TCL: Triplet Consistent Learning for Odometry Estimation of Monocular Endoscope

Hao Yue[1,2,3] and Yun Gu[1,2,3](✉)

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
{yuehao6,geron762}@sjtu.edu.cn
[2] Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[3] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China

Abstract. The depth and pose estimations from monocular images are essential for computer-aided navigation. Since the ground truth of depth and pose are difficult to obtain, the unsupervised training method has a broad prospect in endoscopic scenes. However, endoscopic datasets lack sufficient diversity of visual variations, and appearance inconsistency is also frequently observed in *image triplets*. In this paper, we propose a triplet-consistency-learning framework (TCL) consisting of two modules: Geometric Consistency module(GC) and Appearance Inconsistency module(AiC). To enrich the diversity of endoscopic datasets, the GC module generates synthesis triplets and enforces geometric consistency via specific losses. To reduce the appearance inconsistency in the *image triplets*, the AiC module introduces a triplet-masking strategy to act on photometric loss. TCL can be easily embedded into various unsupervised methods without adding extra model parameters. Experiments on public datasets demonstrate that TCL effectively improves the accuracy of unsupervised methods even with limited number of training samples. Code is available at https://github.com/EndoluminalSurgicalVision-IMR/TCL.

Keywords: Self-supervised monocular pose estimation · Endoscopic images · Data augmentation · Appearance inconsistency

1 Introduction

The technical advances in endoscopes have extended the diagnostic and therapeutic value of endoluminal interventions in a wide range of clinical applications. Due to the restricted field of view, it is challenging to control the flexible endoscopes inside the lumen. Therefore, the development of navigation systems, which

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_14.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 144–153, 2023.
https://doi.org/10.1007/978-3-031-43996-4_14

locates the position of the end-tip of endoscopes and enables the depth-wise visualization, is essential to assisting the endoluminal interventions. A typical task is to visualize the depth-wise information and estimate the six-degree-of-freedom (6DoF) pose of endoscopic camera based on monocular imaging.

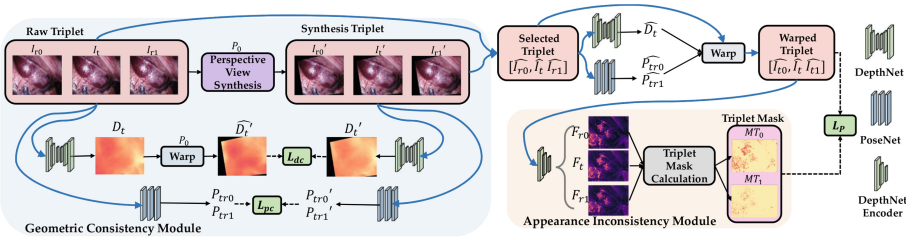

Fig. 1. Overview of the proposed triplet-consistency-learning framework (TCL) with two modules: Geometric Consistency module(GC) and Appearance Inconsistency module(AiC). The GC module utilizes the Perspective View Synthesis technique to produce Synthesis Triplets, while enforcing geometric consistency through the Depth Consistent Loss $\mathcal{L}_{dc}$ and the Pose Consistent Loss $\mathcal{L}_{pc}$. The AiC module generates Triplet Masks based on the Warped Triplet to apply to the photometric loss $\mathcal{L}_{p}$. TCL can be easily embedded into unsupervised SfM methods without adding extra model parameters.

Due to the clinical limitations, the ground truth of depth and pose trajectory of endoscope imaging is difficult to acquire. Previous works jointly estimated the depth and the pose via the unsupervised frameworks [2,8,10,12,17]. The main idea is modeling the differences of video frames with the Structure-from-Motion (SfM) mechanisms. In this framework [17], the *image triplet*, including a specific frame (i.e. *target frame*) and its temporal neighborhood (i.e. *reference frames*), is fed into the model to estimate the pose and depth. The model is optimized by minimizing the warping loss between the target frame and reference frames. Following the basic SfM method [17], scale-consistency [2], auto-masking [5], cost-volume [14] and optical flows [15] are also introduced to further improve the performance. These methods have also been applied to endoscopic scenarios with specific designs of attention modules [10] and priors from sparse depth [8]. Although the performance of SfM methods is promising, the intrinsic challenges of endoscopic data still require further consideration.

