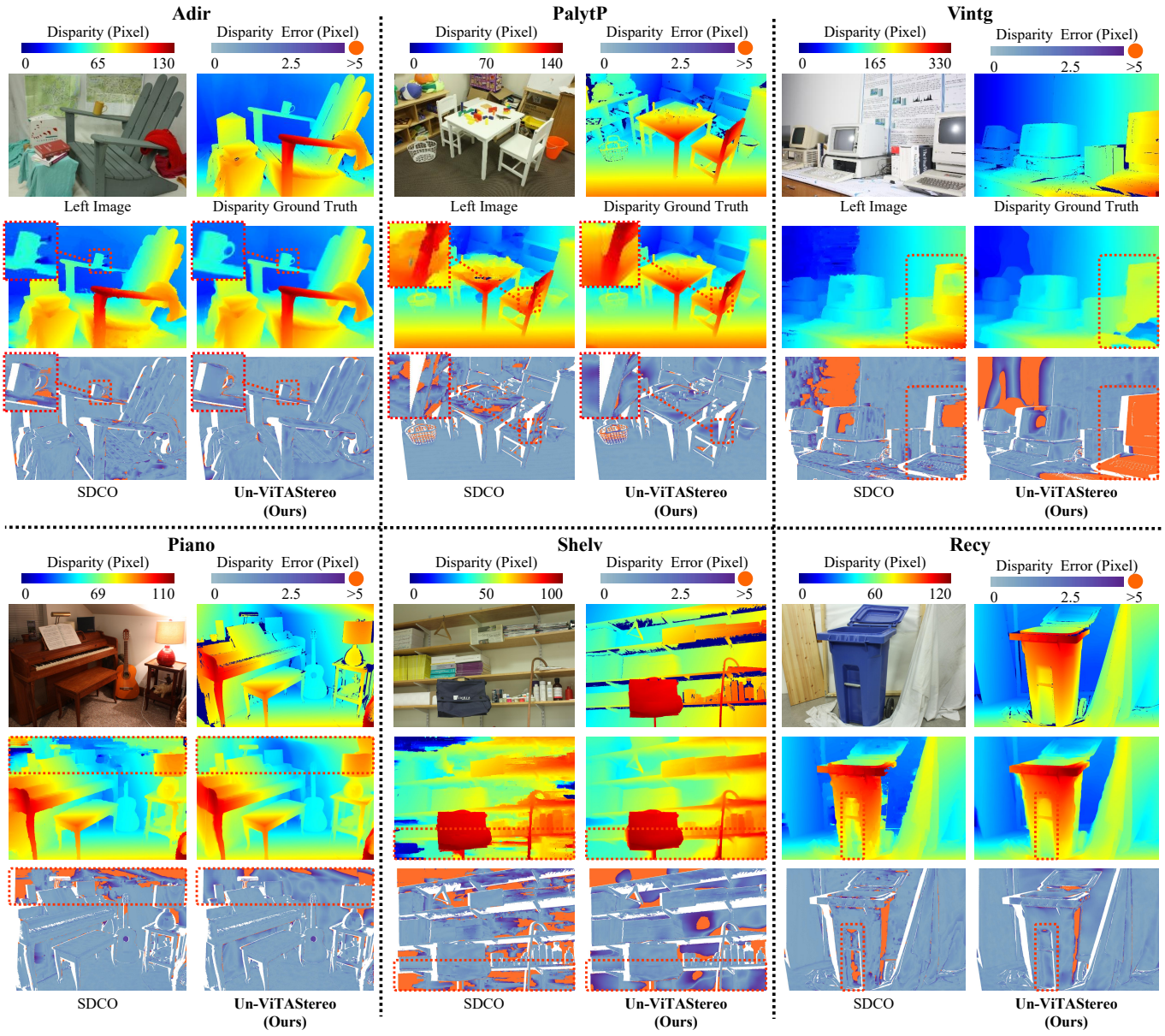

Fig. 1. Qualitative results of Un-ViTAStereo and SDCO on the Middlebury dataset.

