Diving into the Fusion of Monocular Priors for Generalized Stereo Matching

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

Abstract

The matching formulation makes it naturally hard for the stereo matching to handle ill-posed regions like occlusions and non-Lambertian surfaces. Fusing monocular priors has been proven helpful for ill-posed matching, but the biased monocular prior learned from small stereo datasets constrains the generalization. Recently, stereo matching has progressed by leveraging the unbiased monocular prior from the vision foundation model (VFM) to improve the generalization in ill-posed regions. We dive into the fusion process and observe three main problems limiting the fusion of the VFM monocular prior. The first problem is the misalignment between affine-invariant relative monocular depth and absolute depth of disparity. Besides, when we use the monocular feature in an iterative update structure, the over-confidence in the disparity update leads to local optima results. A direct fusion of a monocular depth map could alleviate the local optima problem, but noisy disparity results computed at the first several iterations will misguide the fusion. In this paper, we propose a binary local ordering map to guide the fusion, which converts the depth map into a binary relative format, unifying the relative and absolute depth representation. The computed local ordering map is also used to re-weight the initial disparity update, resolving the local optima and noisy problem. In addition, we formulate the final direct fusion of monocular depth to the disparity as a registration problem, where a pixel-wise linear regression module can globally and adaptively align them. Our method fully exploits the monocular prior to support stereo matching results effectively and efficiently. We significantly improve the performance from the experiments when generalizing from SceneFlow to Middlebury and Booster datasets while barely reducing the efficiency. https://github.com/YaoChengTang/Diving-into-the-Fusion-of-Monocular-Priors-for-Generalized-Stereo-Matching

1. Introduction

Stereo matching provides dense depth for various downstream applications, such as autonomous driving, robotics, AR/MR, etc. These applications require stereo matching to generalize across different scenes from wild worlds. However, the generalization of stereo matching becomes poor in ill-posed regions due to occlusion, texture-less, and non-Lambertian surfaces (e.g., reflective or transparent surfaces). Fusion of monocular priors is proven to help correct the ill-posed binocular matching results [9, 15, 18, 22, 23, 33, 47, 54]. But the monocular prior trained on the limited data distribution of stereo datasets is susceptible to domain bias and can only capture significantly biased monocular features for certain scenes [13, 30].

Taking advantage of large-scale scenes and the easily collected ground truth of monocular depth, the vision foundation model can provide an unbiased monocular prior [16, 48, 49]. Recently, some methods have made great progress in fusing the unbiased monocular prior into the stereo matching to improve the generalization in ill-posed regions [3, 8, 46]. In this paper, we dive into the fusion mechanism and find three main problems limiting a full exploration of the unbiased monocular prior. The first problem lies in the natural gap between the affine-invariant relative depth from monocular depth and absolute depth from disparity. Although we can forcibly align the two kinds of depth with a complex mutual refinement, these alignments could involve heavy computation and greatly harm the efficiency [8, 46]. The other problem exists in the fusion with monocular feature maps in an iterative refinement structure [7, 43, 47]. The implicit feature fusion makes the fusion more biased to the binocular information due to the iterative update training scheme, where the over-confidence of the disparity update causes local optima, as shown in Figure 1. An additional fusion of monocular depth could alleviate the local optima, but the direct fusion of the depth map is easily affected by the noisy depth results. Even with un-

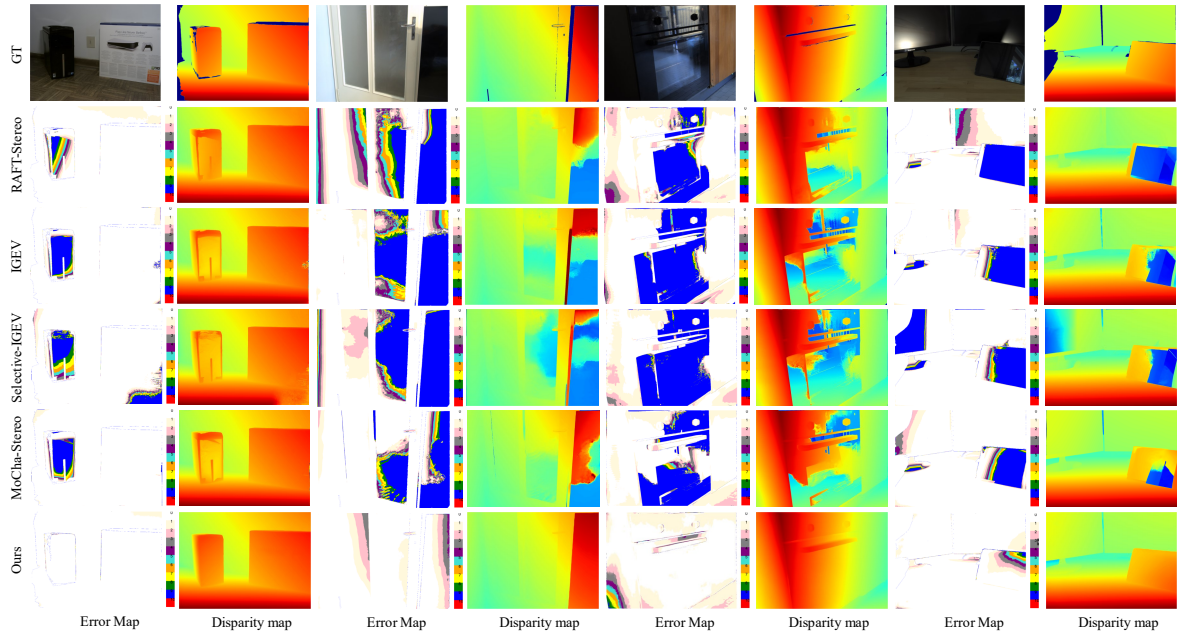

Figure 1. The visualization of different ill-posed regions in the Booster dataset. Our method achieves an overwhelming advantage in all kinds of regions.

biased and smooth monocular depth from the VFM, noisy disparity at early iterations slows down a good fusion.

In this paper, we present a new depth representation called the local binary ordering map that indicates whether two pixels are farther or closer. It converts the depth into a binary relative depth representation, unifying the monocular depth and binocular disparity. The local binary order map also guides fusion in an explicit manner, which restricts the influence of the large noise from outliers. Furthermore, we formulate the binocular disparity map as a noisy version of monocular depth registered by specific pixel-wise scale and shift. Therefore, the alignment between monocular depth and binocular disparity can be deemed a noisy linear regression problem about the registration parameters. The registration formulation globally and adaptively aligns the two kinds of depth in an efficient manner.

Our network can be divided into three modules. The monocular encoder extracts unbiased monocular priors, including monocular depth and context features, using a large pre-trained monocular network like [48–50]. Then, the fusion can be realized by an iterative local fusion module and a global fusion module to fully exploit the usage of the monocular priors with matching information. The iterative local fusion module uses a two-stream architecture to update the disparity iteratively. The first stream computes two binary ordering maps from monocular depth and binocular disparity through a series of LBP-like convolution blocks. Then, we compute the differences between the two binary ordering maps to form a local guidance for fusion. At the same time, the second stream predicts an initial disparity update result through a multi-level GRU using cost volume and monocular context features. The local guidance is used to re-weight the initial disparity update result, resolving the local optima. After local fusion, the global fusion module realizes the optimization of the disparity map by registering to monocular depth. We first compute two parameters to register the relative depth to the absolute depth globally. It solves the noisy linear regression problem between optimized disparity and monocular depth through a series of convolutions. Then, we compute a confidence map using the cost volume, the hidden state of GRU, and the local guidance from the last iteration. The confidence map guides the fusion of the optimized binocular disparity and the registered monocular depth as the final prediction.

We compare our model with state-of-the-art methods using the standard setting, training on the SceneFlow dataset, and testing on five real-world datasets with various ill-posed areas, including KITTI 2012, KITTI 2015, Middlebury, ETH3D, and Booster. The results demonstrate that our method significantly enhances the performance of state-of-the-art approaches, as shown in Figure 1. Experiments show that our method achieves a 10-point improvement in the bad2 metric for transparent regions on the Booster dataset and reduces errors by more than 50% on Middlebury and ETH3D, where we do not use additional stereo data or specific data augmentation. Meanwhile, even involving a VTF model, our method barely raises the time cost, benefiting from the elegant explicit designs.

2. Related Work

2.1. Generalized Stereo Matching

Generalized stereo matching aims to produce a reliable dense disparity map when the target domain (e.g., real-

world data) differs from the source domain (e.g., synthetic data). Some methods focus on the learning of domain invariant features [4, 9, 17, 22, 33, 40, 54]. MS-PSMNet [4] replaces the learning-based features with hand-crafted features to force the stereo network focus on the matching space. MS-PSMNet has achieved great improvement, but its hand-crafted features limit the performance of the stereo network. Thus, many methods turn to improve the training process by transforming learning [22], meta-learning [17, 40], contrastive learning [33, 54], and fisher information [9].

The above feature-based methods significantly improve the generalization of stereo matching. However, it is difficult for them to eliminate domain gaps due to the complexity of real-world scenarios. Then, some researchers integrate other modalities to enrich the features of RGB images [42, 52]. They achieve impressive performance but require additional devices. Instead, other researchers propose to generate more and better data for training [3, 5, 39, 41, 46]. AdaStereo [39] and HVT-RAFT [5] augment the training data in color space to enrich the domain distribution in the synthetic dataset. They improve greatly, but the rendered images in the synthetic dataset are unrealistic to the real world. Thus, NerfStereo [41] turns to reconstruct the real-world scenes from Nerf and re-renders stereo images to improve the quality of training data.

In addition to improving generalization from features and data, some methods focus on architecture design with specific knowledge of stereo matching [7, 12, 14, 20, 43, 47, 53]. DSMNet [53] uses long-range matching information in cost aggregation to correct the mismatched points. The improvement in cost aggregation is remarkable, but the additional operations in 3D space are time-consuming. Many methods then turn to incorporating global information when constructing cost volume. STTR [20] and CSTR [14] use transformers to capture long-range matching information. Other methods [7, 43, 47] build auxiliary volume to augment the original cost volume. There are also some methods resolving the generalization problem from uncertainty learning [15, 23].