The first issue is the insufficient visual diversity of endoscopic datasets. Compared with the large-scaled KITTI dataset [4] for general vision tasks, the collection of endoscopic datasets is challenged by the limited freedom of instruments and the scenarios. In this case, the visual variations of lumen structures and texture appearances cannot be fully explored. While road datasets mostly exhibit 3DOF motion(2DOF translation and 1DOF rotation in the road plane), endoscopy involves 6DOF motion within 3D anatomical structures. Therefore,

SfM algorithms for endoscopic imaging are designed to estimate complicated trajectories with limited data diversity. General data augmentation methods, including random flipping and cropping, cannot generate synthesis with sufficient variance of camera views, and the realistic camera motion cannot be fully guaranteed. Recent works have tried to mimic the camera motion to improve the diversity of samples. For example, PDA [16] generated new samples for supervised depth estimation, and 3DCC [6] used synthetic samples to test the robustness of the models. However, the perspective view synthesis for unsupervised SfM framework, especially the transformation on *image triplets* is still underexplored.

The second issue is the appearance inconsistency in *image triplets*. Due to the complicated environment of endoluminal structures, the illumination changes, motion blurness and specular artefacts are frequently observed in endoscopy images. The appearance-inconsistent area may generate substantial photometric losses even in the well-aligned adjacent frames. These photometric losses caused by the inconsistent appearance impede the training process and remain unable to optimize. To handle this problem, AF-SfMLearner [12] adopted the flow network to predict appearance flow to correct inconsistency between consecutive frames. RNNSLAM [9] used an encoder network to predict masks with supervised HSV signals from the original images. Consequently, these methods adopted auxiliary modules to handle the visual inconsistency, involving more parameters to learn.

In this paper, we propose a triplet-consistency-learning framework (TCL) for unsupervised depth and pose estimation of monocular endoscopes. To improve the visual diversity of *image triplets*, the perspective view synthesis is introduced, considering the geometric consistency of camera motion. Specifically, the depth-consistency and pose-consistency are preserved via specific losses. To reduce the appearance inconsistency in the *image triplets*, a triplet-masking strategy is proposed by measuring the differences between the triplet-level and the frame-level representations. The proposed framework does not involve additional model parameters, which can be easily embedded into previous SfM methods. Experiments on public datasets demonstrate that TCL can effectively improve the accuracy of depth and pose estimation even with small amounts of training samples.

2 Methodology

2.1 Unsupervised SfM with Triplet Consistency Learning

Unsupervised SfM Method. The unsupervised SfM methods adopt DepthNet and PoseNet to predict the depth and the pose, respectively. With depth and pose prediction, the reference frames $I_{ri}(i = 0, 1)$ are warped to the warped frames $I_{ti}(i = 0, 1)$. The photometric loss, denoted by $\mathcal{L}_P$, is introduced to measure the differences between $I_{ti}(i = 0, 1)$ and the target frame I_t. The loss can be implemented by L_1 norm [17] and further improved with SSIM metrics [13].

In addition to $\mathcal{L}_P$, auxiliary regularization loss functions, such as depth map smoothing loss [5] and depth scale consistency loss [2], are also introduced to

improve the performance. In this work, these regularization terms are denoted by $\mathcal{L}_{Reg}$. Therefore, the final loss functions $\mathcal{L}$ of unsupervised methods can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_P + \mathcal{L}_{Reg} \tag{1}$$

Previous unsupervised methods [2,5,17] have achieved excellent results on realistic road datasets such as KITTI dataset [4]. However, endoscopic datasets lack sufficient diversity of visual variations, and appearance inconsistency is also frequently observed in *image triplets*. Therefore, the unsupervised SfM methods based on $\mathcal{L}_p$ require further considerations to address the issues above.

Framework Architecture. As shown in Fig. 1, we propose a triplet-consistency-learning framework (TCL) based on unsupervised SfM with *image triplets*. TCL can be easily embedded into SfM variants without adding model parameters. To enrich the diversity of endoscopic samples, the Geometric Consistency module (GC) performs the perspective view synthesis method to generate synthesis triplets. Additionally, we introduce the Depth Consistent Loss $\mathcal{L}_{dc}$ and the Pose Consistent Loss $\mathcal{L}_{pc}$ to preserve the depth-consistency and the pose-consistency between raw and synthesis triplets. To reduce the affect of appearance inconsistency in the triplet, we propose Appearance Inconsistency module (AiC) where the Triplet Masks of reference frames are generated to reduce the inconsistent warping in the photometric loss.

