# Multi-view Depth Estimation using Epipolar Spatio-Temporal Networks: Supplementary Materials

Xiaoxiao Long[1] Lingjie Liu[2] Wei Li[3] Christian Theobalt[2] Wenping Wang[1,4]
[1]The University of Hong Kong [2]Max Planck Institute for Informatics
[3]Inceptio [4]Texas A&M University

## 1. More results

**Depth evaluation** Figure 3 shows more results of our method, MVDepth [7], DPS [2], NAS [3], Neuralrgbd [4] and DELTAS [6]. Compared with other methods, our estimated depth maps are more accurate and have fewer noises even in texture-less regions, such as table, wall, floor and kitchen cabinets. Moreover, from Table 2 and Table 3, we can see that our model not only performs better than other methods on ScanNet [1] dataset, but also shows superior generalization ability on unseen 7scenes [5] dataset.

**Temporal coherence** To measure the temporal consistency of the estimated depth maps, we adopt standard deviation of the mean absolute error of the estimated depth maps for evaluation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N}} \tag{1}$$

where $i$ is the index of estimated depth map, $e$ is the mean absolute error of estimated depth map and ground truth depth map, and $N$ is the total number of estimated depth maps.

As shown in Table 1, the estimated depth maps of our methods are more accurate and temporally consistent than those of other methods. The Plot of absolute errors of depth maps in Figure 1 shows that our method could generate more temporal coherent estimated depth maps than other methods. Moreover, as shown in Figure 4 and Figure 2, by utilizing temporal coherence, our model generates more temporally consistent depth maps of continuous video frames than other methods.

Table 1. Comparison of temporal coherence over ScanNet and 7scenes datasets with evaluation depth range $0 \sim 5m$.

| Method | ScanNet | | 7scenes | |
|---|---|---|---|---|
| | Abs | Std | Abs | Std |
| DPS | 0.1887 | 0.2243 | 0.2907 | 0.3271 |
| NAS | 0.1823 | 0.2177 | 0.2874 | 0.3252 |
| Neuralrgbd | 0.1642 | 0.1848 | 0.4051 | 0.4600 |
| DELTAS | 0.1650 | 0.1886 | 0.2874 | 0.3003 |
| Ours | **0.1432** | **0.1673** | **0.2498** | **0.2743** |

## 2. Discussions

**Epipolar Spatio-Temporal transformer** We consider several variants of our method for ablation studies. We denote the depth estimated by the model without EST transformer as *independent depth*, the depth jointly estimated by the model with EST transformer as *joint depth*, and the depth sequentially estimated by the model using ESTM operation as *ESTM depth*. As shown in Table 4, when using the hybrid cost regularization network, both *joint depth* and *ESTM depth* are better than *independent depth*. However, when adopting the pure 3D regularization network (without *ContextNet*), *joint depth* is worse than *independent depth*, and the improvement of *ESTM depth* is trivial.

In fact, there are two key factors that may influence the effect of EST transformer: the quality of estimated depth before being enforced temporal coherence and the accuracy of calibrated camera poses of video frames. For models without *ContextNet*, their estimated depth is not accurate enough, the joint estimation will propagate wrong information across the multiple depth maps. But ESTM suffers from it less because the errors can be alleviated gradually as more frames are processed sequentially.

Moreover, as shown in Table 4, when *ContextNet* adopts ResNet-18 as backbone, *ESTM depth* is a bit better than *Joint depth*. When *ContextNet* adopts ResNet-50 as backbone, *Joint depth* outperforms *ESTM depth* a bit. This may provide extra evidences that when the estimated depth is not accurate enough, ESTM inference operation performs better than joint estimation due to long-term temporal coherence it unitizes.

Overall, the combination of the hybrid regularization network and EST transformer can boost the best performance.

**The backbone of ContextNet** Unlike a single 3D cost regularization network used in prior works, we adopt a hybrid network to learn 2D global context information and 3D local matching information separately. Table 4 shows that models with *ContextNet* outperforms that without *ContextNet*. To fully demonstrate the effect of *ContextNet*, we further test our models with *ContextNet* adopting ResNet-18 and ResNet-50 as backbone. As shown in Table 5, replacing ResNet-18 by ResNet-50 as the backbone of *ContextNet* can lead to better results on ScanNet [1] and 7scenes [5] datasets.

