

Consistent Time-of-Flight Depth Denoising via Graph-Informed Geometric Attention

Supplementary Material

In this supplementary material, we provide the derivation of the data fidelity term in MAP formulation based on ToF depth noise distribution in Sec. 8. Then we evaluate the sensitivity to frame time step in Sec. 9. Next, we summarize unrolling of cross-frame graph fusion based ToF depth denoising algorithm in Sec. 10. More visualization results are provided in Sec. 11, with a video demonstrating the estimation accuracy and temporal consistency.

8. Data Fidelity Term in MAP Problem

In this section, we derive the data fidelity term based on ToF depth noise distribution. As assumed in Sec. 4.3, \mathbf{x}_i and \mathbf{x}_q are corrupted by additive white Gaussian noise (AWGN) [12, 13], and the pixels in $\mathbf{y}_i^t, \mathbf{y}_q^t$ are independent and identically distributed with multivariate Gaussian distribution. The joint probability density function of $\mathbf{y}_i^t, \mathbf{y}_q^t$ is given as:

$$P(\mathbf{y}_i^t, \mathbf{y}_q^t | \mathbf{x}_i^t, \mathbf{x}_q^t) = \frac{1}{(2\pi\sigma^2)^N} \times \exp\left(-\frac{(\mathbf{n}_i^t)^\top \mathbf{n}_i^t + (\mathbf{n}_q^t)^\top \mathbf{n}_q^t}{2\sigma^2}\right), \quad (18)$$

$$\mathbf{n}_i^t = \mathbf{y}_i^t - \mathbf{x}_i^t, \quad \mathbf{n}_q^t = \mathbf{y}_q^t - \mathbf{x}_q^t, \quad (19)$$

where σ is the noise variance.

Since the final target is to reconstruct depth, we further investigate depth noise distribution based on (18). Based on (4) and (18), the distribution of depth noise \mathbf{n}_d^t is derived in [12, 13] as,

$$P(\mathbf{n}_d^t) = \prod_{m=1}^N \frac{\cos(4\pi f_m \mathbf{n}_d^t(m)/c)}{2\gamma^t(m)\sqrt{2\pi}} [1 + \operatorname{erf}(\frac{\cos(4\pi f_m \mathbf{n}_d^t(m)/c)}{\gamma^t(m)\sqrt{2}})] \times \exp(-\frac{\sin^2(4\pi f_m \mathbf{n}_d^t(m)/c)}{2\gamma^t(m)^2}) + \frac{1}{2\pi} \exp(-\frac{1}{2\gamma^t(m)^2}), \quad (20)$$

where $\gamma^t(m) = \sigma/\mathbf{y}_a^t(m)$, $\mathbf{y}_a^t(m)$ is noisy amplitude, erf is the Gaussian error function. Under normal noise level, i.e., $\gamma^t(m) \ll 1$, we have erf output equal to 1 and last term equal to 0 in (20), then (20) is approximated with

$$P(\mathbf{n}_d^t) \approx \prod_{m=1}^N \left(\frac{\cos(4\pi f_m \mathbf{n}_d^t(m)/c)}{\gamma^t(m)\sqrt{2\pi}} \times \exp(-\frac{\sin^2(4\pi f_m \mathbf{n}_d^t(m)/c)}{2(\gamma^t(m))^2}) \right). \quad (21)$$

Based on (21), the log of likelihood $P(\mathbf{y}_d^t | \mathbf{x}_d^t)$ is

$$\ln P(\mathbf{y}_d^t | \mathbf{x}_d^t) \approx \sum_{m=1}^N (\ln(\cos(4\pi f_m \mathbf{n}_d^t(m)/c)) - \sin^2(4\pi f_m \mathbf{n}_d^t(m)/c)/(2\gamma^t(m)^2)), \quad (22)$$

where the irrelevant term $-\ln(\gamma^t(m)\sqrt{2\pi})$ is removed. Both terms in (22) minimize n_d , and with $\gamma \ll 1$, the second term dominates. Thus, we remove the first term and compute the likelihood as a function of $\mathbf{x}_i^t, \mathbf{x}_q^t$ as follows:

$$\ln P(\mathbf{y}_d^t | \mathbf{x}_d^t) \approx \sum_{m=1}^N -\frac{\sin^2(\phi(m) - \phi'(m))}{2\gamma^t(m)^2} = \sum_{m=1}^N -\frac{(\sin \phi(m) \cos \phi'(m) - \cos \phi(m) \sin \phi'(m))^2}{2\gamma^t(m)^2}, \quad (23)$$

where $\phi' = 4\pi f_m \mathbf{y}_d^t/c$ is the noisy phase.

