Global Occlusion-Aware Transformer for Robust Stereo Matching Supplementary Material

Zihua Liu[1], Yizhou Li[2], and Masatoshi Okutomi[3]
Tokyo Institute of Technology, Japan
{zliu[1],yli[2]}@ok.sc.e.titech.ac.jp,mxo@ctrl.titech.ac.jp[3]

1. More Implementation Details

1.1. Details About the Context Adjustment Layer.

The context adjustment layer in *GOAT* is designed to refine the disparity map from a mono-depth aspect. We employ a similar architecture adopted in STTR [7], which is a simple refinement module comprised of multiple ResBlocks [3]. The architecture of the context adjustment layer is demonstrated in Figure 1. It can recover disparity details simply by using the left image I_{left} and the current disparity I_{init} as the guidance to regress the disparity residual D_{res} and derive the final disparity D_{final}. Such an image-based refinement module can help refine the disparities in extremely large occluded regions where no matching clues can be utilized.

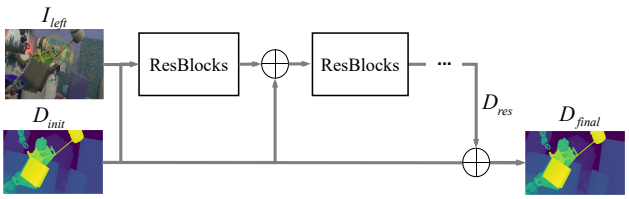

Figure 1. Context Adjustment Layer

1.2. Occlusion Mask Generation for Supervision

SceneFlow: Because the SceneFlow [9] dataset's ground-truth disparity for the left view and right view are both annotated for all pixels, it becomes feasible to generate dense ground-truth occlusion mask M_{occ} by directly applying left-right consistency check [4]. The process can be described as follows:

$$M_{occ} = \begin{cases} 1 & \text{if } D_{gap} \geq 1, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

$$D_{gap} = |D_L(x,y) - D_R(x + D_L(x,y), y)|, \quad (2)$$

where D_{gap} is the disparity difference between the corresponding pixels at the left and right views, and D_L and D_R are ground-truth disparity maps for left and right views, respectively.

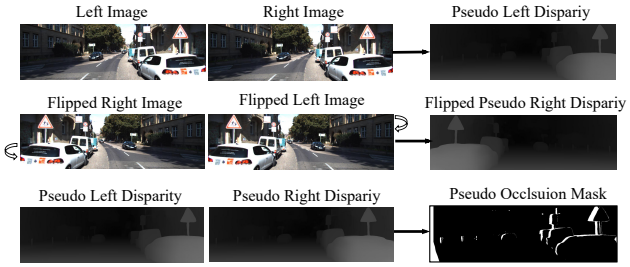

Figure 2. Flipped inference consistency check for pseudo occlusion mask generation.

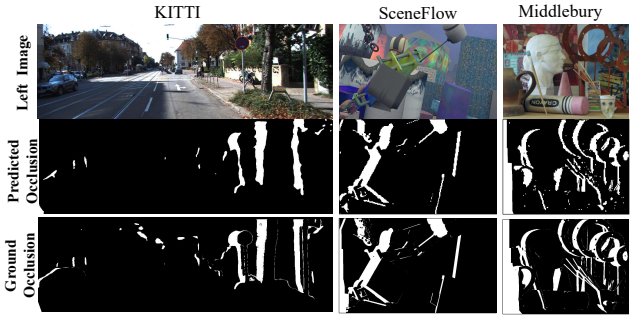

Figure 3. Predicted occlusion masks and corresponding ground truth on different datasets.