## References


[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 2, 3

[2] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: end-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 1

[3] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199, 2020. 1

[4] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019. 1

[5] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013. 1, 2, 3

[6] Ayan Sinha, Zak Murez, James Bartolozzi, Vijay Badrinarayanan, and Andrew Rabinovich. Depth estimation by learning triangulation and densification of sparse points for multi-view stereo. *arXiv preprint arXiv:2003.08933*, 2020. 1

[7] Kaixuan Wang and Shaojie Shen. Mvdepthnet: real-time multiview depth estimation neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018. 1

[8] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1625–1632, 2013. 3

Table 2. Comparison of depth estimation over ScanNet [1] dataset.

| Range | Method | Abs Rel | Abs | Sq Rel | RMSE | RMSE log | $\sigma < 1.25$ | $\sigma < 1.25^2$ | $\sigma < 1.25^3$ |
|---|---|---|---|---|---|---|---|---|---|
| 10m | MVDepth | 0.1167 | 0.2301 | 0.0596 | 0.3236 | 0.1610 | 84.53 | 96.39 | 99.06 |
| | MVDepth-FT | 0.1116 | 0.2087 | 0.0763 | 0.3143 | 0.1500 | 88.04 | 97.34 | 99.19 |
| | DPS | 0.1200 | 0.2104 | 0.0688 | 0.3139 | 0.1604 | 86.40 | 96.12 | 98.60 |
| | DPS-FT | 0.0986 | 0.1998 | 0.0459 | 0.2840 | 0.1348 | 88.80 | 97.85 | 99.20 |
| | NAS | 0.0941 | 0.1928 | 0.0417 | 0.2703 | 0.1269 | 90.09 | 98.26 | 99.47 |
| | CNM | 0.1102 | 0.2129 | 0.0513 | 0.3032 | 0.1482 | 86.88 | 97.22 | 99.32 |
| | DELTAS | 0.0915 | 0.1710 | 0.0327 | 0.2390 | 0.1226 | 91.47 | **98.72** | **99.70** |
| | Ours-EST(concat) | 0.0818 | 0.1536 | 0.0301 | 0.2234 | 0.1130 | 92.99 | 98.70 | 99.67 |
| | Ours-EST(adaptive) | **0.0812** | **0.1505** | **0.0298** | **0.2199** | **0.1104** | **93.13** | 98.71 | 99.68 |
| 5m | Neuralrgbd | 0.1013 | 0.1657 | 0.0502 | 0.2500 | 0.1315 | 91.60 | 97.90 | 99.27 |
| | Ours-EST(concat) | 0.0811 | 0.1469 | 0.0279 | 0.2066 | 0.1109 | 93.19 | 98.77 | 99.70 |
| | Ours-EST(adaptive) | **0.0805** | **0.1438** | **0.0275** | **0.2029** | **0.1083** | **93.33** | **98.78** | **99.71** |

Table 3. Comparison of depth estimation over 7Scenes [5] dataset.

| Range | Method | Abs Rel | Abs | Sq Rel | RMSE | RMSE log | $\sigma < 1.25$ | $\sigma < 1.25^2$ | $\sigma < 1.25^3$ |
|---|---|---|---|---|---|---|---|---|---|
| 10m | MVDepth | 0.2213 | 0.4055 | 0.2401 | 0.5154 | 0.2492 | 67.33 | 89.34 | 96.15 |
| | MVDepth-FT | 0.1905 | 0.3304 | 0.1319 | 0.4260 | 0.2221 | 71.93 | 92.75 | 97.92 |
| | DPS | 0.1963 | 0.3471 | 0.1970 | 0.4625 | 0.2297 | 72.51 | 91.25 | 96.95 |
| | DPS-FT | 0.1675 | 0.2970 | 0.1071 | 0.3905 | 0.2061 | 76.03 | 93.40 | 98.08 |
| | NAS | 0.1631 | 0.2885 | 0.1023 | 0.3791 | 0.1997 | 77.12 | 83.94 | 98.31 |
| | CNM | 0.1602 | 0.2751 | 0.0819 | 0.3602 | 0.2030 | 76.81 | 94.49 | 98.64 |
| | DELTAS | 0.1548 | 0.2671 | 0.0889 | 0.3541 | 0.1860 | 79.66 | 95.28 | 98.77 |
| | Ours-EST(concat) | 0.1458 | 0.2554 | 0.0745 | 0.3436 | 0.2065 | 79.82 | 95.19 | 98.77 |
| | Ours-EST(adaptive) | **0.1465** | **0.2528** | **0.0729** | **0.3382** | **0.1967** | **80.36** | **95.52** | **98.86** |
| 5m | Neuralrgbd | 0.2334 | 0.4060 | 0.2163 | 0.5358 | 0.2516 | 68.03 | 89.94 | 96.47 |
| | Ours-EST(concat) | 0.1458 | 0.2554 | 0.0745 | 0.3435 | 0.2065 | 79.82 | 95.20 | 98.77 |
| | Ours-EST(adaptive) | **0.1465** | **0.2528** | **0.0729** | **0.3382** | **0.1967** | **80.36** | **95.52** | **98.86** |

