Collaborative Learning of Depth Estimation, Visual Odometry and Camera Relocalization from Monocular Videos

Haimei Zhao*[1,3], **Wei Bian**[2], **Bo Yuan**[1] and **Dacheng Tao**[3]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Center for Artificial Intelligence, University of Technology Sydney
[3]UBTECH Sydney AI Centre, School of CS, Faculty of Engineering, The University of Sydney, Australia
zhaohm17@mails.tsinghua.edu.cn, Wei.Bian@uts.edu.au, yuanb@sz.tsinghua.edu.cn,
dacheng.tao@sydney.edu.au

Abstract

Scene perceiving and understanding tasks including depth estimation, visual odometry (VO) and camera relocalization are fundamental for applications such as autonomous driving, robots and drones. Driven by the power of deep learning, significant progress has been achieved on individual tasks but the rich correlations among the three tasks are largely neglected. In previous studies, VO is generally accurate in local scope yet suffers from drift in long distances. By contrast, camera relocalization performs well in the global sense but lacks local precision. We argue that these two tasks should be strategically combined to leverage the complementary advantages, and be further improved by exploiting the 3D geometric information from depth data, which is also beneficial for depth estimation in turn. Therefore, we present a collaborative learning framework, consisting of DepthNet, LocalPoseNet and GlobalPoseNet with a joint optimization loss to estimate depth, VO and camera localization unitedly. Moreover, the Geometric Attention Guidance Model is introduced to exploit the geometric relevance among three branches during learning. Extensive experiments demonstrate that the joint learning scheme is useful for all tasks and our method outperforms current state-of-the-art techniques in depth estimation and camera relocalization with highly competitive performance in VO.

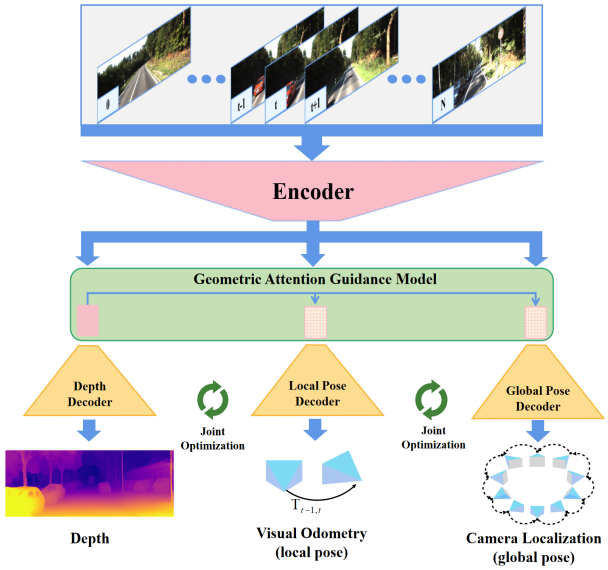

Figure 1: The collaborative learning framework architecture.

1 Introduction

As the basis of various key applications such as autonomous driving, VR/AR, robot vision and drones, researches on scene perceiving and understanding including depth estimation, visual odometry and camera relocalization have attracted significant attention from the community. Due to the development of deep learning, many approaches have been proposed recently to solve these three individual tasks by using CNNs or RNNs to replace traditional methods in supervised or unsupervised manners.

Depth Estimation has been extensively studied recently due to its crucial role in 3D scene understanding. Supervised monocular learning is first proposed as a regression problem to learn a mapping from RGB images to per-pixel depth maps using labelled datasets [Eigen *et al.*, 2014]. DORN [Fu *et al.*, 2018] turns the regression procedure into a multi-class classification problem with discrete depth values. Although these supervised methods can achieve impressive results, collecting a dataset with ground truth is both challenging and expensive, especially for outdoor scenarios. Unsupervised approaches [Zhou *et al.*, 2017; Wang *et al.*, 2018] make use of the photometric consistency between adjacent frames to provide self-supervision. This pipeline relies on a spatial transformer network [Jaderberg *et al.*, 2015] to synthesize reference frames using target frames, which can simultaneously optimize the pose transformation between them. Stereo methods [Godard *et al.*, 2017; Zhan *et al.*,

*The work was done during Haimei Zhao's visit at UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney.

2018] employ the consensus between the left and right cameras in image or feature level. These self-supervised methods tend to suffer from the violation of moving vehicles or people. To tackle this issue, several works [Casser *et al.*, ; Xu *et al.*, 2019] use the Mask R-CNN [He *et al.*, 2017] to separate dynamic objects from the scene and deal with them separately to improve the estimation accuracy. MD2 [Godard *et al.*, 2019] employs an auto-masking strategy to handle static camera and moving objects.