KITTI and Middlebury: Generating the ground-truth occlusion mask for the KITTI dataset [10] poses two significant challenges: Firstly, KITTI's ground-truth disparity maps are derived through LiDAR, resulting in sparse disparity annotation. Secondly, the KITTI dataset only provides the ground-truth disparity maps for the left images, which makes it difficult to directly apply the left-right consistency check to generate the occlusion masks. The STTR [7] attempts to mitigate these challenges by solely focusing on the out-of-bound occlusion mask, which comprises only the pixels that fall outside the field of view (FOV) of the right image. However, this method does not provide sufficient supervision for the occlusion mask estimation. To tackle this problem, we design a pipeline named flipped inference consistency check to generate the pseudo occlusion masks.

As demonstrated in Figure 2, we leverage a model pre-trained on the KITTI dataset without occlusion supervision as the disparity generator. Initially, the left-right image pair is employed as input to produce a dense pseudo-left disparity. Subsequently, the left view image is horizontally flipped to create a new right view, and similarly, the right view image is flipped to generate a new left view. The new left-right image pair will be sent to the identical disparity generator, where the flipped pseudo-right disparity is obtained. Finally, the left-right consistency check between the pseudo disparity maps for the left view and the right view is applied to generate a pseudo occlusion mask. For Middlebury dataset [11], we use the same strategy for occlusion mask generation. Figure 3 shows some examples of estimated occlusion masks by proposed *GOAT* and their corresponding ground truth on the SceneFlow, KITTI, and Middlebury datasets.

2. More Training Details

In this section, we present further information regarding the training process on different datasets.

2.1. Data Augmentation

A domain gap exists between synthetic SceneFlow datasets and real-world KITTI and Middlebury datasets in terms of color and disparity distribution. This poses a challenge for fine-tuning with SceneFlow pre-trained models, which is further compounded by limited annotated training data in real-world KITTI and Middlebury datasets. To enhance the network's robustness and mitigate overfitting, we employ augmentations as follows.

Asymmetric and Symmetric Chromatic Augmentations: To address the issue of diverse lighting and exposure conditions in real-world stereo images, we adopted a method similar to that used in HSM [18]. This involved modifying the brightness, contrast, and gamma of both left and right images with random adjustment parameters from intervals of $[0.8, 1.2]$, with the option of using different parameters for the left and right images. This allowed us to simulate color and exposure variations commonly observed in real scenes.

Color Domain Adaption: To alleviate the difference in color distribution between synthetic data and real-world data, we used color domain adaptation augmentation following [13]. This method utilizes normalization techniques in the LAB color space to reduce the distribution gap between the two types of data.

Vertical y-offset and Flip: To simulate the disparity drift problem caused by imperfect calibration, we applied the y-offset augmentation from [18], which randomly shifts the y-direction pixels in the right image by an offset within [-2,2] pixels. We also utilized symmetric vertical flipping for both left and right views to improve disparity estimation accuracy across all image regions and prevent location bias.

Asymmetric Masking: Similar to [18], we replaced the random patches of the right images with mean values of the whole images. By applying this, it will increase the proportion of occluded areas, making the Occlusion-Aware Global Aggregation Module (*OGA*) more effective. The size of the patch to be replaced was randomly sampled between [40,40] and [120,180].

2.2. Training Setup

SceneFlow: As described in the paper, for training on the SceneFlow dataset, we use all three sub-sets (Flyingthings3D, Driving, and Monkaa) within a total of 35K images. We consider a random crop of 320×640 with a batch size of 8 and a maximum disparity of 192. The whole training process on SceneFlow is performed with 4 NVIDIA RTX 3090 GPUs without data augmentation.

KITTI: In regards to the fine-tuning on the KITTI 2015 dataset, the color domain adaption described in Section 2.1 was initially employed to fine-tune the pre-trained model for an additional 40 epochs on the SceneFlow dataset, utilizing a learning rate of 1e-4. Following this, we used mixed datasets containing KITTI 2012 and KITTI 2015 with in total of 400 pairs of images for the training of the first 400 epochs. We chose the model with the best performance on the validation set, followed by another 200 epochs of fine-tuning on the KITTI 2015 training set to obtain the final model. The whole training process was conducted on 2 NVIDIA RTX6000 GPUs, with a patch size of 320×1088.

