Diving into the Fusion of Monocular Priors for Generalized Stereo Matching Supplementary Material

Chengtang Yao[1,2], Lidong Yu[3], Zhidan Liu[1,2], Jiaxi Zeng[1,2], Yuwei Wu[1,2]*, Yunde Jia[2,1]*
[1]Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
[3]NVIDIA

HuggingFace Demo | GitHub Repo | Model Weights Download

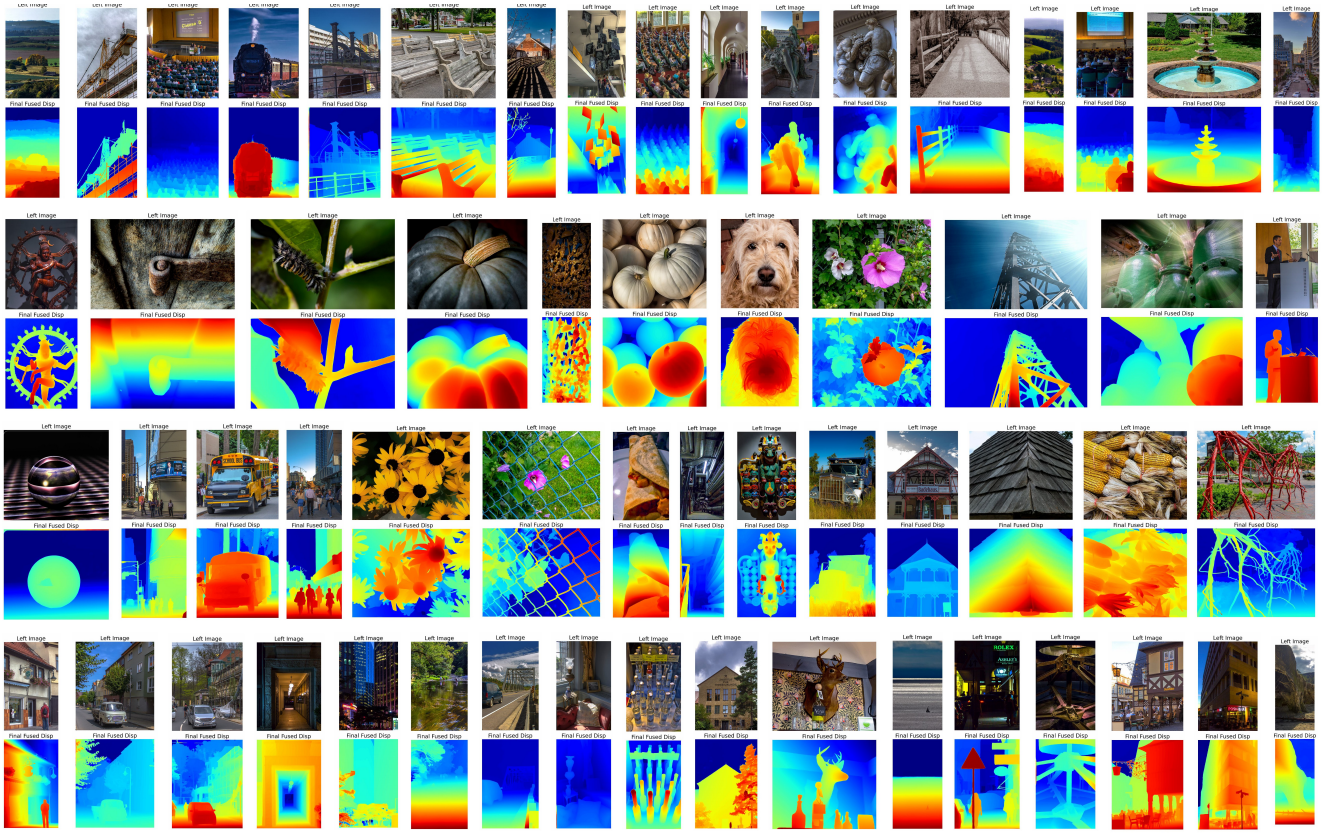

Figure 1. The visualization of results on the Flicker1024 dataset.

Contents

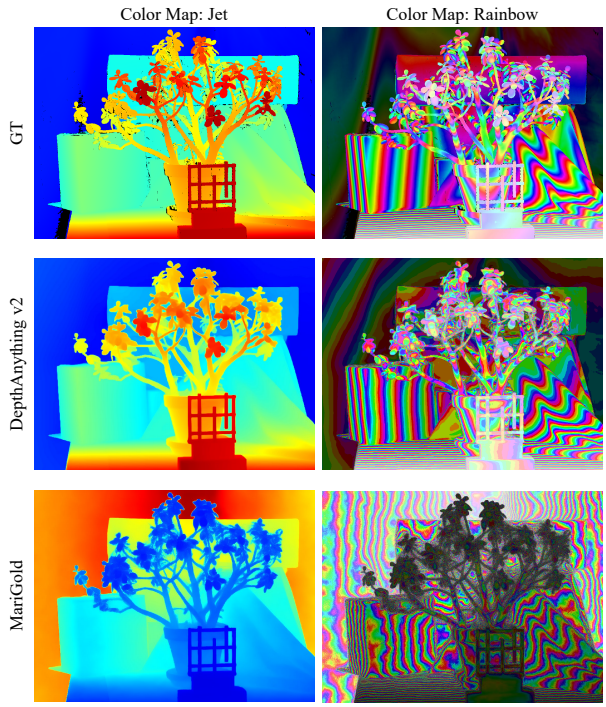

Figure 2. The visualization of results. We use two kinds of colormap to visualize the disparity map.

The training and testing codes for all experiments, including the ablation study, are available in our project. For reproducibility, we strongly recommend referring to our project.

1. Visualization On Flicker1024

We present visualization results demonstrating the generalization capability of our model from the synthetic SceneFlow dataset to the real-world Flickr1024 dataset [11]. As shown in Figure 1, our model performs robustly across diverse scenarios, including large outdoor and indoor scenes, thin and small objects, strong lighting interference and low-light conditions, as well as challenging materials such as glass windows, walls, and bottles.

2. Intuition behind Monocular Depth Model

We choose DepthAnything v2 [13] over Marigold [3] because of the superior continuity of its depth maps. As shown in Figure 2, DepthAnything v2 provides depth maps with better continuity than Marigold, especially in fine-grained regions. The depth maps from Marigold contain considerable noise, while those from DepthAnything v2 are much cleaner.

3. More Results on Booster

We provide additional results on the Booster dataset across various material types. From class 0 to 3, the materials become increasingly transparent and/or specular. As shown in Tables 1 and 2, our method outperforms state-of-the-art approaches on transparent and/or specular objects (classes 1 to 3), while achieving comparable results in normal regions (class 0). The normal regions of the Booster dataset mainly consist of regular objects, flat surfaces, or highly textured areas. Consequently, NerfStereo, which incorporates additional stereo data, performs particularly well in these regions. This indicates that stereo matching effectively captures fine-grained details, whereas monocular depth estimation excels in perceiving coarse shapes. As illustrated in Figures 3 and 4, binocular disparity provides greater detail compared to monocular depth. Our method disentangles monocular depth and binocular disparity, allowing the model to leverage both monocular and stereo data, and explore the fusion of monocular priors effectively.

4. Additional Training Data

We evaluate the scalability of our model by incorporating additional training data from the TranScene dataset [6], a synthetic dataset specifically designed for multi-label transparent scenes. In our experiments, we use labels with the largest disparity in transparent regions. It should be noted that, this time, our model is trained end-to-end using weights pretrained on the SceneFlow dataset, without using any multi-stage training strategy. As shown in Tables 3 and 4, incorporating the additional data leads to consistent performance improvements across all evaluated metrics, with particularly notable gains in transparent regions. Furthermore, our model's performance on common scenes (e.g., non-transparent regions) not only remains stable but also shows slight improvement. These results highlight the scalability potential of our model when augmented with additional large data.

5. More Analysis about Memory

We also compare our model to state-of-the-art methods in terms of memory consumption across different resolutions. To ensure a fair comparison of backbones during inference, we exclude the feature encoder module when evaluating each model's memory consumption. Notably, the memory consumption of IGEV becomes extremely high on the A40 GPU as the maximum disparity range increases. We suspect this may be a bug; therefore, we used a borrowed 4090 GPU for evaluations under the first four resolutions, while the evaluation under the last resolution was conducted on the A40 GPU.

As shown in Table 5, our method, along with RAFT-Stereo [5], maintains a slower growth rate in memory con-

Method	Additional Data/Aug	Booster							
		Class 0				Class 1			
		EPE	bad 2.0	bad 3.0	bad 5.0	EPE	bad 2.0	bad 3.0	bad 5.0
Mocha-Stereo 192[1]		1.30	6.93	5.54	4.18	2.91	23.05	17.67	13.45
Mocha-Stereo 320[1]		1.20	6.18	4.84	3.53	2.88	22.83	17.34	12.98
ELFNet [7]		2.97	14.08	11.38	8.80	5.67	24.68	19.00	14.42
Selective-RAFT [10]		1.35	8.06	6.01	4.01	3.37	27.37	21.87	17.19
Selective-IGEV 192[10]		1.46	8.03	6.19	4.66	3.61	25.57	20.05	15.93
Selective-IGEV 320[10]		1.31	7.27	5.39	3.81	3.51	25.05	19.39	15.18
IGEV 192[12]		1.17	6.67	4.84	3.46	3.76	25.46	20.26	16.39
IGEV 320[12]		1.00	6.07	4.37	2.82	3.60	24.69	19.46	15.70
NMRF [2]		2.76	17.43	13.21	9.51	4.60	32.81	26.08	19.84
NerfStereo [8]	✓	**0.73**	**4.07**	**2.55**	**1.47**	2.41	18.67	13.92	10.56
RAFTstereo [5]		1.14	5.84	4.39	3.08	3.66	25.34	19.35	14.37
RAFT-Stereo + ME		0.96	6.57	5.24	3.93	1.81	13.68	8.77	5.98
Ours		0.79	5.90	4.57	3.17	**1.53**	**12.67**	**7.80**	**4.88**

Table 1. Generalization from SceneFlow dataset to Booster dataset in quarter resolution and balanced set. ME represents our monocular encoder module. All results are evaluated in the same metrics and settings. The 192 and 320 represent the maximum disparity range used in each model.