SUPPLEMENTARY MATERIAL

A. Comparisions on the Middlebury Dataset

We have further compared our proposed unsupervised ViTAStereo [3] (Un-ViTAStereo) with SDCO [1] on the Middlebury dataset [2] in terms of end-point error (EPE). The qualitative and quantitative experimental results are presented in Fig. 1 and Table I, respectively. It can be observed that our proposed Un-ViTAStereo demonstrates superior stereo matching accuracy in 10 out of the 15 scenes. However, ViTAStereo also exhibits notably higher EPE in two scenes, the *Jade* and *Vintg*. The disparity maps visualizations in Fig. 1 suggest that this performance degradation is caused by the excessively large ground-truth disparities that exceed the maximum disparity range supported by Un-ViTAStereo, which has been set to 192 during the network training process. While in scenes with typical disparity ranges, Un-ViTAStereo exhibits superior stereo matching accuracy in areas with both smooth and discontinuous disparities.

B. Ablation Study on Relative Depth Maps

High-quality monocular depth estimation results help improve the accuracy and efficiency of our proposed loss functions in transferring 3D geometric knowledge to a stereo matching network. In response to your suggestion, we have conducted ablation studies on the SceneFlow dataset [6] by taking input as relative depth maps generated from various Depth Anything models and ViT sizes. The quantitative results are presented in Table II. It can be observed that acquiring monocular relative depth results from different monocular depth estimation VFMs results in an EPE variation of less than

TABLE I
COMPARISONS WITH SDCO [1] ON THE MIDDLEBURY DATASET [2]. ALL RESULTS ARE IN EPE (PIXEL). THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Method	Adir	ArtL	Jade	Motor	MotorE	Piano	PianoL	Pipe	Playr	Playt	PlaytP	Recy	Shelv	Teddy	Vintg
SDCO [1]	1.49	3.62	**8.14**	**2.42**	**2.43**	3.27	8.34	**4.81**	4.57	32.0	2.14	2.57	9.29	1.15	**6.51**
Un-ViTAStereo	**1.23**	**3.42**	56.4	2.75	2.57	**1.53**	**3.55**	4.99	**2.46**	**2.81**	**1.41**	**1.40**	**4.34**	**1.02**	16.6