The aforementioned approaches have achieved great performance but still rely on biased monocular priors. Our method introduces unbiased monocular priors from a pre-trained large model and uses effective fusion mechanisms to fuse them, achieving impressive generalization ability.

2.2. Fusing Monocular and Stereo Estimation

Inspired by the human vision system that fuses binocular disparity and monocular depth cues [3, 8, 10, 34, 44–46], researchers are exploring the fusion mechanism for machine vision. Tradition method [35] achieves it via an MRF optimization objective that relies on the binocular disparity and monocular cues. Deep learning methods mainly focus on volume fusion or depth map fusion [3, 8, 46]. The volume fusion methods introduce monocular priors into cost volume [3, 8, 19, 46, 51]. These methods require a fixed disparity range and depend on domain-biased monocular priors. Our method assumes no limit on the disparity range and introduces unbiased monocular priors from a pre-trained large model. The depth map fusion methods fuse the monocular depth and the binocular disparity in one-stage post-processing [1, 2, 6, 24, 55]. These methods mainly predict affine-invariant monocular depth, which is not aligned with binocular disparity and results in noise in fusion. Instead, our method uses local ordering maps to make compatibility between monocular depth and binocular disparity, reducing noise in the iterative matching process. Based on the optimized binocular disparity, the monocular depth is further globally aligned with it by learning two parameters to solve the scale ambiguity in monocular results.

3. Method

Our network structure is illustrated in Figure 2. First, we extract features from the left and right images to construct a cost volume. Meanwhile, the monocular encoder module extracts initial hidden states, context features, and monocular depth from the left image using a pre-trained large monocular model [49]. Then, the local fusion module iteratively optimizes the disparity estimation with monocular priors using the local binary ordering map. Finally, the global fusion module registers the optimized disparity with the monocular depth as the final result.

3.1. Monocular Encoder

The monocular priors learned by the stereo-matching model are heavily biased due to the scarcity of wild-world stereo data [13, 30]. This paper uses the widely used DepthAnything v2[49] to extract unbiased monocular priors to mitigate the domain gap, including monocular context features and depth. However, it is flexible to use other VTFs as long as the monocular prior is not biased to specific scenarios.

As shown in Figure 2, given an image with a resolution of $H \times W$, we pre-process the image as DepthAnything v2 [49] by resizing the longest side of the image to 512 pixels. The resized image is then fed into a frozen DepthAnything v2 to extract intermediate features before the DPTHead [32] and monocular depth after the DPTHead. These intermediate features and monocular depth are subsequently resized to a $H/4 \times W/4$ resolution using a bilinear function to interact with the stereo-matching pipeline. We build a two-stream convolution module to generate the initial hidden state and monocular context features from the intermediate features. Although the output of DepthAnything v2 is inverse depth (disparity under unknown camera parameters), we refer to it as monocular depth for consistency description.

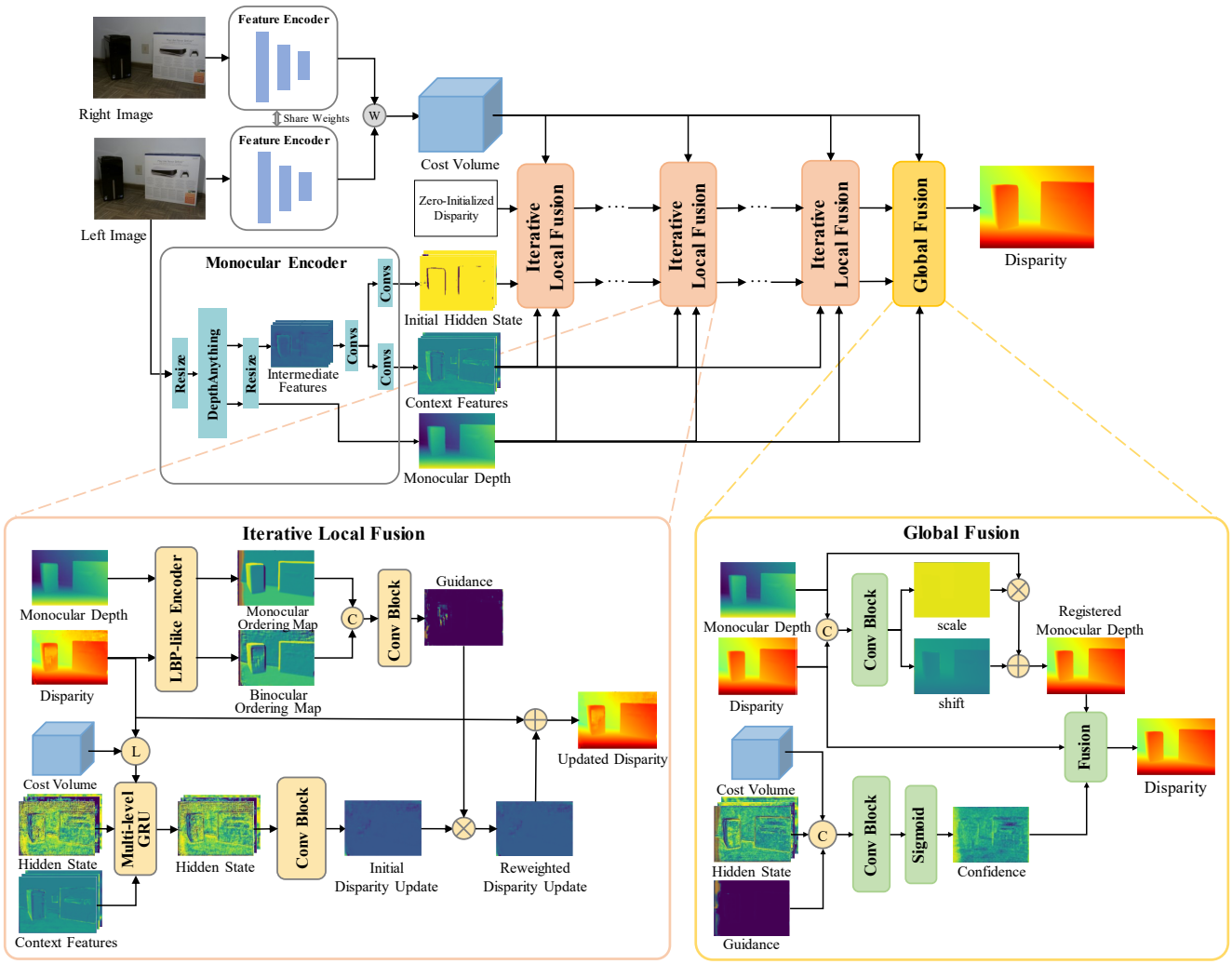

Figure 2. The pipeline of our method. Ⓦ represents the warping operation when constructing the cost volume. Ⓛ is look up operation used to sample cost volume. Ⓒ represents concatenation. ⊕ represents add operation, while ⊗ represents multiplication.

3.2. Iterative Local Fusion

The iterative local fusion module leverages the binary local ordering map to update disparity with monocular priors iteratively. The binary local ordering map helps mitigate the impact of outlier noises by converting absolute values into ordering relationships, which is much more robust than the pixel-wise depth value. Besides, it also unifies the affine-invariant depth and absolute disparity to be compatible with the order relationship.

To compute the binary local ordering map, we use a series of LBP-like operations [28, 29] with varying window sizes to extract local ordering features. Each LBP-like operation consists of a convolution with fixed weights followed by a sigmoid function, which measures the relative depth relationships between the center pixel of the window and its neighboring pixels, indicating which pixels are closer or farther. The formulation of the local ordering map M_O at pixel (u, v) is as follows:

$$M_O(u,v) = \{\sigma(D(u',v') - D(u,v)) \mid (u',v') \in \mathcal{N}_{(u,v)}\}, \tag{1}$$

where D is depth or disparity map, σ is the sigmoid function and $\mathcal{N}$ is the neighborhood. To employ the binary local ordering map into the iterative refinement structure, we use the LBP-like encoder to extract local ordering maps from both monocular depth and binocular disparity in the previous iteration, as shown in Figure 2. These two kinds of local ordering maps are concatenated to predict the monocular guidance. The guidance G is then used to re-weight the initial disparity update Δ_d to avoid local optima. We represent the monocular guidance as a Beta distribution $Beta(\alpha, \beta)$. During training, G is sampled via reparameterization trick. At test time, we compute the distribution expectation to ob-

tain it: $G = \alpha/(\alpha + \beta)$. G is used to reweight the disparity update, as defined in Equation (1) of the main text.

As we mentioned, the first several disparity predictions are noisy, especially during training. The local ordering map may still have many wrong relative depth values, leading to wrong guidance and slow training convergence. Therefore, we propose to gradually release the influence of guidance to the initial disparity update results as

$$\tilde{\Delta}_d = \Delta_d(1 + G \cdot r \cdot t/T). \tag{2}$$

Here, r is the manually specified amplitude parameter that controls the influence of the guidance. t represents the current iteration number, and T is the total number of iterations. The initial disparity update Δ_d is predicted by a multi-level GRU followed by a convolution block. Finally, the disparity is updated by adding the re-weighted disparity update to the disparity from the previous iteration:

$$D_d^t = D_d^{t-1} + \tilde{\Delta}_d. \tag{3}$$

3.3. Global Fusion

After all iterations of disparity update, we use a global fusion module to incorporate fine-grained 3D shape priors from the monocular depth map into the disparity map, as shown in Figure 2. Here, we formulate the optimized binocular disparity as a registered version of monocular depth with minor noise by specific intrinsic parameters. Therefore, monocular depth can be globally registered to binocular disparity. The registration can be deemed a linear regression problem with noise between monocular depth and binocular disparity. To this end, we first align the monocular depth D_m with the optimized disparity D_d by estimating two registration parameters, a, b by

$$\begin{aligned} \tilde{D}_m &= a \cdot D_m + b, \\ a, b &= \mathcal{F}(D_m, D_d^T), \end{aligned} \tag{4}$$

where $\mathcal{F}$ represents a network with a series of convolution layers and ReLU activation, which take the concatenation of the monocular depth D_m and the optimized disparity D_d^T as input. Simultaneously, we use the sampled cost volume, hidden state, and weights from the previous iteration to predict a confidence map. This confidence, c, is then used to fuse the aligned monocular depth $\tilde{D}_m$ and the optimized disparity D_d as follows

$$D_f = c \cdot D_d^T + (1 - c) \cdot \tilde{D}_m, \tag{5}$$

where c is a confidence map. D_f is the final disparity prediction.