Method	Additional Data/Aug	Booster							
		Class 2				Class 3			
		EPE	bad 2.0	bad 3.0	bad 5.0	EPE	bad 2.0	bad 3.0	bad 5.0
Mocha-Stereo 192[1]		15.68	53.56	46.23	37.77	9.45	66.44	57.96	45.73
Mocha-Stereo 320[1]		15.05	53.88	46.63	37.62	9.21	65.88	57.30	44.65
ELFNet [7]		22.74	78.89	74.81	69.70	9.03	72.07	62.73	49.82
Selective-RAFT [10]		16.12	55.66	49.87	43.04	10.34	69.84	61.64	49.55
Selective-IGEV 192[10]		20.41	57.55	49.78	42.86	9.50	66.85	58.9	47.15
Selective-IGEV 320[10]		19.81	57.35	49.27	42.10	9.29	66.02	57.91	45.86
IGEV 192[12]		18.55	54.64	46.45	37.79	10.00	68.96	61.14	49.51
IGEV 320[12]		18.00	54.50	46.05	37.72	9.74	68.55	60.49	48.22
NMRF [2]		17.36	56.34	48.33	38.18	10.36	70.92	60.93	47.16
NerfStereo [8]	✓	17.92	45.67	40.39	35.19	8.88	62.67	53.35	41.79
RAFTstereo [5]		18.58	54.00	47.52	40.44	9.79	67.69	59.31	47.40
RAFT-Stereo + ME		**5.16**	24.38	19.01	14.58	8.97	64.84	56.05	43.95
Ours		5.32	**23.34**	**17.62**	**13.50**	**7.93**	**59.83**	**50.36**	**38.44**

Table 2. Generalization from SceneFlow dataset to Booster dataset in quarter resolution and balanced set. ME represents our monocular encoder module. All results are evaluated in the same metrics and settings. The 192 and 320 represent the maximum disparity range used in each model.

sumption compared to IGEV [12], Selective IGEV [10], and Mocha [1]. Compared to RAFTStereo, our method exhibits a similar memory consumption increase across resolutions due to the resizing operation required by DepthAnything v2.

6. More Visualization

We provide additional visualizations of generalized stereo matching in Figures 5, 6, 7, 8, and 9. The visualizations span a variety of environments, ranging from open outdoor scenes (e.g., driving scenarios), to semi-open outdoor scenes (e.g., playgrounds), and to enclosed indoor scenes (e.g., rooms, tables). The results demonstrate that our method generalizes effectively to the wild world, achieving strong performance even when trained only on a limited amount of synthetic stereo data.

Metric	ALL		Trans		NoTrans	
	Ours	Ours+TranScene	Ours	Ours+TranScene	Ours	Ours+TranScene
EPE	2.26	**1.24**	7.93	**5.67**	1.52	**0.75**
RMSE	5.60	**4.19**	11.03	**8.42**	3.93	**3.07**
2px	11.02	**7.91**	59.83	**46.78**	6.98	**4.77**
3px	8.59	**5.97**	50.36	**38.55**	4.97	**3.23**
5px	6.60	**4.52**	38.44	**28.65**	3.64	**2.29**
6px	6.00	**4.08**	33.87	**25.41**	3.27	**2.01**
8px	5.35	**3.44**	27.56	**21.30**	2.89	**1.59**

Table 3. Generalization from the SceneFlow dataset to the Booster dataset in quarter resolution and balanced set. 'All', 'Trans', and 'NonTrans' represent all regions, transparent regions, and nontransparent regions, respectively.

Metric	Class 0		Class 1		Class 2		Class 3	
	Ours	Ours+TranScene	Ours	Ours+TranScene	Ours	Ours+TranScene	Ours	Ours+TranScene
EPE	0.79	**0.75**	1.53	**1.40**	5.32	**1.62**	7.93	**5.67**
RMSE	3.02	**2.99**	4.70	**4.74**	6.39	**2.26**	11.03	**8.42**
2px	5.90	**5.15**	12.67	**9.17**	23.34	**13.51**	59.83	**46.78**
3px	4.57	**4.08**	7.80	**5.63**	17.62	**10.23**	50.36	**38.55**
5px	3.17	**3.00**	4.88	**3.80**	13.50	**7.40**	38.44	**28.65**
6px	2.58	**2.59**	3.96	**3.37**	12.80	**6.50**	33.87	**25.41**
8px	1.45	**1.73**	3.14	**2.86**	12.15	**4.93**	27.56	**21.30**

Table 4. Generalization from the SceneFlow dataset to the Booster dataset in various regions.

7. Ablation Study

7.1. More Analysis of Backbone

In addition to replacing the context network with the pretrained DepthAnything v2 [13], we also experimented with replacing the feature extractor for cost volume construction using DepthAnything v2 [13] and MASt3R [4, 9]. As shown in Table 6, the results become worse after replacing the feature extractor for cost volume construction with DepthAnything v2 or MASt3R. Moreover, a bug with the A40 GPU causes memory issues when converting the alternate correlation function from dot product to Euclidean distance during training. Therefore, the model with MASt3R was trained using the original correlation function with dot product, where additional learnable convolution layers are further used after MASt3R for feature extraction.

7.2. More Analysis of Iterative Local Fusion

We provide additional visualizations of the intermediate results from the iterative local fusion process in Figures 10, 11, 12, 13, 14, 15, 16, and 17. As the iterations progress, the ordering maps generated from binocular disparity gradually become smoother. The convolution layers learn the differences between ordering maps generated from binocular disparity and monocular depth, allowing the guidance to focus more effectively on non-smooth regions, thereby significantly affecting disparity update.

7.3. More Analysis of Components in Global Fusion

We present more visualization for the intermediate results of global fusion in Figure 10, 11, 12, 13, 14, 15, 16, and 17. The visualization shows that the registration of monocular depth is different for each pixel, particularly on different objects. Since the monocular depth from DepthAnything is scale ambiguity but not absolute depth before registration, the visualization of it is not alinged to the ground truth range, other wise its visualization is almost a single color. The implicit learned confidence also filters out the noise of monocular depth, especially in Figure 4.

We provide additional visualizations of the intermediate results from global fusion in Figures 10, 11, 12, 13, 14, 15, 16, and 17. These visualizations illustrate the varying registration of monocular depth across individual pixels, particularly across different objects. Given that the monocular depth obtained from DepthAnything is scale ambiguous and does not represent absolute depth before registration, we do not align it with the ground truth range in visualization; otherwise, it would appear almost uniformly as a single color. The implicitly learned confidence also effectively filters out noise in the monocular depth as demonstrated in Figure 4.

8. Future Work Discussion

We present failure cases in Figures 18 and 19. In the first failure case, our method is confused by the glass door and glass window, where both the transparent surfaces and the behind scene are significant. Unlike simple transparent objects (e.g., a glass bottle), transparent scenes raise a new challenge for robotics, as they need to perceive both the transparent surface and the scene behind it. Failure to do so may cause robots to get stuck, for instance, when trying to reach an apple behind a glass window. If the robot

	750×2484	1125×3726	1500×4968	1688×5589	1875×6210
RAFTStereo reg [5]	2268.35	6023.82	10795.02	14299.5	19666.78
RAFTStereo alt [5]	1715.8	4151.8	6466.7	8157.96	11177.66
IGEV 384 [12]	2816.46	7290.82	14484.61	18810.14	-
IGEV 640 [12]	3167.46	8475.43	17366.83	-	-
Selective IGEV 384 [10]	2960.34	7608.44	15035.55	19505.5	-
Selective IGEV 640 [10]	3311.84	8793.07	18701.57	-	-
Mocha-Stereo 384 [1]	5525.56	12986.73	24665.95	-	-
Mocha-Stereo 640 [1]	6136.18	15056.66	29476.45	-	-
ours reg	5031.07	8609.63	14088.98	17782.53	22279.12
ours alt	3452.42	6745.23	9761.22	11641.82	13790.82

Table 5. Memory comparison across different resolutions. We evaluate the memory consumption of each model, excluding the feature encoder module, to ensure a fair comparison of backbones during inference. reg: pre-computation of the entire cost volume, allowing for look-up operations at each iteration, alt: dynamically computing a thin cost volume at each iteration. 384/640: the maximum disparity range used for the resolution of 750×2484. '-': out of memory in our GPU.

Exp	Middlebury (H)	
	epe	bad 2.0
Baseline + FE-DepthAnything	3.26±0.03	28.73±0.28
Baseline + FE-MASt3R	4.41±0.40	26.83±0.57
Baseline + ME + ILF + GF	1.15±0.01	8.35±0.04

Table 6. The effectiveness of each module. Baseline: RAFT-Stereo, ME: our monocular encoder, ILF: iterative local fusion, GF: our global fusion. FE-DepthAnything: replacing the original feature extractor with DepthANything v2. FE-MASt3R: replacing the original feature extractor with MASt3R.

perceives only the glass window, it will miss the apple entirely, while perceiving only the apple means the glass acts as an unrecognized and insurmountable barrier. Therefore, a novel representation for depth estimation is necessary to allow for multiple depths at a single pixel.

In the second failure case, our method is confused by the very close black screen and the very dark tunnel. In these scenes, registering monocular depth with binocular disparity is highly challenging due to excessive and concentrated noise in the disparity, along with pixel-wise differences in monocular depth registration, particularly across different objects. Consequently, information from video streams and segmentation becomes essential, like video stereo matching or simultaneously learning segmentation.