Table 4. The usefulness of ContextNet and EST transformer. We test models with various settings on SUN3D [8] dataset.

| ContextNet | Transformer | Inference type | Abs Rel | Abs | Sq Rel | RMSE | RMSE log | $\sigma < 1.25$ | $\sigma < 1.25^2$ | $\sigma < 1.25^3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | Independent | 0.1338 | 0.3333 | 0.0994 | 0.4897 | 0.1881 | 80.89 | 94.64 | 98.28 |
| ✗ | ✓ | Joint | 0.1391 | 0.3429 | 0.1291 | 0.4927 | 0.1877 | 81.36 | 94.32 | 97.80 |
| ✗ | ✓ | ESTM | 0.1345 | 0.3319 | 0.1073 | 0.4822 | 0.1858 | 81.43 | 94.73 | 98.12 |
| ✓ ResNet-18 | ✗ | Independent | 0.1253 | 0.3213 | 0.0873 | 0.4623 | 0.1759 | 83.31 | 95.91 | 98.49 |
| ✓ ResNet-18 | ✓ | Joint | 0.1269 | 0.3180 | 0.0933 | 0.4605 | 0.1758 | 83.61 | 95.58 | 98.25 |
| ✓ ResNet-18 | ✓ | ESTM | 0.1262 | 0.3160 | 0.0897 | 0.4580 | 0.1756 | 83.63 | 95.68 | 98.31 |
| ✓ ResNet-50 | ✗ | Independent | 0.1258 | 0.3220 | 0.0897 | 0.4657 | 0.1894 | 82.82 | 95.55 | 98.33 |
| ✓ ResNet-50 | ✓ | Joint | 0.1243 | 0.3133 | 0.0883 | 0.4556 | 0.1910 | 83.52 | 95.60 | 98.30 |
| ✓ ResNet-50 | ✓ | ESTM | 0.1254 | 0.3137 | 0.0884 | 0.4554 | 0.1913 | 83.43 | 95.68 | 98.33 |

Table 5. The effect of different backbone for *ContextNet*. We run our models using ResNet-18 and ResNet-50 as *ContextNet* respectively on SUN3D dataset by ESTM inference operation.

| Method | ScanNet | | | | | 7scenes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Abs Rel | Abs | Sq Rel | RMSE | $\sigma < 1.25$ | Abs Rel | Abs | Sq Rel | RMSE | $\sigma < 1.25$ |
| Ours (ResNet-18) | 0.0869 | 0.1600 | 0.0393 | 0.2313 | 92.56 | 0.1522 | 0.2572 | 0.0852 | 0.3398 | 80.36 |
| Ours (ResNet-50) | **0.0812** | **0.1505** | **0.0298** | **0.2199** | **93.13** | **0.1465** | **0.2528** | **0.0729** | **0.3382** | **80.36** |