Based on (23) and (4), the log of likelihood of \mathbf{n}_d^t is given as a function of $\mathbf{x}_i^t, \mathbf{x}_q^t$:

$$\ln P(\mathbf{n}_d^t) \approx -\frac{1}{2\sigma^2} \|(\mathbf{X}_a^t)^{-1}(\mathbf{x}_q^t \odot \mathbf{y}_i^t - \mathbf{x}_i^t \odot \mathbf{y}_q^t)\|_2^2, \quad (24)$$

where $\mathbf{X}_a^t = \operatorname{diag}(\mathbf{x}_a^t)$ is the amplitude, \odot is Hadamard product.

9. Analysis of Frame Time Step

Following [9], to investigate the effect of time step between reference and current frames, we test GIGA-ToF on DVToF dataset with different time steps. For small time steps $\Delta t = 1, 2$, the performances are similar. When time steps become larger, $\Delta t = 4, 8$ reduces due to the limited similarity of graph structures in the neighboring pixels in the reference frame. Nevertheless, the performance still surpasses that of single-frame processing, validating the necessity of multi-frame processing and the temporal self-similarity of graph structures despite large frame gaps.

Table 3. Comparison of quantitative evaluation on DVToF testing dataset with different frame rate

Time step	MAE(m)↓	AbsRel↓	δ_1 ↑	TEPE(m)↓
1	0.0190	0.0060	0.9974	0.0634
2	0.0192	0.0060	0.9973	0.0608
4	0.0194	0.0062	0.9972	0.0647
8	0.0210	0.0071	0.9970	0.0725

10. Algorithm Summary

Based on the algorithm unrolling of graph Laplacian regularization, we obtain the solution to (14), which is summarized in Algorithm 1.

Algorithm 1 Unrolling of Cross-frame Graph Fusion based ToF Depth Denoising Algorithm

Require: Noisy ToF raw data $\mathbf{y}_i^t, \mathbf{y}_q^t$, intra-frame graph adjacency matrices $\mathbf{W}_i^t, \mathbf{W}_q^t, \mathbf{W}_i^{t-1}, \mathbf{W}_q^{t-1}$, inter-frame graph adjacency matrix $\mathbf{W}^{t,t-1}$ and fusion weight $\Phi^{t,t-1}$, GLR prior weight Λ_i^t, Λ_q^t , iteration number R, T

Ensure: Denoised output $\mathbf{x}_i^t, \mathbf{x}_q^t$

- 1: Map reference frame graphs $\mathbf{W}_i^{t-1}, \mathbf{W}_q^{t-1}$ to current frame to obtain mapped graphs $\tilde{\mathbf{W}}_i^{t-1}, \tilde{\mathbf{W}}_q^{t-1}$ using (5)
 - 2: Fuse the mapped graphs with current frame graphs to obtain fused graphs $\tilde{\mathbf{W}}_i^t, \tilde{\mathbf{W}}_q^t$ using (6)
 - 3: Obtain corresponding $\tilde{\mathbf{D}}_i^t, \tilde{\mathbf{D}}_q^t$ from $\tilde{\mathbf{W}}_i^t, \tilde{\mathbf{W}}_q^t$ using (10)
 - 4: Initialize $\mathbf{x}_i^{t,0} = \mathbf{y}_i^t, \mathbf{x}_q^{t,0} = \mathbf{y}_q^t$
 - 5: **for** $r = 0 : R - 1$ **do**
 - 6: Update $\Lambda_i^{t,r-1}$ with $\mathbf{X}_a^{t,r-1}$, fix $\mathbf{x}_q^{t,r}$ and optimize $\mathbf{x}_i^{t,r}$
 - 7: **for** $p = 0 : P - 1$ **do**
 - 8: Transform $\mathbf{x}_i^{t,r,p}$ with convolutional kernel $\tilde{\mathbf{W}}_i^t$
 - 9: Fuse with $\mathbf{x}_i^{t,r-1}$ with weight $\Lambda_i^{t,r-1}$ as specified in (14) to update $\mathbf{x}_i^{t,r,p+1}$
 - 10: **end for**
 - 11: Fix $\mathbf{x}_i^{t,r}$ and repeat steps 7-10 to optimize $\mathbf{x}_q^{t,r}$
 - 12: **end for**
 - 13: **Output**
-

11. More Visualization

We provide more results for the qualitative comparison of ToF depth denoising methods. In particular, we demonstrate results on synthetic DVToF dataset in Fig. 11 and Fig. 12, and DVToF dataset with noise augmentation in Fig. 13. To further demonstrate the generalization ability to real Kinectv2 data, we shown results in Figs. 8, 9, 10. Note that we directly apply the model trained on original DVToF dataset to the noise-augmented DVToF dataset and Kinectv2 dataset without fine-tuning, which validates its generalization ability. Please kindly refer to the supplementary video for better temporal visualizations.