Middlebury: To address the limited amount of training data in the Middlebury dataset, we followed the similar strategy utilized in [6] by augmenting the Middlebury dataset to the 20% amount of the SceneFlow dataset with techniques mentioned in 2.1. We conducted mixed training for 100 epochs and then fine-tuned the model on the Middlebury dataset alone for another 500 epochs to get the final model.

3. Intermediate Outputs for the OGA Module

Similar to [8], the proposed *OGA* module is a GRU-based iterative refinement module. To further demonstrate how the *OGA* module uses global correlation to refine the disparities in the occluded area, we illustrate intermediate disparity outputs of the *OGA* module of each iteration in the KITTI dataset in Figure 4.

The *OGA* module applies a global attention mechanism to aggregate features within occluded regions, thereby enhancing the accuracy of disparity estimations. Our findings indicate that as the number of iterations of the *OGA* module increases, the estimated disparity progressively improves. It is worth noting that applying the *OGA* module only once can already yield considerable enhancement in the disparity

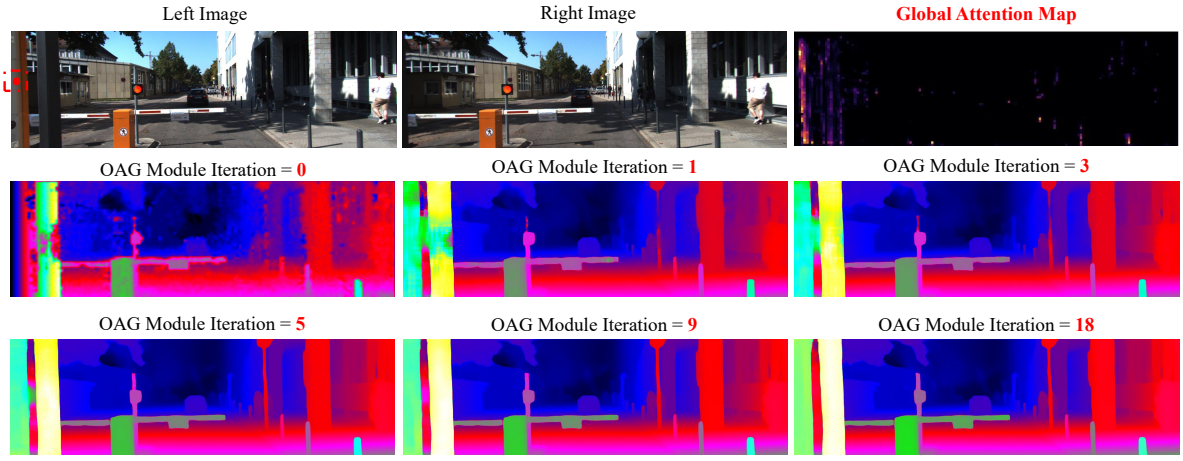

Figure 4. **Intermediate Outputs for *OGA* module.** The third image of the first row illustrates the global attention map of the location marked by a red dot in the left image. Images in the second row and the third row demonstrate the disparity estimation results at the different iterations using the proposed *OGA* module. Through iterative refinement using the *OGA* module, the disparity estimation improves progressively, particularly in occluded regions.

estimation compared with the preliminary results as shown in Figure 4.

4. Additional Experiments Results

4.1. More Comparisons on the SceneFlow dataset

'In this subsection, we will provide more comparisons with the proposed *GOAT* and other SOTA methods on the SceneFlow dataset. Based on the results presented in Table 1, it is apparent that the proposed *GOAT* surpasses all other evaluated approaches in regards to both the overall end-point-error (EPE) and the EPE at the occluded regions(EPE-Occ). Furthermore, the proposed approach demonstrates a lower rate of outliers, as indicated by P1 and P3 values.