Visual Odometry is one of the most essential techniques in computer vision and robotic localization. Traditional methods generally follow a pipeline [Fraundorfer and Scaramuzza, 2012] including camera calibration, feature detection, feature matching, outlier rejection, motion estimation, scale estimation and Bundle Adjustment. Recently, deep learning based methods have been presented, which replace the original visual odometry (VO) process with an end-to-end neural network. DeepVO [Wang *et al.*, 2017] combines CNNs and Long Short-Term Memory Networks (LSTMs) to obtain pose estimation from image sequences by conducting sequential modelling. UndeepVO [Li *et al.*, 2018] employs a framework similar to [Zhou *et al.*, 2017] to estimate VO with recovered scales from stereo sequences. CTCNet [Iyer *et al.*, 2018] adds a set of transformation constraints across a series of frames to enforce the geometric consistency of the trajectory. At present, VO estimation is relatively accurate locally but it is still confronted with the issue of long-distance drift.

Camera Relocalization aims to infer the global pose of a camera from visual scene representations, which is crucial for navigation applications. PoseNet [Kendall *et al.*, 2015] is one of the early attempts to train CNNs in the end-to-end manner to infer the camera's 6-DoF pose from a single image. MapNet [Brahmbhatt *et al.*, 2018] combines camera relocalization with traditional VO and GPS data to improve the estimation accuracy. With the self-attention mechanism, AtLoc [Wang *et al.*, 2019] devotes to helping CNNs focus on geometrically robust objects or features, producing more robust estimation. Currently, the quality of camera relocalization is reasonably satisfactory in the global range but there is still a big margin for improvement concerning local precision.

In previous work, the three tasks (depth estimation, VO and camera relocalization) are accomplished using separate neural networks from a single image or video clips. Since they all fall into the domain of scene understanding from input images, it is reasonably to argue that the abundant inter-task geometric correlations should be exploited to bring extra benefits to each other.

In this paper, we propose a collaborative learning network to jointly conduct the tasks of depth estimation, VO and camera relocalization from monocular videos. We leverage three branches (DepthNet, LocalPoseNet and GlobalPoseNet) to perceive the environment from different perspectives, which are particularly suitable for exploiting 3D geometry structure, local pose transformation and global pose, respectively. With a purposefully designed joint optimization loss, our three branches are able to provide complementary benefits and overcome the individual defects (e.g., drift in long distance for LocalPoseNet and low precision in local range for GlobalPoseNet) during learning. In addition, the Geometric Attention Guidance Model (GAGM) is introduced to extract valuable 3D geometric information from the depth estimation branch to enhance the VO and camera relocalization branches, which is also helpful for the improvement of depth estimation in turn. The visualization of the attention maps shows that our GAGM does have the ability to learn meaningful guidance from the depth information.

The main contributions of our work are as follows:

1) We present a collaborative learning framework to jointly conduct the tasks of depth estimation, visual odometry and camera relocalization to leverage the complementary advantages among tasks. To the best of our knowledge, this is the first work to jointly solve these three problems.

2) We introduce a Geometric Attention Guidance Model (GAGM) as an inter-task interaction mechanism to help LocalPoseNet and GlobalPoseNet acquire valuable geometric information from the depth data to improve the accuracy.

3) We conduct extensive experiments on KITTI and our method outperforms SOTA methods by 10.4% in depth estimation and 15.2% and 27.4% in camera relocalization according to two major performance metrics, respectively.

2 Methodology

This section presents our collaborative learning framework for depth estimation, visual odometry and camera relocalization from monocular videos. Taking a sequence of video frames as input, the aim of our framework is to produce depth map D_t and camera global pose P_t from each frame and the ego-motion of each frame pair $T_{t+1\rightarrow t}$ simultaneously. For this purpose, our framework consists of three main network branches DepthNet, LocalPoseNet and GlobalPoseNet with an interaction mechanism GAGM.

2.1 Depth and Visual Odometry

The estimation of depth and VO employs the self-supervised pipeline [Zhou *et al.*, 2017], which is built based on the photometric consistency among adjacent frames.

As shown in Figure 2, our DepthNet and LocalPoseNet both adopt encoder-decoder architecture, while skip connections are used in DepthNet to utilize both shallow information and deep abstract features to get accurate depth maps. To ensure a fair comparison with other methods, we modified the ResNet18 network pretrained on ImageNet [Russakovsky *et al.*, 2015] as our encoder, E_D and E_L. The DepthNet Θ takes frame sequence $< I_{t-n}, I_t, I_{t+n} >$ as input and outputs corresponding depth maps $< D_{t-n}, D_t, D_{t+n} >$, $f(I, \Theta) = D$. Meanwhile, from continuous frames, the LocalPoseNet Φ produces the 6-DoF relative pose transformation $< T_{t+n\rightarrow t}, T_{t-n\rightarrow t} >$ of adjacent frame pairs, $g(< I_t, I_{t\pm n} >, \Phi) =< T_{t\pm n\rightarrow t} >$.