References

[1] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 27768–27777, 2024. 3, 5

[2] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2024. 3

[3] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2

[4] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 4

[5] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision*, pages 218–227. IEEE, 2021. 2, 3, 5

[6] Zhidan Liu, Chengtang Yao, Jiaxi Zeng, Yuwei Wu, and Yunde Jia. Multi-label stereo matching for transparent scene depth estimation. *ArXiv*, 2025. 2

[7] Jieming Lou, Weide Liu, Zhuo Chen, Fayao Liu, and Jun Cheng. Elfnet: Evidential local-global fusion for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 17784–17793, 2023. 3

[8] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 855–866, 2023. 3

[9] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 4

[10] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 3, 5

[11] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *International Conference on Computer Vision Workshops*, pages 3852–3857, 2019. 2

[12] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 3, 5

[13] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2, 4

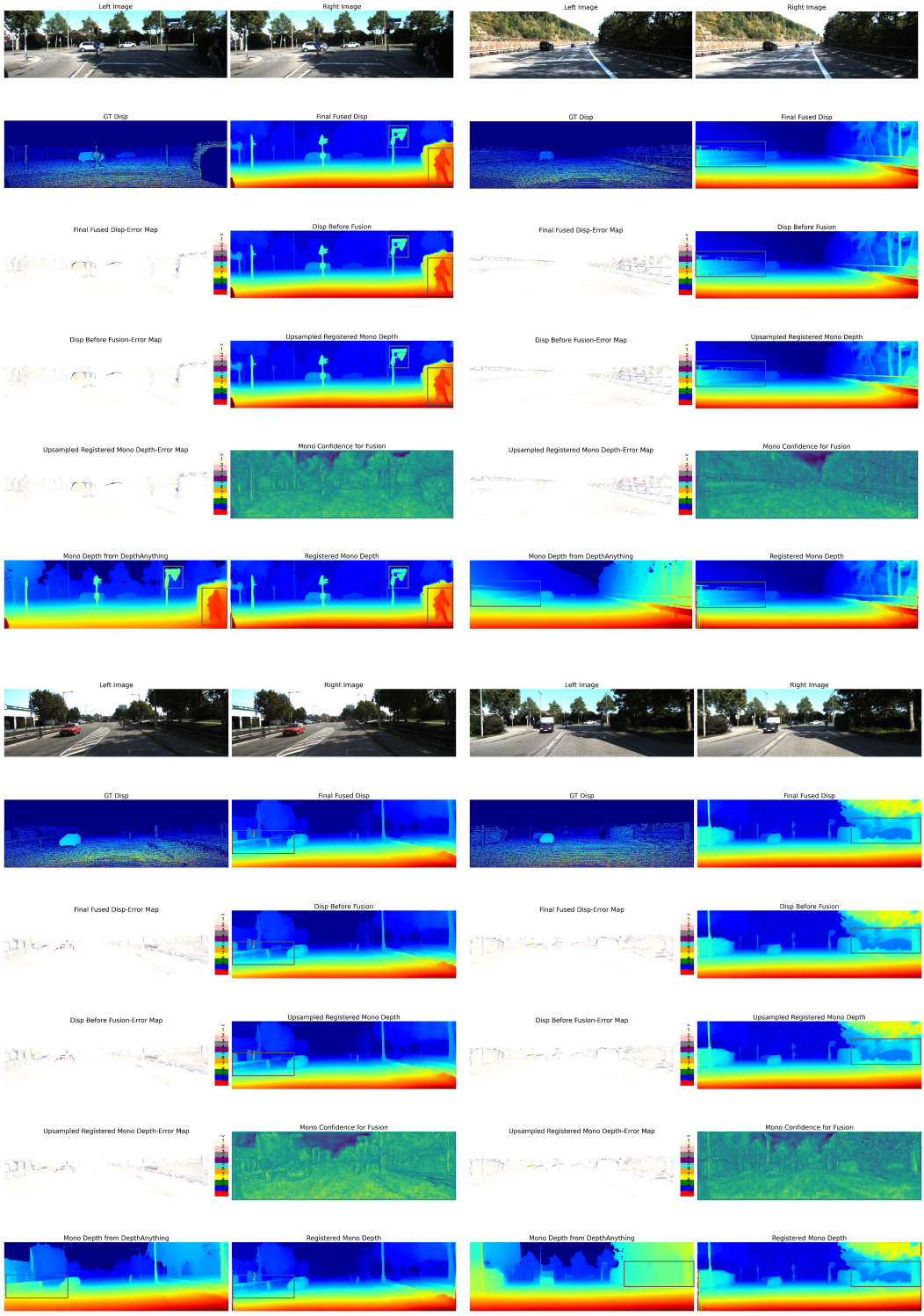

Figure 3. The visualization of binocular disparity and monocular depth. The regions highlighted with gray boxes demonstrate that stereo matching excels at capturing fine-grained details, whereas monocular depth estimation performs better in perceiving overall shapes. The mono depth from DepthAnything is scale ambiguity but not absolute depth before registration.

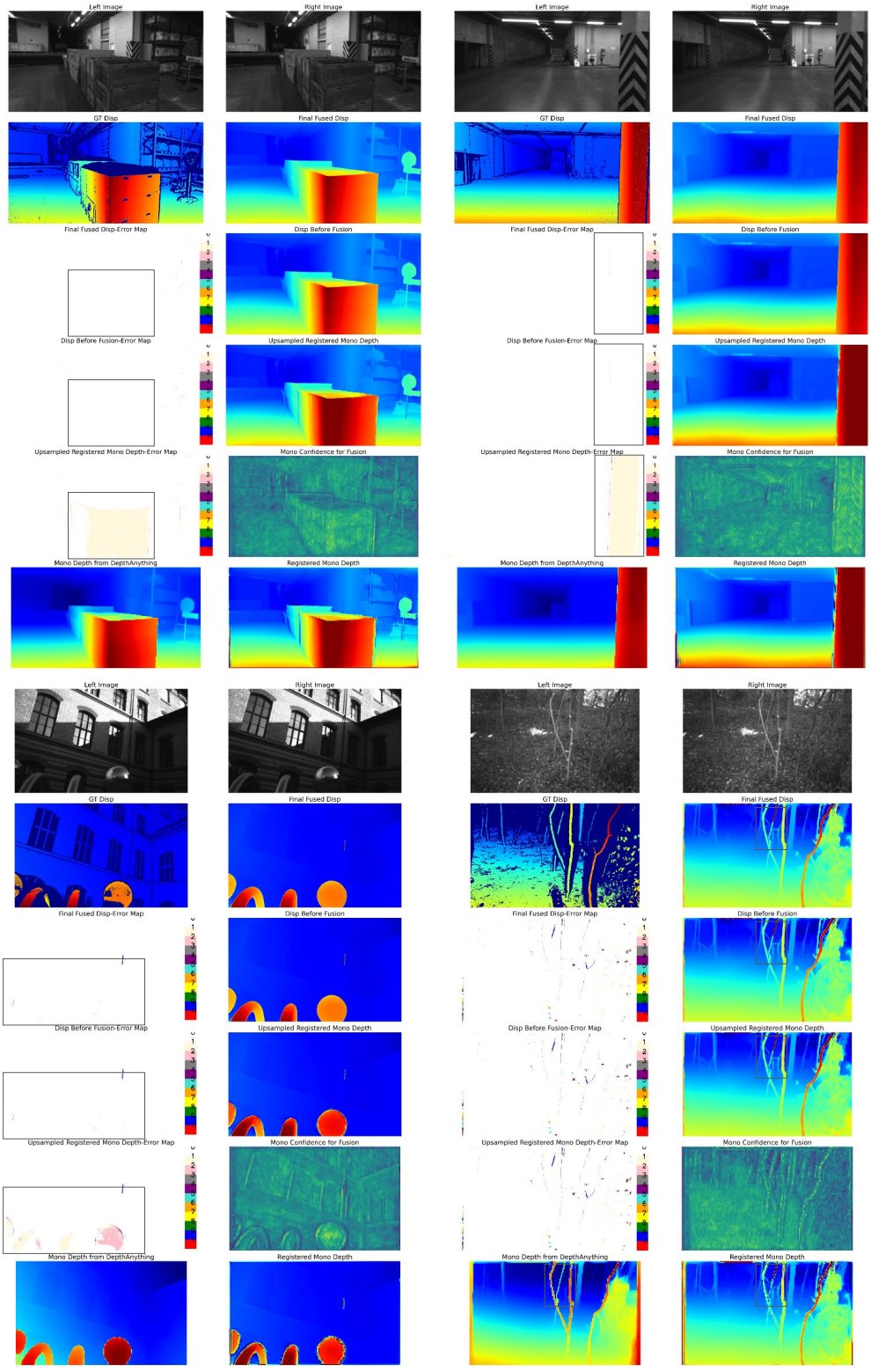

Figure 4. The visualization of binocular disparity and monocular depth. The regions highlighted with gray boxes demonstrate that stereo matching excels at capturing fine-grained details, whereas monocular depth estimation performs better in perceiving overall shapes. The mono depth from DepthAnything is scale ambiguity but not absolute depth before registration.

