SVDC: Consistent Direct Time-of-Flight Video Depth Completion with Frequency Selective Fusion

Supplementary Material

This supplementary material provides additional information to complement the main paper. It contains the following sections:

- More experimental results in Sec. A.
- More implementation details in Sec. B.
- Network architecture details in Sec. C.
- More qualitative results in Sec. D.

A. More Experimental Results

In this section, we present additional experimental results.

A.1. Ablation Study on Kernel Sizes

We conducted an ablation study on the kernel size within the Adaptive Frequency Selective Fusion (AFSF) module. The detailed results are shown in Tab. 1. Considering both accuracy and temporal consistency, we ultimately selected the combination of 1×1 and 3×3 convolutional kernels as our experimental configuration.

Kernel Sizes	TartanAir[8]			Dynamic Replica[4]		
	RMSE(m)	REL	OPW	RMSE(m)	REL	OPW
$1\times1 + 5\times5$	0.173	0.025	0.163	**0.082**	**0.020**	0.175
$3\times3 + 5\times5$	**0.164**	**0.024**	0.172	0.084	0.021	0.201
$1\times1 + 3\times3$	**0.164**	**0.024**	**0.159**	0.086	**0.020**	**0.171**

Table 1. Comparison of different kernel sizes on TartanAir and Dynamic Replica datasets.

A.2. Computational Cost of Methods

We evaluated the parameter count and computational cost of different completion methods, as detailed in Tab. 2. It can be observed that our proposed baseline model for multi-frame fusion, DVDC, achieves the smallest parameter count and FLOPs. Building on this baseline, the SVDC model, which incorporates CSEA and AFSF, increases the parameter count by only 0.1M and the FLOPs by 3.4 GFLOPs, demonstrating the lightweight characteristics of our proposed design.

A.3. More Quantitative Comparisons

In the accuracy comparison between our method and the SOTA methods, only RMSE and REL are used. Additional results on the TartanAir and Dynamic Replica datasets are shown in Tab. 3 and Tab. 4.

	CFormer	BPNet	DVDC	SVDC
FLOPs (G)	184.1	247.9	48.2	51.6
Params (M)	82.5	89.9	22.7	22.8

Table 2. Comparison of computational cost and the parameters.

Methods	TartanAir				
	RMSE↓ (m)	REL↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
BPNet	0.337	0.051	0.965	0.976	0.983
CFormer	0.352	0.052	0.963	0.975	0.982
DVDC	0.183	0.030	0.994	0.998	**0.999**
SVDC	**0.164**	**0.024**	**0.995**	**0.999**	**0.999**

Table 3. Quantitative results on the TartanAir dataset.

Methods	Dynamic Replica				
	RMSE↓ (m)	REL↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
BPNet	0.126	0.031	0.987	0.993	0.995
CFormer	0.127	0.030	0.986	0.993	0.995
DVDC	0.095	0.026	0.993	0.997	**0.998**
SVDC	**0.086**	**0.020**	**0.994**	**0.998**	**0.998**

Table 4. Quantitative results on the Dynamic Replica dataset.

B. More Implementation Details

B.1. Sparse dToF Data

When simulating actual dToF data from ground truth depth, several steps are taken to make the simulated sparse dToF depth closely resemble those collected by real-world devices. The field of view (FOV) is set to 70°, and a uniform sampling of 30×40 pixels is applied. Barrel distortion is introduced, along with global rotation and translation transformations. Points with low reflectance are dropped based on their RGB values. Random noise and dropout are also added to the data. The visualized results of the simulated sparse dToF depth are shown in Fig. 1.

These perturbations significantly degrade the quality of the sparse dToF depth. The RMSE and REL of the valid depth points returned by the dToF simulation are summarized in Tab. 5. On the TartanAir dataset, the REL is 0.060, and the RMSE is 0.494, while on the Dynamic Replica dataset, the REL is 0.058, and the RMSE is 0.292.

