Consistent Direct Time-of-Flight Video Depth Super-Resolution Supplementary Material

Zhanghao Sun[1]
[1]Stanford University, zhsun@stanford.edu

Wei Ye[2], Jinhui Xiong[2], Gyeongmin Choe[2], Jialiang Wang[3], Shuochen Su[2], Rakesh Ranjan[2]
[2] Meta Reality Labs, [3]Meta Research

1. Dataset Comparisons

In Table 1, we summarize several recent RGB-D datasets and compare with the proposed DyDToF dataset. We mark the drawbacks in previous datasets with red color.

2. Performance with Hardware Imperfectness

In the main text, we are assuming an ideal dToF imaging model. Here we discuss more general situations where hardware imperfectness presents.

We assume two types of noises in the histogram: signal-dependent shot noise, which originates from the dToF active illumination signal itself, and signal-independent Gaussian noise, which originates from the ambient light or sensor noise[1]. We modify the image formation model (Eqn. 1 in the main text) to be

$$\mathbf{h}_0[k] = \int_{iFoV} \int_{kt_0}^{(k+1)t_0} r[x,y] g(t - 2d[x,y]/c) \; dxdydt \tag{1}$$

$$\mathbf{h}[k] = \mathcal{P}[\mathbf{h}_0[k]] + \mathcal{N}(0,1)\sigma_g, \quad k = 1,2,...,K$$

Where $\mathcal{P}$ applies the photon shot noise and σ_g denotes the standard deviation of Gaussian noise. We define the signal-background ratio (SBR) [3] as the expectation value of the per-histogram SBR

$$\text{SBR} = \mathbb{E}[\sum_k \mathbf{h}[k] / \sqrt{K\sigma_g^2 + \sum_k \mathbf{h}[k]}] \tag{2}$$

As shown in Table. 2, with a moderate amount of noise, DVSR and HVSR have performance close to noise-free condition, while the per-frame estimation quality drops significantly. This is due to two reasons: First, since noise is random in each frame, by aggregating information from multiple frames, the network obtains denoising capability. Similar effects are also demonstrated in RGB video/burst denoising [9, 10, 13]. Second, our histogram matching module is inherently robust to noise. Instead of attending to all the details in the full histogram, peak detection, and rough scale rebinning only pick out the strongest signal and average out most of the fluctuations.

We also evaluate the network performances with lower dToF depth resolution (we use 1024 in the main text and 128 in this ablation study). This has limited influence on all the algorithms.

3. Performance with Multi-path Interference

Multi-path interference (MPI) is a long-standing problem in time-of-flight depth sensing. Much progress has been made to alleviate this effect in indirect time-of-flight (iToF) sensors leveraging modern neural networks and high-quality synthetic datasets [5, 6, 12]. However, limited by the temporal resolution of iToF sensor, handling strong MPI (e.g. at a corner) remains challenging. On the other hand, dToF sensor provides high temporal resolution and thus is more favorable in handling MPI [6]. Here we use the transient renderer (TR) [8] to validate the fidelity of our dToF simulator and evaluate our HVSR network with MPI. Note that when the light propagation medium (air) is transparent and MPI is disabled, the TR (Eqn.1 in [8]) falls back to Eqn.5 in our paper, as shown in Fig.1(a). When strong MPI presents, the TR generated histogram contains additional peaks (red bonding box). However, our HVSR network still performs reasonably well without finetuning, as shown in Fig.1(b), (c) (MPI indeed introduces minor artifacts close to intersections). We attribute this to the temporal resolving capability of dToF sensors. As shown in Fig.1(a), the MPI-induced peaks are generally much weaker than the main peak and separated temporally (different from iToF). We believe the performance can be further enhanced by incorporating MPI (or data augmentation) in training [5].

[1] We assume the non-negative background induced by ambient photon is subtracted, while leaving the zero-mean photon shot noise.

Dataset	# Scenes	# Frames	Data Modalities	Format	Quality	Environment	Dynamic
ScanNet	1513	2.5M	RGB-D + Semantics	Video	MQ	Indoor	None
Matterport3D	90	194k	RGB-D + Semantics	Video	MQ	Indoor	None
Replica	18	–	Colored Meshes	–	HQ	Indoor	None
TarTanAir	18	185k	RGB-D + Semantics + Flow	Video	HQ	Outdoor (15) Indoor (3)	Very few
Sintel	20	1k	RGB-D + Flow	Video	Cartoon	Outdoor	Yes
HyperSim	400	75k	RGB-D + Albedo + Surface Normal + Semantics	Image	HQ	Indoor	None
DyDToF (ours)	100	50k	RGB-D + Albedo + Surface Normal	Video	HQ	Indoor	Yes

Table 1. RGB-D dataset comparisons.