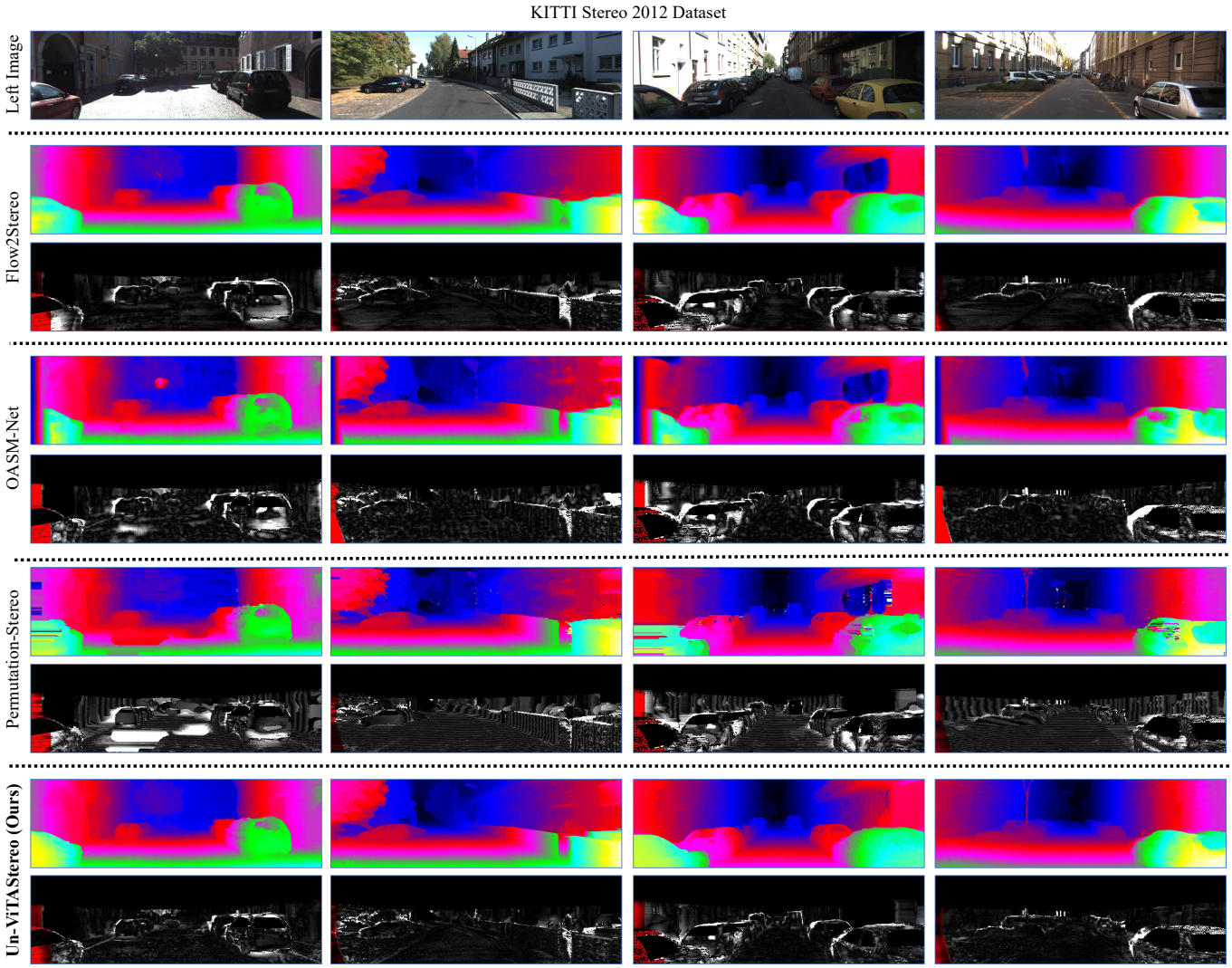

Fig. 2. Additional comparisons with SoTA unsupervised stereo matching networks, including OASM-Net [7], Flow2Stereo [8], and Permutation-Stereo [9], published on the KITTI Stereo 2012 benchmark [10]. The images in the first row of each method represent the estimated disparity maps and images in the second row denote the visualizations of D1 error.

TABLE II
ABLATION STUDY ON THE MONOCULAR DEPTH ESTIMATION VFMS.

VFM	Depth Anything V1 [4]		Depth Anything V2 [5]	
	ViT Large	ViT Base	ViT Large	ViT Base
EPE (pixel)	2.78	2.86	2.75	2.90

6%, demonstrating the robustness of our knowledge transfer-based loss functions to relative depth maps of varying quality.

REFERENCES

[1] X. Cheng *et al.*, "A novel cell structure-based disparity estimation for unsupervised stereo matching," *IET Image Processing*, vol. 16, no. 6, pp. 1678–1693, 2022.
[2] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition: 36th German Conference (GCPR)*. Springer, 2014, pp. 31–42.
[3] C.-W. Liu *et al.*, "Playing to vision foundation model's strengths in stereo matching," *IEEE Transactions on Intelligent Vehicles*, 2024, DOI:10.1109/TIV.2024.3467287.
[4] L. Yang *et al.*, "Depth Anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 371–10 381.