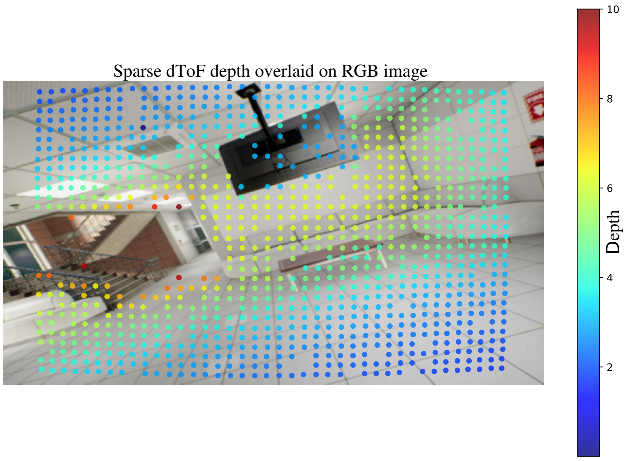

Figure 1. Sparse dToF depth on RGB image

Input data	TartanAir		Dynamic Replica	
Sparse	RMSE(m)	REL	RMSE(m)	REL
dToF depth	0.494	0.060	0.292	0.058

Table 5. Sparse dToF depth metrics

B.2. Definition of Evaluation Metrics

We provide the definitions of the metrics used during our testing. The temporal consistency metric OPW[9] has already been mentioned in the main text of the paper. Here, we supplement it with detailed explanations of the accuracy metrics RMSE, REL, and Accuracy with threshold t, as well as the temporal consistency metric TEPE[6].

- **Accuracy Metrics**

  Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2}$$

where $\hat{d}_i$ represents the predicted depth, d_i represents the ground truth depth, and N is the number of valid pixels.

  Mean Absolute Relative Error (REL):

$$\text{REL} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|}{d_i}$$

where $\hat{d}_i$ represents the predicted depth, d_i represents the ground truth depth, and N is the number of valid pixels.

  Accuracy with threshold t: Percentage of d_i such that

$$\max\left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\right) = \delta < t, \quad t \in \{1.25, 1.25^2, 1.25^3\},$$

where $\hat{d}_i$ and d_i are the predicted depth and ground truth depth of pixel i.

- **Temporal Consistency Metric**

  Temporal End-Point Error (TEPE):

$$\text{TEPE} = \|\big(\mathcal{W}(d_i) - d_{i+1}\big) - \big(\mathcal{W}(\hat{d}_i) - \hat{d}_{i+1}\big)\|_1$$

where $\mathcal{W}(\cdot)$ represents the optical flow warping operation from frame i to frame $i+1$. We use the optical flow predicted by the GMFlow[10] to perform this warping.

C. Network Architecture Details

C.1. Multi-frame Fusion

The multi-frame fusion network architecture is shown in Fig. 2. Multi-frame features are aligned using a flow-guided network and then sent to a bidirectional propagation module, where feature fusion is performed using a Res-block[3]. Taking the alignment of features between the t-th and $(t-1)$-th frames as an example, the optical flow-guided alignment network first inputs RGB_t and RGB_{t-1} into the pre-trained optical flow model SpyNet[5] to obtain the coarse optical flow $O_{t\to t-1}$. Then, $O_{t\to t-1}$ and features f_t, f_{t-1} are concatenated, sent into a deformable convolutional network[2] to derive the refined optical flow $\overline{O}_{t\to t-1}$. Due to the diversity of the deformable convolution network, we can obtain 8 different offsets to flexibly extract features near the corresponding pixels. Finally, we warp the feature f_t with the fine optical flow $\overline{O}_{t\to t-1}$, obtaining the feature $\tilde{f}_t$, aligned with f_{t-1}.

$$O_{t\to t-1} = SpyNet(RGB_t, RGB_{t-1}) \tag{1}$$

$$\overline{O}_{t\to t-1} = DCN(concat(f_t, f_{t-1}), O_{t\to t-1}) \tag{2}$$

$$\tilde{f}_t = \mathcal{W}(f_t, \overline{O}_{t\to t-1}) \tag{3}$$

C.2. DepthHead

We employ the method proposed in AdaBins[1], replacing its miniViT module with a lightweight convolutional module as our depth head, which maps the feature representations to the depth. Unlike directly regressing depth, we predict the depth as a linear combination of different depth bins. Specifically, for each image, we predict its bin-width vector b, which is used to derive the depth bin centers $c(b)$. For each pixel, we predict its probabilities p of belonging to different bins. Assuming the depth range is divided into N different bins, the final predicted depth $\hat{d}$ for each pixel can be expressed as follows:

$$\hat{d} = \sum_{k=1}^{N} c(b_k)p_k \tag{4}$$

D. More Qualitative Results

In this section, we provide additional visual comparisons on the TartanAir and Dynamic Replica datasets. We plotted