During learning, the predicted depth map D_t and pose transformation T are used to warp the source frame I_{t+n} to obtain the synthesized target frame $\hat{I}_{t+n\rightarrow t}$. By adopting the differential bilinear sampling mechanism [Jaderberg *et al.*, 2015], we can obtain corresponding reprojected pixel coordinate $\hat{p}$ in $\hat{I}_{t+n\rightarrow t}$ for each pixel p in frame I_t:

$$\hat{p} \sim K T_{t+n\rightarrow t} D_t(p) K^{-1} h(p) \tag{1}$$

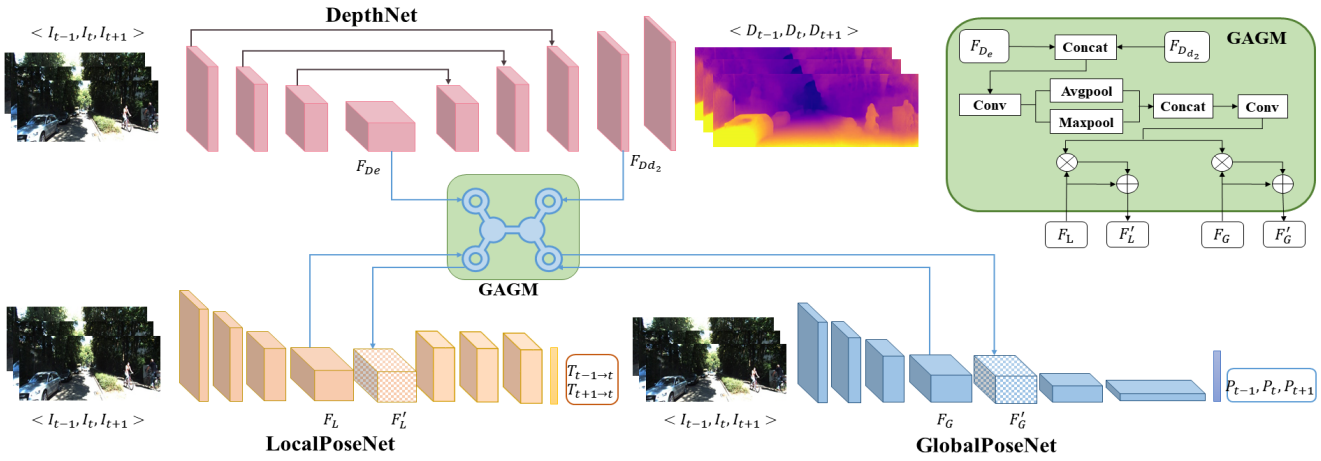

Figure 2: Network architecture of the collaborative learning framework.

where K is the known camera intrinsics and $h(p) = (x, y, 1)$ means the homogeneous coordinates of pixel p. By constraining the difference between the synthetic frame $\hat{I}_{t+n\to t}$ and the original target frame I_t, depth and visual odometry can be optimized in an self-supervised manner:

$$\ell = \sum_{p\in I_t} |I_t(p) - \hat{I}_{t+n\to t}(p)| \tag{2}$$

The difference measurement ℓ is a weighted combination of l_1 loss and SSIM following prior work [Godard *et al.*, 2017]:

$$\begin{aligned}\ell_r(I_t, \hat{I}_{t+n\to t}) = \lambda \ell^{l_1}(I_t, \hat{I}_{t+n\to t}) \\ + (1-\lambda)\ell^{SSIM}(I_t, \hat{I}_{t+n\to t})\end{aligned} \tag{3}$$

To overcome the effect of occlusion and dynamic vehicles moving at the same speed as the camera, the minimum dissimilarities and the stationary mask κ are adopted as proposed in [Godard *et al.*, 2019]:

$$\ell_e = \kappa \min_{n\in N} \ell_r(I_t, \hat{I}_{t+n\to t}) \tag{4}$$

where, N is the collection of the reference frames.

An image-aware smoothing item is used to regularize the depth discontinuity as previous work [Godard *et al.*, 2017] :

$$\ell_m = |\partial_x \mu_{D_t}| e^{-|\partial_x I_t|} + |\partial_y \mu_{D_t}| e^{-|\partial_y I_t|} \tag{5}$$

where μ_{D_t} is the normalized inverse depth by mean value.

