Supplementary Material – SAGE: SLAM with Appearance and Geometry Prior for Endoscopy

Xingtong Liu, Zhaoshuo Li, Masaru Ishii, Gregory D. Hager, *Fellow, IEEE,* Russell H. Taylor, *Life Fellow, IEEE,* and Mathias Unberath

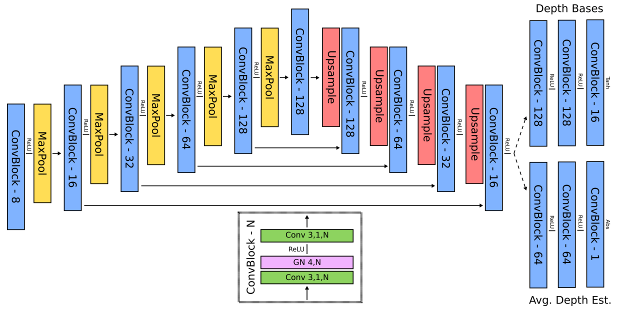

Fig. 1: **Network architecture for optimizable depth estimation.** Each ConvBlock consists of two partial convolution layers with kernel size as 3 and stride as 1, one group normalization layer with a group size of 4, and one ReLU activation, which are arranged in the way as the figure above. The number after the ConvBlock means the size of the output channel dimension. Two output branches exist in the network for the average depth estimate and the depth bases. Hyperbolic tangent and absolute functions are used as output activation in these branches.

I. Network Architecture

The network architecture of the depth network is shown in Fig. 1 and that of the discriminator used for depth training is shown in Fig. 2.

II. Soft Cumulative Density Function (CDF) Computation

The value of k^{th} bin in the histogram $\boldsymbol{h}_i^{\text{src}}$ can be written as follows

$$\boldsymbol{h}_i^{\text{src}}(k) = \frac{1}{|\Omega^{\text{src}}|} \sum_{\boldsymbol{x} \in \Omega^{\text{src}}} \left(\sigma \left(\frac{\boldsymbol{I}_i^{\text{src}}(\boldsymbol{x}) - \mu_k + 1/K}{\beta} \right) - \sigma \left(\frac{\boldsymbol{I}_i^{\text{src}}(\boldsymbol{x}) - \mu_k - 1/K}{\beta} \right) \right), \tag{1}$$

where the center value of k^{th} bin is $\mu_k = -1 + (2k+1)/K \in \mathbb{R}$; the kernel function is $\sigma(a) = 1/(1 + \mathrm{e}^{-a})$. The values used in μ_k are because the descriptor map has a value range of $(-1, 1)$. Ω^{src} consists of all 2D locations within the source video mask; $\beta \in \mathbb{R}$ is a bandwidth parameter. All other notations are defined in the main paper. The histograms for target and source images are the same as above except the corresponding descriptor maps are used for calculation instead of the source one.

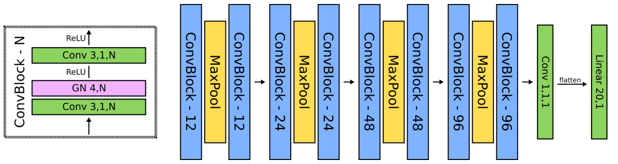

Fig. 2: **Network architecture of discriminator for depth estimation learning.** The input is the RGB image and the normalized depth map, concatenated along the channel dimension, with a resolution of 64 × 80. Each ConvBlock consists of two normal convolution layers with kernel size as 3 and stride as 1, one group normalization layer with a group size of 4, and two ReLU activation layers, which are arranged in the way as the figure above. The number after the ConvBlock means the size of the output channel dimension. The final convolution layer, with kernel size as 1, stride as 1, and output channel size as 1, and linear layer, with input channel size as 20 and output channel size as 1, converts the feature map to a scalar value used to indicate the predicted validity of the input sample. Note that before being fed to the final linear layer, the output map from the final convolution layer is first flattened along the sample-wise dimensions.