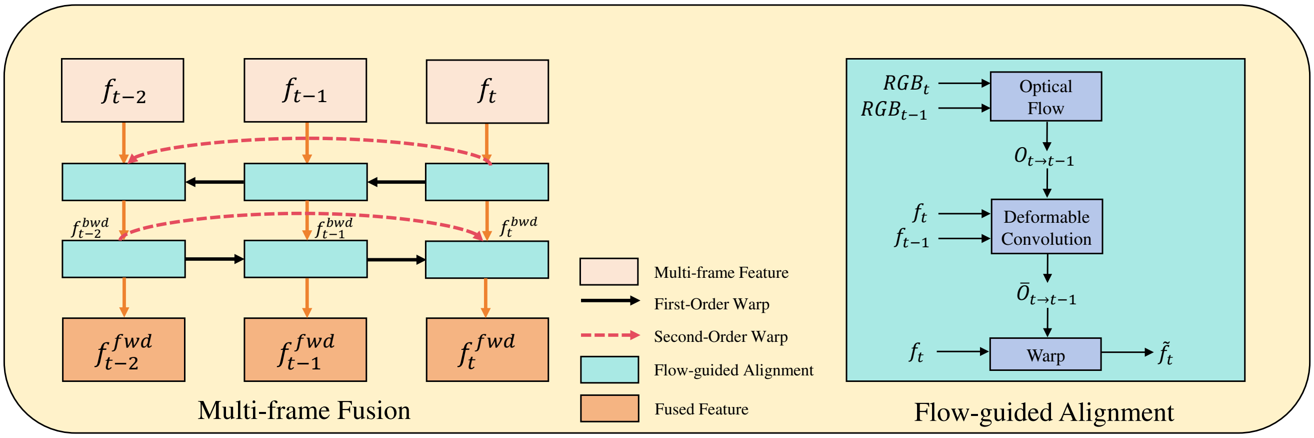

Figure 2. Multi-frame fusion network details

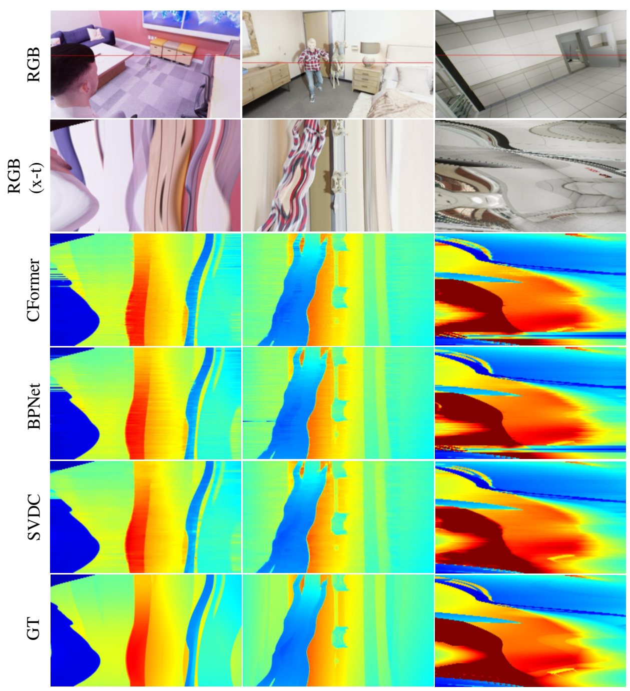

Figure 3. Qualitative results of scanline slice over time

a scanline slice over time to illustrate the temporal consistency of different methods. Moreover, we also present comparisons of the predictions made by various methods[7, 11] in object edges(high-frequency) and smooth regions(low-frequency), highlighting their differences.