Furthermore, an important aspect of efficient neural network architecture is the training parameters (Params) as well as their computing complexities, which are often quantified by the number of multi-accumulate operations (Macs). As shown in Table 2, our *GOAT-T* with 1/8 resolution can achieve a competitive performance with the small-est Macs and rather small parameters compared with the latest PCWNet [12] and IGEVStereo [16]. However, transformer-based methods inherently suffer from quadratic computational complexity where the *GOAT-B* has 4 times larger Macs when increasing the resolution to 1/4.

We present a comprehensive analysis with visualization of the estimated disparity on the SceneFlow dataset which can be inferred in Figure 6. Compared with other notable networks, the proposed *GOAT* can generate better disparity estimation results around the thin structures and in the texture-less regions with the assistance of the proposed parallel disparity and occlusion estimation module (*PDO*). Besides, our proposed method displays strong robustness in

Table 1. we compare the performance of our proposed methods with other notable works on the **SceneFlow** dataset. We present the end-point-error (EPE) results for disparities in overall (All) regions, as well as occluded regions (Occ), and report P1 and P3 errors. The best-performing method is highlighted in **Bold**.

Method	EPE		P1 Error	P3 Error
	EPE-All	EPE-Occ		
DispNetC [9]	1.68	-	-	-
StereoNet [5]	1.07	3.31	13.7%	5.3%
AANet++ [17]	0.72	2.44	10.4%	4.0%
PSMNet [1]	1.09	3.14	11.1%	4.6%
GANet [19]	0.84	2.83	8.8%	3.8%
GwcNet [2]	0.77	2.47	8.7%	3.9%
RAFT-Stereo [8]	0.69	2.14	7.9%	3.3%
STTR-light [7]	4.14	23.9	16.4%	10.7%
ACVNet [15]	0.48	1.65	**6.2%**	2.9%
EDNet [20]	0.63	2.08	8.2%	3.9%
IGEVStereo [15]	**0.47**	1.62	6.6%	3.3%
PCW-Net [12]	0.86	2.54	9.1%	4.0%
GOAT (Ours)	**0.47**	**1.53**	**5.6%**	**2.7%**

Table 2. Numbers of training parameters (Params) and multi-accumulate operations (Macs) compared with other lastest methods. We use an input resolution of 320×640 for Mac's computation.

Method	Params	Macs	EPE
ACVNet [15]	7.1M	465.1G	0.48
RAFTStereo [8]	11.1M	654.8G	0.69
PCW-Net [12]	35.8M	768.6G	0.86
IGEVStereo [16]	12.6M	541.6G	0.47
GOAT-T (Ours)	10.0M	192.0G	0.56
GOAT B(Ours)	12.1M	858.3G	0.47

Table 3. Ablation study of our proposed *GOAT* on the FallingThings [14] Dataset. *"PDO"* is short for Parallel Disparity and Occlusion Estimation Module. *"OGA"* is short for Iterative Occlusion-Awareness Global Aggregation Module. We calculated the EPE and P1(outliers) both in the overall and the occluded regions, separately."*" means a higher resolution.

Method	Disparity Estimation		Update Module		CA Layer		EPE		P1(%)		Occ mIOU	Res
	Cost Volume	PDO	RAFT	OGA			All	Occ	All	Occ		
Baseline	✓		✓				0.53	2.30	5.9%	28.9%	-	1/8
PDO		✓	✓				0.41	1.65	5.2%	26.7%	0.905	1/8
PDO + OGA		✓		✓			0.31	1.22	3.6%	20.2%	0.905	1/8
PDO + OGA + CA		✓		✓	✓		0.29	1.18	3.4%	19.1%	0.906	1/8
PDO + OGA + CA*		✓		✓	✓		0.25	1.11	2.7%	17.6%	0.913	1/4