Methods	SBR	AE (mm) ↓
NLSPN	noise free	48.8
NLSPN	50	61.2
DVSR	noise free	40.2
DVSR	50	43.2
HVSR	noise free	27.5
HVSR	50	28.1

Table 2. Ablation studies on noise.

Methods	# Time bin	AE (mm) ↓
NLSPN	128	45.6
DVSR	128	40.3
HVSR	128	28.9

Table 3. Ablation studies with lower depth resolution.

DVSR Variants	AE (mm) ↓	TEPE (mm) ↓
Single stage	52.4	22.1
Full model	40.2	15.6

Table 4. Ablation studies on per-frame network design.

HVSR Variants	# Frames	AE (mm) ↓
Forward only w/o hist-conf	5	34.4
Forward only	5	32.2
w/o hist-conf	5	30.6
Full model	5	29.4
w/o hist-conf	30	28.2
Full model	30	27.5

Table 5. Ablation studies on histogram information.

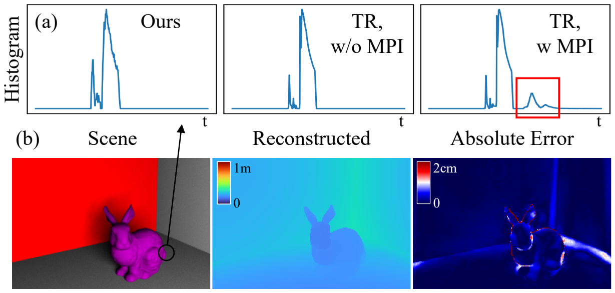

Figure 1. (a) Histograms generated with our simulator and TR [8]. (b) HVSR reconstruction with MPI (without finetuning). Despite minor erros, the reconstruction maintains high accuracy due to the high temporal resolution in dToF sensor.

4. Real-world Generalization to Apple ARKit

Different from the raw dToF data that our networks are trained on, the dToF data provided by Apple's ARKit [1, 2] is pre-processed with a closed-source depth densification algorithm. To compare our DVSR network with ARKit without introducing additional information, we naively downsample the pre-processed ARKit dToF data as input to our DVSR network. As shown in Fig.2, DVSR (without finetuning) not only achieves sharper edges but also corrects minor errors in the input (red box) using multi-frame cues. This demonstrates the generalizability of our model, even to pseudo dToF depth. We use both static scene from the official ARKitscenes dataset [1] (first row) and self-captured dynamic scene [2] (second row) for evaluations. Please refer to the supplementary video or project page for video comparisons.

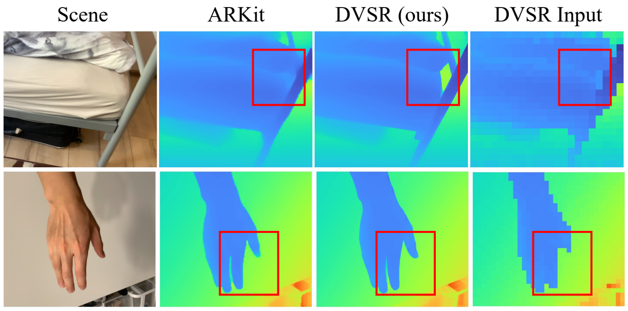

Figure 2. Qualitative comparison with Apple ARKit. Our DVSR network (without finetuning) not only achieves sharper edges, but also corrects minor errors in the pre-processed ARKit depth with multi-frame information aggregation.

5. Extention: Sparse Depth Completion

The proposed depth video super-resolution (DVSR) framework can be adequately applied in other video depth estimation tasks. As an example, we retrain the network on the conventional depth completion task (with small modifications at the input stage to accommodate the different data modalities). The task converts high-resolution, sparse depth maps into high-resolution, dense depth maps. We use a random dot sampling pattern at each frame in the video clip, with density $\sim 1/16^2 = 0.4\%$. We use the same settings to train the network on the TarTanAir dataset. We name this new model depth video sparse-to-dense (DVS2D).