2.2 Camera Relocalization

Our GlobalPoseNet Ψ aims to learn the global camera pose from each image on the training set $C = (I, P^*)$ in a supervised manner, $g(I; \Psi) = P$. Given an input frame I_t, the pretrained ResNet34 network is first used in GlobalPoseNet to extract valuable feature F_{Pt}. The following Multilayers Perceptrons (MLPs) then map feature F_{Pt} to global pose P.

$$P = MLPs(F_{Pt}) \tag{6}$$

We represent the camera pose P as $[p, q]$ with $p \in R^3$ for position and a unit quaternion $q \in R^4$ for orientation to regress it with l_1 or l_2 norm properly. Because any rotations in 3D space can be effectively mapped to valid and unique unit quaternions by normalizing 4D quaternions to unit length and restricted to the same hemisphere according to [Brahmbhatt *et al.*, 2018; Wang *et al.*, 2019]. To regularize the network, we adopt the l_1 loss between the predicted $[p, q]$ and the label $[p^*, q^*]$:

$$\ell_g = \|p - p^*\|_1 e^{-\eta} + \eta + \|\log q - \log q^*\|_1 e^{-\varphi} + \varphi \tag{7}$$

where η and φ are weights for balancing the position loss and rotation loss, learning from initial values η_0 and φ_0 during training simultaneously. $\log q$ is the logarithmic form of q:

$$\log q = \begin{cases} \dfrac{v}{\|v\|} \cos^{-1} u, & \text{if } \|v\| \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

v and u are unit quaternion q's real and imaginary parts.

2.3 Collaborative Learning

Geometric Attention Guidance Model We propose a self-learning Geometric Attention Guidance Model (GAGM) to learn and leverage the latent geometric correlation between DepthNet and two pose networks LocalPoseNet and GlobalPoseNet to acquire more accurate estimation. Our GAGM takes the deepest feature of depth encoder F_{De} and the second last feature of depth decoder F_{Dd_2} as input and outputs two attention scale maps for LocalPoseNet and GlobalPoseNet, respectively. As shown in Figure 2 (green area), in GAGM, we first combine F_{De} and F_{Dd_2} to get the depth feature F_D via a concatenation and a convolution layer, to utilize both abstract and explicit 3D geometric features from the depth learning branch. Next, F_D is passed to an average pooling and a max pooling operations along the channel axis and then concatenated to obtain a highly compressed expression, since pooling layers can help feature be concentrated as shown in prior studies [Woo *et al.*, 2018]. Lastly, a subsequent convolution layer is employed to get the final attention scale maps. During collaborative learning, the corresponding learned attention maps in GAGM are introduced

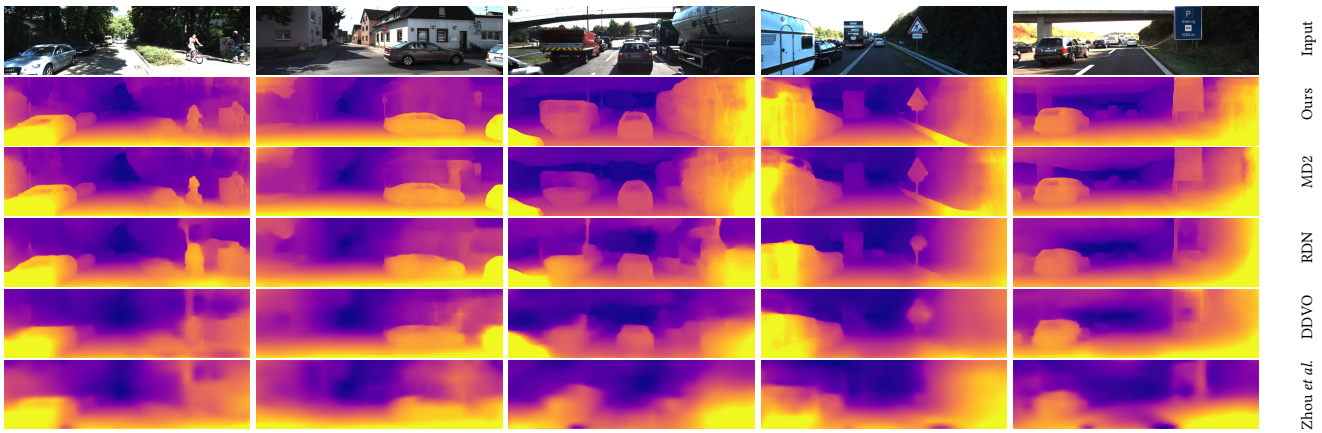

Figure 3: Qualitative performance on KITTI test set.