In Fig. 3, we present a scanline slice over time, where the first row corresponds to RGB images and the second row represents the scanline patterns over time. Fewer zigzag patterns indicate better temporal consistency. Compared to other methods, our approach demonstrates fewer zigzag patterns, showcasing superior temporal consistency.

In Fig. 4, we display qualitative results on the TartanAir dataset. It can be observed that our SVDC method achieves smoother estimations in low-frequency regions, demonstrating the effectiveness of our frequency-selective fusion strategy in suppressing high-frequency noise in low-frequency areas.

In Fig. 5, we present qualitative results on the Dynamic Replica dataset. The results show that our SVDC method achieves more accurate estimations in high-frequency regions, highlighting its capability to preserve high-frequency details effectively.

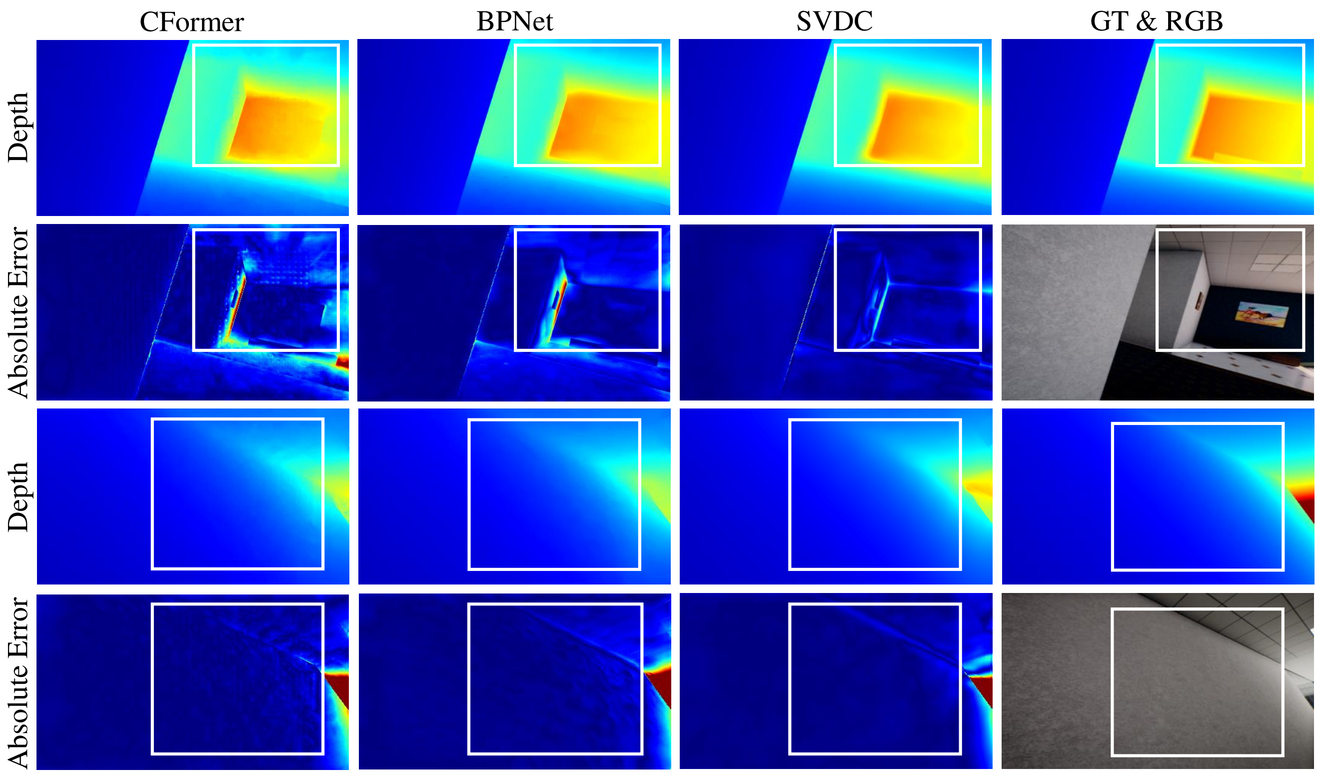

Figure 4. More qualitative results on the TartanAir dataset