3.4. Loss

We use L_1 loss to supervised the learning of each updated disparity D_d^t, registered monocular depth $\tilde{D}_m$, and the final output of our method D_f:

$$\begin{aligned} \mathcal{L} = &\sum_{t=1}^{T} \gamma^{T+2-t} ||D_d^t - D_G||_1 \\ &+ \gamma ||\tilde{D}_m - D_G||_1 + ||D_f - D_G||_1. \end{aligned} \tag{6}$$

D_G is the ground-truth disparity. γ is the balancing scalar.

4. Experiments

4.1. Implementation Details

For the stereo part, our pipeline is built on the classical iterative structure of RAFT-Stereo [21], which is widely used and flexible to deploy without stacking network tricks to raise the computation burden. As for the monocular part, we use DepthAnything V2 [48, 49] to extract unbiased monocular priors. Still, it is flexible to use other VTFs as long as the monocular prior is not overfitted. We set parameters set as $r = 1$ and $\gamma = 0.9$. The window sizes for the LBP-like operations are configured to $5, 3$. The training was conducted on 4 NVIDIA A40 GPUs using the AdamW optimizer with a one-cycle learning rate schedule. During training, the DepthAnything V2 module remains frozen. Specifically, we first train the model without the global fusion module on the SceneFlow dataset, using a maximum learning rate of 0.0002, a batch size of 8, and for 100k steps, maintaining the consistency of matching parts with the total data used in RAFT-Stereo. Then, we train the monocular registration of the global fusion module while keeping the other modules frozen, using a maximum learning rate of 0.0005 and a batch size of 32 for 100k steps on the SceneFlow dataset. Finally, we train the entire global fusion module while keeping the other modules frozen, using a maximum learning rate of 0.0005, a batch size of 32, and 100k steps on the SceneFlow dataset. Our results are not sensible to the hyper-parameters of the training process. With the training and testing codes provided in the supplementary materials, all the evaluation results can simply be reproduced.

4.2. Evaluation

Datasets. Domain generalized stereo matching is typically trained on the SceneFlow dataset [25] and evaluated on the training sets of various real-world datasets. We select five real-world datasets, each containing different ill-posed regions, to evaluate the in-the-wild generalization ability of the models, including KITTI 2012 [11], KITTI 2015 [26, 27], Middlebury [36], ETH3D [37], and Booster [31].
Metrics. (1) We use two metrics: EPE, which measures the mean absolute disparity error in pixels, and Bad x, which represents the percentage of pixels where the predicted disparity deviates from the ground truth by at least x pixels. (2) It is important to note that many recent methods report

Method	Year	Additional Data/Aug	KITTI 2015		KITTI 2012		Middlebury (H)						ETH3D	
							All		NonOcc		Occ			
			EPE	bad 3.0	EPE	bad 3.0	EPE	bad 2.0	EPE	bad 2.0	EPE	bad 2.0	EPE	bad 1.0
FC-PSMNet [54]	2022		1.58	7.50	1.42	7	4.14	18.3	-	-	-	-	1.25	12.8
ITSA-PSMNet [9]	2022		1.39	5.80	1.09	5.2	3.25	12.7	-	-	-	-	0.94	9.8
Graft-PSMNet [22]	2022		1.32	5.30	1.09	5	2.34	10.9	-	-	-	-	1.16	10.7
Mask-CFNet [33]	2023		-	5.80	-	4.8	-	13.7	-	-	-	-	-	5.7
STTR* [20]	2021		2.14	9.5	2.51	9.62	9.13	21.76	5.03	13.49	35.98	78.84	-	-
PCWNet [38]	2022		-	5.60	-	4.2	-	15.8	-	15.8	-	-	3.8	14.4
RAFTStereo* [21]	2021		1.13	5.69	0.9	4.35	1.92	12.6	1.09	8.65	3.31	26.39	0.36	3.3
IGEV* [47]	2023		1.21	6.03	1.03	5.13	2.63	11.93	2.27	9.49	5.02	26.04	0.33	4
ELFNet* [23]	2023		2.31	7.68	1.36	5.85	5.16	17.5	2.16	10.14	-	-	-	-
Mocha-Stereo* [7]	2024		1.29	5.97	1.02	4.83	2.66	10.18	2.49	7.96	3.84	**24.16**	0.28	3.47
NMRF* [12]	2024		1.17	5.31	0.92	4.63	2.91	13.36	2.73	10.90	-	-	0.31	3.8
Selective-RAFT* [43]	2024		1.27	6.68	1.08	5.19	2.34	12.04	2.05	9.45	4.17	27.4	0.34	4.36
Selective-IGEV* [43]	2024		1.25	6.06	1.08	5.64	2.59	11.79	2.31	9.22	4.35	28.10	0.33	4.05
HVT-RAFT [5]	2023	✓	1.12	**5.20**	0.87	3.7	1.37	10.40	-	-	-	-	0.29	3.00
NerfStereo* [41]	2023	✓	1.14	5.41	**0.84**	**3.6**	1.42	9.67	0.91	6.39	4.09	29.89	0.29	2.94
RAFT-Stereo + ME			1.18	6.18	0.87	4.19	1.42	9.73	1.11	7.00	3.06	26.50	0.26	2.31
Ours			**1.12**	5.60	**0.87**	4.10	**1.15**	**8.39**	**0.85**	**5.67**	**2.89**	26.50	**0.25**	**1.88**

Table 1. Generalization from SceneFlow dataset to KITTI2015, KITTI 2012, Middlebury (H), and ETH3D dataset. 'ME' represents our monocular encoder module. * represents the results evaluated in our metrics and settings using official models and weights. 'All', 'NonOcc', and 'Occ' represent all regions, non-occluded regions, and occluded regions, respectively.

Method	Additional Data/Aug	Booster (Q)											
		ALL				Trans				NonTrans			
		EPE	bad 2.0	bad 3.0	bad 5.0	EPE	bad 2.0	bad 3.0	bad 5.0	EPE	bad 2.0	bad 3.0	bad 5.0
Mocha-Stereo [7]		3.88	16.82	14.31	11.84	9.45	66.44	57.96	45.73	2.89	12.31	10.19	8.38
ELFNet [23]		6.05	24.51	20.43	16.40	9.03	72.07	62.73	49.82	5.33	20.85	17.18	13.84
Selective-RAFT [43]		4.14	19.52	16.69	13.63	10.34	69.84	61.64	49.55	2.99	14.99	12.44	10.00
Selective-IGEV [43]		4.62	19.28	16.58	13.92	9.50	66.85	58.9	47.15	3.60	14.74	12.34	10.27
IGEV [47]		4.26	17.58	15.21	12.89	10.00	68.96	61.14	49.51	3.25	12.99	10.94	9.24
NMRF [12]		5.05	26.22	21.31	16.58	10.36	70.92	60.93	47.16	4.00	22.43	17.77	13.50
NerfStereo [41]	✓	3.48	13.40	11.13	9.22	8.88	62.67	53.35	41.79	2.49	9.06	7.19	5.89
RAFTstereo [21]		4.18	17.64	14.92	12.23	9.79	67.69	59.31	47.40	3.23	13.13	10.75	8.70
RAFT-Stereo + ME		2.40	11.44	9.17	7.30	8.97	64.84	56.05	43.95	**1.45**	**6.96**	5.08	3.89
Ours		**2.26**	**11.02**	**8.59**	**6.6**	**7.93**	**59.83**	**50.36**	**38.44**	1.52	6.98	**4.97**	**3.64**

Table 2. Generalization from SceneFlow dataset to Booster dataset in quarter resolution and balanced set. 'ME' represents our monocular encoder module. 'All', 'Trans', and 'NonTrans' represent all regions, transparent regions, and nontransparent regions, respectively.

Method	Additional Data/Aug	DrivingStereo			
		Sunny	Cloudy	Rainy	Foggy
RAFTStereo		1.01	0.97	1.8	0.95
IGEV		1.11	1.11	2.32	1.14
MoCha-Stereo		1.01	0.99	1.35	0.98
Selective-IGEV		1.18	1.13	2.22	1.12
NerfStereo	✓	**0.90**	**0.91**	1.46	1.01
Ours		**0.93**	**0.92**	**1.29**	**0.93**

Table 3. Generalization from SceneFlow to DrivingStereo. EPE is used as the evaluation metric.

their results with some implicit assumptions, such as evaluating only pixels with ground truth disparity less than 192 or only evaluating non-occluded regions. In our experiments, unless otherwise specified, both ours and compared methods consider all regions as the classical metric does without limitations. For the Middlebury dataset, we evaluate

Method	All		NonOcc		Occ	
	EPE ↓	bad 2.0 ↓	EPE ↓	bad 2.0 ↓	EPE ↓	bad 2.0 ↓
DA V2 - M	205.04	99.99	207.51	99.99	196.96	99.98
DA V2 - GA	5.83	69.28	5.61	69.16	6.95	69.34
Metric3D	33.14	97.18	33.05	97.06	34.54	98.09
Ours	**1.15**	**8.39**	**0.85**	**5.67**	**2.89**	**26.50**

(a) Metric disparity space

Method	All		NonOcc		Occ	
	δ^1 ↑	RMS ↓	δ^1 ↑	RMS ↓	δ^1 ↑	RMS ↓
DA V2 - M	0.022	6.356	0.024	6.171	0.008	7.189
DA V2 - GA	0.923	1.487	0.934	1.342	0.852	2.022
Metric3D	0.288	4.097	0.298	4.014	0.223	4.588
Ours	**0.985**	**0.677**	**0.991**	**0.492**	**0.948**	**1.282**

(b) Metric depth space

Table 4. Comparison to DepthAnything V2 and Metric3D on Middlebury. 'M': the fine-tuned metric version. 'GA': alignment using the same registration parameters from GT for all pixels.

both all regions and non-occluded regions. For the Booster dataset, we evaluate all regions, as well as transparent and

Exp	Middlebury	
	epe	bad 2.0
Baseline	2.11±0.16	14.12±0.64
Baseline w/o mono feature	1.83±0.11	12.45±0.86
Baseline + ME	1.42±0.01	9.81±0.18
Baseline + ME + IDF	1.41±0.04	10.34±0.19
Baseline + ME + PF	1.41±0.00	9.71±0.00
Baseline + ME + ILF	1.20±0.08	9.06±0.70
Baseline + ME + ILF + GF	1.15±0.01	8.35±0.04

Table 5. The effectiveness of each module. The baseline is RAFT-Stereo, while ME is our monocular encoder, DF is iterative direct fusion, ILF is iterative local fusion, PF is post-fusion, and GF is global fusion. w/o mono feature means removing the context features from RAFT-Stereo.