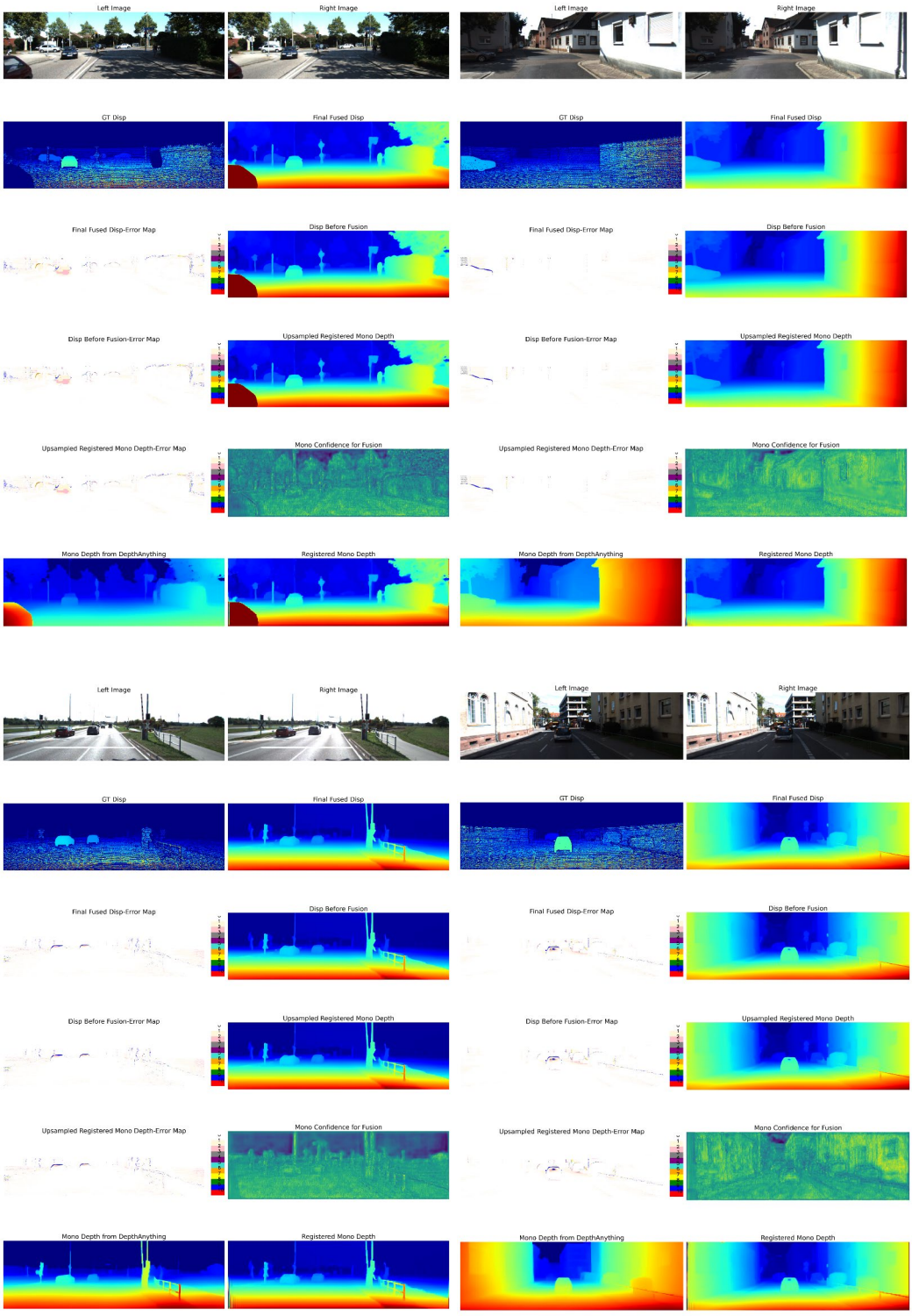

Figure 5. The visualization for generalized stereo matching.

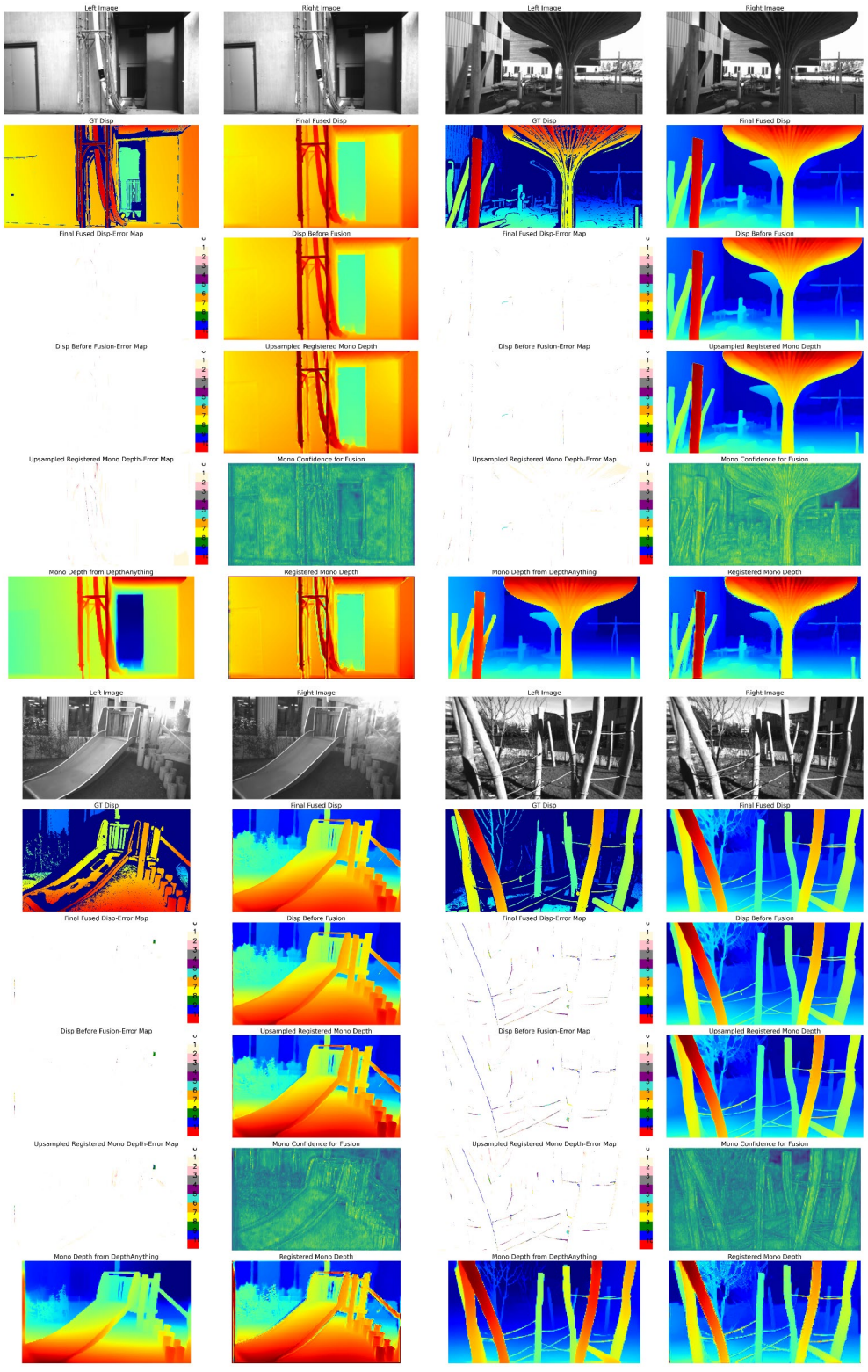

Figure 6. The visualization for generalized stereo matching.

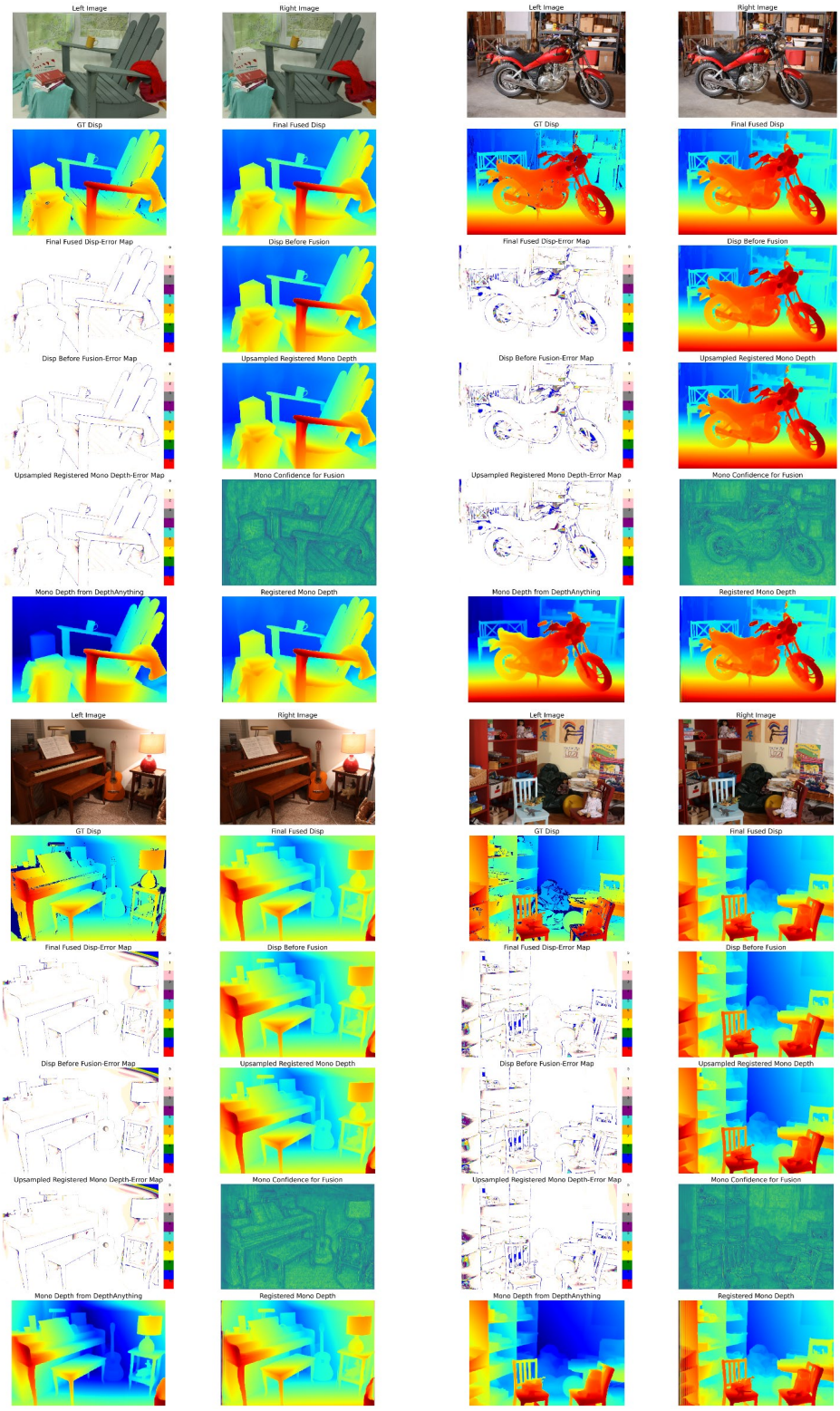

Figure 7. The visualization for generalized stereo matching.