As shown in Fig. 3 (a), we compare the DVS2D performance with the per-frame processing NLSPN baseline [11]. Due to the low sampling density, per-frame prediction results miss important but not sampled details (red bounding boxes), while DVS2D has the capability of maintaining all structures even if they are not sampled in the current frame. This intuitively demonstrates the effectiveness of our temporal fusion module.

In Fig. 3 (b), we further conduct a cross-dataset evaluation on the KITTI dataset [4]. A significant improvement in depth edges can also be observed. What's more, in real-world captured data (e.g., KITTI dataset), incorrect depth values induced by misalignment and transparent objects (e.g., car windows) are unavoidable. However, DVS2D is generally stable to these artifacts despite not encountering them in the training process. We again attribute this advantage to our temporal fusion module through comparison to the per-frame baseline model.

6. Network Architecture Details

We show our detailed network architecture in Fig. 4. Our code and dataset are open-source at https://github.com/facebookresearch/DVSR/. The DVSR and HVSR network runs at $\sim$ 10FPS, with $\sim$ 280MB memory consumption per frame with output resolution 480×640.

7. More Ablation Studies

In the additional ablation studies, we use the same training/testing split as in the main text unless mentioned.

7.1. Double stage vs. Single stage

In the main text, we conduct ablation studies on the multi-frame fusion module. In Table. 4, we conduct an ablation study on another key design choice, the double-stage processing framework. Double-stage processing is widely applied in computer vision tasks, including depth prediction [7, 11, 16], object detection [17], flow estimation [14, 15], etc. It generally involves an encoder-decoder-based initial prediction stage and a refinement stage based on spatial propagation [11], recurrent modules [14], or simply another encoder-decoder network [7,17]. We choose the last approach since it is the most general and most compatible with our multi-frame fusion module. It is evident that the double-stage design is important to the network performance.

7.2. dToF Histogram Processing

We further analyze how utilizing the histogram information facilitates the depth estimation, as shown in Table 5. Inputting the rebinned histogram instead of a single depth map boosts the network performance significantly, simply due to more information in the processing. When temporal fusion is insufficient (e.g., in the case of shorter video clips or forward-only operations), histogram matching-based confidence helps identify errors and contributes more to the estimation quality.

8. More Results Visualization

We show more qualitative comparisons in Fig. 5 (Replica dataset) and Fig. 6 (DyDToF dataset). Please refer to our supplementary video for temporal stability comparisons.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes–a diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 2

[2] Ilya Chugunov, Yuxuan Zhang, Zhihao Xia, Xuaner Zhang, Jiawen Chen, and Felix Heide. The implicit values of a good hand shake: Handheld multi-frame neural depth refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2852–2862, 2022. 2