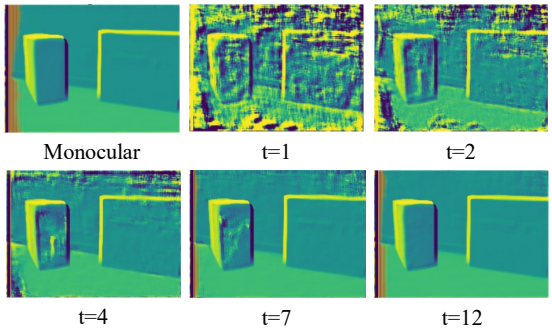

Figure 3. The visualization of local ordering map. The monocular represents the results from monocular depth. $t = x$ represents the results from binocular disparity.

non-transparent regions. (3) Additionally, we observe fluctuations in model performance when trained with different numbers of steps. To fully analyze the improvement contributed by each model component, we calculate the mean and standard deviation (std) of results from the last 100k, 90k, and 80k training steps. We use $mean \pm std$ to measure the accuracy and robustness of our model.

4.3. In-the-wild Generalization Ability

As shown in Table 1, our method achieves state-of-the-art results across all datasets, with particularly strong performance on Middlebury and ETH3D. Compared to other methods that do not use additional data or augmentation, we almost double their performance. Even when compared to methods incorporating additional data or augmentation, our approach leverages limited stereo data to achieve superior results. Furthermore, as presented in Table 2, our method demonstrates substantial improvements in the Booster dataset. Compared to methods without additional data or augmentation, we nearly double the improvement on EPE and Bad 5.0 across all regions, achieve more than a 10-point improvement on Bad x.0 in transparent regions, and show double or even triple the improvement in non-transparent regions. For more detailed quantitative results and analysis, please refer to our supplementary materials.

Exp	LBP Kernel	S	OP	r	Middlebury	
					epe	bad 2.0
L(1)	1		L	1	1.38±0.06	10.20±0.53
L(2)	3		L	1	1.36±0.04	10.10±0.25
L(3)	3	✓	L	1	1.44±0.06	9.65±0.26
L(4)	5,3	✓	L	1	1.20±0.08	9.06±0.70
L(5)	9,7,5,3	✓	L	1	1.32±0.07	9.57±0.62
L(6)		✓	C	1	1.32±0.14	9.53±1.03
L(7)		✓	DC	1	1.39±0.03	9.71±0.29
L(8)	13,11,9,7,5,3	✓	L	1	1.31±0.02	9.89±0.04
L(9)	5,3,	✓	L	2	1.32±0.03	9.45±0.14
L(10)	5,3	✓	L	3	1.26±0.09	9.42±0.42

Table 6. Ablation study on iterative local fusion. 'S' is sigmoid function in LBP-like operation. 'OP' represents the type operation, 'L' is the LBP-like operation, 'C' is the convolution, 'DC' is the deeper convolution, and 'r' is the amplitude parameter.

Exp	Reg	Confidence	Middlebury	
			epe	bad 2.0
G(1)		Cost	1.23±0.03	9.69±0.38
MonoDepth	✓		1.19±0.02	8.72±0.20
G(2)	✓	Cost	1.18±0.02	8.77±0.07
G(3)	✓	Hybrid	1.15±0.01	8.35±0.04

Table 7. Ablation study on the global fusion. 'Reg' means registration for monocular depth. 'Cost' means we estimate the confidence from the sampled cost volume. 'Hybrid' means we estimate the confidence from the concatenation of sampled cost volume, hidden state, and guidance from the last iteration. 'MonoDepth' means evaluation of the registered monocular depth.

We also provide visualization results on the Booster dataset to show the zero-shot generalization ability of our method in the wild world. As illustrated in Figure 1, our method significantly improves performance in various challenging regions, such as areas with occlusion, textureless surfaces, reflections, and transparent regions. Due to space limitations, additional visualization results are available in the supplementary materials.

4.4. Ablation Study and Analysis

We conduct comprehensive ablation studies to analyze the impact of each module and illustrate the construction process of our model. It is important to note that each ablation study involves training the model from scratch rather than removing a component from an already well-trained model.

The Effectiveness of Each Module. As shown in Table 5, the baseline model performs better without context features, indicating that monocular priors are susceptible to domain bias when data is limited. By incorporating less-biased monocular priors from a pre-trained large monocular network in the monocular encoder (ME), generalization performance is significantly improved, highlighting the importance of robust monocular priors in the wild world. Comparing the Baseline + ME with Baseline in Table 1, its performance becomes worse than Raft-Stereo, showing that it is easy to suffer over-confidence when simply fusing

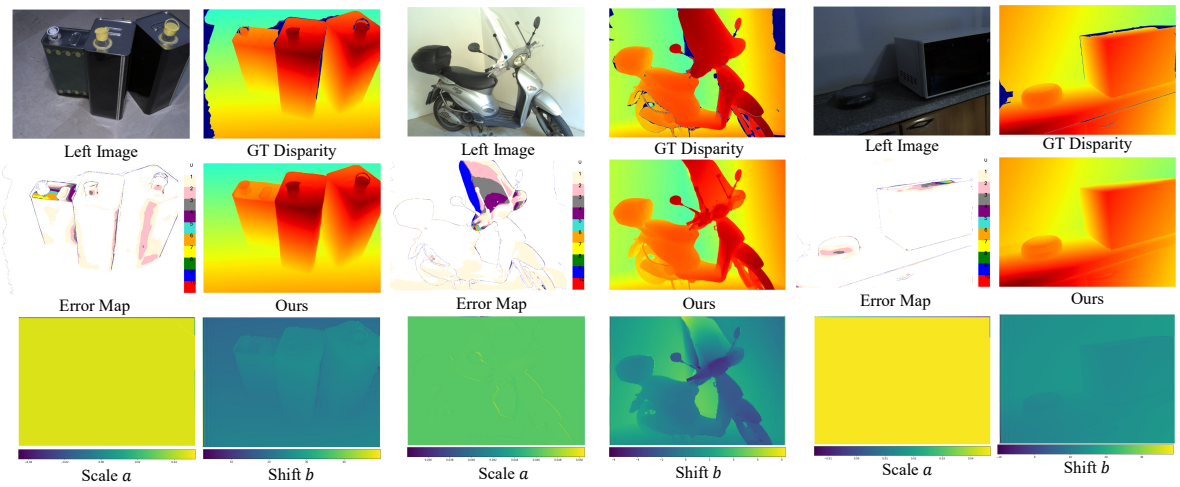

Figure 4. The visualization of registration parameters, scale a and shift b.

monocular features with disparities during iterative disparity update. The iterative direct fusion method (IDF) fuses monocular depth and binocular disparity through direct concatenation and convolution at each iteration. Compared to this approach, our iterative local fusion (ILF) is more robust to noise in binocular disparity, resulting in superior performance. The post-fusion method (PF) fuses monocular depth with the optimized binocular disparity from the previous iteration without registration. Compared to this approach, our global fusion (GF) achieves better compatibility between monocular depth and binocular disparity, mitigating the noise caused by scale ambiguity during fusion. Our iterative local fusion and global fusion modules further enhance performance and improve model robustness when combined with the monocular encoder. It is also noted that the time cost of the Baseline is 0.32s while our model is 0.4s. Even though it involves a VTF model, benefiting from the elegant and controllable design, our model barely raises the time cost.

The Analysis of Iterative Local Fusion. We also analyze the specific configurations of iterative local fusion. As shown in Table 6, the fixed weights in the LBP-like operation have a slight impact on performance, with a kernel size of $3, 5$ providing optimal results. We also try to use convolutions with learnable weights to replace LBP-like convolutions. Comparing L(4), L(9-10) with L(6-7), we find that fixed-weight convolutions are more robust than learnable convolutions, and deeper learnable convolutions produce worse results. This is because limited data makes monocular-related learning unreliable for generalization, whereas manually designed convolutions incorporate prior knowledge and are less affected by data bias. Using a sigmoid function after LBP-like convolutions further improves overall performance. The amplitude parameter does not show a significant influence. We also visualize the local ordering map in Figure 3. The local ordering maps of predicted disparity gradually become similar to the result of monocular depth as the iteration increases. For more visualizations, please refer to our supplemental materials.

The Analysis of Components in Global Fusion. We analyze the specific configurations of global fusion, as shown in Table 7. Comparing G(1) with G(2), global fusion achieves nearly a 1-point improvement in the Bad 2.0 metric after registration. Comparing G(2) with G(3), learning confidence with more information enhances overall performance. Comparing MonoDepth and G(3), the fused results are more robust to monocular depth. We also visualize the registration parameters $\{a, b\}$ in Figure 4. $\{a, b\}$ are changed in different areas but remain inconsistent for every pixel. Due to page limitations, please refer to our supplementary materials for additional failure case analysis and future work discussion.