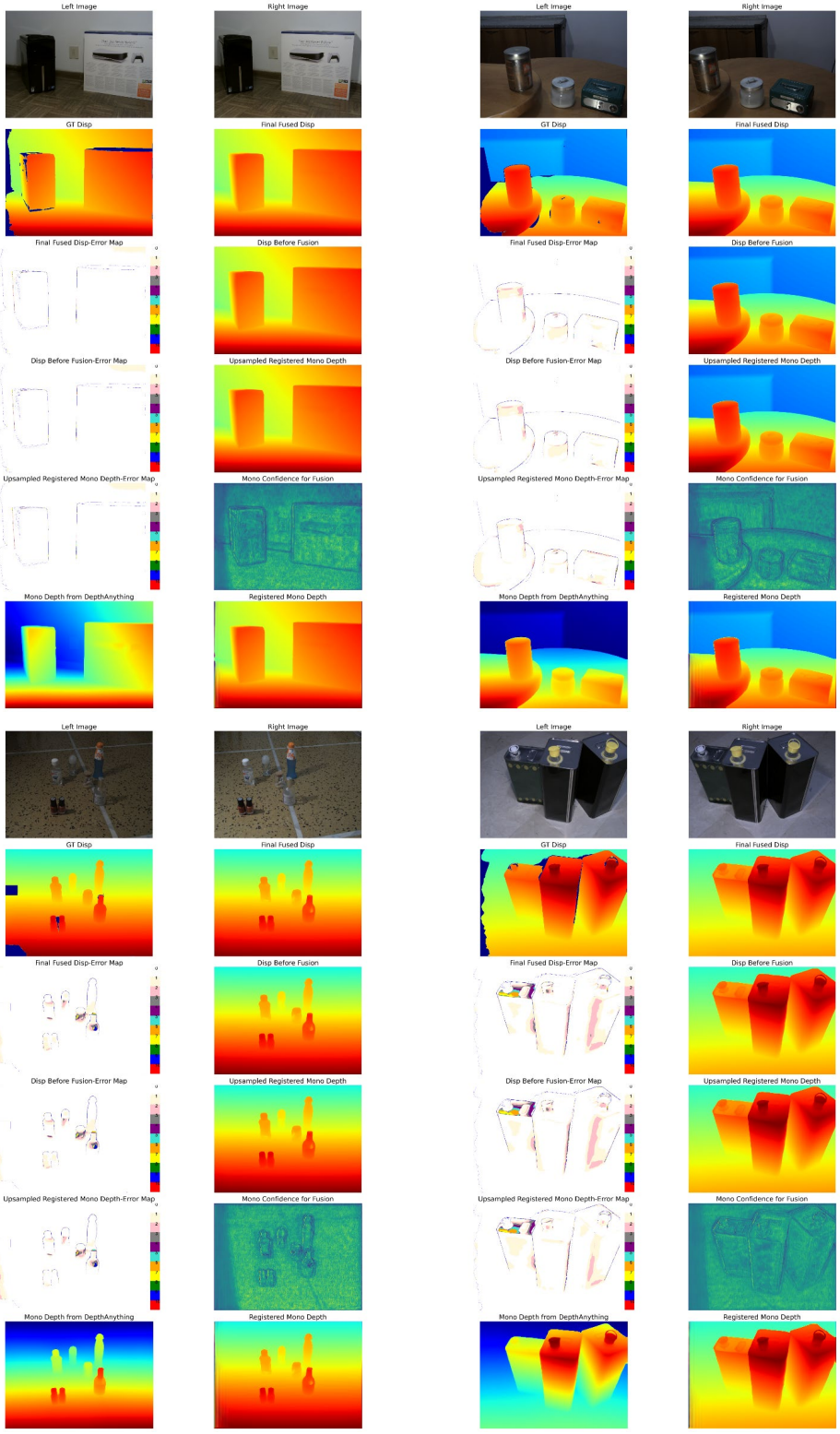

Figure 8. The visualization for generalized stereo matching.

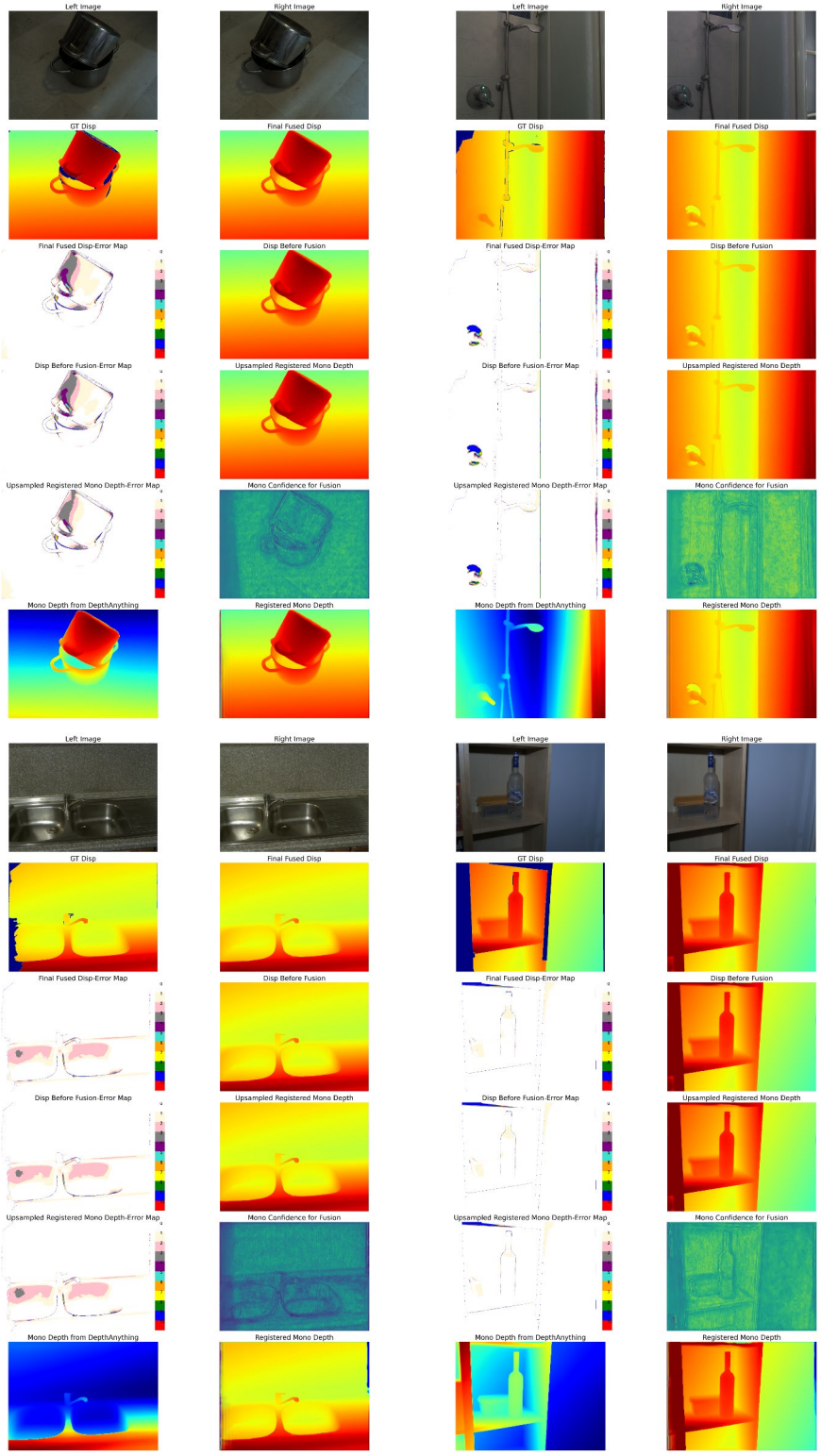

Figure 9. The visualization for generalized stereo matching.