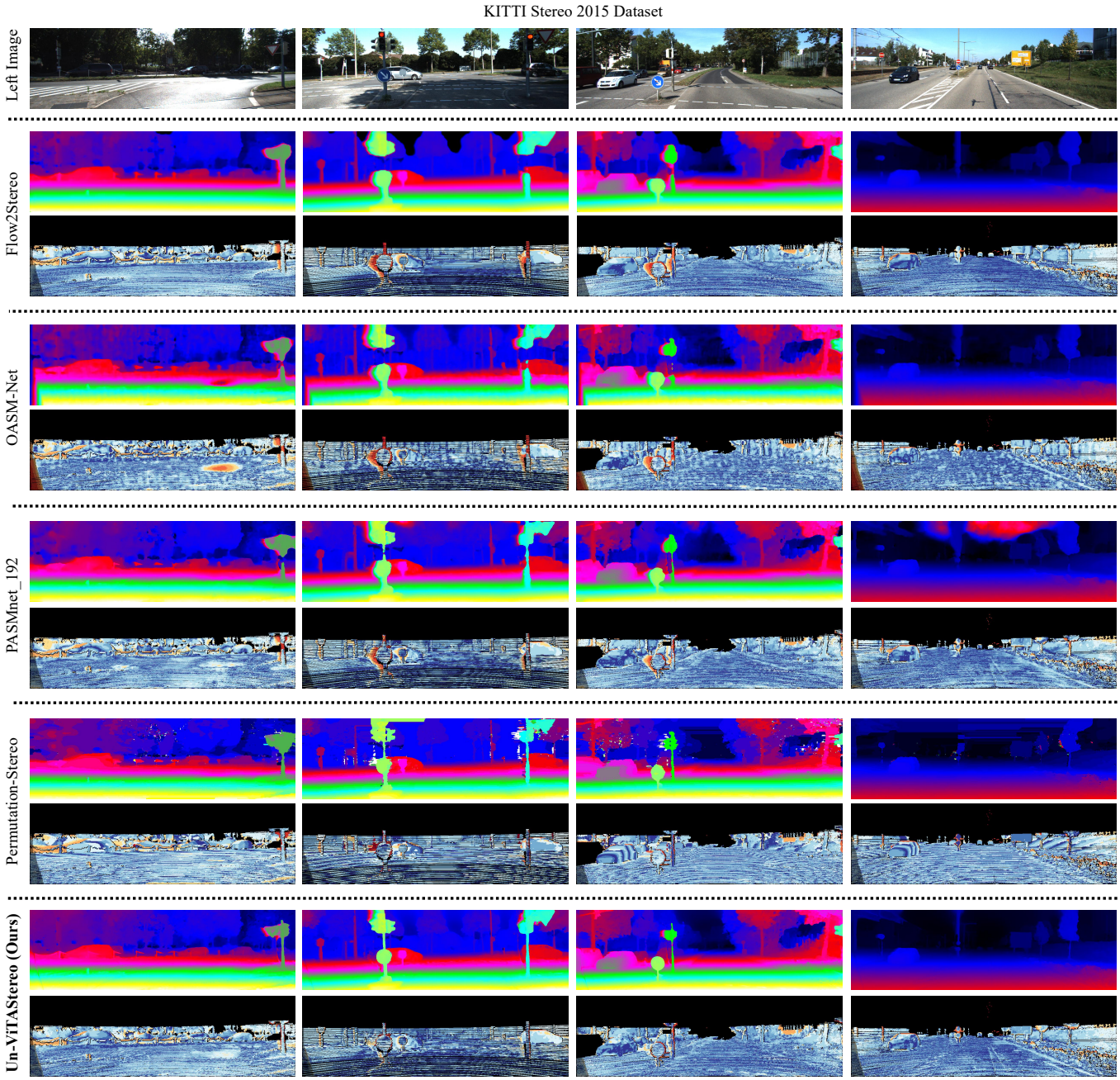

Fig. 3. Additional comparisons with SoTA unsupervised stereo matching networks, including OASM-Net [7], Flow2Stereo [8], PASMnet_192 [11] and Permutation-Stereo [9], published on the KITTI Stereo 2015 benchmark [12]. The images in the first row of each method represent the estimated disparity maps and images in the second row denote the visualizations of D1 error.

[5] L. Yang *et al.*, "Depth Anything V2," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 21 875–21 911, 2024.

[6] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[7] A. Li and Z. Yuan, "Occlusion aware stereo matching via cooperative unsupervised learning," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 197–213.

[8] P. Liu *et al.*, "Flow2Stereo: Effective self-supervised learning of optical flow and stereo matching," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6648–6657.