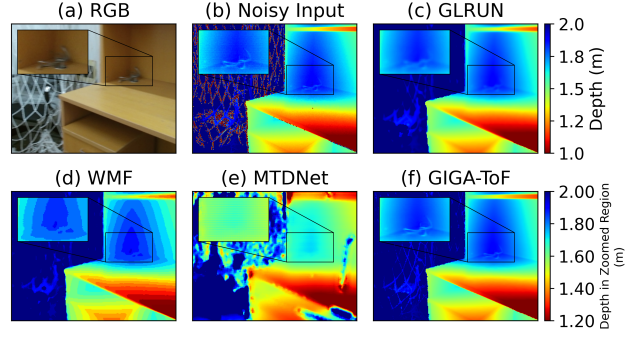


Figure 8. Visual results of ToF depth denoising on real data captured by Kinect v2 sensor: (a) RGB and (b) noisy depth captured by Kinectv2 camera, and results of (c) GLRUN, (d) WMF, (e) MTDNet and (f) GIGA-ToF, where GIGA-ToF shows accurate and smooth estimation.

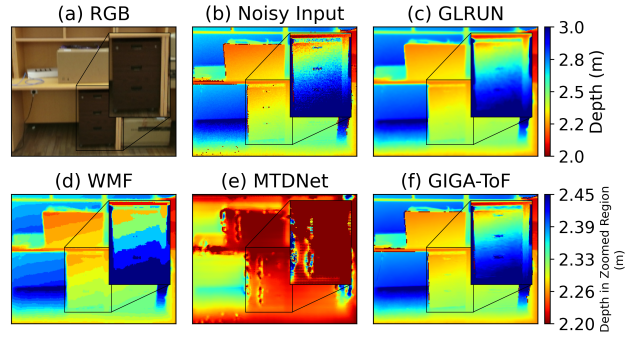


Figure 9. Visual results of ToF depth denoising on real data captured by Kinect v2 sensor: (a) RGB and (b) noisy depth captured by Kinectv2 camera, and results of (c) GLRUN, (d) WMF, (e) MTDNet and (f) GIGA-ToF, where GIGA-ToF shows robustness to real noise and recovers accurate details.

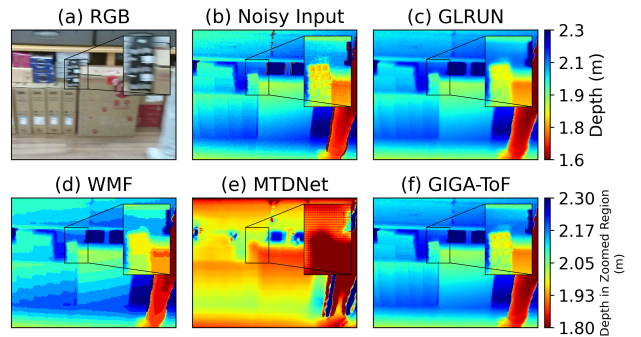


Figure 10. Visual results of ToF depth denoising on real data captured by Kinect v2 sensor: (a) RGB and (b) noisy depth captured by Kinectv2 camera, and results of (c) GLRUN, (d) WMF, (e) MTDNet and (f) GIGA-ToF, where GIGA-ToF exhibits better spatial sharpness than other competing schemes.

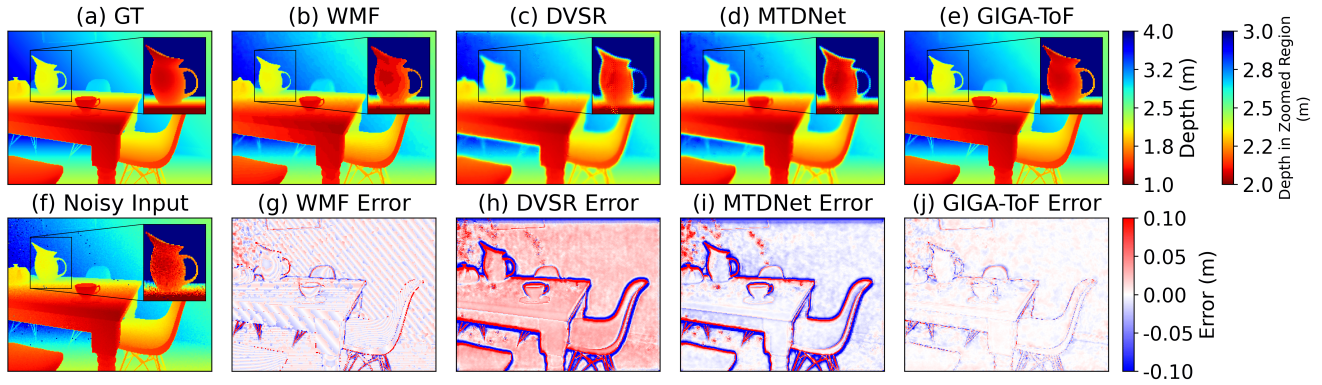


Figure 11. Depth results and error maps of ToF depth denoised on DvToF dataset: (a) GT, results of (b) WMF [23], (c) DVSR [35], (d) MTDNet [9] and (e) proposed GIGA-ToF. Corresponding error maps are in the second row. GIGA-ToF shows more accurate depth estimation, maintaining global smoothness with edge preservation, *e.g.*, the teapot handle highlighted in the zoom-in block.