as guidance to scale the deepest feature of LocalPoseNet and GlobalPoseNet, F_L and F_G, by multiplication. The scaled features are regarded as a residual item to be added to the original feature F_L and F_G, respectively. By introducing the condensed feature from DepthNet, our framework allows LocalPoseNet and GlobalPoseNet branches to acquire useful information from not only 2D input images but also 3D geometry to improve accuracy, which is also beneficial for the depth estimation in turn. GAGM can be formulated as:

$$\begin{aligned} F_D &= W_c[F_{De}; F_{Dd_2}] \\ F'_L &= F_L(W_l[AVP(F_D); MAP(F_D)]) + F_L \\ F'_G &= F_G(W_g[AVP(F_D); MAP(F_D)]) + F_G \end{aligned} \tag{9}$$

here, F'_L and F'_G denote the guided feature of LocalPoseNet and GlobalPoseNet, respectively. W_c, W_l and W_g are learnable weights of the corresponding convolution layers.

The Joint Optimization Loss We design a joint optimization loss ℓ_c for LocalPoseNet and GlobalPoseNet to regularize them mutually. We can calculate a local pose transformation $T_{t+n\to t-n}$ within input images window from pairwise pose transformation learned from LocalPoseNet:

$$T_{t+n\to t-n} = T_{t+n\to t} T^{-1}_{t-n\to t} \tag{10}$$

Besides, the local pose transformation can also be produced from camera pose $[p, q]$ estimated in GlobalPoseNet:

$$\begin{aligned} [\tilde{p}, \tilde{q}] &= [p_{t+n} - p_{t-n}, q_{t+n} - q_{t-n}] \\ \tilde{T}_{t+n\to t-n} &= \Gamma([\tilde{p}, \tilde{q}]) \end{aligned} \tag{11}$$

where Γ is the transmutation function to change the representation mode for pose.

Therefore, we can use the l_1 loss to regularize the consistency between the poses predicted from LocalPoseNet and GlobalPoseNet to optimize them jointly within the time window whose size is $2n + 1$, taking n as 1 in our work:

$$\ell_c = \|T_{t+n\to t-n} - \tilde{T}_{t+n\to t-n}\|_1 \tag{12}$$

To summarize, the integrated loss for our collaborative learning framework is:

$$\ell = \frac{1}{S}\sum_s \alpha\ell_e + \beta\ell_m + \gamma\ell_g + \delta\ell_c \tag{13}$$

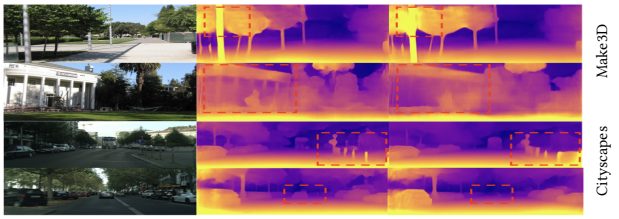

Figure 4: The visual results evaluated on Cityscapes and Make3D.

where s denotes different scales and S means the number of the scales (4 in our work).

Training Details We trained our models using the KITTI dataset [Geiger *et al.*, 2012], which is one of the most commonly used dataset in autonomous driving. We took sequence 00-08 from the KITTI Odometry dataset as training data, sequence 09 and 10 as test data following prior works [Zhou *et al.*, 2017; Zou *et al.*, 2018]. Our DepthNet and LocalPoseNet were first pretrained with input images resized to 640×192, a batch size of 12 and parameter α and β set to 1 and 0.05. After that, we started to train the collaborative framework with input images resized to 1024×320, a batch size of 4 and the loss weight α, β, γ and δ set to 0.5, 0.05, 0.85 and 0.1, respectively. The learning rate was first set to 10^{-4} for the first 10 epochs and then dropped to 10^{-5} for the remaining 40 epochs. We took three consecutive frames as input. With the ability of locally optimizing relative pose and global pose within a time window during learning, our framework should perform better if given longer sequences as input. Our method was implemented in Pytorch.

3 Experiments

Comprehensive evaluations were implemented towards all three tasks (depth estimation, VO and camera relocalization). Our method achieved excellent results in the comparison with existing SOTA researches, which demonstrates the effectiveness of our collaborative learning method.

