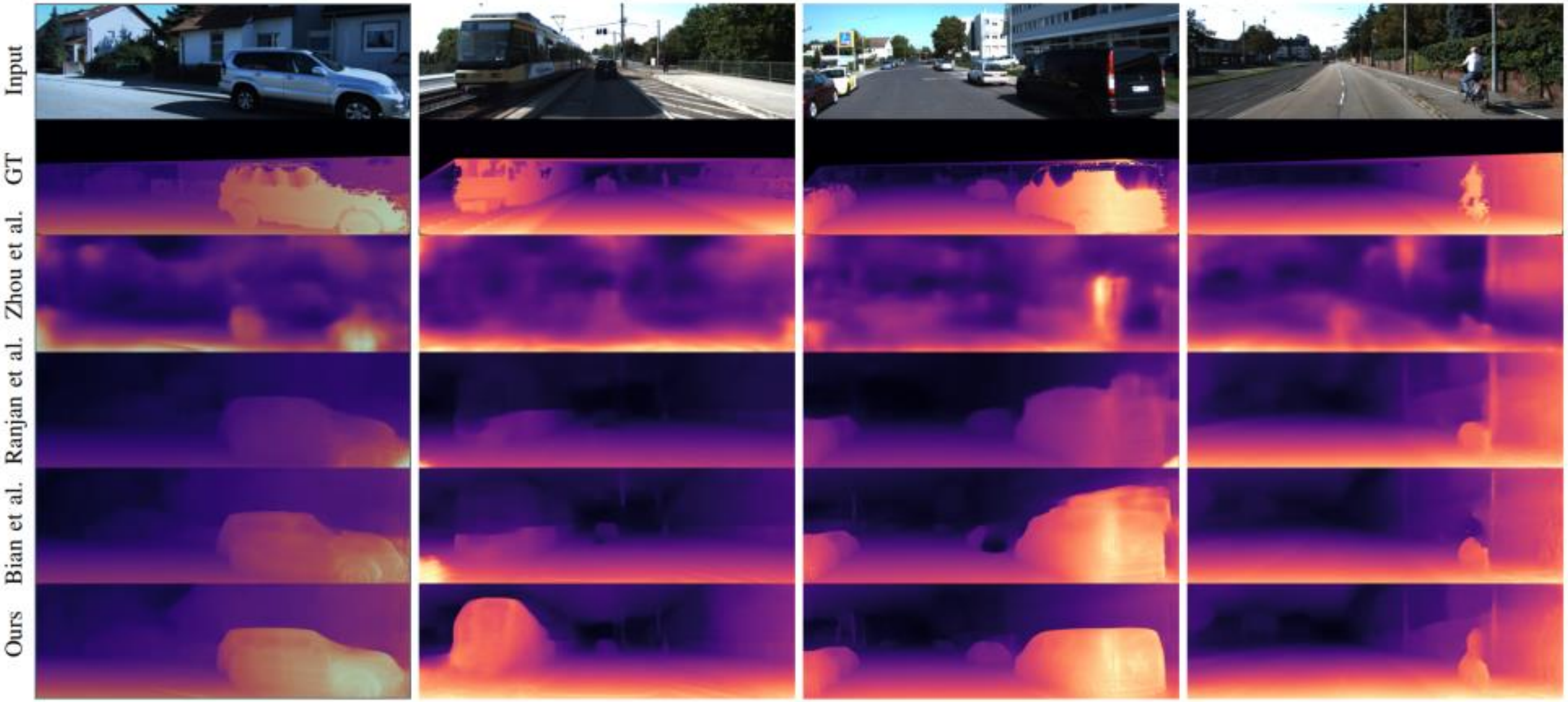
Method	Data	Cap (m)	Resolutions	Error↓ AbsRel	SqRel	RMSE	RMSE log	Accuracy↑ $\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al.	K	80	128×416	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou et al.	K+CS	80	128×416	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Ranjan et al.	K	80	256×832	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Ranjan et al.	K+CS	80	256×832	0.139	1.032	5.199	0.213	0.827	0.943	0.977
Bian et al.	K	80	256×832	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Bian et al.	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**

Tab. 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.

SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Experiment Results

2.2 Qualitative Comparison of Example Results



Experiment Results

2.3 Ablation studies

Method	Cap (m)	Error↓				Accuracy↑		
		AbsRel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
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

Tab. II. 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.

SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Thank You For Your Attention