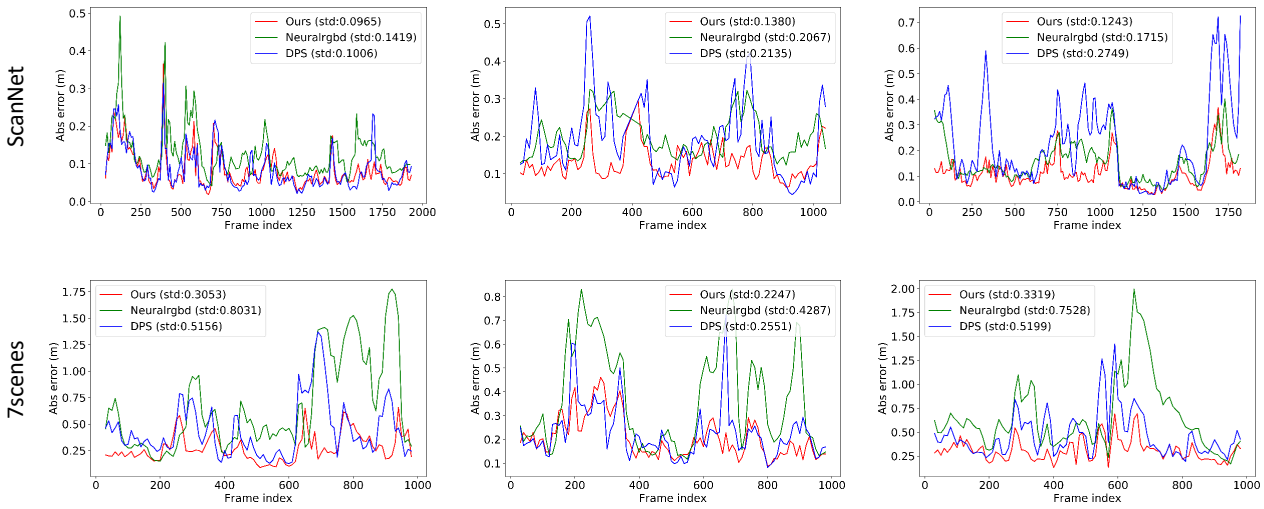


Figure 1. Depth accuracy diagram of whole video frames. We plot the mean absolute errors of the estimated depth maps of several videos in ScanNet and 7scenes. According to the curves and standard deviation, our model generates more temporally consistent depth maps.

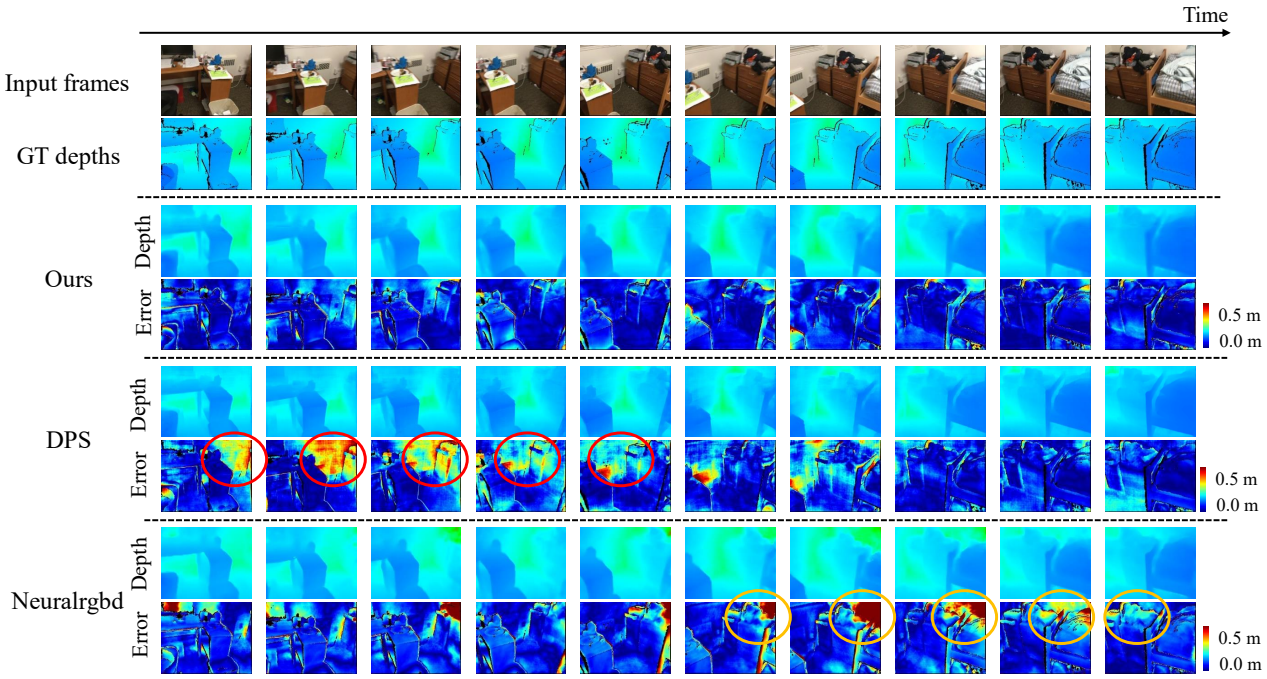


Figure 2. Depth comparisons of ten consecutive video frames. The estimated depth maps of model method are more temporally coherent than those of other models.