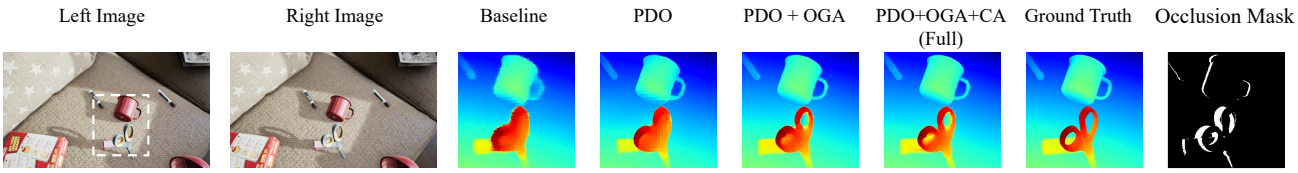

Figure 5. Visualizations of ablation studies on FallingThings Datset. We cropped and enlarge the selected part of the disparity map for easier viewing.

the occluded regions where other notable approaches fail to yield a satisfactory result in such ill-conditioned regions.

4.2. Supplementary Ablation Studies on the FallingThings Dataset.

Besides ablation studies shown in the main paper, we also conduct the ablation studies on the FallingThings [14] dataset. Compared to random floating objects in SceneFlow dataset, FallingThings dataset contains scenes with carefully placed objects, thus it has more realistic semantics and occlusions. The related results can be shown in Table 3.

Compared with the Baseline, the model integrates with the *PDO* module (designated as PDO) is able to improve the overall performance by a big margin from 0.53 to 0.41. As demonstrated in Figure 5, applying the *PDO* shows better structural disparity performance where the baseline shows blur and ambiguous disparity values.

Furthermore, Table 3 exemplifies the efficacy of the *OGA* module. It demonstrates an enhancement in the performance of disparity estimation within occluded regions, reducing the disparity from 1.65 to 1.22, resulting in a 26% improvement. This pattern is also observable in Figure 5, where the PDO + OGA clearly shows better disparities in the occluded regions.

Finally, the complete model incorporated with *PDO* and *OGA* modules witnesses the best performance by showing an EPE-Occ of 1.18.

4.3. More Comparisons on the KITTI dataset.

In this section, we will provide more visualization comparison results on the KITTI 2015 test set. As illustrated in Figure 7, the proposed *GOAT* has more continuous disparity estimation results in the occluded areas (regions within the red bounding boxes.). But other most advanced methods, such as PCWNet [12] and IGEVStereo [16], have obvious outliers and disparity discontinuities. Furthermore, our proposed *GOAT* model demonstrates superior robustness compared to alternative networks, as evidenced by its ability to produce more accurate structural modeling of street scenes with fewer artifacts. This characteristic is highly advantageous for autonomous driving applications.

4.4. More Comparisons on the Middlebury dataset.

In addition to the generalization evaluation visualization mentioned in our paper, we also fine-tuned the SceneFlow pre-trained model on the Middlebury dataset with half resolution (H) following the training scheme outlined in 2.2. More visualization results can be inferred from Figure 8, where our proposed method demonstrates better structural disparity estimation and fewer artifacts.

References

[1] J. Chang and Y. Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 3

[2] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, 2019. 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[4] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994. 1

[5] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[6] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16263–16272, 2022. 2

[7] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6197–6206, 2021. 1, 3

[8] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 3

[9] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 1, 3

[10] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 1

[11] D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nesic, and P. Westling X Wang. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*, 2014. 2

[12] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–297. Springer, 2022. 3, 4

[13] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: A simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10328–10337, 2021. 2

[14] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018. 4

[15] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Acvnet: Attention concatenation volume for accurate and efficient stereo matching. *arXiv preprint arXiv:2203.02146*, 2022. 3

[16] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 3, 4

[17] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020. 3

[18] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2019. 2

[19] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[20] Songyan Zhang, Zhicheng Wang, Qiang Wang, Jinshuo Zhang, Gang Wei, and Xiaowen Chu. Ednet: Efficient disparity estimation with cost volume combination and attention-based spatial residual. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5433–5442, 2021. 3