[9] P.-A. Brousseau and S. Roy, "A permutation model for the self-supervised stereo matching problem," in *2022 19th Conference on Robots and Vision (CRV)*. IEEE, 2022, pp. 122–131.

[10] A. Geiger *et al.*, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[11] L. Wang *et al.*, "Parallax attention for unsupervised stereo correspondence learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2108–2125, 2020.

[12] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070.

[13] T. Schops *et al.*, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp.

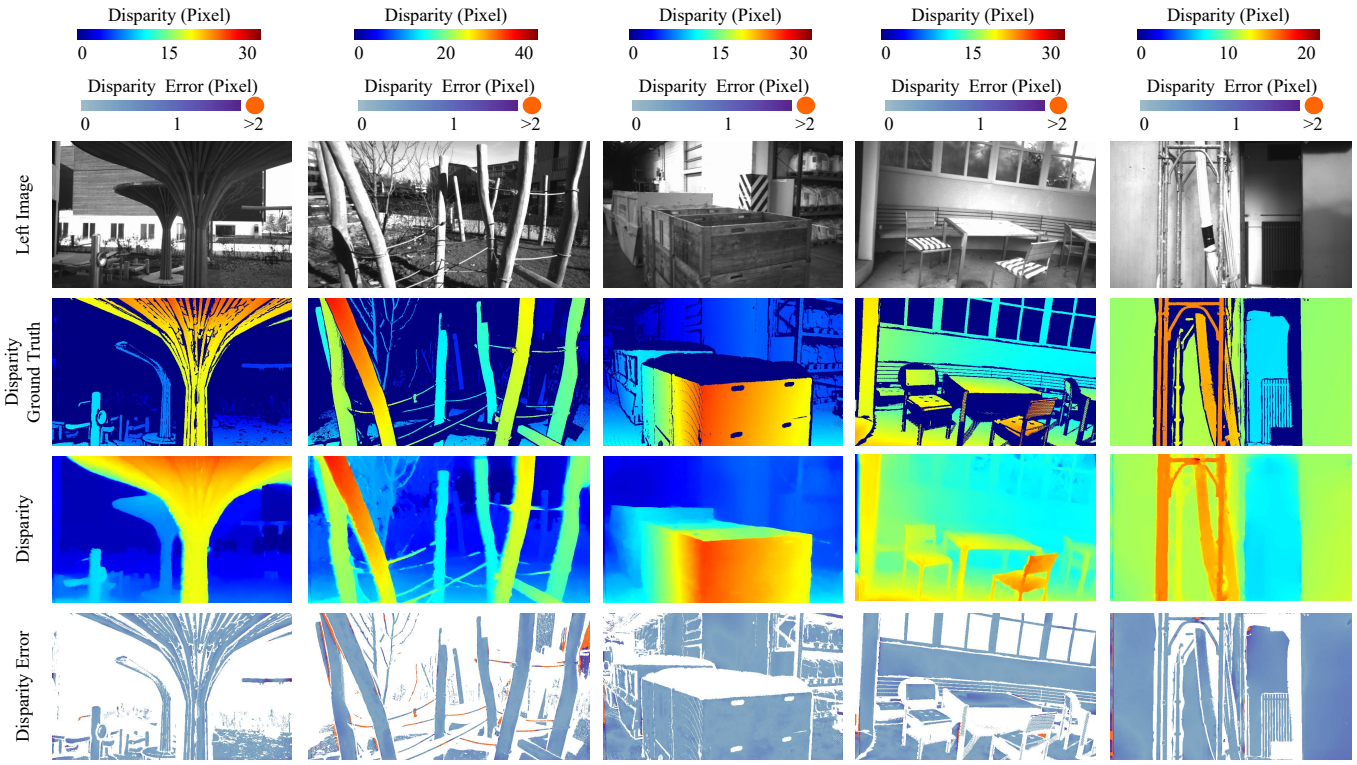

Fig. 4. Visualizations of disparity estimation results of Un-VITAStereo on the ETH3D dataset. [13]

3260–3269.