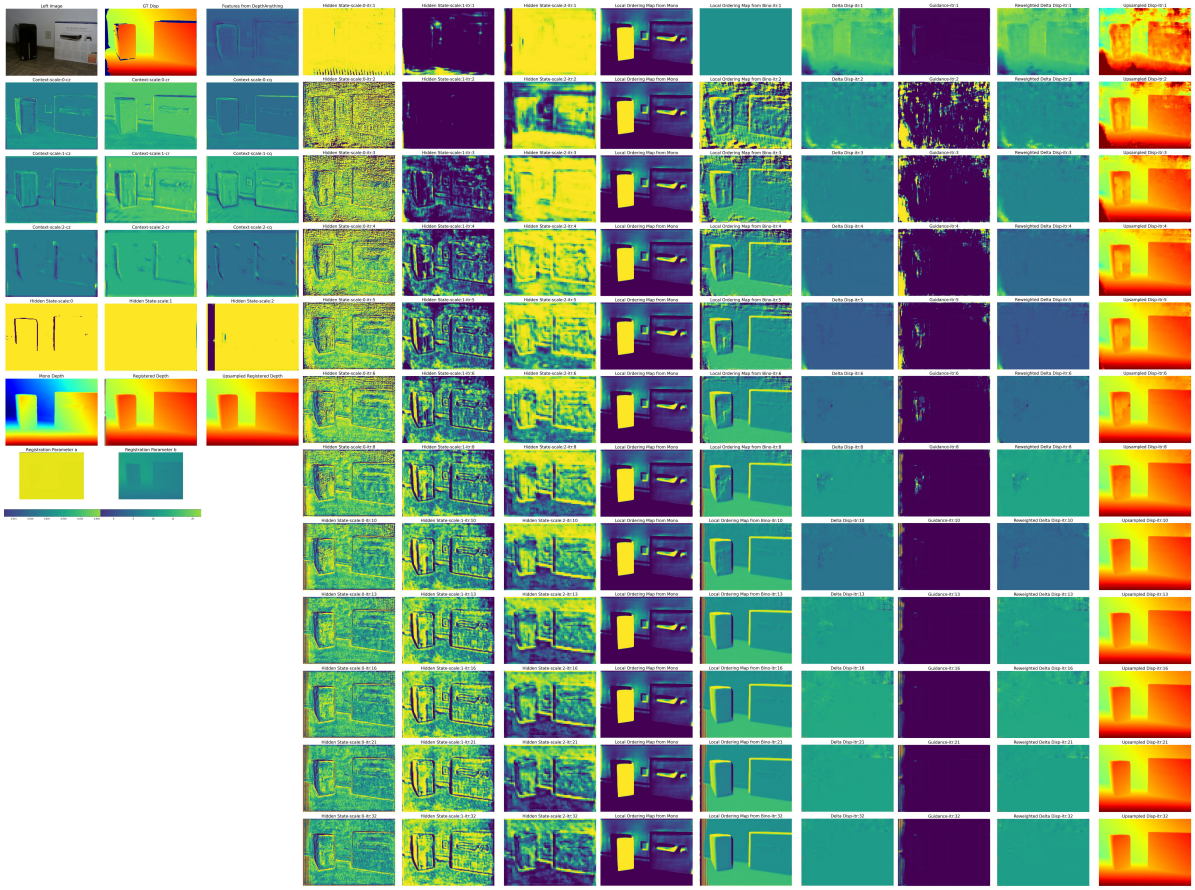

Figure 10. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

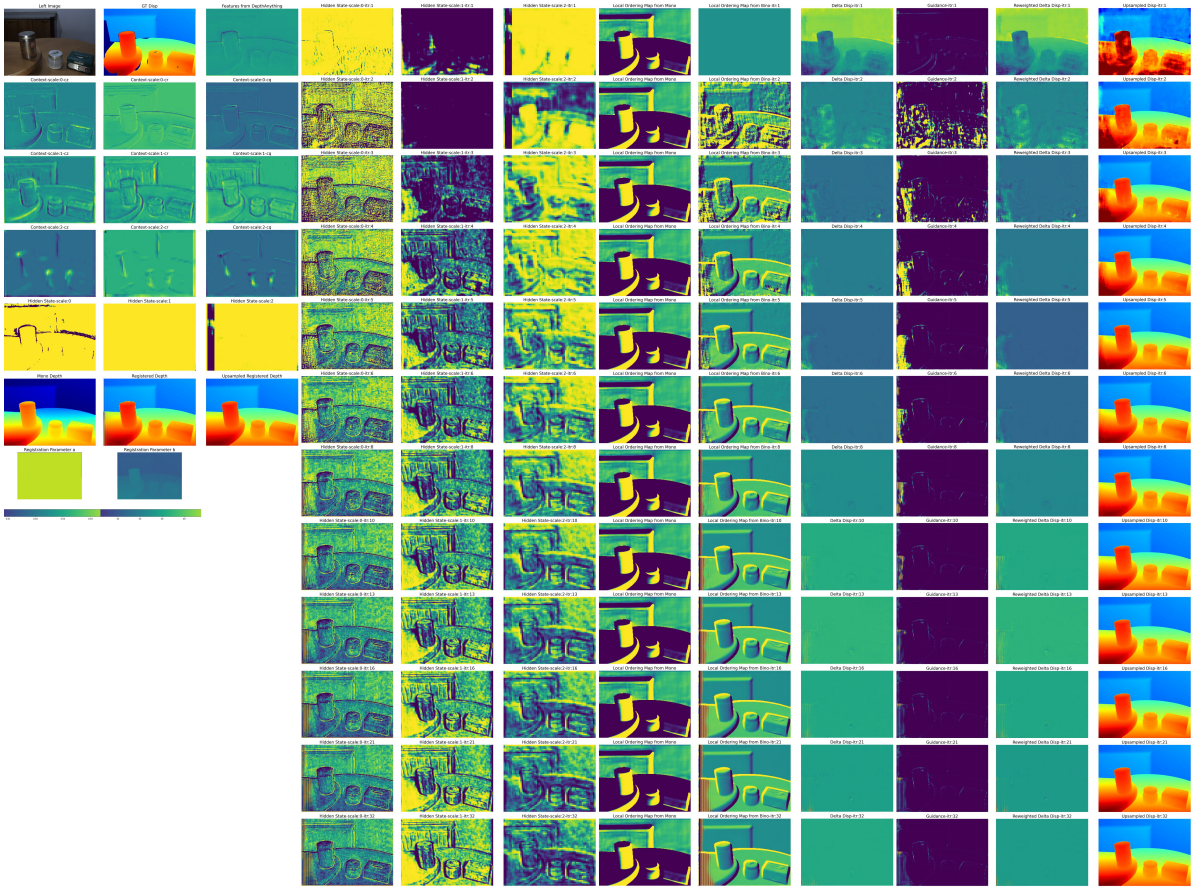

Figure 11. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

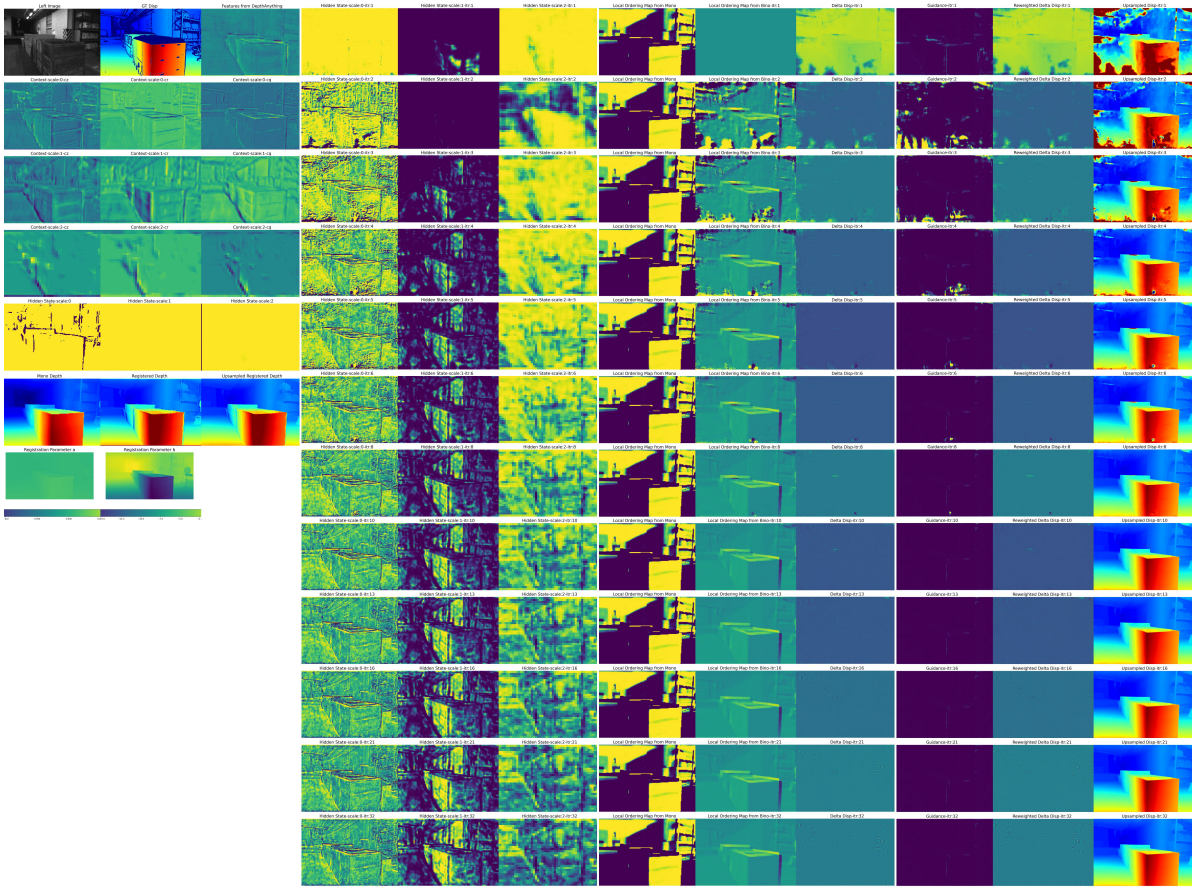

Figure 12. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

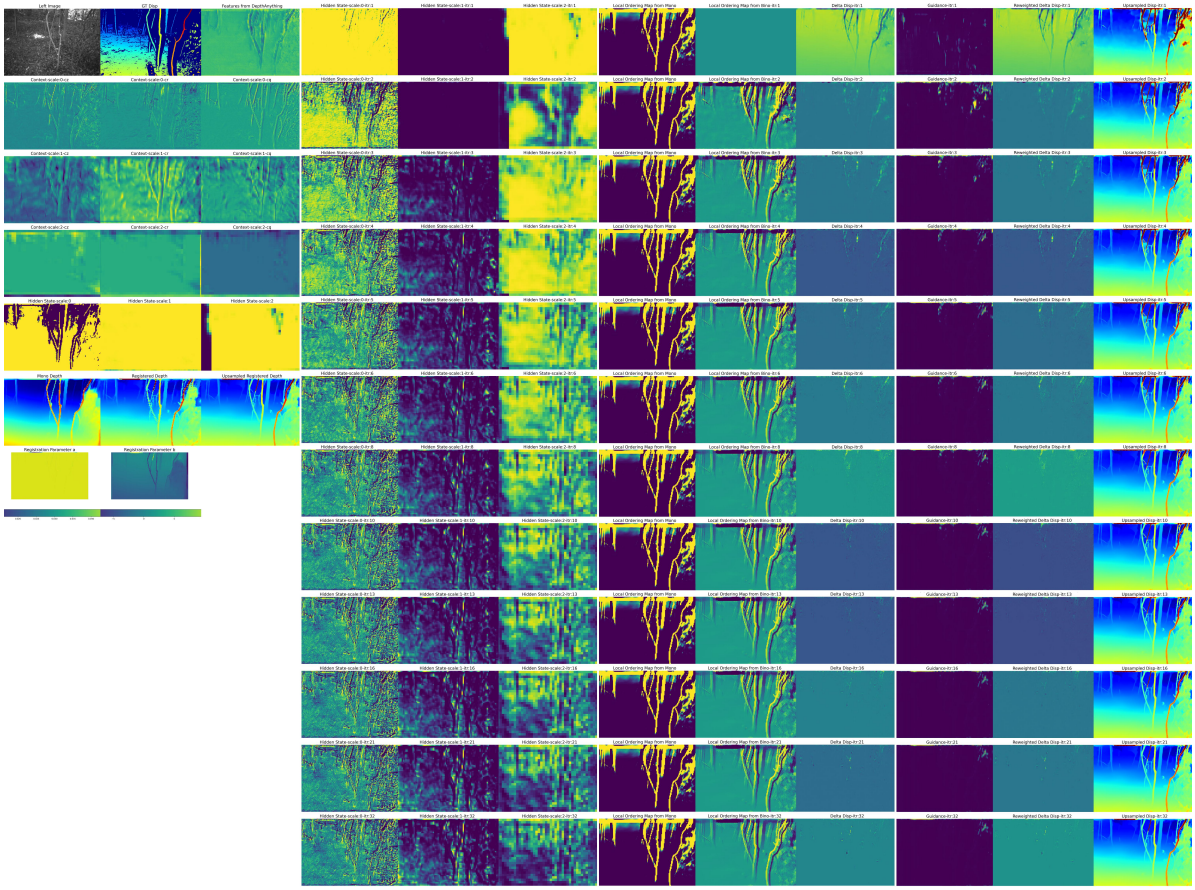