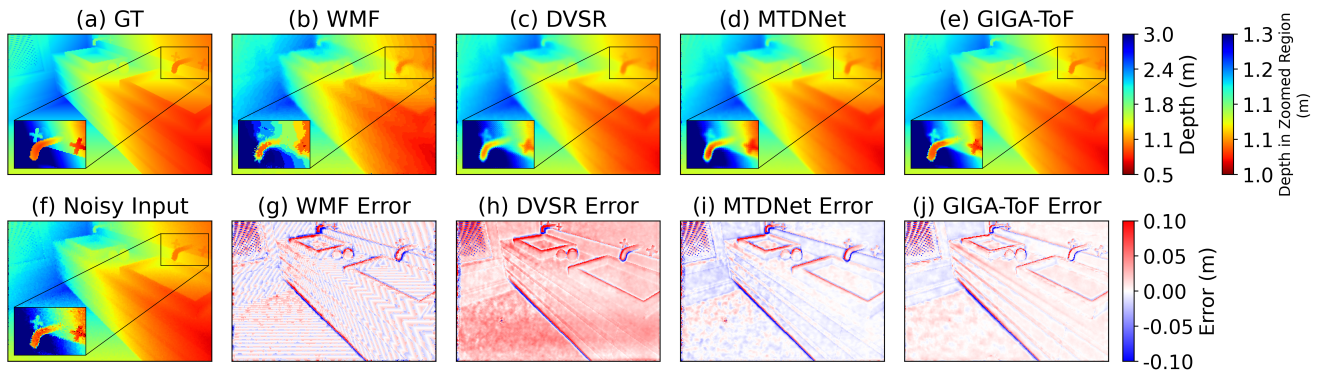


Figure 12. Depth results and error maps of ToF depth denoised on DvToF dataset: (a) GT, results of (b) WMF [23], (c) DVSR [35], (d) MTDNet [9] and (e) proposed GIGA-ToF. Corresponding error maps are in the second row. While MTDNet shows competing results, the details are blurred as highlighted in the zoom-in block, while GIGA-ToF generates sharp edges due to utilization of motion-invariant graph structure fusion.

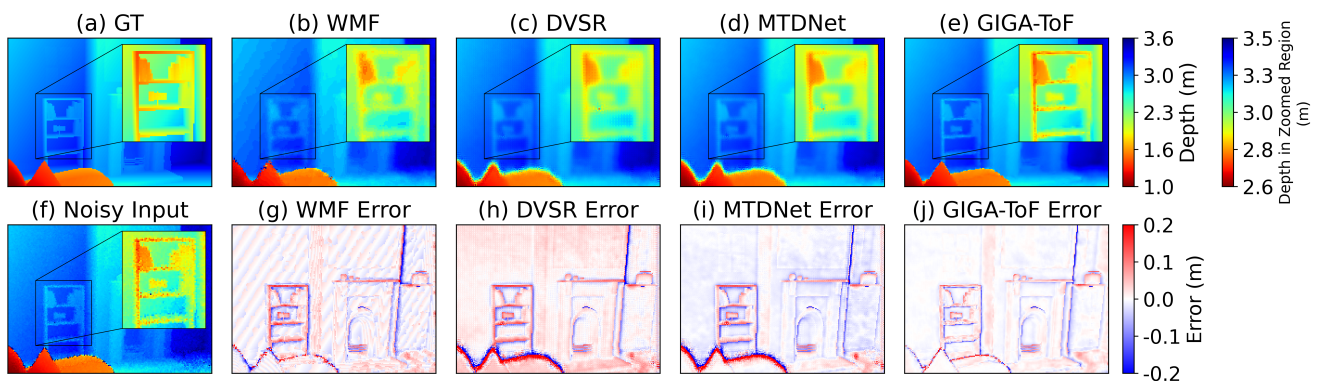


Figure 13. Depth results and error maps of ToF depth denoised on DvToF dataset with augmented edge noise: (a) GT, results of (b) WMF [23], (c) DVSR [35], (d) MTDNet [9] and (e) proposed GIGA-ToF. Corresponding error maps are in the second row. GIGA-ToF shows strong generalization to unseen edge noise and generates accurate details, *e.g.*, in the bookshelf in the zoom-in block.