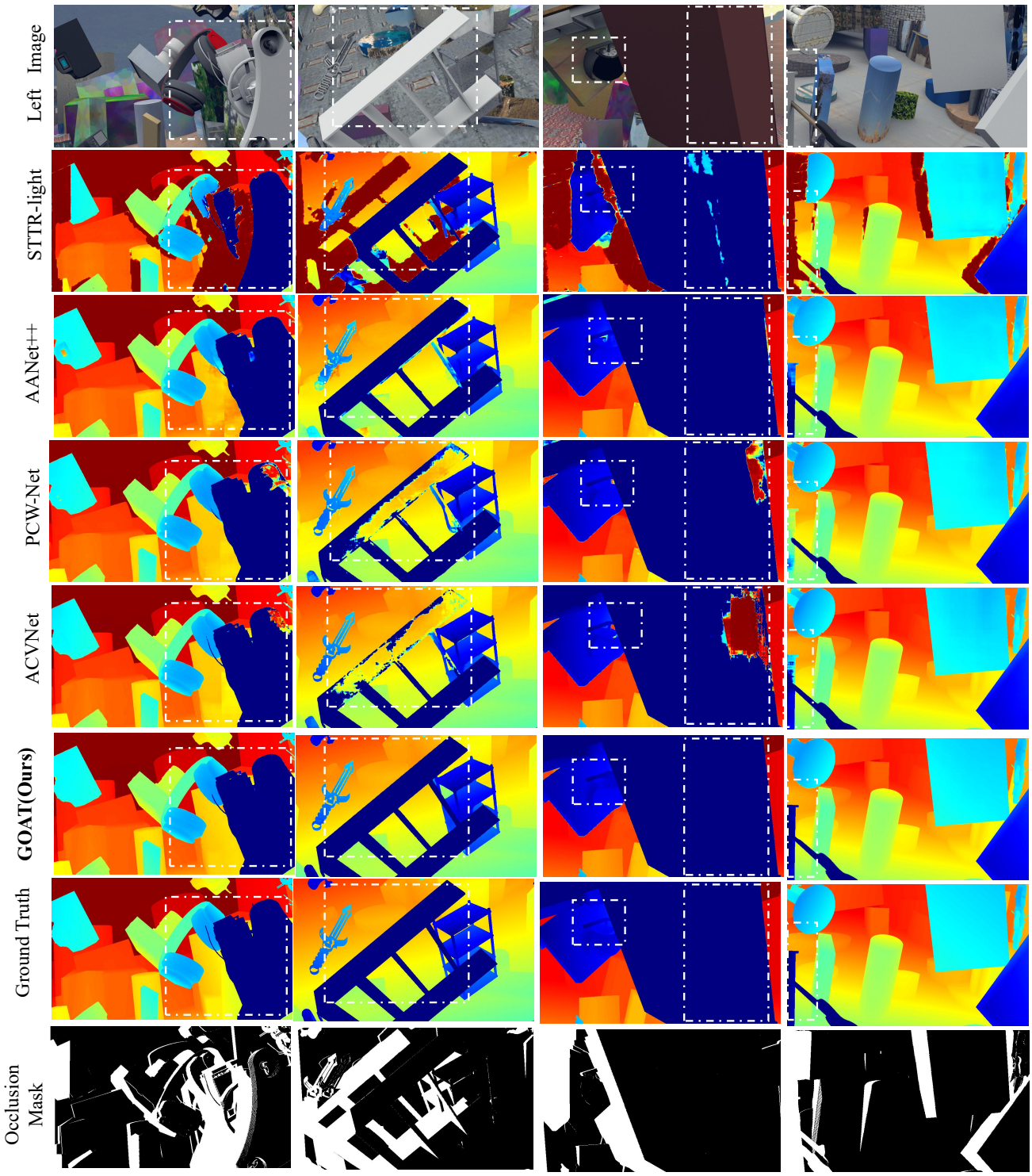

Figure 6. Visualization comparison of estimated disparities on the SceneFlow dataset. Our proposed *GOAT* demonstrates more structured and continuous disparity results in the white bounding box.