III. Verification for Loop Detection

For the local loop detection, because the temporal window is set to be small, the trajectory drifting error will not be large. The camera pose of each keyframe can thus still be roughly relied on for filtering candidates. For this reason, the spatial distance between keyframe pairs is first used. For the verification, the query keyframe and the closest one within its temporal connections are used as the reference pair. If the spatial distance between the candidate pair is smaller than that between the reference pair multiplying a constant factor, the pair will be kept. For pairs being kept after distance filtering, the appearance verification will be run, where the feature match inlier ratio is computed. The candidate pair will be kept if the inlier ratio is larger than that of the reference pair multiplying a constant factor and a specified constant inlier ratio. Lastly, a geometric verification is applied, where a pair-wise optimization similar to *Camera Tracking* is run. The difference in terms of factors is that SMG is used in place of RP. The local connection will only be accepted

Xingtong Liu, Zhaoshuo Li, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath are with the Computer Science Department, Johns Hopkins University, Baltimore, MD 21287 USA (e-mail: xliu89@jh.edu; zli122@jhu.edu; hager@cs.jhu.edu; rht@jhu.edu; unberath@jhu.edu).

Masaru Ishii is with Johns Hopkins Medical Institutions, Baltimore, MD 21224 USA (e-mail: mishii3@jhmi.edu).

if the overlap ratio and flow magnitude, computed in the geometric verification, are larger and smaller than those of the reference pair multiplying a constant factor, respectively. After verification, only the best candidate, in terms of overlap ratio and flow magnitude, will be used to build the local connection.

Global loop detection searches for keyframe pairs whose interval is beyond a specified temporal range and uses the appearance for the initial candidate selection Whenever a keyframe is created, the bag-of-words descriptor will be added to a database. When a global loop connection is searched for a query keyframe, the database will be searched through with the extracted bag-of-words descriptor. A specified number of keyframes that are similar to the query keyframe in terms of bag-of-words descriptor will be selected as candidates. The candidates are then filtered so that the similarities between the query keyframe and candidates are larger than the one between the reference pair multiplying a specified constant factor. Candidates should not be temporally close to the query keyframe, opposite to the local loop connection. After that, the same verification as the local loop detection is used to verify the global loop candidates. The verified candidates are ranked based on feature match inlier ratio and, from high to low, each candidate that is temporally far enough from the selected candidates is added to avoid connection redundancy.

IV. Experiment Setup

The endoscopic videos used in the experiments were acquired from seven consenting patients and four cadavers under an IRB-approved protocol. The anatomy captured in the videos is the nasal cavity. The total time duration of videos is around 40 minutes. The input images to both networks are 8-time spatially downsampled, resulting in a resolution of 128×160; the output maps of both networks have a resolution of 64×80. Note that the binary masks with the same resolution are also fed, together with images, into the networks to exclude contributions of invalid pixels. SGD optimizer with cyclic learning rate scheduler [**?**] is used for network training, where the learning rate range is $[1.0\mathrm{e}^{-4}, 5.0\mathrm{e}^{-4}]$. Full-range rotation augmentation is used for input images to the networks during training. The first stage of training lasts for 40 epochs and the second stage lasts until the loss curves plateau, where each epoch consists of 300 iterations with the batch size of 1. Image pairs are selected so that the groundtruth ratio of scene overlap is larger than 0.6; the initialized relative pose is randomized so that the initial ratio of scene overlap is larger than 0.4. The weights for scale-invariant loss, RR loss, flow loss, histogram loss, generator adversarial loss, and discriminator adversarial loss are 20.0, 4.0, 10.0, 4.0, 1.0, and 1.0. In terms of the hyperparameters related to loss design, ϵ is $1.0\mathrm{e}^{-4}$; η_{hist} is 0.3; β is $\dfrac{4}{5K}$; K is 100; C is 16; H is 64; W is 80;

For SLAM system running, in cases where post-operative processing in a SLAM system is allowed, the *Mapping* and *Loop Closure* modules can be run for an additional amount of time after all frames have been tracked. The *Mapping* module will continue refining the full factor graph. The maximum number of iterations and consecutive no-relinearization iterations are 20 and 5, respectively. In the meantime, the *Loop Closure* module will search for loop pairs for the query keyframes that have not been processed before. When the *Mapping* module finishes, the entire system run will end. For the other settings in terms of the SLAM system running, please refer to the supplementary material.