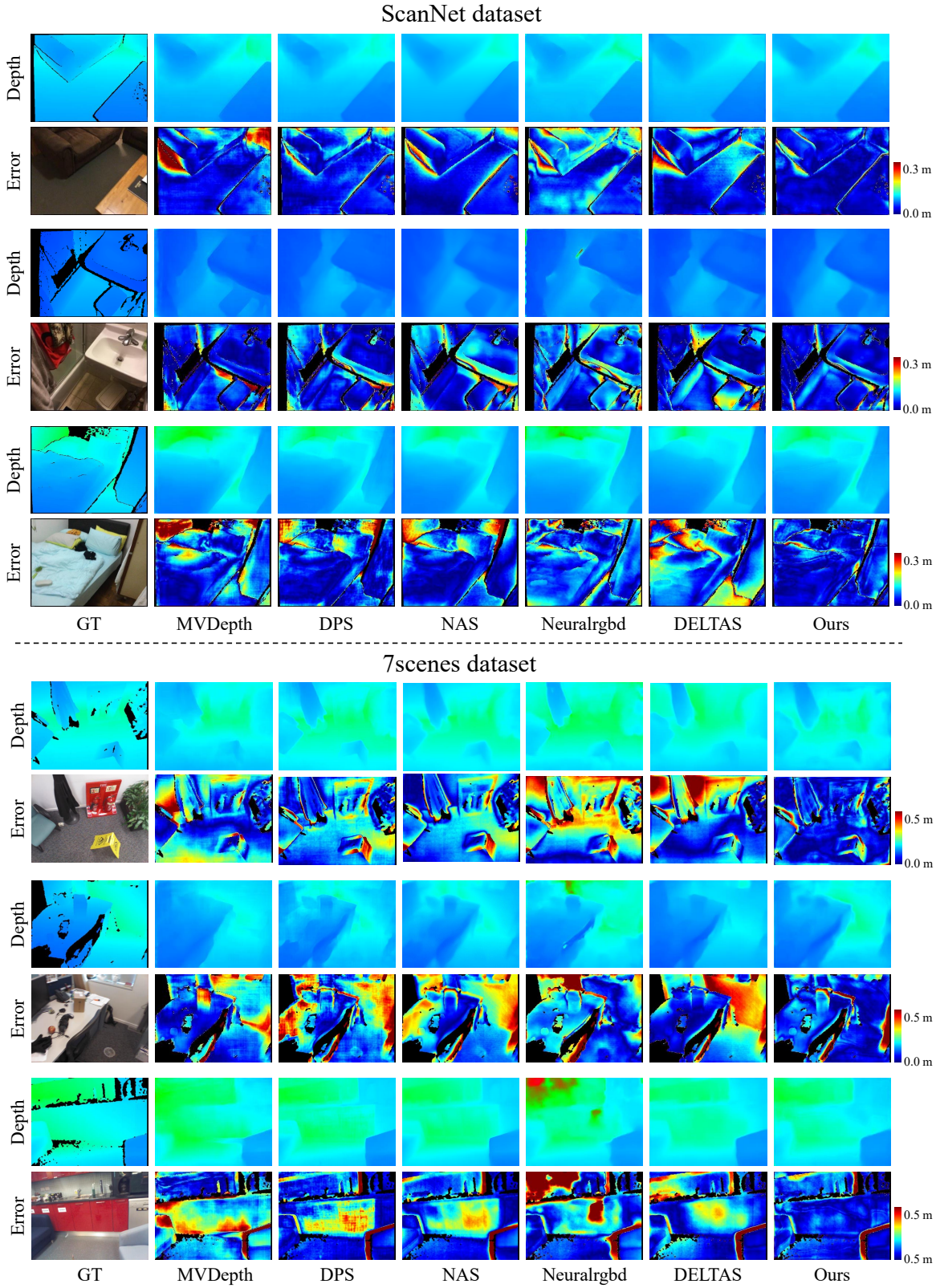


Figure 3. Qualitative depth comparisons.