2.2 Learning with Geometric Consistency

Since the general data augmentation methods cannot generate synthesis with sufficient variance of camera views, 3DCC [6] and PDA [16] mimic the camera motion to generate the samples by applying perspective view synthesis. However, raw and novel samples are used separately in the previous works. To enrich the diversity of endoscopic datasets, we perform the synthesis on triplets. Furthermore, the loss functions are introduced to preserve the depth-consistency and the pose-consistency.

Synthesis Triplet Generation. The perspective view synthesis method aims to warp the original image I to generate a new image I'. The warping process is based on the camera intrinsic matrix K, the depth map D of the original image and perturbation pose P_0. For any point q in I, its depth value is denoted as z in D. The corresponding point q' on the new image I' is calculated by Eq.(2):

$$q' \sim K P_0 z K^{-1} q \tag{2}$$

Given the depth maps generated from a pre-trained model, we perform perspective view synthesis with the same pose transformation P_0 on the three frames of raw triplet respectively. Then we obtain the synthesis triplet $[I'_{r0}, I'_t, I'_{r1}]$. Selected triplet $[\widehat{I_{r0}}, \widehat{I_t}, \widehat{I_{r1}}]$ is randomly selected from raw and synthesis triplets as the unsupervised training triplet to calculate $\mathcal{L}$ in Eq.(1).

Depth Consistent Loss. Figure 1 illustrates that the depth prediction of raw target frame I_t is D_t, and the depth prediction of synthesis target frame I'_t is D'_t. The relative pose between the raw and synthesis triplets is randomly generated as P_0. With P_0, we can warp the depth prediction D_t to $\widehat{D'_t}$. To preserve depth-consistency between target frames of raw and synthesis triplet, we propose the Depth Consistent Loss function $\mathcal{L}_{dc}$, defined as the L1 loss function of D'_t and $\widehat{D'_t}$, written as

$$\mathcal{L}_{dc} = |D'_t - \widehat{D'_t}| \tag{3}$$

Pose Consistent Loss. As shown in Fig. 1, the inner pose predictions of the adjacent frames of the raw and synthesis triplet are $P_{tri}(i = 0, 1)$ and $P'_{tri}(i = 0, 1)$. Since the three frames of the synthesis triplet are warped from the same pose perturbation P_0, the inner poses of the synthesis triplet remain the same as the raw triplet. To preserve inner pose-consistency between raw and synthesis triplets, we propose Pose Consistent Loss $\mathcal{L}_{pc}$, defined as the weighted L1 loss function of the inner pose prediction of the raw and synthesis triplets. Specifically, we use the translational $t_{tri}, t'_{tri} \in \mathcal{R}^{3\times1}(i = 0, 1)$ and rotational $R_{tri}, R'_{tri} \in \mathcal{R}^{3\times3}(i = 0, 1)$ components of the pose transformation matrix to calculate $\mathcal{L}_{pc}$. We weight the translation component by λ_{pt}, Pose Consistent Loss is written as Eq.(4).

$$\mathcal{L}_{pc} = \sum_{i=0,1} (|R_{tri} - R'_{tri}| + \lambda_{pt}|t_{tri} - t'_{tri}|) \tag{4}$$

2.3 Learning with Appearance Inconsistency

The complicated environment of endoluminal structures may cause appearance inconsistency in *image triplets*, leading to the misalignment of reference and target frames. Previous works [9,12] proposed auxiliary modules to handle the appearance inconsistency, involving more parameters of models.

Since the movement of camera is normally slow, the same appearance inconsistency is unlikely to exist multiple times within an endoscopic *image triplet*. Therefore, we can measure the differences between the triplet-level and the frame-level representations to eliminate the appearance inconsistency.