Figure 13. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

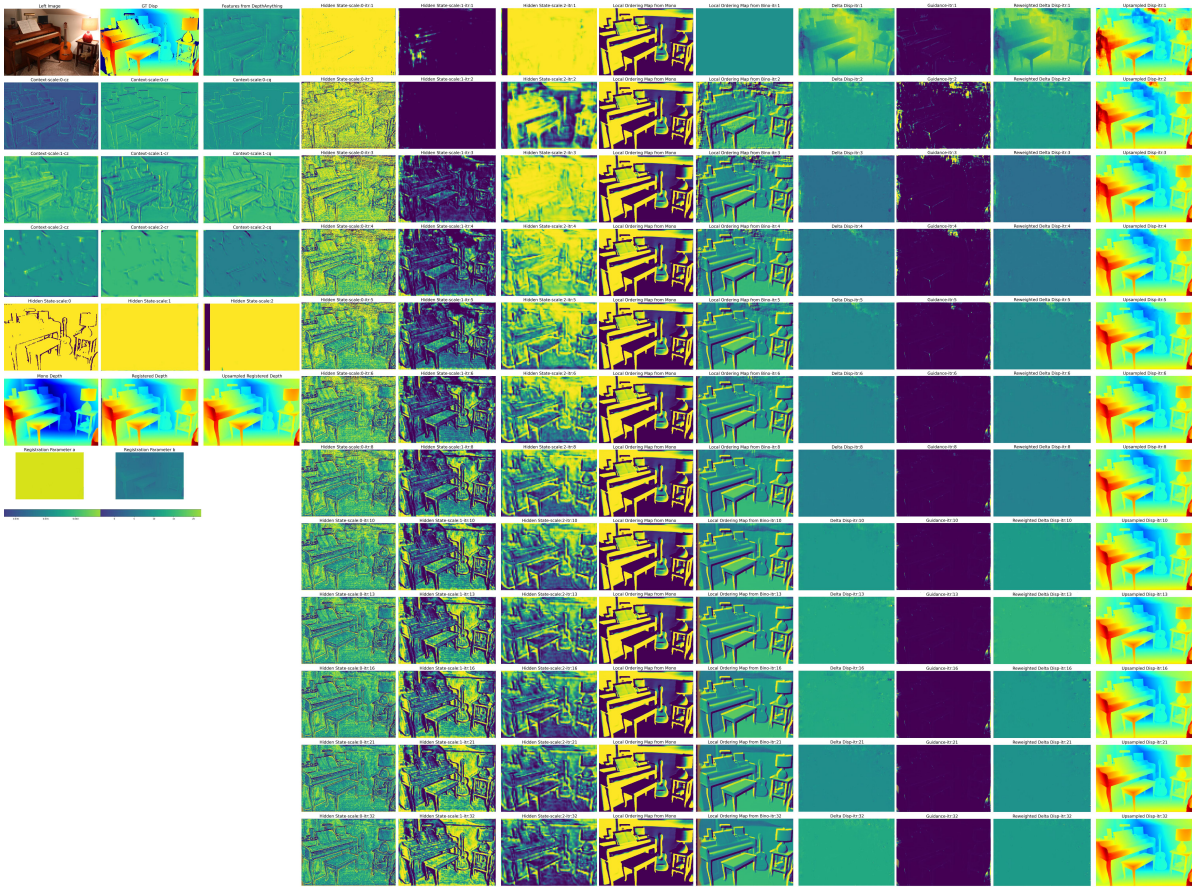

Figure 14. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

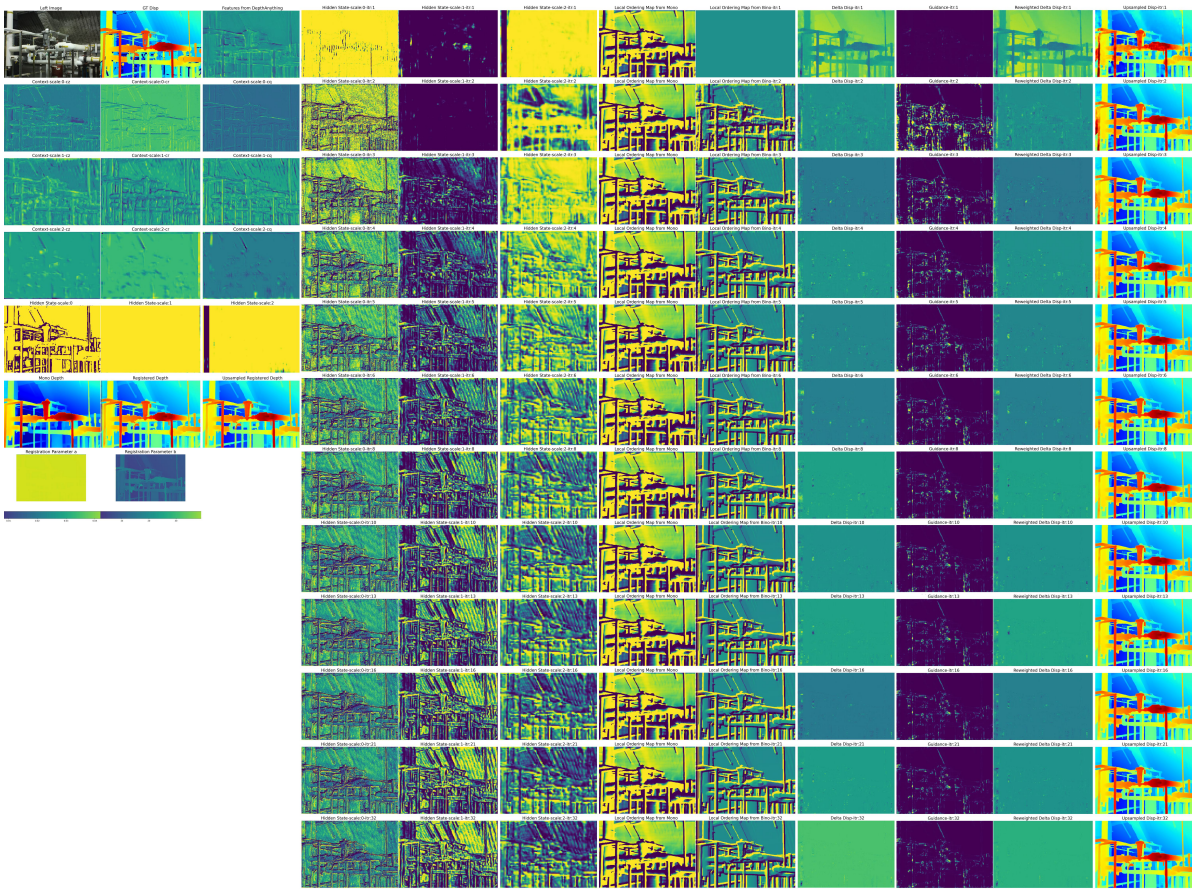

Figure 15. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

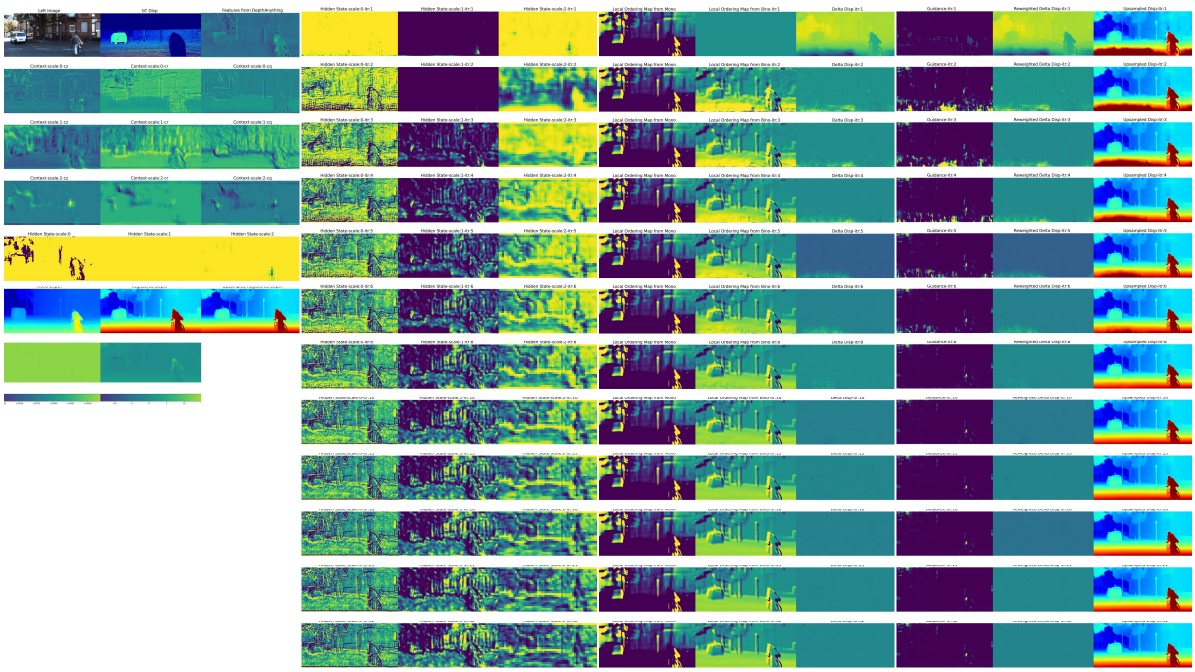