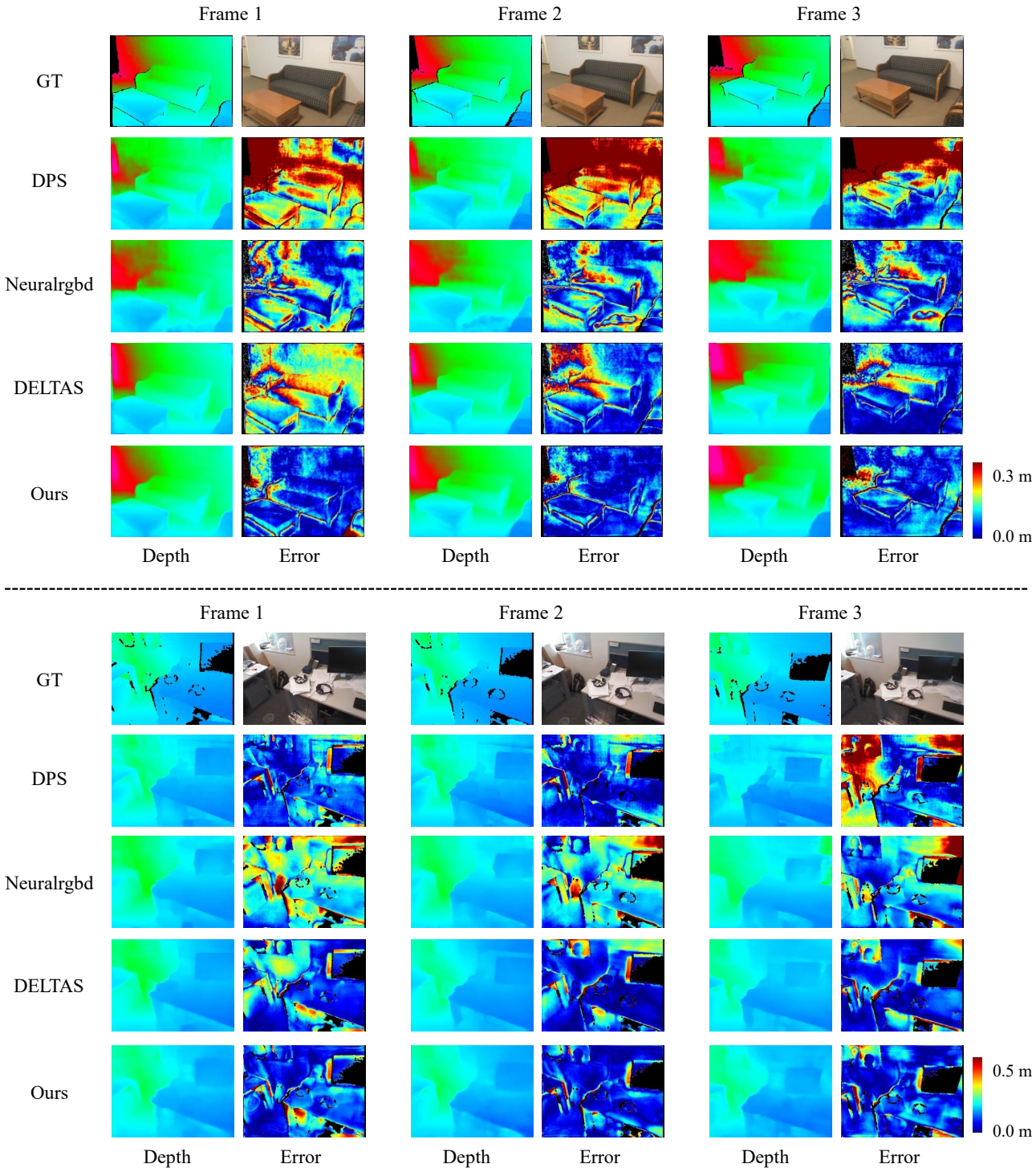


Figure 4. Depth comparisons of three consecutive video frames. The estimated depth maps of model method are more temporally coherent than those of other models.