5. Conclusion

In this paper, we dived into the fusion of monocular priors from VTF stereo matching and found three main problems limiting the fusion process. We proposed a binary local ordering map to unify the relative monocular depth and absolute disparity map. It also guided the fusion between monocular and binocular depth information in an explicit and controllable manner. Besides, we formulated the optimization of the disparity map as a registration process to monocular depth, which can adaptively and globally align the two kinds of depth maps. We designed a network to extract the unbiased monocular priors from the VFM and and leveraged the above two modules to fully exploit the unbiased monocular prior to the stereo matching pipeline to improve generalization in the ill-posed regions. Benefiting from the explicit design, our method barely increased the computation cost. Experimental results demonstrated the effectiveness of our method, with a significant improvement of 10 points on Booster and an error reduction of more than half on Middlebury and ETH3D, without using additional stereo data or data augmentation.

References

[1] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: Self-supervised deep stereo through enhanced monocular distillation. In *European Conference on Computer Vision*, pages 614–632, 2020. 3

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2842–2851, 2022. 3

[3] Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. *arXiv preprint arXiv:2412.04472*, 2024. 1, 3

[4] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *Proceedings of the International Conference on 3D Vision*, pages 364–373. IEEE, 2020. 3

[5] Tianyu Chang, Xun Yang, Tianzhu Zhang, and Meng Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9559–9568, 2023. 3, 6

[6] Zhi Chen, Xiaoqing Ye, Wei Yang, Zhenbo Xu, Xiao Tan, Zhikang Zou, Errui Ding, Xinming Zhang, and Liusheng Huang. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15529–15538, 2021. 3

[7] 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. 1, 3, 6, 12, 13, 14

[8] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. *arXiv preprint arXiv:2501.08643*, 2025. 1, 3

[9] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13022–13032, 2022. 1, 3, 6

[10] James E Cutting and Peter M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pages 69–117. Elsevier, 1995. 3

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[12] 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, 6, 13

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1, 3

[14] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *Proceedings of the European Conference on Computer Vision*, pages 263–279. Springer, 2022. 3

[15] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal. Uncertainty guided adaptive warping for robust and efficient stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3318–3327, 2023. 1, 3

[16] 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. 1, 12

[17] Kwonyoung Kim, Jungin Park, Jiyoung Lee, Dongbo Min, and Kwanghoon Sohn. Pointfix: Learning to fix domain bias for robust online stereo adaptation. In *European Conference on Computer Vision*, pages 568–585. Springer, 2022. 3

[18] Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Los: Local structure-guided stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19746–19756, 2024. 1

[19] Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21539–21548, 2023. 3

[20] 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 International Conference on Computer Vision*, pages 6197–6206, 2021. 3, 6

[21] 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. 5, 6, 12, 13, 14

[22] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13012–13021, 2022. 1, 3, 6

[23] 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. 1, 3, 6, 13

[24] Diogo Martins, Kevin Van Hecke, and Guido De Croon. Fusion of stereo and still monocular depth estimates in a self-supervised learning context. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 849–856. IEEE, 2018. 3

[25] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 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*, pages 4040–4048, 2016. 5

[26] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 5

[27] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 5

[28] Timo Ojala, Matti Pietikainen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th international conference on pattern recognition*, pages 582–585. IEEE, 1994. 4

[29] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. 4

[30] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 1, 3

[31] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21168–21178, 2022. 5

[32] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 12179–12188, 2021. 3

[33] Zhibo Rao, Bangshu Xiong, Mingyi He, Yuchao Dai, Renjie He, Zhelun Shen, and Xing Li. Masked representation learning for domain generalized stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5435–5444, 2023. 1, 3, 6

[34] Rebekka S Renner, Boris M Velichkovsky, and Jens R Helmert. The perception of egocentric distances in virtual environments-a review. *ACM Computing Surveys (CSUR)*, 46(2):1–40, 2013. 3

[35] Ashutosh Saxena, Jamie Schulte, Andrew Y Ng, et al. Depth estimation using monocular and stereo cues. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2197–2203, 2007. 3

[36] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition)*, pages 31–42, 2014. 5

[37] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 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*, pages 3260–3269, 2017. 5

[38] 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*, pages 280–297. Springer, 2022. 6

[39] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Yuexin Ma, Zhe Wang, and Jianping Shi. Adastereo: An efficient domain-adaptive stereo matching approach. *International Journal of Computer Vision (IJCV)*, pages 1–20, 2022. 3

[40] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2019. 3

[41] 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, 6, 13

[42] Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, and Felix Heide. Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13252–13262, 2023. 3

[43] 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. 1, 3, 6, 12, 13, 14

[44] Andrew E Welchman. The human brain in depth: how we see in 3d. *Annual review of vision science*, 2(1):345–376, 2016. 3

[45] Andrew E Welchman, Arne Deubelius, Verena Conrad, Heinrich H Bülthoff, and Zoe Kourtzi. 3d shape perception from combined depth cues in human visual cortex. *Nature neuroscience*, 8(6):820–827, 2005.

[46] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *arXiv preprint arXiv:2501.09898*, 2025. 1, 3

[47] 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. 1, 3, 6, 12, 13, 14

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1, 2, 5

[49] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 3, 5, 12

[50] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 2

[51] Fanqi Yu and Xinyang Sun. Multi-view stereo by fusing monocular and a combination of depth representation methods. In *International Conference on Neural Information Processing (NeurIPS)*, pages 298–309. Springer, 2023. 3

[52] Chenghao Zhang, Kun Tian, Bolin Ni, Gaofeng Meng, Bin Fan, Zhaoxiang Zhang, and Chunhong Pan. Stereo depth estimation with echoes. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022. 3

[53] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Proceedings of the European Conference on Computer Vision*, pages 420–439. Springer, 2020. 3

[54] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13001–13011, 2022. 1, 3, 6

[55] Zhengming Zhou and Qiulei Dong. Two-in-one depth: Bridging the gap between monocular and binocular self-supervised depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9411–9421, 2023. 3

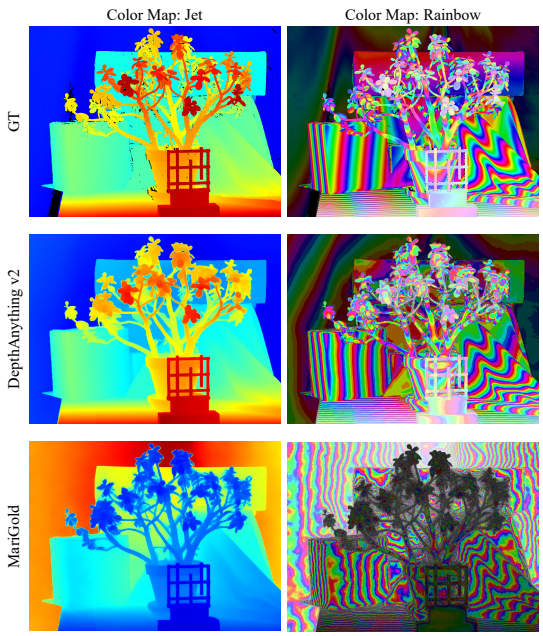

Figure 5. The visualization of results. We use two kinds of colormap to visualize the disparity map.

A. Intuition behind Monocular Depth Model

We choose DepthAnything v2 [49] over Marigold [16] because of the superior continuity of its depth maps. As shown in Figure 5, 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.

B. 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 8 and 9, 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 6 and 7, binocular disparity provides greater detail compared to monocular depth. Our method disentangles monocular depth and binocular disparity, allowing model to leverage both monocular and stereo data, and explore the fusion of monocular priors effectively.

C. 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 10, our method, along with RAFT-Stereo [21], maintains a slower growth rate in memory consumption compared to IGEV [47], Selective IGEV [43], and Mocha [7]. Compared to RAFTStereo, our method exhibits a similar memory consumption increase across resolutions due to the resizing operation required by DepthAnything v2.

D. More Visualization for Generalized Stereo Matching

We provide additional visualizations of generalized stereo matching in Figures 8, 9, 10, 11, and 12. 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.

E. Ablation Study

E.1. More Analysis of Backbone

In addition to replacing the context network with the pre-trained DepthAnything v2 [49], we also experimented with replacing the feature extractor for cost volume construction using DepthAnything v2 [49] and MASt3R [**?** **?**]. As shown in Table 11, the results become worse after replacing the feature extractor for cost volume construction with DepthAnything v2 or MASt3R. Moreover, there is a bug with the A40 GPU that 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.

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[7]		1.30	6.93	5.54	4.18	2.91	23.05	17.67	13.45
Mocha-Stereo 320[7]		1.20	6.18	4.84	3.53	2.88	22.83	17.34	12.98
ELFNet [23]		2.97	14.08	11.38	8.80	5.67	24.68	19.00	14.42
Selective-RAFT [43]		1.35	8.06	6.01	4.01	3.37	27.37	21.87	17.19
Selective-IGEV 192[43]		1.46	8.03	6.19	4.66	3.61	25.57	20.05	15.93
Selective-IGEV 320[43]		1.31	7.27	5.39	3.81	3.51	25.05	19.39	15.18
IGEV 192[47]		1.17	6.67	4.84	3.46	3.76	25.46	20.26	16.39
IGEV 320[47]		1.00	6.07	4.37	2.82	3.60	24.69	19.46	15.70
NMRF [12]		2.76	17.43	13.21	9.51	4.60	32.81	26.08	19.84
NerfStereo [41]	✓	**0.73**	**4.07**	**2.55**	**1.47**	2.41	18.67	13.92	10.56
RAFTstereo [21]		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 8. 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[7]		15.68	53.56	46.23	37.77	9.45	66.44	57.96	45.73
Mocha-Stereo 320[7]		15.05	53.88	46.63	37.62	9.21	65.88	57.30	44.65
ELFNet [23]		22.74	78.89	74.81	69.70	9.03	72.07	62.73	49.82
Selective-RAFT [43]		16.12	55.66	49.87	43.04	10.34	69.84	61.64	49.55
Selective-IGEV 192[43]		20.41	57.55	49.78	42.86	9.50	66.85	58.9	47.15
Selective-IGEV 320[43]		19.81	57.35	49.27	42.10	9.29	66.02	57.91	45.86
IGEV 192[47]		18.55	54.64	46.45	37.79	10.00	68.96	61.14	49.51
IGEV 320[47]		18.00	54.50	46.05	37.72	9.74	68.55	60.49	48.22
NMRF [12]		17.36	56.34	48.33	38.18	10.36	70.92	60.93	47.16
NerfStereo [41]	✓	17.92	45.67	40.39	35.19	8.88	62.67	53.35	41.79
RAFTstereo [21]		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 9. 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.