Figure 16. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

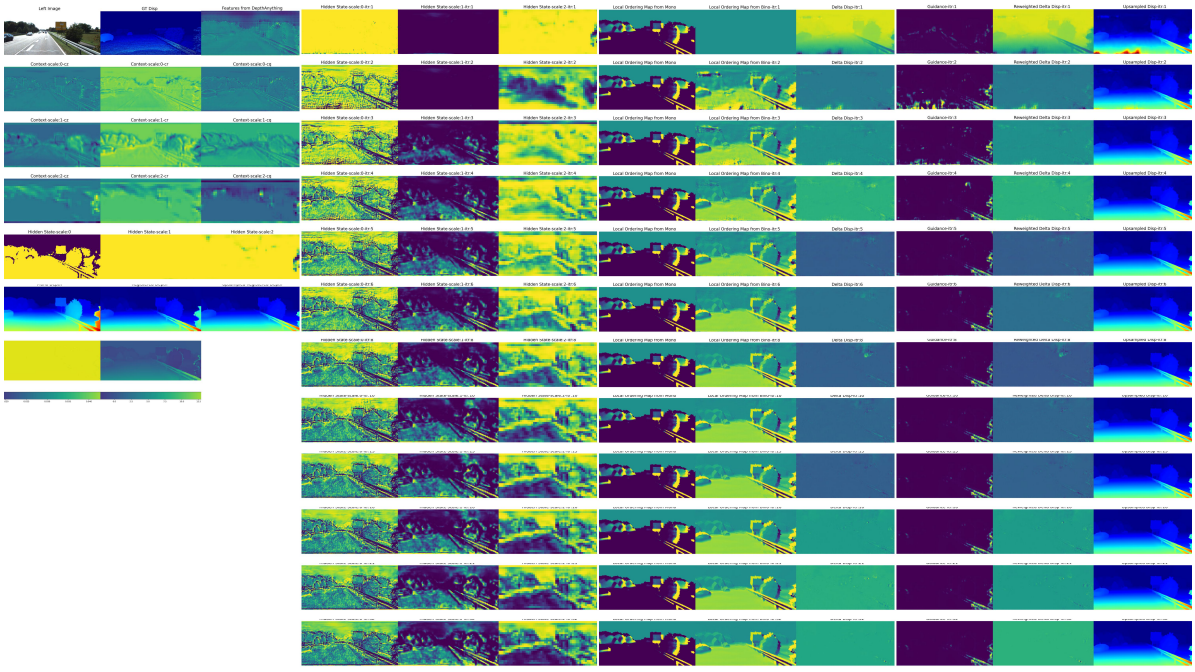

Figure 17. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

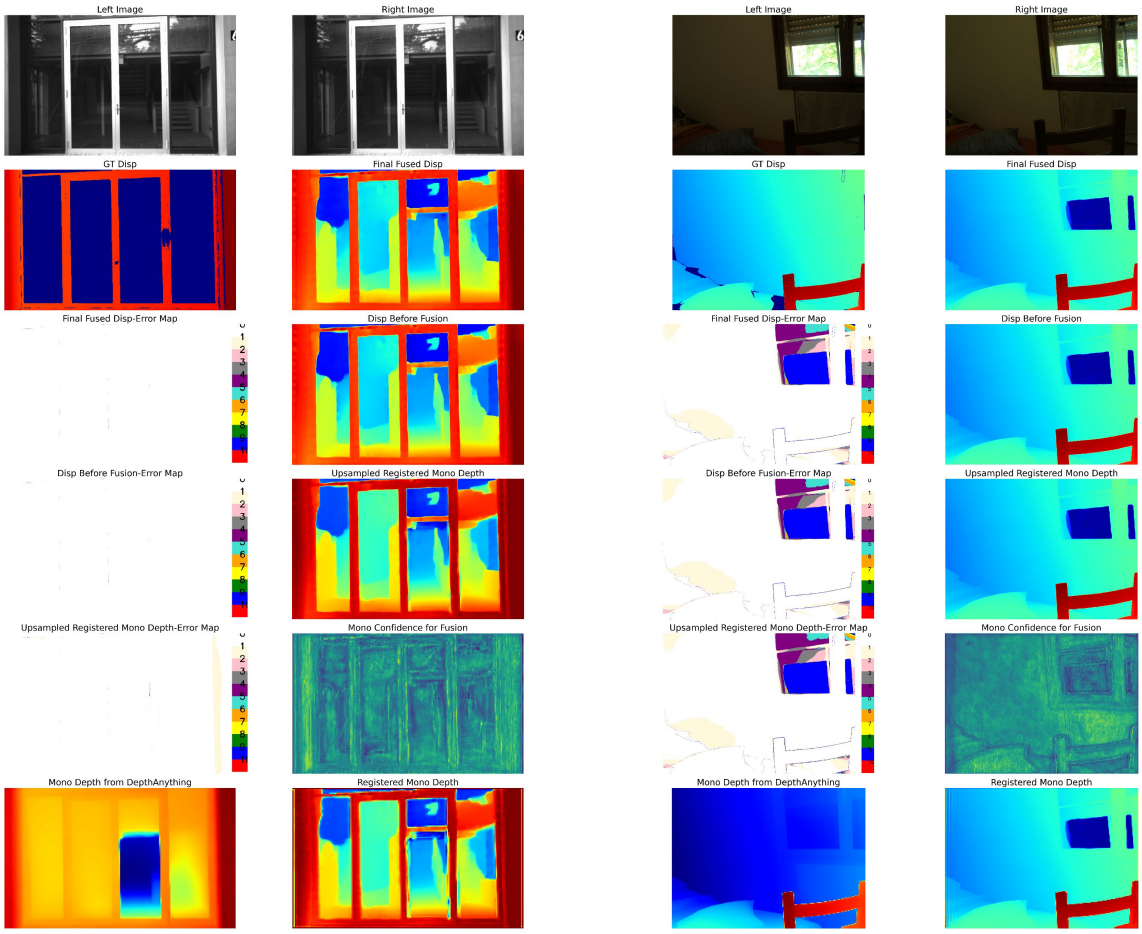

Figure 18. The visualization for failure case analysis.

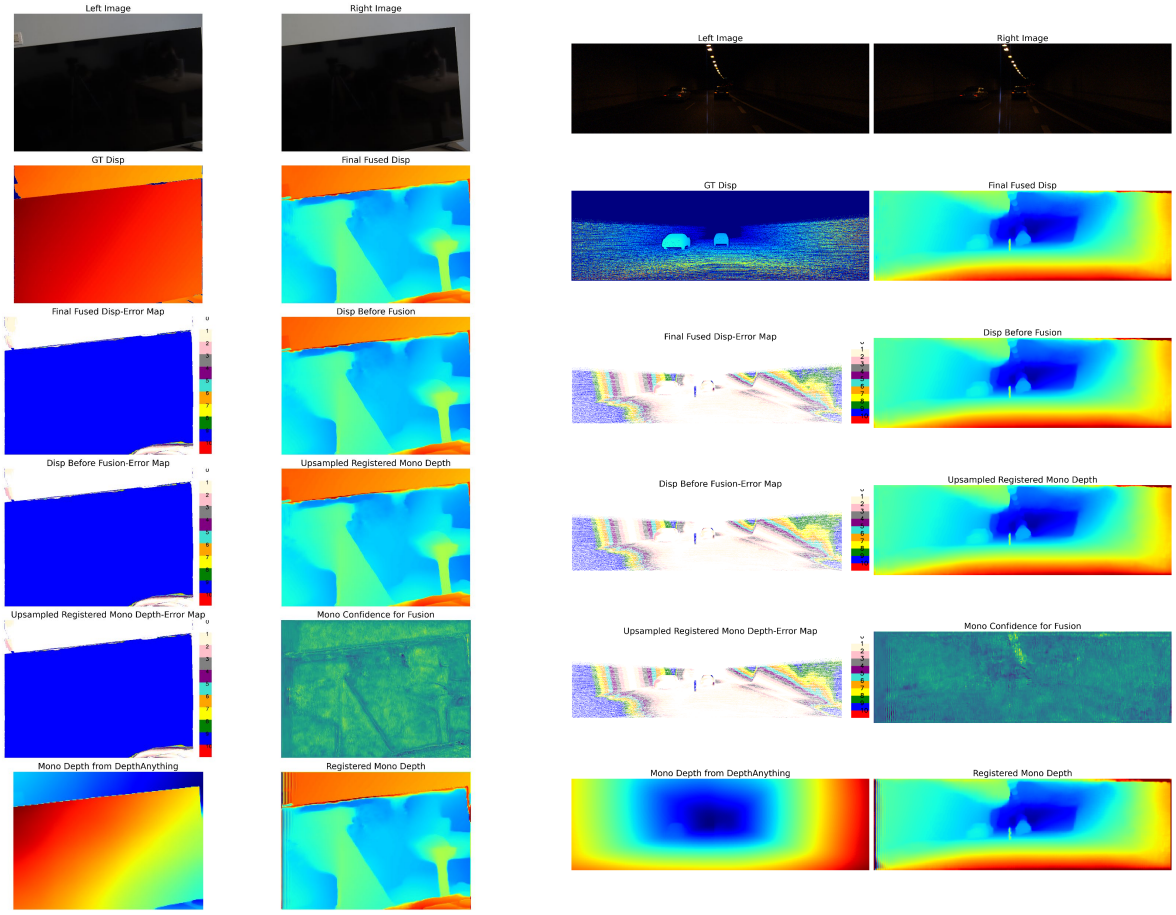

Figure 19. The visualization for failure case analysis.