In terms of the hyperparameters of the differentiable LM optimization, damp value range is $[1.0\mathrm{e}^{-6}, 1.0\mathrm{e}^{-2}]$, with $1.0\mathrm{e}^{-4}$ as the initial value. The increasing and decreasing multiplier of the damp value is 11.0 and 9.0, respectively. LM optimization terminates when one of the three below is met: 1) number of iterations reaching 40, 2) maximum gradient smaller than $1.0\mathrm{e}^{-4}$, 3) maximum parameter increment ratio smaller than $1.0\mathrm{e}^{-2}$. Factors involved have the same parameter setting as the SLAM system, which will be described below.

Below are the hyperparameters of the SLAM system. For the *Camera Tracking* module, the multiplying factor used for the reference keyframe selection is 0.6; the maximum number of iterations in the optimization is 40; the damp value range is $[1.0\mathrm{e}^{-6}, 1.0\mathrm{e}^{-2}]$, with $1.0\mathrm{e}^{-4}$ as the initial value; the increasing and decreasing multipliers are 100.0 and 10.0, respectively; the jacobian matrix recompute condition is when the error update between steps is larger than $1.0\mathrm{e}^{-2}$ of the current error. As for factors in the *Camera Tracking* module, settings are as follows. In FM, all samples within the video mask are used for computation; the weights for all 4 pyramid levels (from high resolution to low one) are 10.0, 9.0, 8.0, and 7.0. In RP, the factor weight and σ_{rp} are 0.1 and 0.03, respectively. In SMG, the factor weight and σ_{smg} are 0.1 and 0.1, respectively; the number of feature match candidates before filtering is 256; in terms of the Teaser++ filtering, the maximum clique time limit, rotation maximum iterations, rotation graph, inlier selection mode, and noise bound multiplier are 50ms, 20, chain mode, no inlier selection, and 2.0, respectively; Other parameters of Teaser++ are set to the default ones.

For the *Keyframe Creation* module, settings are as below. The maximum ratios of scene overlap in terms of the area and the number of point inliers within the video mask for a new keyframe are 0.8 and 0.9, respectively; the maximum feature match inlier ratio is 0.4; the minimum average magnitude of 2D flow is 0.08 of the image width. For the temporal connection building in the *Keyframe Creation* module, the maximum number of temporal connections per keyframe is 3; the minimum feature match inlier ratio to connect a previous keyframe is 0.7.

For the *Loop Closure* module, settings are shown as follows. For the local loop detection, the temporal window for searching is 9; the rotation and translation weights to compute pose distance for candidate filtering are both set to 1.0; the spatial distance multiplier for candidate filtering is 5.0; the metric multiplier for verification is 0.7; the minimum constant inlier ratio for verification is 0.2, which is the same in global loop detection; the minimum ratios of scene overlap for verification in terms of the area and the number of point inliers within the video mask are 0.5 and 0.5, respectively.

Regarding the global loop detection, only keyframes that are at least 10 keyframes away are considered; the multiplier of description similarity for verification is 0.7; the metric multiplier for verification is 0.7; a global loop candidate will be selected if it is at least 10 keyframes away from the ones already selected in a single global loop closure process. In the pose-scale graph optimization for loop closure, the weights of RPS for non-global and global connections are 1.0 and 5.0, respectively; within this factor, the weights of rotation and scale component, which are ω_{rot} and ω_{scl}, are 5.0 and 0.5, respectively; the weight of SC within the loop closure optimization is 10.0; the number of maximum iterations of such optimization is 200; the number of maximum iterations with no relinearization is 5; the relinearization thresholds for pose and scale are $3.0\mathrm{e}^{-3}$ and $1.0\mathrm{e}^{-2}$.

For the *Mapping* module, settings are as follows. In terms of hyperparameters of factors used in the full factor graph, the weights for PS and SC of the first keyframe are $1.0\mathrm{e}^{4}$, which are used to anchor the graph in terms of camera pose and depth scale; The FM and GC use all samples within the video mask for computation; FM has the same weight as the one in camera tracking; GC has the factor weight of 0.1 and σ_{gc} as 0.03; the weight of CD is $1.0\mathrm{e}^{-4}$. In terms of the hyperparameters in factor graph optimization algorithm ISAM2 [**?**], the relinearization thresholds for camera poses, depth scales, and depth codes are $1.0\mathrm{e}^{-3}$, $1.0\mathrm{e}^{-3}$, and $1.0\mathrm{e}^{-2}$, respectively; partial relinearization check and relinearization skipping are not used; Other parameters in ISAM2 are set to the default ones.