E.2. More Analysis of Iterative Local Fusion

We provide additional visualizations of the intermediate results from the iterative local fusion process in Figures 13, 14, 15, 16, 17, 18, 19, and 20. 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.

E.3. More Analysis of Components in Global Fusion

We present more visualization for the intermediate results of global fusion in Figure 13, 14, 15, 16, 17, 18, 19, and 20. The visualization shows that the registration of monocular

	750×2484	1125×3726	1500×4968	1688×5589	1875×6210
RAFTStereo reg [21]	2268.35	6023.82	10795.02	14299.5	19666.78
RAFTStereo alt [21]	1715.8	4151.8	6466.7	8157.96	11177.66
IGEV 384 [47]	2816.46	7290.82	14484.61	18810.14	-
IGEV 640 [47]	3167.46	8475.43	17366.83	-	-
Selective IGEV 384 [43]	2960.34	7608.44	15035.55	19505.5	-
Selective IGEV 640 [43]	3311.84	8793.07	18701.57	-	-
Mocha-Stereo 384 [7]	5525.56	12986.73	24665.95	-	-
Mocha-Stereo 640 [7]	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 10. Memory comparison. We evaluate the memory consumption of each model, excluding the feature encoder module, to ensure a fair comparison of backbones during inference. The evaluation is performed across different resolutions. 'reg' denotes pre-computation of the entire cost volume, allowing for look-up operations at each iteration, while 'alt' refers to dynamically computing a thin cost volume at each iteration. The 384 and 640 represent the maximum disparity range used for the resolution of 750×2484. '-' indicates that the GPU does not support the model at the given resolution.

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 11. The effectiveness of each module. Baseline is RAFT-Stereo, while ME is our monocular encoder, ILF is our iterative local fusion, and GF is our global fusion. FE-DepthAnything means using DepthANything v2 to replace the original feature extractor. FE-MASt3R means using MASt3R to replace the original feature extractor.

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 7.

We provide additional visualizations of the intermediate results from global fusion in Figures 13, 14, 15, 16, 17, 18, 19, and 20. 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 7.

F. Future Work Discussion

We present failure cases in Figures 21 and 22. 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 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.

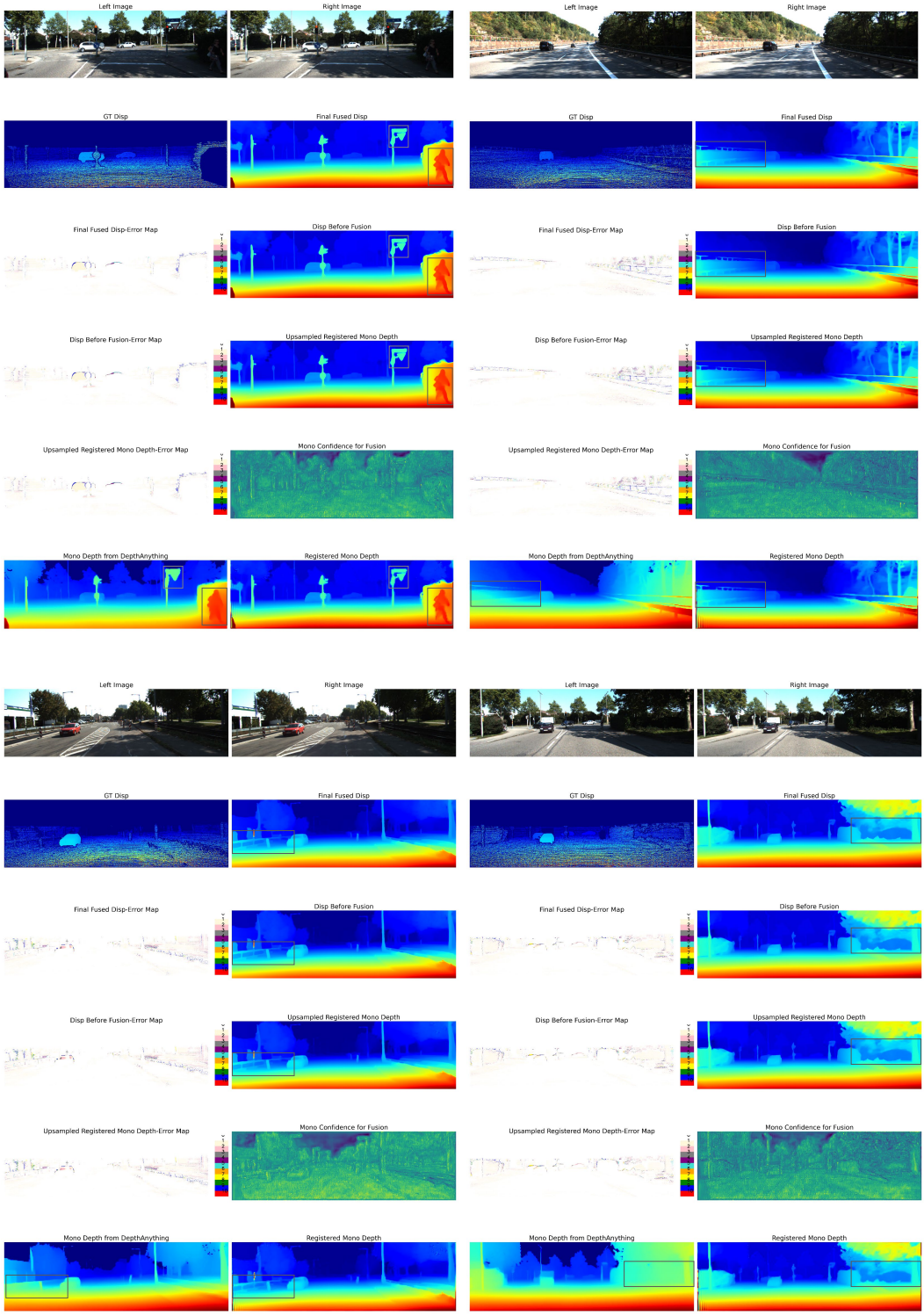

Figure 6. The visualization of binocular disparity and monocular depth. The regions highlighted with gray boxes demonstrate that stereo matching excels at capturing fine-grained details, whereas monocular depth estimation performs better in perceiving overall shapes. The mono depth from DepthAnything is scale ambiguity but not absolute depth before registration.

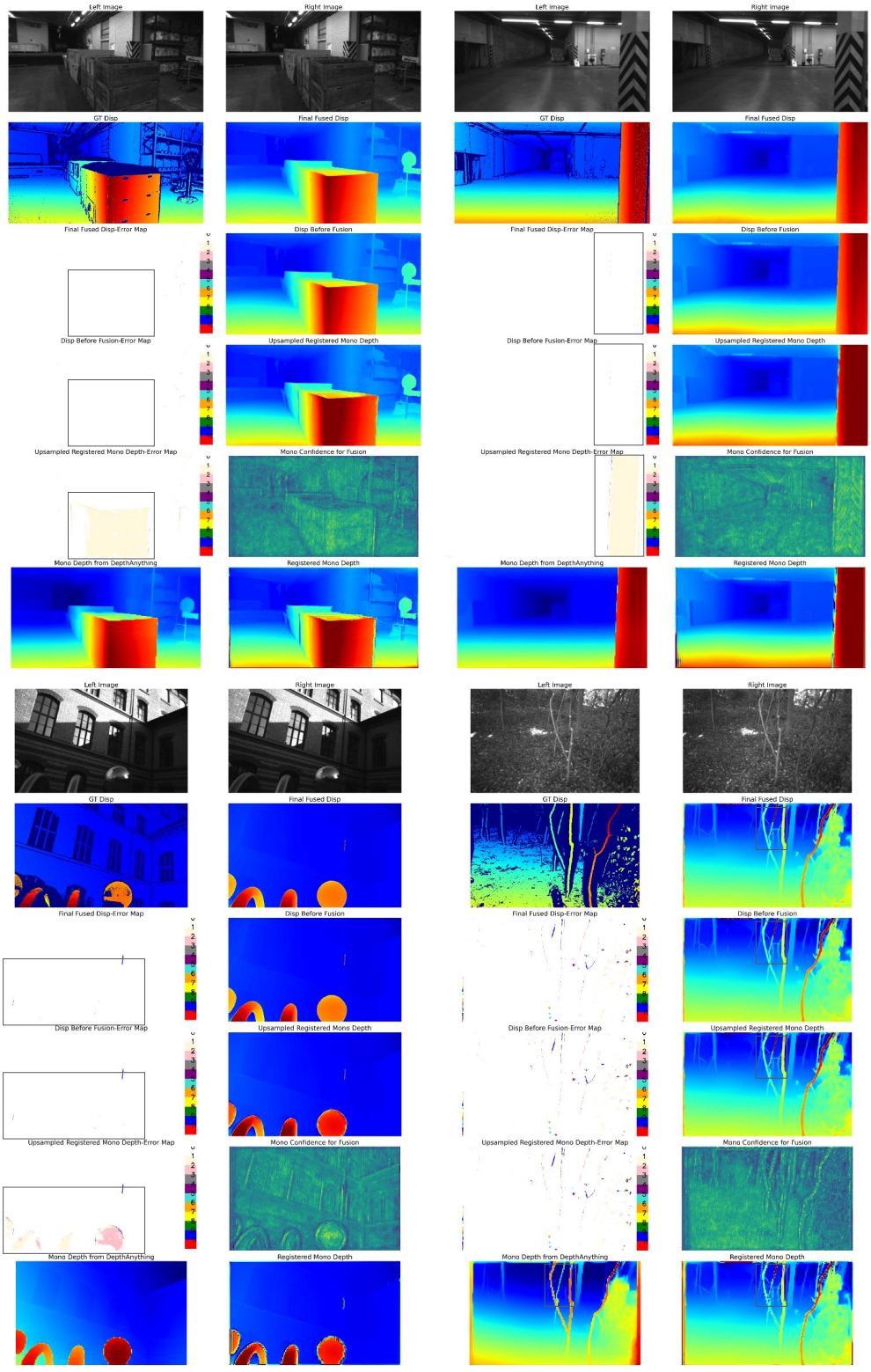

Figure 7. The visualization of binocular disparity and monocular depth. The regions highlighted with gray boxes demonstrate that stereo matching excels at capturing fine-grained details, whereas monocular depth estimation performs better in perceiving overall shapes. The mono depth from DepthAnything is scale ambiguity but not absolute depth before registration.