[3] Agata M. Pawlikowska et al. *Optics express*, 25(10):11919–11931, 2017. 1

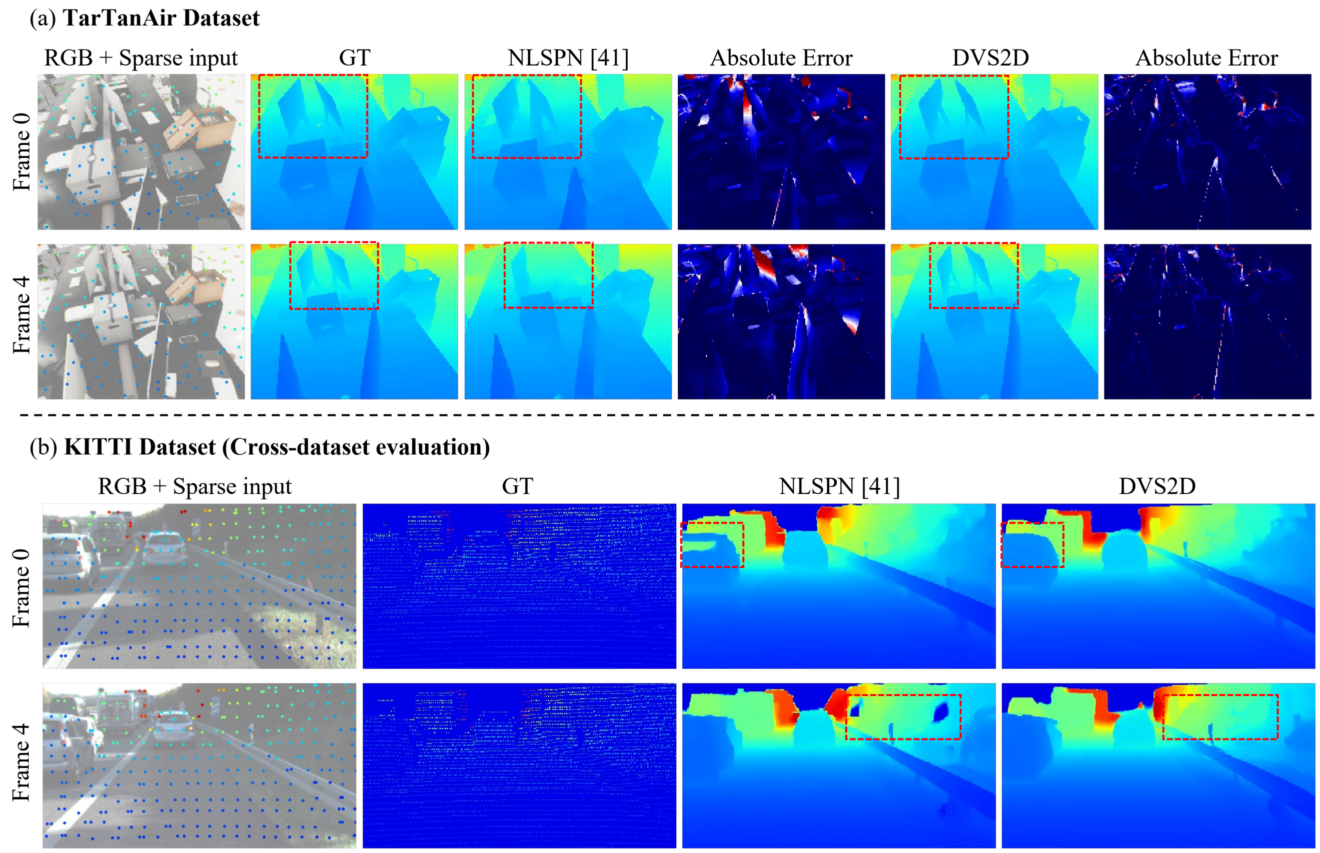

Figure 3. Extension to depth completion task on (a) TarTanAir dataset, (b) KITTI dataset. Note that the video processing network has the capabilities of: predicting structures missing in the current frame with the assistance from adjacent frames (bounding boxes in (a)), mitigating errors induced by misalignment and transparent objects (e.g., windows) (bounding boxes in (b))

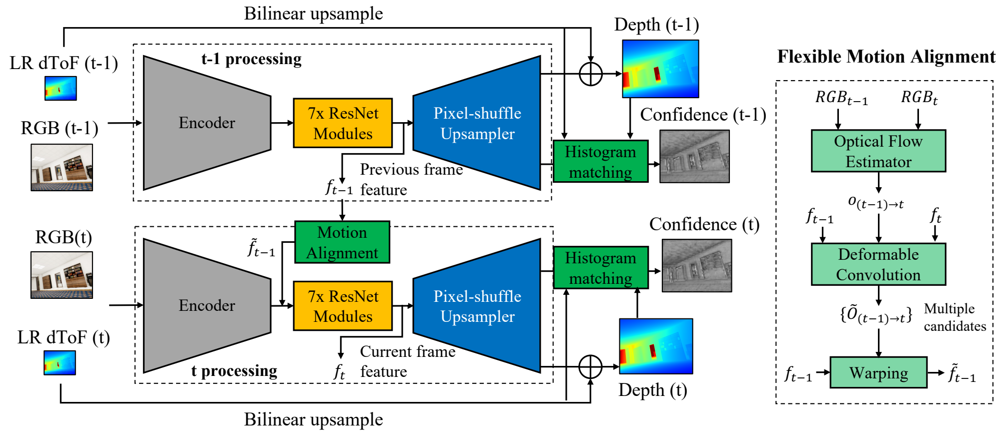

Figure 4. Network architecture details.