In cases where post-operative processing in a SLAM system is allowed, the *Mapping* and *Loop Closure* modules can be run for an additional amount of time after all frames have been tracked. The *Mapping* module will continue refining the full factor graph. The maximum number of iterations and consecutive no-relinearization iterations are 20 and 5, respectively. In the meantime, the *Loop Closure* module will search for loop pairs for the query keyframes that have not been processed before. When the *Mapping* module finishes, the entire system run will end.

V. Evaluation Metrics

The metrics used for camera trajectory evaluation are Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) [**?**]. Note that only the frames that are treated as keyframes by the SLAM system will be evaluated in terms of both trajectory error and depth error. Therefore, synchronization needs to be done to first associate the trajectory estimate with the groundtruth one. The trajectory estimate will then be spatially aligned with the groundtruth trajectory, where a similarity transform is estimated with the method in [**?**].

ATE is used to quantify the whole trajectory and the Root Mean Square Error (RMSE) is used. The rotation and translation components of this metric are defined as

$$\begin{aligned}\mathrm{ATE}_{\mathrm{rot}} &= \left(\frac{1}{N}\sum_{i=0}^{N-1}\left\|\log\left(\boldsymbol{R}_i^{\mathrm{ATE}}\right)\right\|_2^2\right)^{1/2} \quad \text{and} \\ \mathrm{ATE}_{\mathrm{trans}} &= \left(\frac{1}{N}\sum_{i=0}^{N-1}\left\|\boldsymbol{t}_i^{\mathrm{ATE}}\right\|_2^2\right)^{1/2} \quad ,\end{aligned} \tag{2}$$

where $\boldsymbol{R}_i^{\mathrm{ATE}} = \tilde{\boldsymbol{R}}_i^{\mathrm{wld}}\left(\boldsymbol{R}_i^{\mathrm{wld}}\right)^{\mathsf{T}}$ and $\boldsymbol{t}_i^{\mathrm{ATE}} = \tilde{\boldsymbol{t}}_i^{\mathrm{wld}} - \boldsymbol{R}_i\boldsymbol{t}_i^{\mathrm{wld}}$. $\tilde{\boldsymbol{R}}_i^{\mathrm{wld}} \in \mathrm{SO}\left(3\right)$ and $\tilde{\boldsymbol{t}}_i^{\mathrm{wld}} \in \mathbb{R}^3$ are the groundtruth rotation and translation components of the i^{th} pose in the trajectory, respectively, while $\boldsymbol{R}_i^{\mathrm{wld}} \in \mathrm{SO}\left(3\right)$ and $\boldsymbol{t}_i^{\mathrm{wld}} \in \mathbb{R}^3$ are the estimated ones. $N \in \mathbb{R}$ is the number of poses in the synchronized and aligned trajectory estimate.

RPE measures the local accuracy of the trajectory over a fixed frame interval $\Delta \in \mathbb{R}$. This measures the local drift of the trajectory, which is less affected by the loop closure and emphasizes more on the other components of the system. The rotation and translation components of this metric are defined as

$$\begin{aligned}\mathrm{RPE}_{\mathrm{rot}} &= \left(\frac{1}{N-\Delta}\sum_{i=0}^{N-\Delta-1}\left\|\log\left(\boldsymbol{R}_i^{\mathrm{RPE}}\right)\right\|_2^2\right)^{1/2} \quad \text{and} \\ \mathrm{RPE}_{\mathrm{trans}} &= \left(\frac{1}{N-\Delta}\sum_{i=0}^{N-\Delta-1}\left\|\boldsymbol{t}_i^{\mathrm{RPE}}\right\|_2^2\right)^{1/2} \quad ,\end{aligned} \tag{3}$$