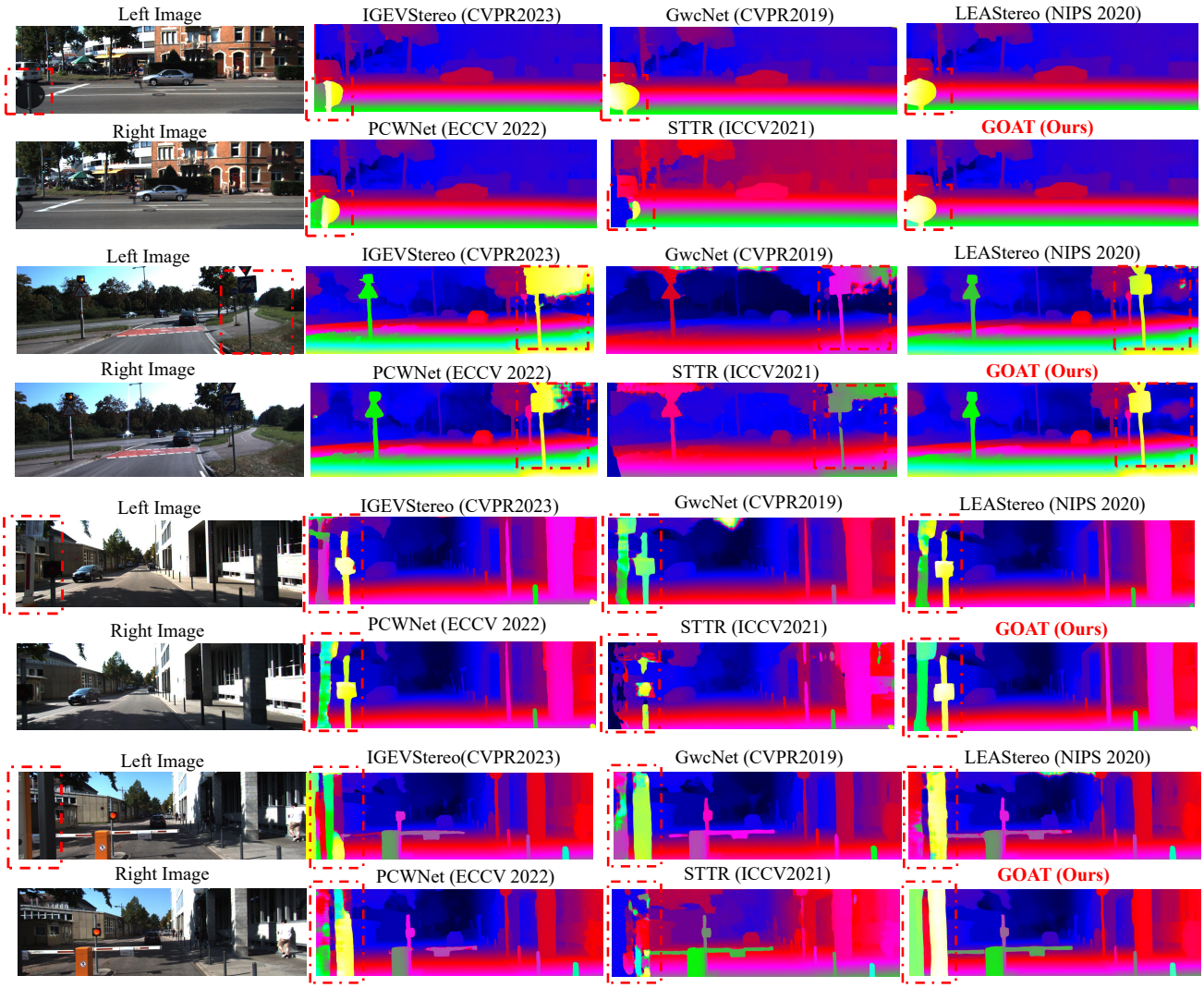

Figure 7. Visualization comparison of estimated disparities on the KITTI 2015 dataset. Note our proposed *GOAT* can generate more detailed disparities outputs, especially in the occluded regions compared with other SOTA networks.