Methods	Train	Resolution	Error metric↓				Accuracy metric↑		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
†[Garg *et al.*, 2016]	S	608×176	0.152	1.226	5.849	0.246	0.784	0.921	0.967
MD1 R50† [Godard *et al.*, 2017]	S	512×256	0.133	1.142	5.533	0.230	0.830	0.936	0.970
monoResMatch [Tosi *et al.*, 2019]	S	640×192	0.116	0.986	5.098	0.214	0.847	0.939	0.972
MD2 [Godard *et al.*, 2019]	S	1024×320	0.107	0.849	4.764	0.201	0.874	0.953	0.977
†[Zhou *et al.*, 2017]	M	416×128	0.183	1.595	6.709	0.270	0.734	0.902	0.959
†GeoNet [Yin and Shi, 2018]	M	416×128	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [Wang *et al.*, 2018]	M	416×128	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [Zou *et al.*, 2018]	M	576×160	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Struct2depth [Casser *et al.*,]	M	416×128	0.141	1.026	5.142	0.210	0.845	0.845	0.948
RDN [Xu *et al.*, 2019]	M	832×256	0.138	1.016	5.352	0.217	0.823	0.943	0.976
HR [Zhou *et al.*, 2019]	M	1248×384	0.121	0.873	4.945	0.197	0.853	0.955	0.982
MD2(R18) [Godard *et al.*, 2019]	M	1024×320	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Ours (baseline w/o GAGM)	M	1024×320	0.116	0.891	4.758	0.192	0.866	0.956	0.981
Ours (baseline)	M	1024×320	**0.108**	**0.766**	**4.562**	**0.183**	**0.887**	**0.964**	**0.983**
Ours (R18)	M	1024×320	**0.103**	**0.725**	**4.466**	**0.179**	**0.893**	**0.964**	**0.983**

Table 1: Quantitative results of single depth estimation over KITTI test set [Eigen *et al.*, 2014]. For a fair comparison, all the results were evaluated taking 80 m as the maximum depth threshold. The resolution column means the size of input images and the "S" and "M" in train column denote using stereo or monocular images for training. "†" means updated result after publication.

Sequences	PoseNet[Kendall *et al.*, 2015]		MapNet [Brahmbhatt *et al.*, 2018]		AtLoc [Wang *et al.*, 2019]		Ours	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
09	22.80m,6.99°	18.60m,4.99°	31.44m,9.55°	26.03m,7.82°	10.32m,6.51°	9.42m, 5.26°	**8.74m,4.49°**	**7.06m,3.14°**
10	28.53m,7.47°	23.06m,5.88°	57.22m,8.52°	47.38m,7.37°	9.23m,5.60°	7.50m,4.81°	**7.84m, 4.31°**	**6.92m, 3.86°**
Average	25.67m,7.23°	20.83m,5.44°	44.33m,18.07°	36.71m,7.60°	9.78m,6.06	8.46m,5.04	**8.29m,4.4°**	**6.99m,3.5°**

Table 2: Results of camera relocalization on KITTI Odometry. Our method attained the best test results in both mean and median value.

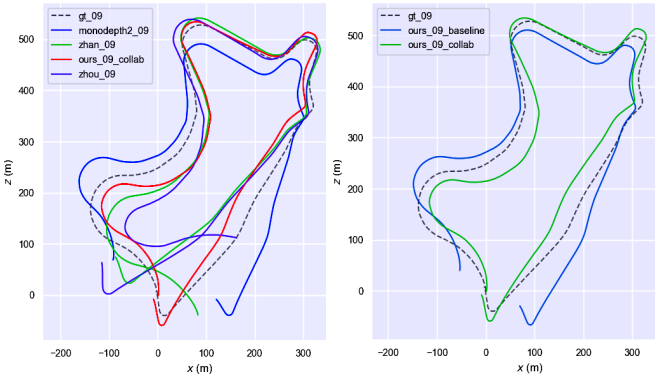

Figure 5: Visualization of VO trajectories using Evo [Grupp, 2017].

3.1 Evaluation of Depth Estimation

We conducted extensive evaluation experiments to compare the depth estimation performance of our method with previous works. The quantitative results are reported in Table 1, which clearly demonstrate that our model outperforms current SOTA approaches trained in self-supervised manner. Moreover, although being trained with monocular sequences only, our model surpassed other methods learned from stereo videos. As shown in Figure 3, our method can generate satisfying depth maps with clear instance boundary and perform properly in some challenging situations including delicate structures (e.g. traffic signs) and low-texture regions (e.g. the surface of tankers or carriages).

Ablation To ensure the fairness of the comparisons, we list ablation study in Table 1. Ours (baseline) means method without collaborative learning with GlobalPoseNet and the joint optimization loss ℓ_c, but with GAGM between DepthNet and LocalPoseNet, which is totally self-supervised. It is clear that our method is superior to other methods even without the supervision of poses for GlobalPoseNet. Ours (baseline w/o GAGM) means further removing the GAGM between DepthNet and LocalPoseNet. From the comparisons, the effect of our collaborative learning framework and GAGM for depth estimation can be demonstrated obviously.