Specifically, for the selected triplet $[\widehat{I_{r0}}, \widehat{I_t}, \widehat{I_{r1}}]$, two warped frames $[\widehat{I_{t0}}, \widehat{I_{t1}}]$ are generated from two reference frames. To obtain the frame-level representations, we used the encoder of DepthNet to extract the feature maps of the three frames $[\widehat{I_{t0}}, \widehat{I_t}, \widehat{I_{t1}}]$ respectively. The feature maps are upsampled to the size of the original image, denoted by $[F_{r0}, F_t, F_{r1}]$. As in Eq.(5), the triplet-level representations F_R are generated by the weighted aggregation of the feature maps, which is dominated by the feature of the target frame with weight λ_t. To measure the differences between the triplet-level and the frame-level representations, we calculate feature difference maps $Df_i(i = 0, 1)$ by weighting direct subtraction

Table 1. Quantitative results on the SCARED and SERV-CT Dataset. The best results among all methods are in **bold**. The best results among each series are underlined.

Series	Methods	**SCARED Dataset**									
		Pose Metrics			Depth Metrics						
		ATE(mm)↓	t_{RPE}(mm)↓	r_{RPE}(deg)↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
	SfMLearner	6.308	$0.356_{\pm 0.16}$	$\mathbf{0.217}_{\pm 0.12}$	$0.472_{\pm 0.07}$	$7.870_{\pm 3.32}$	$14.024_{\pm 5.55}$	$0.499_{\pm 0.05}$	$0.365_{\pm 0.10}$	$0.636_{\pm 0.10}$	$0.810_{\pm 0.06}$
MonoDepth2-Based	MonoDepth2	4.848	$0.479_{\pm 0.20}$	$0.459_{\pm 0.22}$	$0.454_{\pm 0.06}$	$7.311_{\pm 2.87}$	$13.583_{\pm 5.02}$	$0.487_{\pm 0.05}$	$0.368_{\pm 0.10}$	$0.650_{\pm 0.11}$	$0.818_{\pm 0.06}$
	AF-SfMLearner	3.506	$\underline{\mathbf{0.161}}_{\pm 0.10}$	$0.265_{\pm 0.14}$	$\underline{0.446}_{\pm 0.06}$	$\underline{7.153}_{\pm 3.02}$	$13.517_{\pm 5.16}$	$0.481_{\pm 0.05}$	$0.371_{\pm 0.10}$	$0.651_{\pm 0.11}$	$0.825_{\pm 0.06}$
	MonoDepth2+Ours	$\underline{\mathbf{3.110}}$	$\underline{\mathbf{0.161}}_{\pm 0.11}$	$\underline{0.250}_{\pm 0.15}$	$\underline{0.446}_{\pm 0.06}$	$7.185_{\pm 3.30}$	$\underline{\mathbf{13.405}}_{\pm 5.20}$	$\underline{0.480}_{\pm 0.05}$	$\underline{\mathbf{0.373}}_{\pm 0.11}$	$\underline{\mathbf{0.655}}_{\pm 0.12}$	$\underline{0.828}_{\pm 0.06}$
