Temporally Consistent Stereo Matching Supplementary Materials

Jiaxi Zeng[1,2], Chengtang Yao[1,3], Yuwei Wu[1,2⋆], and Yunde Jia[2,1]

[1] Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2] Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
[3] Horizon Robotics
{jiaxi,yao.c.t,wuyuwei,jiayunde}@bit.edu.cn

1 Architecture of Temporal Disparity Completion Module

Our temporal disparity completion module is depicted in Fig. 1. It takes the semi-dense disparity projected from the previous timestamp, a binary mask indicating the sparsity in the disparity map, and the multi-level context features from the feature extractor as the inputs.

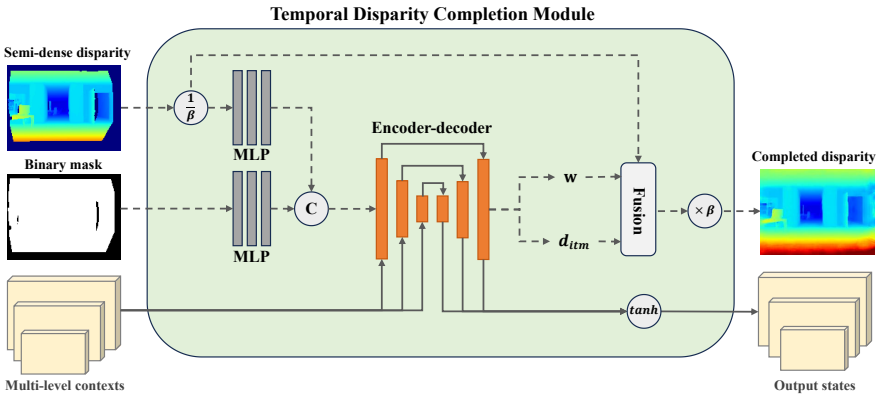

Fig. 1: Architecture of the temporal disparity completion module.

The semi-dense disparity map undergoes a preprocessing step where the disparity values are scaled down by a factor of $1/\beta$ before being encoded by a lightweight MLP. This is done to compress the range of disparities into a smaller interval, aiming to obtain a globally well-initialized disparity map. We set β to 10 in practice. Afterward, features of disparity and binary mask are concatenated

⋆ Corresponding author

and passed into an encoder-decoder network, along with multi-level context features. This network outputs a weight map and an intermediate disparity map d_{itm}. The weight $w \in (0, 1)$ is used to fuse the intermediate disparity d_{itm} with the semi-dense disparity d_{semi}, which is formulated as:

$$d_{fused} = w \times d_{itm} + (1 - w) \times d_{semi}.$$

The fused disparity map is scaled up by the factor of β to produce the completed disparity map. Additionally, the multi-level features from the decoder network are passed through a *tanh* function to provide state features for further temporal state fusion.

2 Training and Testing Split of Ablation Study

The TartanAir [3] dataset is categorized into Easy mode and Hard mode based on motion patterns, with the former having fixed pitch and roll angles, whereas the latter encompasses 6 DoF motion, along with higher translation and rotation speeds. Unlike the strategy of TemporalStereo [4] training exclusively on data with hard motion patterns, our ablation study involves both Easy and Hard data for training and testing. This provides a comprehensive evaluation of our method with different motion patterns. Specifically, we sampled 17 videos from the entire dataset, encompassing a variety of scenes, for evaluation, as shown in Table 1. The remaining data was utilized for training.

Video ID	Scene	Motion Pattern	Part	Video ID	Scene	Motion Pattern	Part
1	abandonedfactory	Easy	P002	10	hospital	Hard	P042
2	abandonedfactory	Hard	P002	11	office	Easy	P006
3	amusement	Easy	P007	12	office	Hard	P006
4	amusement	Hard	P007	13	office2	Easy	P004
5	carwelding	Hard	P003	14	office2	Hard	P004
6	endofworld	Easy	P006	15	oldtown	Hard	P006
7	endofworld	Hard	P006	16	soulcity	Easy	P008
8	gascola	Easy	P001	17	soulcity	Hard	P008
9	gascola	Hard	P001				

Table 1: Test set of ablation study on TartanAir dataset.

3 Additional Temporal Metrics of Ablation Study

The additional 3-px error rate of temporal metrics for the ablation study is shown in Table 2. Setting (H) shows a significant reduction in the 3-px error rates compared to Setting (A). Both $|\Delta d|_{>3px}$ and $Relu(\Delta e)_{>3px}$ decreased by over 40%. Additionally, compared to (B), despite setting (C) showing the lower overall accuracy, the $|\Delta d|_{>3px}$ and $Relu(\Delta e)_{>3px}$ metrics are significantly better. This highlights the importance of temporal information for maintaining temporal consistency.