$\boldsymbol{R}_i^{\mathrm{RPE}} \in \mathrm{SO}\left(3\right)$ and $\boldsymbol{t}_i^{\mathrm{RPE}} \in \mathbb{R}^3$ are the rotation and translation components of $\boldsymbol{T}_i^{\mathrm{RPE}} \in \mathrm{SE}\left(3\right)$, respectively; $\boldsymbol{T}_i^{\mathrm{RPE}}$ is the i^{th} RPE matrix, which is defined as

$$\boldsymbol{T}_i^{\mathrm{RPE}} = \left((\tilde{\boldsymbol{T}}_i^{\mathrm{wld}})^{-1}\tilde{\boldsymbol{T}}_{i+\Delta}^{\mathrm{wld}}\right)^{-1}\left((\boldsymbol{T}_i^{\mathrm{wld}})^{-1}\boldsymbol{T}_{i+\Delta}^{\mathrm{wld}}\right) \quad . \tag{4}$$

Δ in Eq. 3 is set to 7 for our results; for ORB-SLAM3, Δ is set so that the number of original video frames between $\boldsymbol{T}_i^{\mathrm{wld}}$ and $\boldsymbol{T}_{i+\Delta}^{\mathrm{wld}}$ is roughly the same as ours.

To evaluate depth estimates, Absolute Relative Difference (ARD) and Threshold [**?**] are used. Before computing metrics, different pre-processing is applied for two sets of metrics, which are $\mathrm{ARD}_{\mathrm{traj}}$ and $\mathrm{Threshold}_{\mathrm{traj}}$, and $\mathrm{ARD}_{\mathrm{frame}}$ and $\mathrm{Threshold}_{\mathrm{frame}}$. For the former, the estimated depth per keyframe is scaled with the scale component in the similarity transform obtained from the trajectory alignment above. For the latter, each depth estimate is scaled with the median value of ratios between the corresponding groundtruth one and the estimate. In terms of the definitions of these metrics, ARD is

$$\mathrm{ARD} = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{|\Omega_i|}\sum_{\boldsymbol{x}\in\Omega}\frac{|\boldsymbol{D}_i\left(\boldsymbol{x}\right) - \tilde{\boldsymbol{D}}_i\left(\boldsymbol{x}\right)|}{\tilde{\boldsymbol{D}}_i\left(\boldsymbol{x}\right)} \quad ; \tag{5}$$

Threshold is

$$\begin{aligned}\mathrm{Threshold} = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{|\Omega_i|}&\sum_{\boldsymbol{x}\in\Omega} \\ &\mathbb{1}\left[\max\left(\frac{\boldsymbol{D}_i\left(\boldsymbol{x}\right)}{\tilde{\boldsymbol{D}}_i\left(\boldsymbol{x}\right)}, \frac{\tilde{\boldsymbol{D}}_i\left(\boldsymbol{x}\right)}{\boldsymbol{D}_i\left(\boldsymbol{x}\right)}\right) < \theta\right] \quad .\end{aligned} \tag{6}$$

Note that Ω_i here is the region where both scaled depth estimate $\boldsymbol{D}_i \in \mathbb{R}^{1\times H\times W}$ and groundtruth depth $\tilde{\boldsymbol{D}}_i \in \mathbb{R}^{1\times H\times W}$, for the i^{th} synchronized keyframe, have valid depths; $\theta \in \mathbb{R}$ is the threshold used to determine if the depth ratio between the estimate and groundtruth is small enough.

VI. Cross-Subject Evaluation Additional Results

To evaluate the performance of the SLAM system on endoscopic videos from unseen subjects, we run a cross-validation study. Four models are trained with different train/test splits

TABLE I: **Cross-subject evaluation on SLAM systems.** Note that $\sim$ is used as the name abbreviation of the comparison method ORB-SLAM3.