SC-SfMLearner-Based	SC-SfMLearner	4.743	$\underline{0.478}_{\pm 0.21}$	$0.466_{\pm 0.24}$	$0.442_{\pm 0.06}$	$7.044_{\pm 2.96}$	$13.580_{\pm 5.42}$	$0.479_{\pm 0.04}$	$\underline{0.368}_{\pm 0.11}$	$0.650_{\pm 0.12}$	$0.826_{\pm 0.06}$
	Endo-SfMLearner	5.013	$0.494_{\pm 0.22}$	$\underline{0.461}_{\pm 0.24}$	$0.438_{\pm 0.06}$	$6.969_{\pm 3.24}$	$13.592_{\pm 5.46}$	$0.478_{\pm 0.05}$	$0.365_{\pm 0.10}$	$0.650_{\pm 0.11}$	$0.826_{\pm 0.06}$
	SC-SfMLearner+Ours	$\underline{4.601}$	$0.490_{\pm 0.22}$	$0.464_{\pm 0.24}$	$\underline{\mathbf{0.437}}_{\pm 0.05}$	$\underline{\mathbf{6.865}}_{\pm 2.93}$	$\underline{13.471}_{\pm 5.32}$	$\underline{\mathbf{0.475}}_{\pm 0.04}$	$\underline{0.368}_{\pm 0.11}$	$\underline{0.653}_{\pm 0.12}$	$\underline{\mathbf{0.831}}_{\pm 0.05}$
Series	Methods	**SERV-CT Dataset**									
		Pose Metrics			Depth Metrics						
		ATE(mm)↓	t_{RPE}(mm)↓	r_{RPE}(deg)↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
	SfMLearner	-	-	-	$0.114_{\pm 0.04}$	$2.005_{\pm 1.42}$	$12.632_{\pm 6.35}$	$0.149_{\pm 0.05}$	$0.864_{\pm 0.12}$	$0.985_{\pm 0.02}$	$\mathbf{1.000}_{\pm 0.00}$
MonoDepth2-Based	MonoDepth2	-	-	-	$0.124_{\pm 0.03}$	$2.298_{\pm 1.38}$	$13.639_{\pm 5.78}$	$0.165_{\pm 0.04}$	$0.843_{\pm 0.10}$	$0.976_{\pm 0.04}$	$0.998_{\pm 0.01}$
	AF-SfMLearner	-	-	-	$0.115_{\pm 0.04}$	$2.101_{\pm 1.56}$	$13.136_{\pm 6.74}$	$0.156_{\pm 0.05}$	$0.860_{\pm 0.10}$	$0.979_{\pm 0.03}$	$0.998_{\pm 0.01}$
	MonoDepth2+Ours	-	-	-	$\underline{\mathbf{0.103}}_{\pm 0.03}$	$\underline{\mathbf{1.694}}_{\pm 1.25}$	$\underline{\mathbf{11.711}}_{\pm 5.89}$	$\underline{\mathbf{0.139}}_{\pm 0.04}$	$\underline{0.886}_{\pm 0.09}$	$\underline{\mathbf{0.986}}_{\pm 0.02}$	$\underline{\mathbf{1.000}}_{\pm 0.00}$
SC-SfMLearner-Based	SC-SfMLearner	-	-	-	$0.113_{\pm 0.04}$	$2.417_{\pm 2.21}$	$13.719_{\pm 8.34}$	$0.177_{\pm 0.08}$	$0.872_{\pm 0.09}$	$0.959_{\pm 0.05}$	$0.980_{\pm 0.03}$
	Endo-SfMLearner	-	-	-	$0.133_{\pm 0.05}$	$3.295_{\pm 2.88}$	$15.974_{\pm 9.23}$	$0.224_{\pm 0.10}$	$0.829_{\pm 0.11}$	$0.928_{\pm 0.07}$	$0.961_{\pm 0.05}$
	SC-SfMLearner+Ours	-	-	-	$\underline{\mathbf{0.103}}_{\pm 0.04}$	$\underline{1.868}_{\pm 1.50}$	$\underline{12.199}_{\pm 6.71}$	$\underline{0.150}_{\pm 0.06}$	$\underline{\mathbf{0.888}}_{\pm 0.09}$	$\underline{0.970}_{\pm 0.04}$	$\underline{0.996}_{\pm 0.01}$