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[5] Qi Guo, Iuri Frosio, Orazio Gallo, Todd Zickler, and Jan Kautz. Tackling 3d tof artifacts through learning and the

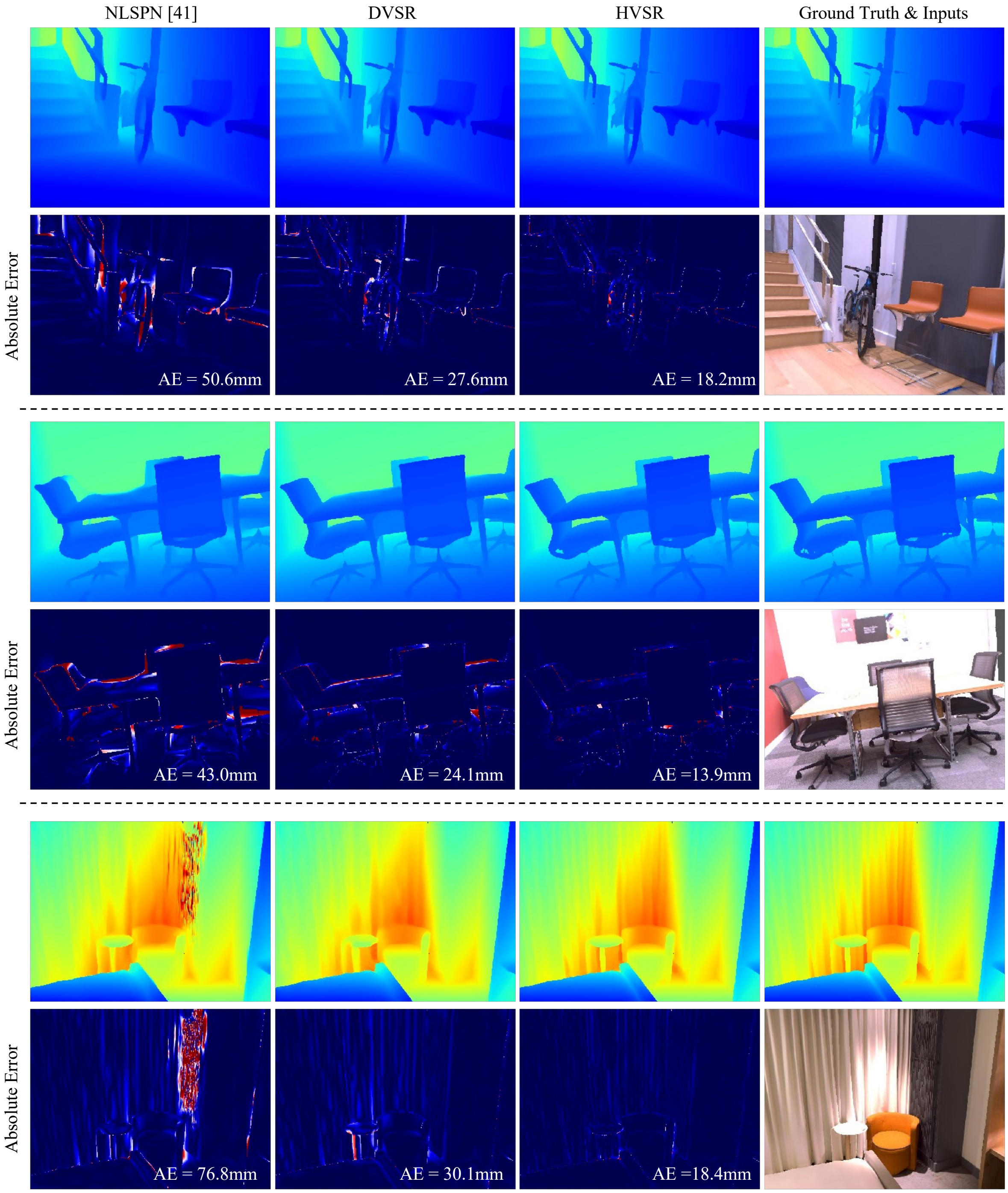

Figure 5. More qualitative comparisons on Replica dataset.

flat dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 368–383, 2018. 1

[6] Felipe Gutierrez-Barragan, Huaijin Chen, Mohit Gupta, Andreas Velten, and Jinwei Gu. itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. *IEEE Transactions on Computational Imaging*, 7:1205–1214, 2021. 1