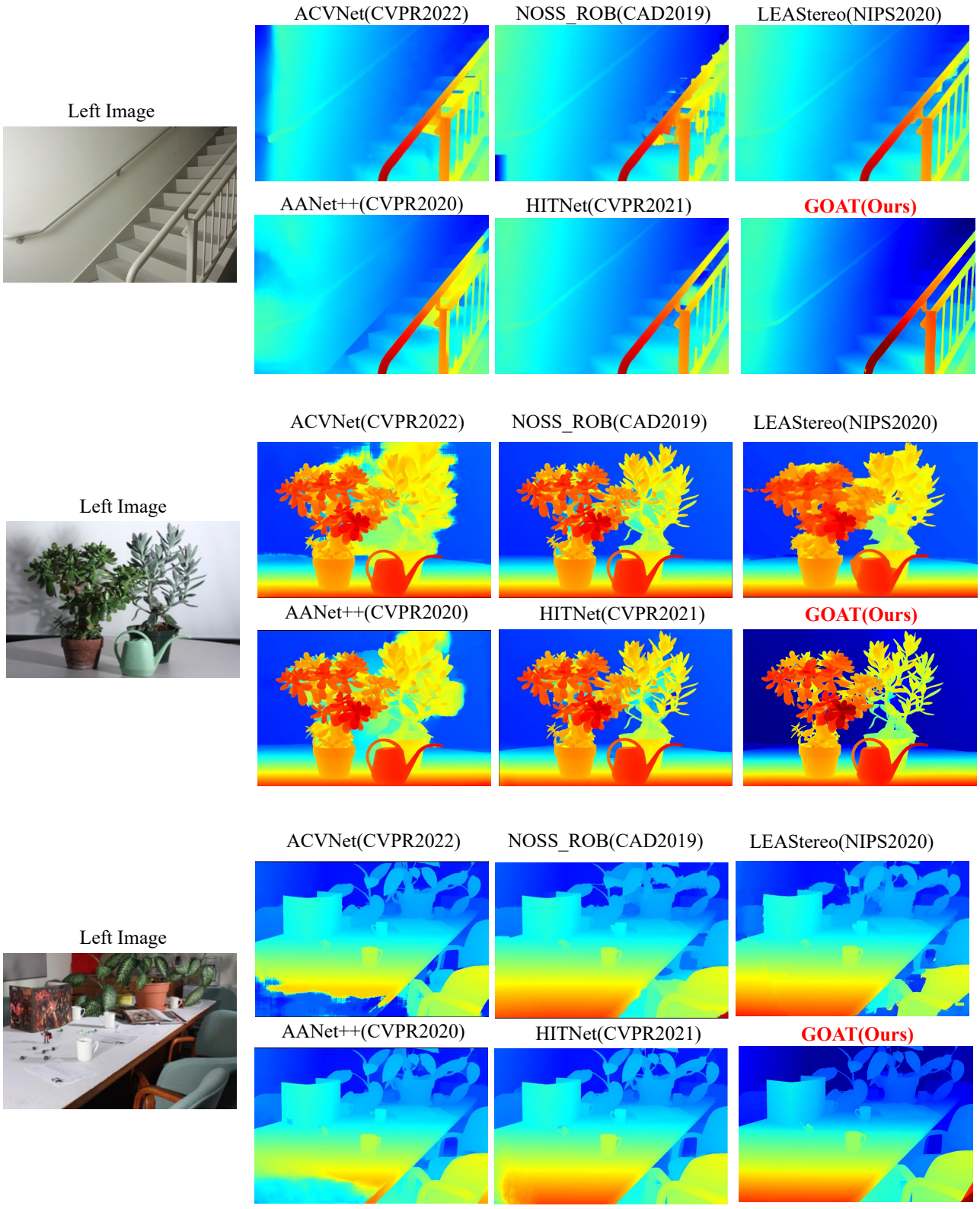

Figure 8. Visualization comparison of estimated disparities on the Middlebury test set.