and SSIM similarity [13] with weight λ_{sub}. The Triplet Mask of each reference frame is generated by reverse normalizing the difference map to $[\beta, 1]$.

$$\begin{aligned} F_R &= \lambda_t F_t + \frac{1}{2}(1-\lambda_t)(F_{r0} + F_{r1}) \\ Df_i &= \lambda_{sub} N_{(0,1)}(|F_R - F_{ri}|) \\ &\quad + (1-\lambda_{sub}) N_{(0,1)}(|1 - \mathrm{SSIM}(F_R, F_{ri})|), (i = 0, 1) \\ MT_i &= N_{(\beta,1)}(1 - Df_i), (i = 0, 1) \end{aligned} \tag{5}$$

where $N_{(a,b)}(\cdot)$ normalizes the input to the range $[a, b]$.

2.4 Overall Loss

The final loss of TCL $\mathcal{L}_t$ is formulated as follows:

$$\mathcal{L}_t = MT \odot \mathcal{L}_P + \mathcal{L}_{Reg} + \lambda_d \mathcal{L}_{dc} + \lambda_p \mathcal{L}_{pc} \tag{6}$$

where $\odot$ denotes that the Triplet Mask $MT_i (i = 0, 1)$ is applied to the photometric loss calculation of the two reference frames respectively. The final photometric loss is obtained by averaging the photometric losses of the two reference frames after applying $MT_i (i = 0, 1)$. λ_d, λ_p are weights of $\mathcal{L}_{dc}$ and $\mathcal{L}_{pc}$. Since the early adoption of $\mathcal{L}_{dc}$ and $\mathcal{L}_{pc}$ may lead to overfitting, the DepthNet and PoseNet are warmed up with N_w epochs before adding the two loss functions. The synthesis method may inherently generate invalid (black) areas in the augmented samples. This arises from the single-image-based augmentation process, which lacks the additional information to fill the new areas generated from the viewpoint transformation. The invalid regions should be masked in the related loss functions.

3 Experiments

Dataset and Implementation Details. The public datasets, including SCARED [1] with ground truth of both depth and pose and SERV-CT [3] with only depth ground truth, were used to evaluate the proposed method. Following the settings in [12], we trained on SCARED and tested on SCARED and SERV-CT. The depth metrics (*Abs Rel, Sq Rel, RMSE, RMSE log and* δ), and pose metrics (*ATE,* t_{RPE}*,* r_{RPE}) were used to measure the difference of predictions and the ground truth[1]. ATE is noted as the weighted average of the RMSE of sub-trajectories, and the rest metrics are noted as Mean ± Standard Deviation of error.

We implemented networks using PyTorch [11] and trained the networks on 1 NVIDIA RTX 3090 GPU with Adam [7] for 100 epochs with a learning rate of $1e^{-4}$, dropped by a scale factor of 10 after 10 epochs. Given the SCARED dataset, we divided 5/2/2 subsets for training/validation/testing. We finally obtained $\sim$ 8k frames for training. The batch size was 12 and all images were downsampled to 256×320. $\lambda_{pt}, \lambda_t, \lambda_{sub}, \beta, \lambda_d, \lambda_p, N_w$ in the loss function were empirically set to $1, 0.5, 0.2, 0.5, 0.001, 0.5, 5$ which were tuned on validation set. For more details of the experiments, please refer to the supplementary material.

Results. To evaluate the effectiveness of the proposed method, TCL was applied to MonoDepth2 [5] and SC-SfMLearner [2] by exactly using the same architectures of DepthNet and PoseNet. For comparisons, SfMLearner [17], MonoDepth2 [5], AF-SfMLearner [12], SC-SfMLearner [2] and Endo-SfMLearner [10] were adopted as baseline methods. Specifically, AF-SfMLearner improved the MonoDepth2 by predicting the appearance flow, while Endo-SfMLearner is the alternative of SC-SfMLearner with better model architectures.

Table 1 presents the quantitative results on SCARED and SERV-CT. After TCL is applied to the series baselines(MonoDepth2 and SC-SfMLearner), most depth and pose metrics are significantly improved, and most metrics achieve the best performance in their series. Our method not only outperforms the series baseline on SERV-CT, but also achieves all the best values of the MonoDepth2-Based and SC-SfMLearner-Based series. For visualization and further ablations, we present the results of TCL applied to MonoDepth2, which is denoted as Ours below. Figure 2(a) presents the comparisons of the depth prediction. In the first image, our method predicts the depth most accurately. In the second image, the ground truth reveals that the dark area in the upper right corner is closer, and only our method accurately predicts this area, resulting in a detailed and accurate full-image depth map. Despite the improvement, our prediction results still remain inaccurate compared to the ground truths. Figure 2(b) visualizes the trajectory and the projection on three planes. Our predicted trajectory is the closest to the ground truth compared with baselines. The SfMlearner predicts almost all trajectories as straight lines, a phenomenon also observed in [10], in which case a low r_{RPE} metric is ineffective.

[1] The detailed implementations of the metrics can be found in [2,10].

Ablations on Proposed Modules. We introduce two proposed modules to the baseline (MonoDepth2) separately. We additionally propose a simple version of AiC that computes Triplet Masks directly using two reference frames without warping, denoted as AiC*. From Table 2, GC and AiC can significantly improve the effect of the baseline when introduced separately. AiC outperforms AiC* as the warping process can provide greater benefits for pixel-level alignment. Figure 3(a) intuitively demonstrates the effect of the proposed Triplet Mask(AiC), which effectively covers pixels with apparent appearance inconsistency between reference and target frames.