Generalization Ability Although being trained on KITTI only, our method can also achieve promising results on unseen datasets without any fine-tuning. We conducted inference experiments on Make3D and Cityscapes to verify the generalization ability as shown in Figure 4. Compared with SOTA approaches, our method is capable to produce more accurate depth maps with sharper object boundary and better perception of distant instances even on unseen dataset.

3.2 Evaluation of Visual Odometry

Previous studies on Visual Odometry (VO) suffer severely from long-distance drift. To overcome it, our collaborative learning framework takes advantage of the rich 3D geometry information from depth estimation branch and the global localization auxiliary from relocalization network. In Table 2, we conducted quantitative evaluation on KITTI Odometry test sequence 09 and 10. The Absolute Trajectory Error (ATE) was calculated on 5-frame snippets and averaged over the full sequences following the protocol of [Zhou *et al.*, 2017].

Ablation We also trained our models without collaborative learning with camera relocalization task, which was taken as the baseline in Table 3 and Figure 5 (right). The

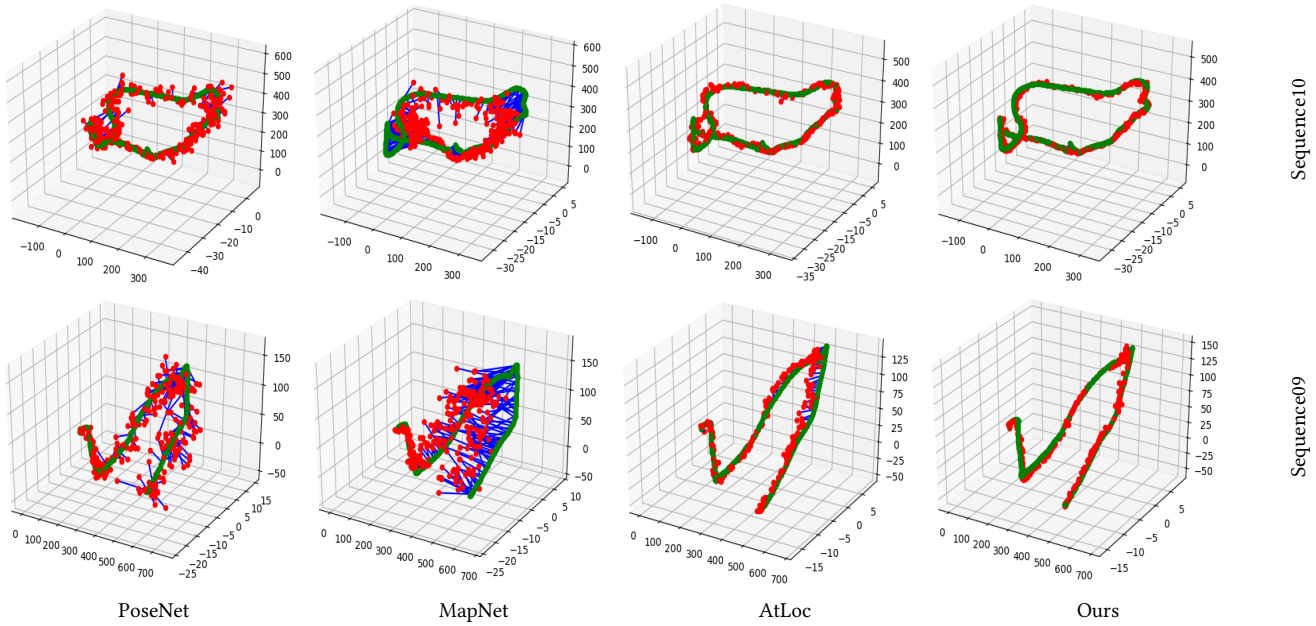

Figure 6: 3D visualization of the camera trajectory of KITTI Odometry sequence 09 and 10 . The units of the axes are meters.

Methods	Sequence09	Sequence10
ORB-SLAM [Mur-Artal *et al.*, 2015]	0.014 ± 0.008	0.012 ± 0.011
Zhou *et al.* [Zhou *et al.*, 2017]	0.021 ± 0.017	0.020 ± 0.015
DDVO [Wang *et al.*, 2018]	0.045 ± 0.108	0.033 ± 0.074
DF-Net [Zou *et al.*, 2018]	0.017 ± 0.007	0.015 ± 0.009
monodepth2 [Godard *et al.*, 2019]	0.017 ± 0.008	0.015 ± 0.010
ours (baseline w/o GAGM)	0.018 ± 0.010	0.017 ± 0.011
ours (baseline)	0.016 ± 0.008	0.016 ± 0.009
ours	$\mathbf{0.014 \pm 0.007}$	$\mathbf{0.014 \pm 0.008}$

Table 3: Results of Visual Odometry on KITTI Odometry dataset.