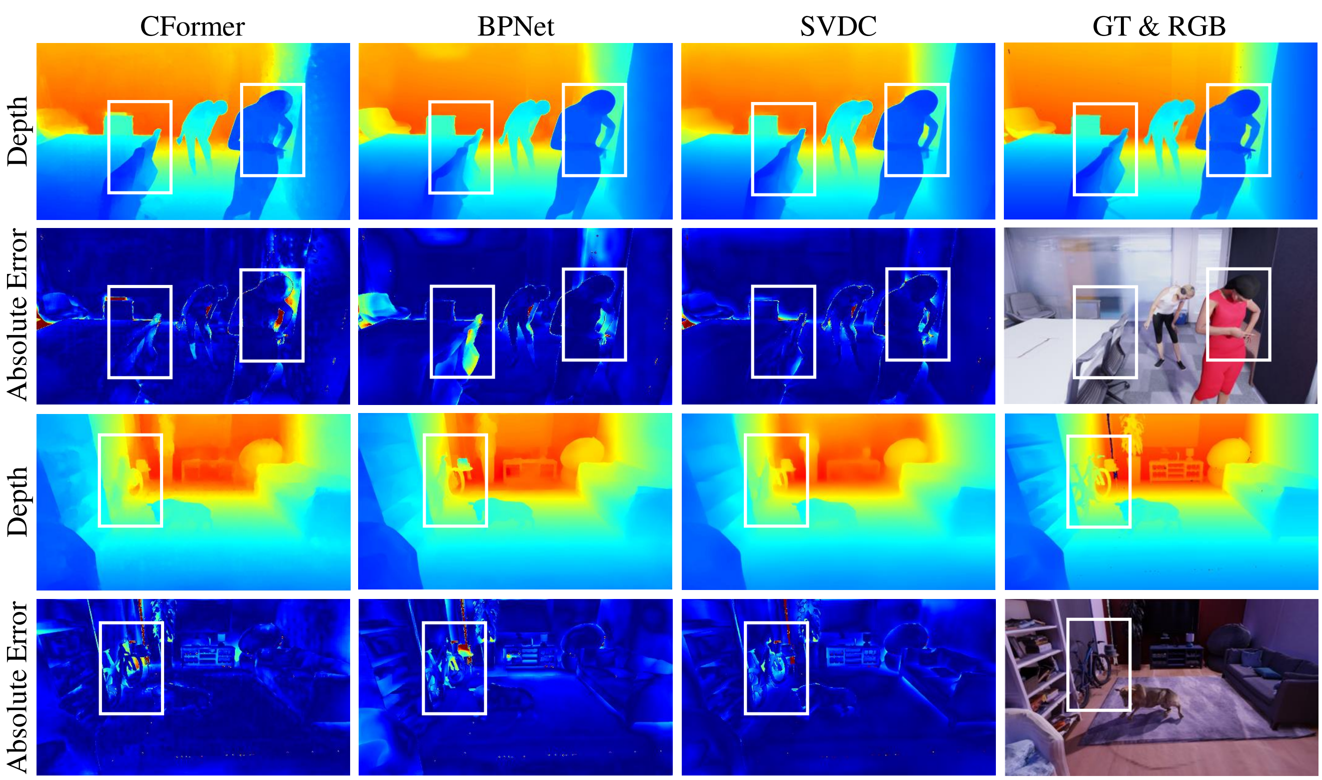

Figure 5. More qualitative results on the Dynamic Replica dataset

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 2

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[4] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13229–13239, 2023. 1

[5] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017. 2

[6] Zhanghao Sun, Wei Ye, Jinhui Xiong, Gyeongmin Choe, Jialiang Wang, Shuochen Su, and Rakesh Ranjan. Consistent direct time-of-flight video depth super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5075–5085, 2023. 2

[7] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9763–9772, 2024. 3

[8] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916, 2020. 1

[9] Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6347–6358, New York, NY, USA, 2022. Association for Computing Machinery. 2

[10] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 2

[11] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2023. 3