Ablations on Different Dataset Amounts. To verify the effect of our proposed method on different amounts of training and validation sets, we utilize the main depth metric RMSE and pose metric ATE for comparison. In Fig. 3(b), our method achieves significant improvements over MonoDepth2 with different dataset amounts. The performance of our approach is almost optimal for depth and pose at 11k and 8k training samples, respectively. Therefore, our proposed framework has the potential to effectively enhance the performance of various unsupervised SfM methods, even with limited training data.

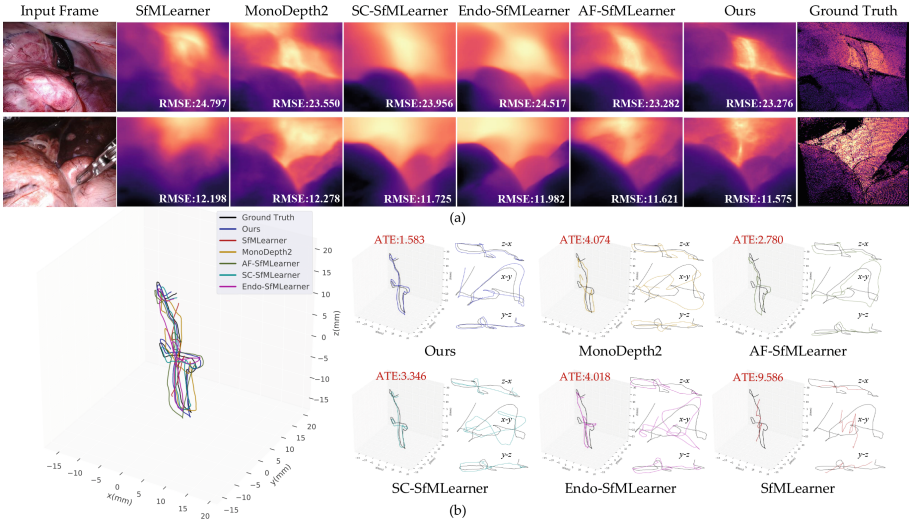

Fig. 2. Qualitative depth and pose results on the SCARED dataset.

Table 2. Ablation results of proposed modules on the SCARED dataset.

Settings	**Module Ablations**									
	Pose Metrics			Depth Metrics						
	ATE(mm)↓	t_{RPE}(mm)↓	r_{RPE}(deg)↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
MonoDepth2	4.848	0.479$_{\pm 0.20}$	0.459$_{\pm 0.22}$	0.454$_{\pm 0.06}$	7.311$_{\pm 2.87}$	13.583$_{\pm 5.02}$	0.487$_{\pm 0.05}$	0.368$_{\pm 0.10}$	0.650$_{\pm 0.11}$	0.818$_{\pm 0.06}$
AF-SfMLearner	3.506	0.161$_{\pm 0.10}$	0.265$_{\pm 0.14}$	**0.446**$_{\pm 0.06}$	**7.153**$_{\pm 3.02}$	13.517$_{\pm 5.16}$	0.481$_{\pm 0.05}$	0.371$_{\pm 0.10}$	0.651$_{\pm 0.11}$	0.825$_{\pm 0.06}$
Ours w/GC Only	3.334	**0.160**$_{\pm 0.11}$	0.266$_{\pm 0.15}$	0.449$_{\pm 0.06}$	7.314$_{\pm 3.25}$	13.519$_{\pm 5.34}$	0.485$_{\pm 0.05}$	0.370$_{\pm 0.10}$	0.651$_{\pm 0.12}$	0.824$_{\pm 0.06}$
Ours w/AiC Only	3.121	0.161$_{\pm 0.10}$	**0.244**$_{\pm 0.14}$	0.447$_{\pm 0.06}$	7.179$_{\pm 2.92}$	13.551$_{\pm 4.91}$	0.482$_{\pm 0.05}$	0.368$_{\pm 0.11}$	0.653$_{\pm 0.11}$	0.824$_{\pm 0.06}$
Ours w/AiC* Only	3.518	0.162$_{\pm 0.10}$	0.269$_{\pm 0.14}$	0.451$_{\pm 0.06}$	7.445$_{\pm 2.93}$	13.623$_{\pm 4.96}$	0.488$_{\pm 0.05}$	0.368$_{\pm 0.10}$	0.649$_{\pm 0.11}$	0.822$_{\pm 0.06}$
Ours	**3.110**	0.161$_{\pm 0.11}$	0.250$_{\pm 0.15}$	**0.446**$_{\pm 0.06}$	7.185$_{\pm 3.30}$	**13.405**$_{\pm 5.20}$	**0.480**$_{\pm 0.05}$	**0.373**$_{\pm 0.11}$	**0.655**$_{\pm 0.12}$	**0.828**$_{\pm 0.06}$