Subjects	$\{1,2,3\}$		$\{4,5,6\}$		$\{7,8,11\}$		$\{8,9,10\}$	
Methods / Metrics	Ours	ORB-SLAM3 [?]	Ours	$\sim$	Ours	$\sim$	Ours	$\sim$
$\mathrm{ATE}_{\mathrm{trans}}(\mathrm{mm})$	$\mathbf{1.4 \pm 1.0}$	3.8 ± 2.7	$\mathbf{1.3 \pm 1.7}$	3.8 ± 4.6	$\mathbf{2.2 \pm 1.2}$	6.3 ± 4.8	$\mathbf{1.6 \pm 1.0}$	5.5 ± 3.0
$\mathrm{ATE}_{\mathrm{rot}}(°)$	$\mathbf{19.7 \pm 7.8}$	66.2 ± 59.5	$\mathbf{22.8 \pm 17.2}$	61.1 ± 68.1	$\mathbf{25.3 \pm 18.4}$	66.9 ± 48.9	$\mathbf{19.4 \pm 9.5}$	55.8 ± 22.4
$\mathrm{RPE}_{\mathrm{trans}}(\mathrm{mm})$	$\mathbf{1.3 \pm 0.4}$	2.5 ± 1.4	$\mathbf{1.4 \pm 0.7}$	2.7 ± 2.1	$\mathbf{1.9 \pm 0.6}$	4.8 ± 3.5	$\mathbf{1.2 \pm 0.5}$	3.6 ± 1.6
$\mathrm{RPE}_{\mathrm{rot}}(°)$	$\mathbf{5.9 \pm 1.7}$	6.4 ± 3.5	4.3 ± 2.0	$\mathbf{3.8 \pm 2.6}$	$\mathbf{7.4 \pm 2.6}$	7.7 ± 3.9	$\mathbf{4.5 \pm 1.1}$	8.5 ± 2.9
$\mathrm{ARD}_{\mathrm{traj}}$	$\mathbf{0.39 \pm 0.17}$	1.73 ± 1.02	$\mathbf{0.34 \pm 0.10}$	2.00 ± 1.82	$\mathbf{0.38 \pm 0.14}$	1.58 ± 1.42	$\mathbf{0.29 \pm 0.09}$	1.56 ± 1.20
$\mathrm{ARD}_{\mathrm{frame}}$	$\mathbf{0.17 \pm 0.04}$	1.73 ± 1.02	$\mathbf{0.17 \pm 0.04}$	2.00 ± 1.82	$\mathbf{0.18 \pm 0.03}$	1.58 ± 1.42	$\mathbf{0.15 \pm 0.02}$	1.56 ± 1.20
$\mathrm{Threshold}_{\mathrm{traj}}$ $(\theta = 1.25)$	$\mathbf{0.39 \pm 0.19}$	0.15 ± 0.13	$\mathbf{0.46 \pm 0.14}$	0.24 ± 0.21	$\mathbf{0.38 \pm 0.15}$	0.14 ± 0.14	$\mathbf{0.49 \pm 0.13}$	0.14 ± 0.15
$\mathrm{Threshold}_{\mathrm{frame}}$ $(\theta = 1.25)$	$\mathbf{0.39 \pm 0.19}$	0.15 ± 0.13	$\mathbf{0.46 \pm 0.14}$	0.24 ± 0.21	$\mathbf{0.38 \pm 0.15}$	0.14 ± 0.14	$\mathbf{0.49 \pm 0.13}$	0.14 ± 0.15
$\mathrm{Threshold}_{\mathrm{traj}}$ $(\theta = 1.25^2)$	$\mathbf{0.70 \pm 0.22}$	0.28 ± 0.22	$\mathbf{0.81 \pm 0.13}$	0.38 ± 0.29	$\mathbf{0.66 \pm 0.16}$	0.27 ± 0.23	$\mathbf{0.84 \pm 0.10}$	0.27 ± 0.22
$\mathrm{Threshold}_{\mathrm{frame}}$ $(\theta = 1.25^2)$	$\mathbf{0.70 \pm 0.22}$	0.28 ± 0.22	$\mathbf{0.81 \pm 0.13}$	0.38 ± 0.29	$\mathbf{0.66 \pm 0.16}$	0.27 ± 0.23	$\mathbf{0.84 \pm 0.10}$	0.27 ± 0.22

TABLE II: **Ablation study for the proposed SLAM system on trajectory-related metrics.**