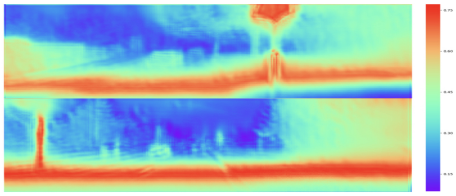

Figure 7: The visualization of learned attention maps in GAGM.

comparisons of our predicted trajectory with other methods and our baseline in Figure 5 can demonstrate our method is valuable for alleviating drift in long range.

3.3 Evaluation of Camera Relocalization

The evaluation of camera relocalization was also conducted on KITTI Odometry sequence 09 and 10. As shown in Table 2, we summarized the mean and median value of position and rotation error with corresponding ground truth. Benefited from the 3D geometry guidance of depth network and local constraints of visual odometry branch, our relocalization results exceed prior SOTA methods in both position and rotation accuracy. We visualized the prediction trajectory of relevant methods using a subsampling factor 5 to show results more clearly, as listed in Figure 6. The green, red and blue lines denote the ground truth, estimation and error, respectively. Compared with other methods, our trajectories are more accurate and noiseless.

Attention Map Analysis The visualization of the attention guidance maps generated by GAGM is shown in Figure 7. It is clear that GAGM is useful to teach the GlobalPoseNet to focus on stable geometric features and regions such as road signs, trees and the ground instead of dynamic cars for remembering and recognizing scenes, which is highly valuable for camera relocalization task.

4 Conclusion

The motivation of this work is to highlight the importance of exploiting the inherent correlation among three classical scene understanding tasks: depth estimation, VO and camera relocalization. With the proposed collaborative learning framework and the joint optimization loss along with the GAGM, we show the great prospect of boosting the performance of the three tasks simultaneously through collaborative learning. In the experimental studies, we show that our method is superior to existing SOTA methods in depth estimation and camera relocalization, and can achieve highly competitive results in VO. As to future work, we will further train and evaluate our method on other datasets to fully investigate its potential in various application scenarios and provide more insights into its working mechanism.

Acknowledgments

The authors are supported by Australian Research Council Projects FL-170100117, IH-180100002, the Natural Science Foundation of China (NSFC) (No. U1713214) and Shenzhen Fundamental Research Fund (No. JCYJ20170412170602564).

References

[Brahmbhatt *et al.*, 2018] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018.

[Casser *et al.*,] 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.

[Eigen *et al.*, 2014] 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*, pages 2366–2374, 2014.

[Fraundorfer and Scaramuzza, 2012] Friedrich Fraundorfer and Davide Scaramuzza. Visual odometry: Part ii: Matching, robustness, optimization, and applications. *IEEE Robotics & Automation Magazine*, 19(2):78–90, 2012.

[Fu *et al.*, 2018] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[Garg *et al.*, 2016] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *The European Conference on Computer Vision*, pages 740–756. Springer, 2016.

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

[Godard *et al.*, 2017] 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*, pages 270–279, 2017.

[Godard *et al.*, 2019] 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*, pages 3828–3838, 2019.

[Grupp, 2017] Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[Iyer *et al.*, 2018] Ganesh Iyer, J. Krishna Murthy, Gunshi Gupta, Madhava Krishna, and Liam Paull. Geometric consistency for self-supervised end-to-end visual odometry. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[Kendall *et al.*, 2015] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[Li *et al.*, 2018] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *IEEE International Conference on Robotics and Automation*, pages 7286–7291. IEEE, 2018.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[Tosi *et al.*, 2019] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

[Wang *et al.*, 2017] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050. IEEE, 2017.

[Wang *et al.*, 2018] 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*, pages 2022–2030, 2018.

[Wang *et al.*, 2019] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. *arXiv preprint arXiv:1909.03557*, 2019.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *The European Conference on Computer Vision*, pages 3–19, 2018.

[Xu *et al.*, 2019] Haofei Xu, Jianmin Zheng, Jianfei Cai, and Juyong Zhang. Region deformer networks for unsupervised depth estimation from unconstrained monocular videos. In *IJCAI*, 2019.

[Yin and Shi, 2018] 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*, pages 1983–1992, 2018.

[Zhan *et al.*, 2018] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[Zhou *et al.*, 2017] 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*, pages 1851–1858, 2017.

[Zhou *et al.*, 2019] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[Zou *et al.*, 2018] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *The European Conference on Computer Vision*, pages 36–53, 2018.