[7] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Con-*

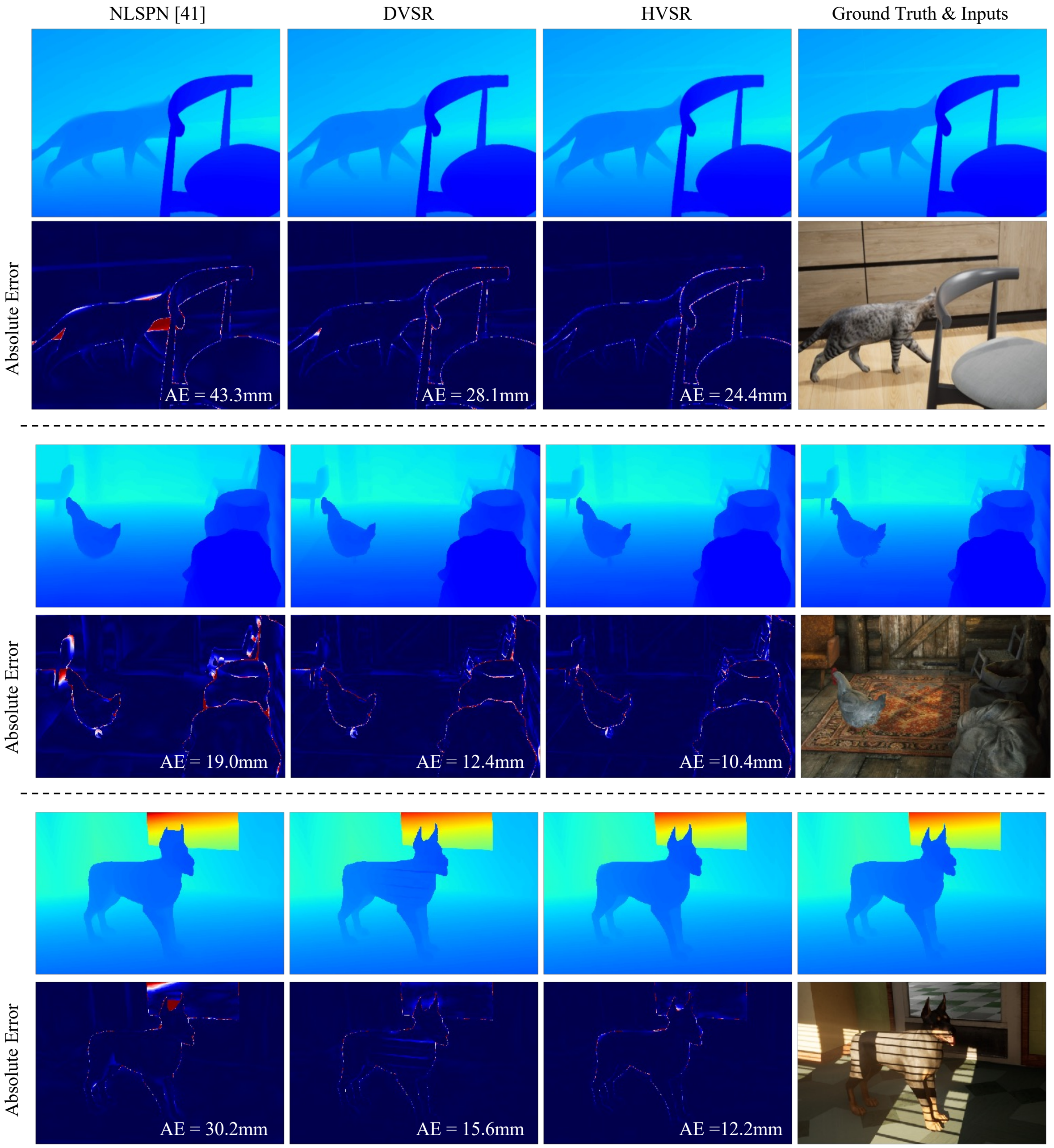

Figure 6. More qualitative comparisons on DyDToF dataset.

ference on Robotics and Automation (ICRA), pages 13656–13662. IEEE, 2021. 3

[8] Adrian Jarabo, Julio Marco, Adolfo Munoz, Raul Buisan, Wojciech Jarosz, and Diego Gutierrez. A framework for transient rendering. *ACM Transactions on Graphics (ToG)*, 33(6):1–10, 2014. 1, 2

[9] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 1

[10] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2502–2510, 2018. 1

[11] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision*,

pages 120–136. Springer, 2020. 3

[12] Di Qiu, Jiahao Pang, Wenxiu Sun, and Chengxi Yang. Deep end-to-end alignment and refinement for time-of-flight rgb-d module. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9994–10003, 2019. 1

[13] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809. IEEE, 2019. 1

[14] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[15] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8375–8384, 2021. 3

[16] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Baobei Xu, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. *arXiv preprint arXiv:2107.13802*, 2021. 3

[17] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3