Setting	Seq length	N	Past disp & state	State fusion	TDC module	Dual-space refinement	ALL $\lvert\Delta d\rvert_{>3px}$ (%)	ALL $Relu(\Delta e)_{>3px}$ (%)	OCC $\lvert\Delta d\rvert_{>3px}$ (%)	OCC $Relu(\Delta e)_{>3px}$ (%)
(A)	2	5					1.78	0.81	7.53	3.06
(B)	2	32					1.48	0.71	6.08	2.60
(C)	2	5	✓				1.34	0.65	5.90	2.07
(D)	2	5	✓	✓			1.29	0.61	5.54	1.96
(E)	2	5	✓	✓	✓		1.05	0.45	4.84	1.77
(F)	2	5	✓	✓		✓	1.03	0.47	4.77	1.80
(G)	2	5	✓	✓	✓	✓	1.00	0.45	4.58	1.74
(H)	4	5	✓	✓	✓	✓	**0.99**	**0.45**	**4.46**	**1.69**
(I)	4	5		✓	✓	✓	1.40	0.67	5.95	2.54

Table 2: 3-px error rate of temporal metrics for ablation study on TartanAir [3].

4 Visualizations on Temporal Consistency

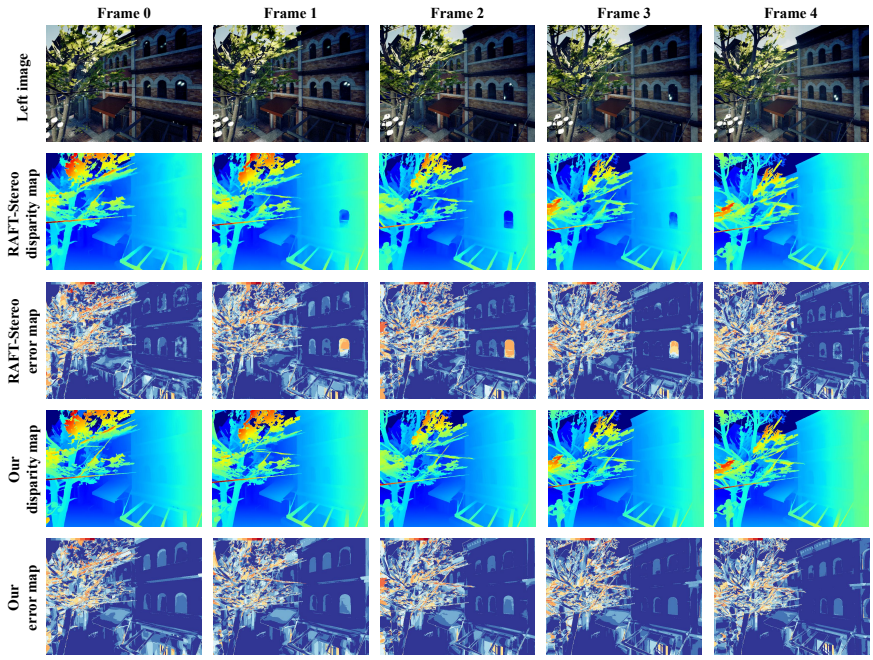

Fig. 2: Visualizations of temporal consistency on the TartanAir dataset. For the error map, blue regions indicate small errors, while red regions indicate large errors.

Fig. 2 illustrates the temporally inconsistent cases of RAFT-Stereo caused by reflective surfaces. As the camera moves, the reflection spots on the window also move, disrupting the matching of corresponding points and leading to temporal inconsistency in the results predicted by RAFT-Stereo [2]. In contrast, our method effectively maintains temporal consistency. This is primarily due to the

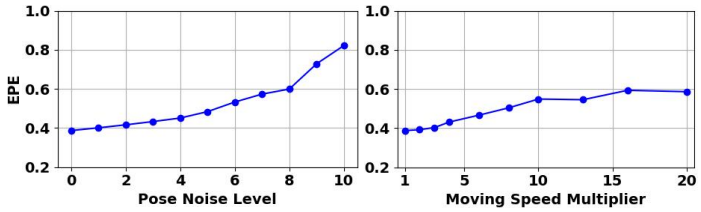

Fig. 3: EPE change with pose noise level (left), and moving speed multiplier (right) on Tartanair.

utilization of temporal information and the dual-space refinement module, which constrains the ill-posed regions.

5 Robustness on Pose and Large Camera Motion

Similar to XR-stereo [1], we evaluate the robustness of our method by examining its performance under varying pose noise levels and moving speed. Fig. 3 illustrates the change in End-Point Error (EPE) with varying pose noise levels (left) and moving speed multipliers (right) on the TartanAir dataset. Each noise level randomly perturbs the rotation by 0.3° and the position by 3% of the baseline length. It can be observed that to a certain extent, our method is robust to pose noise and large motions. However, for larger pose noise, the accuracy of the method decreases rapidly as the noise level increases. This decline is likely due to incorrect initialization disparity caused by erroneous poses, which in turn misguides the model. As for large motion, due to the limited overlap between two frames, the model gradually degrades into regressing disparity from scratch like RAFT-Stereo [2].

References

1. Cheng, Z., Yang, J., Li, H.: Stereo matching in time: 100+ fps video stereo matching for extended reality. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8719–8728. IEEE (2024)
2. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021)
3. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
4. Zhang, Y., Poggi, M., Mattoccia, S.: Temporalstereo: Efficient spatial-temporal stereo matching network. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9528–9535. IEEE (2023)