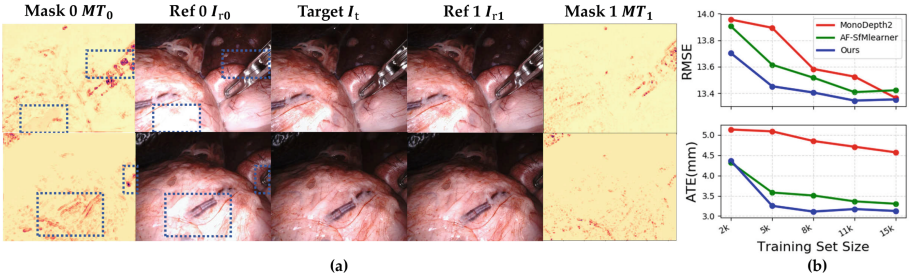

Fig. 3. Qualitative results of Triplet Mask and ablation results on different dataset amounts.

4 Conclusion

We present a triplet-consistency-learning framework (TCL) to improve the effect of monocular endoscopy unsupervised depth and pose estimation. The GC module generates synthesis triplets to increase the diversity of the endoscopic samples. Furthermore, we constrain the depth and pose consistency using two loss functions. The AiC module generates Triplet Mask(MT) based on the triplet information. MT can effectively mask the appearance inconsistency in the triplet, which leads to more efficient training of the photometric loss. Extensive experiments demonstrate the effectiveness of TCL, which can be easily embedded into various SfM methods without additional model parameters.

Acknowledgement. This work is supported in part by the Open Funding of Zhejiang Laboratory under Grant 2021KH0AB03, in part by the Shanghai Sailing Program under Grant 20YF1420800, and inpart by NSFC under Grant 62003208, and in part by Shanghai Municipal of Science and Technology Project, under Grant 20JC1419500 and Grant 20DZ2220400.

References

1. Allan, M., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Bian, J.W., et al.: Unsupervised scale-consistent depth learning from video. Int. J. Comput. Vision **129**(9), 2548–2564 (2021)
3. Edwards, P.E., Psychogyios, D., Speidel, S., Maier-Hein, L., Stoyanov, D.: SERV-CT: a disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction. Med. Image Anal. **76**, 102302 (2022)
4. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the kitti dataset. Int. J. Robot. Res. (IJRR) **32**, 1231–1237 (2013)
5. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828–3838 (2019)
6. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3D common corruptions and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18963–18974 (2022)

7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Liu, X., et al.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Trans. Med. Imaging **39**(5), 1438–1447 (2019)
9. Ma, R., et al.: RNNSLAM: reconstructing the 3D colon to visualize missing regions during a colonoscopy. Med. Image Anal. **72**, 102100 (2021)
10. Ozyoruk, K.B., et al.: EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med. Image Anal. **71**, 102058 (2021)
11. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
12. Shao, S., et al.: Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. Med. Image Anal. **77**, 102338 (2022)
13. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
14. Watson, J., Aodha, O.M., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: self-supervised multi-frame monocular depth. In: Computer Vision and Pattern Recognition (CVPR) (2021)
15. Zhao, W., Liu, S., Shu, Y., Liu, Y.J.: Towards better generalization: joint depth-pose learning without posenet. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9151–9161 (2020)
16. Zhao, Y., Kong, S., Fowlkes, C.: Camera pose matters: improving depth prediction by mitigating pose distribution bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15759–15768 (2021)
17. 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)