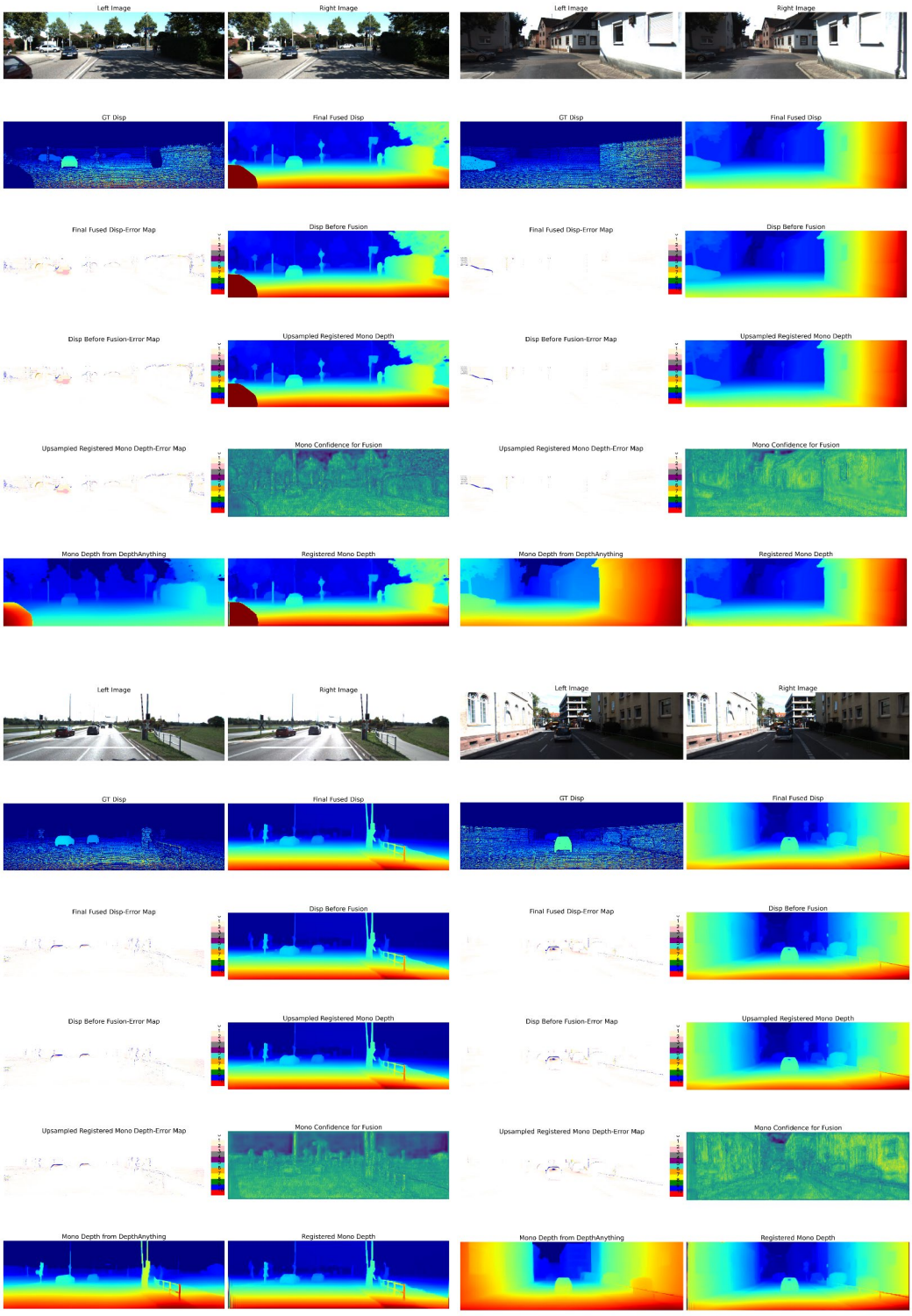

Figure 8. The visualization for generalized stereo matching.

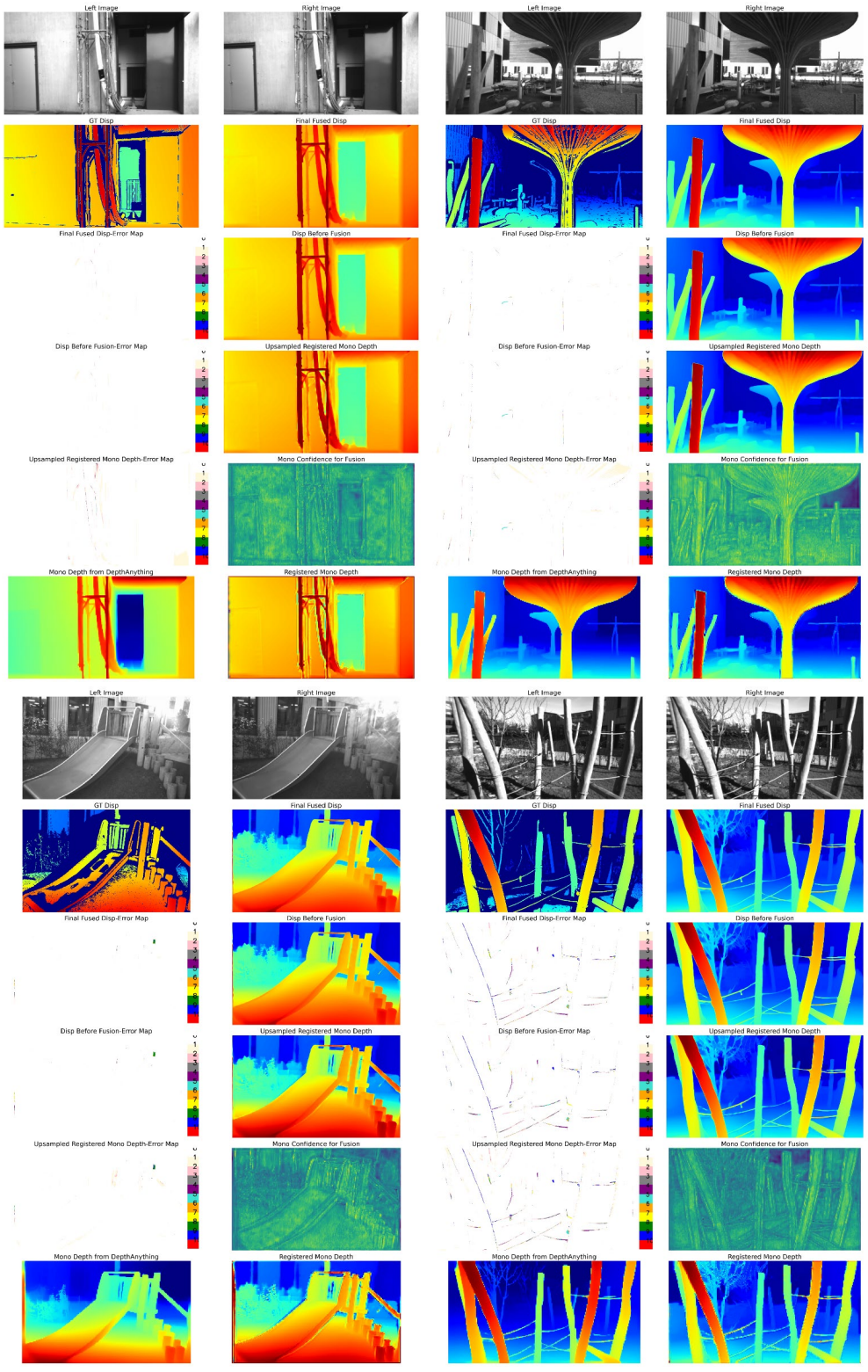

Figure 9. The visualization for generalized stereo matching.