FMT	FMM	RPT	Local	Global	$\mathrm{ATE}_{\mathrm{trans}}(\mathrm{mm})$	$\mathrm{ATE}_{\mathrm{rot}}(°)$	$\mathrm{RPE}_{\mathrm{trans}}(\mathrm{mm})$	$\mathrm{RPE}_{\mathrm{rot}}(°)$
✓	✓	✓	✓	✓	$\mathbf{1.6 \pm 1.4}$	$\mathbf{22.2 \pm 15.1}$	$\mathbf{1.5 \pm 0.6}$	5.5 ± 2.4
	✓	✓	✓	✓	$3.4 \pm 2.7^{***}$	$43.3 \pm 27.9^{***}$	$2.6 \pm 1.4^{***}$	$7.3 \pm 3.0^{***}$
		✓	✓	✓	$3.3 \pm 2.8^{***}$	$40.2 \pm 23.6^{***}$	$2.6 \pm 1.2^{***}$	$7.0 \pm 2.6^{***}$
✓	✓		✓	✓	2.7 ± 5.5	23.8 ± 14.5	2.1 ± 3.2	$\mathbf{5.3 \pm 2.1}$
✓	✓	✓	✓		$2.0 \pm 1.9^{*}$	$26.8 \pm 21.2^{*}$	1.5 ± 0.7	5.5 ± 2.4
✓	✓	✓			$2.0 \pm 1.9^{*}$	$25.5 \pm 18.5^{*}$	1.5 ± 0.7	5.4 ± 2.4

TABLE III: **Ablation study for the proposed SLAM system on depth-related metrics.**

FMT	FMM	RPT	Local	Global	$\mathrm{ARD}_{\mathrm{traj}}$	$\mathrm{ARD}_{\mathrm{frame}}$	$\mathrm{Threshold}_{\mathrm{traj}}$ $(\theta = 1.25)$	$\mathrm{Threshold}_{\mathrm{frame}}$ $(\theta = 1.25)$	$\mathrm{Threshold}_{\mathrm{traj}}$ $(\theta = 1.25^2)$	$\mathrm{Threshold}_{\mathrm{frame}}$ $(\theta = 1.25^2)$
✓	✓	✓	✓	✓	0.36 ± 0.16	$\mathbf{0.17 \pm 0.03}$	0.42 ± 0.17	0.73 ± 0.08	0.74 ± 0.21	$\mathbf{0.95 \pm 0.04}$
	✓	✓	✓	✓	$0.49 \pm 0.19^{***}$	0.17 ± 0.03	$0.29 \pm 0.16^{***}$	0.73 ± 0.08	$0.59 \pm 0.23^{***}$	0.95 ± 0.04
		✓	✓	✓	$0.50 \pm 0.25^{**}$	0.17 ± 0.03	$0.32 \pm 0.17^{**}$	$\mathbf{0.74 \pm 0.08}$	$0.61 \pm 0.24^{**}$	0.95 ± 0.04
✓	✓		✓	✓	$\mathbf{0.35 \pm 0.15}$	0.17 ± 0.03	$\mathbf{0.43 \pm 0.17}$	0.73 ± 0.08	$\mathbf{0.76 \pm 0.18}$	0.95 ± 0.04
✓	✓	✓	✓		0.36 ± 0.16	0.17 ± 0.03	0.42 ± 0.17	0.73 ± 0.08	0.74 ± 0.21	0.95 ± 0.04
✓	✓	✓			$0.35 \pm 0.16^{*}$	0.17 ± 0.03	0.42 ± 0.18	0.73 ± 0.08	0.74 ± 0.22	0.95 ± 0.04

on the 11 subjects in total. With subjects named as consecutive numbers, the test splits for 4 models are $\{1,2,3\}$, $\{4,5,6\}$, $\{7,8,11\}$, and $\{8,9,10\}$, and the train splits for each model are the subjects left. The evaluation metrics, as reported in Table I, are averaged over all the sequences within the corresponding test split for each trained model. Besides, we also compare against a state-of-the-art feature-based SLAM system, ORB-SLAM3 [**?**], which we evaluate on all videos at once and use the same set of metrics for evaluation.

VII. Ablation Study Results

Table II and III show the ablation study results on trajectory-related and depth-related metrics, respectively. As can be seen, FM has a large impact on both trajectory and trajectory-scaled depth metrics; RP mainly affects trajectory metrics; the *Loop Closure* module mainly affects the trajectory metrics $\mathrm{ATE}_{\mathrm{trans}}$ and $\mathrm{ATE}_{\mathrm{rot}}$.