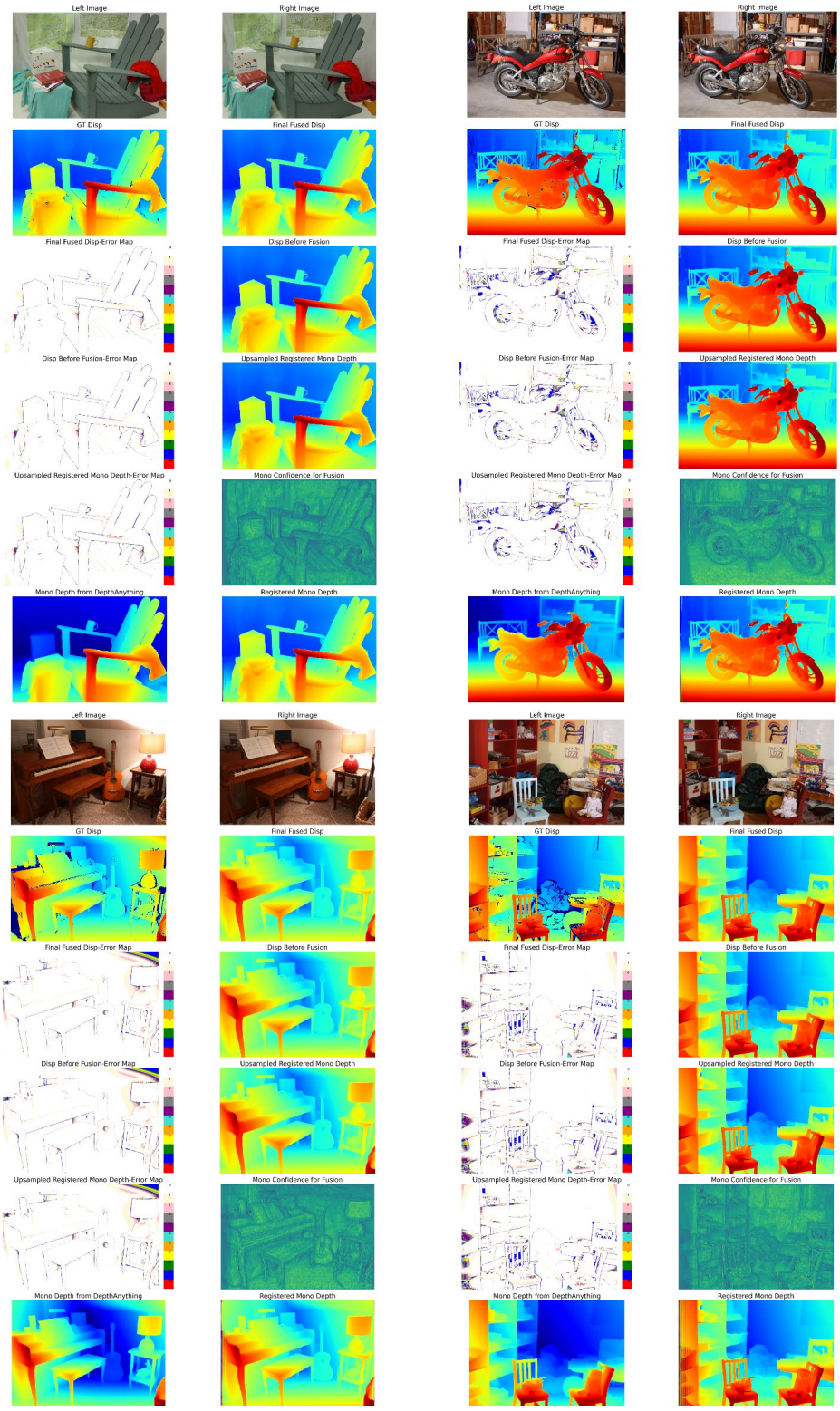

Figure 10. The visualization for generalized stereo matching.

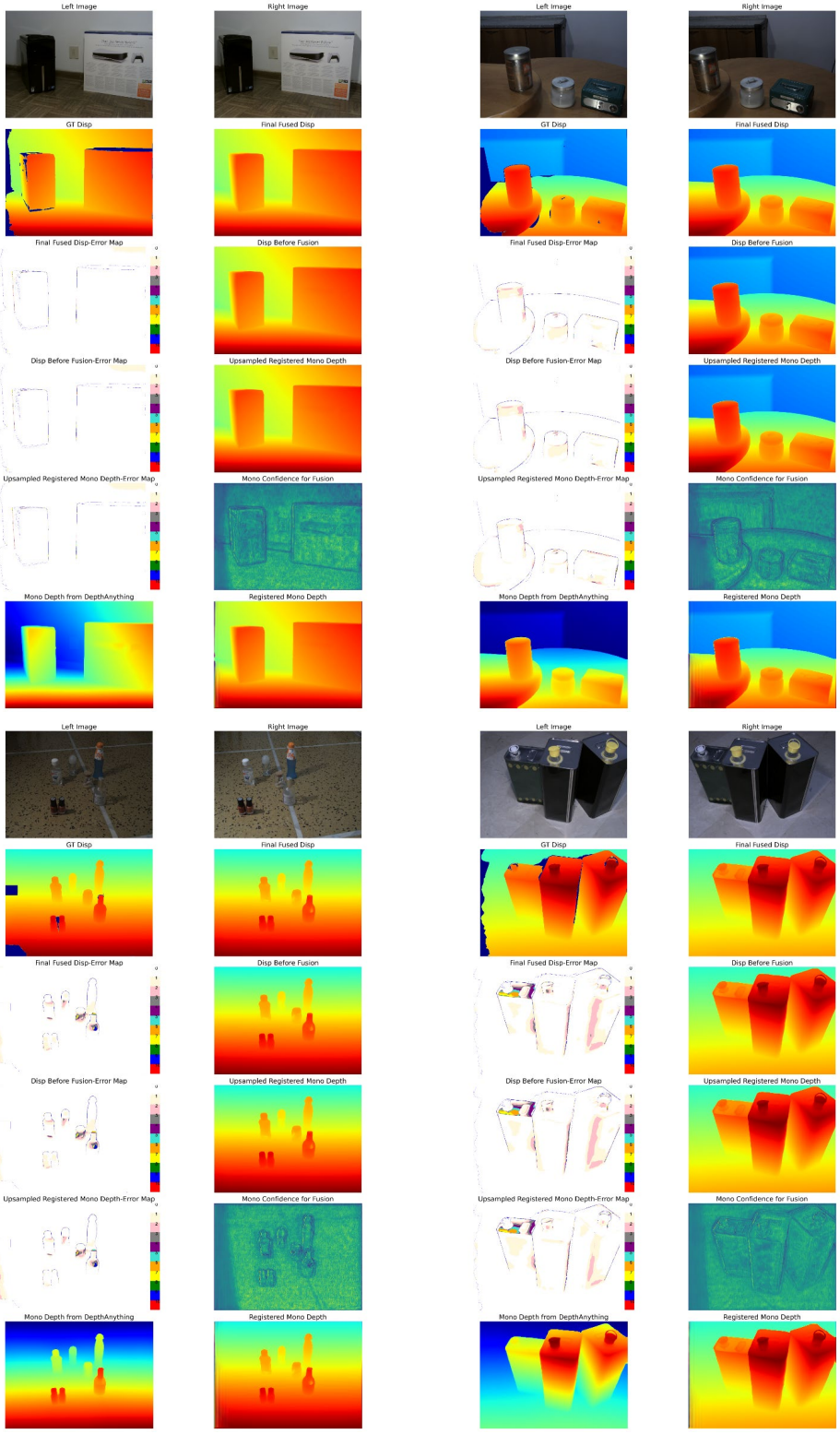

Figure 11. The visualization for generalized stereo matching.

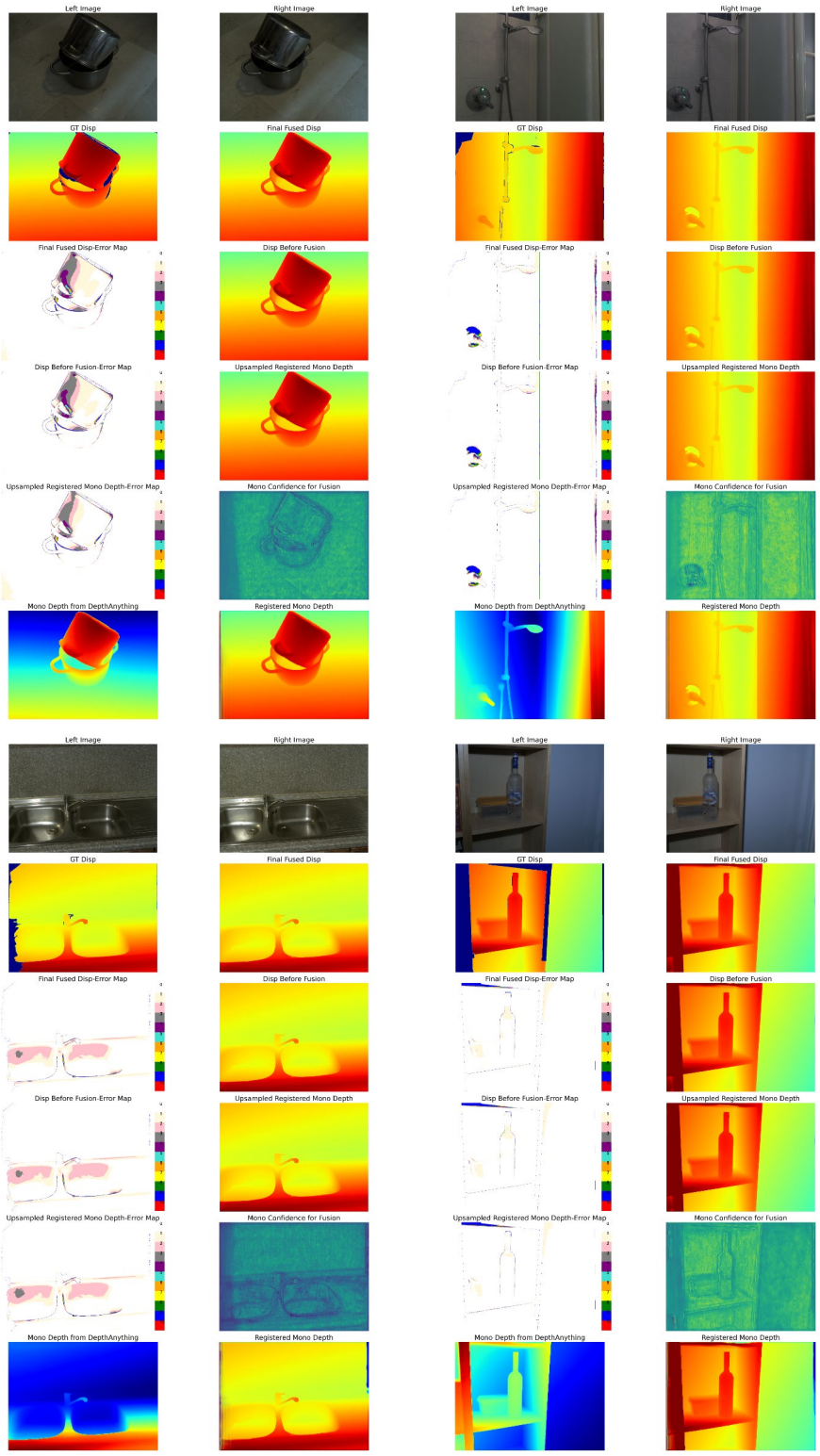

Figure 12. The visualization for generalized stereo matching.

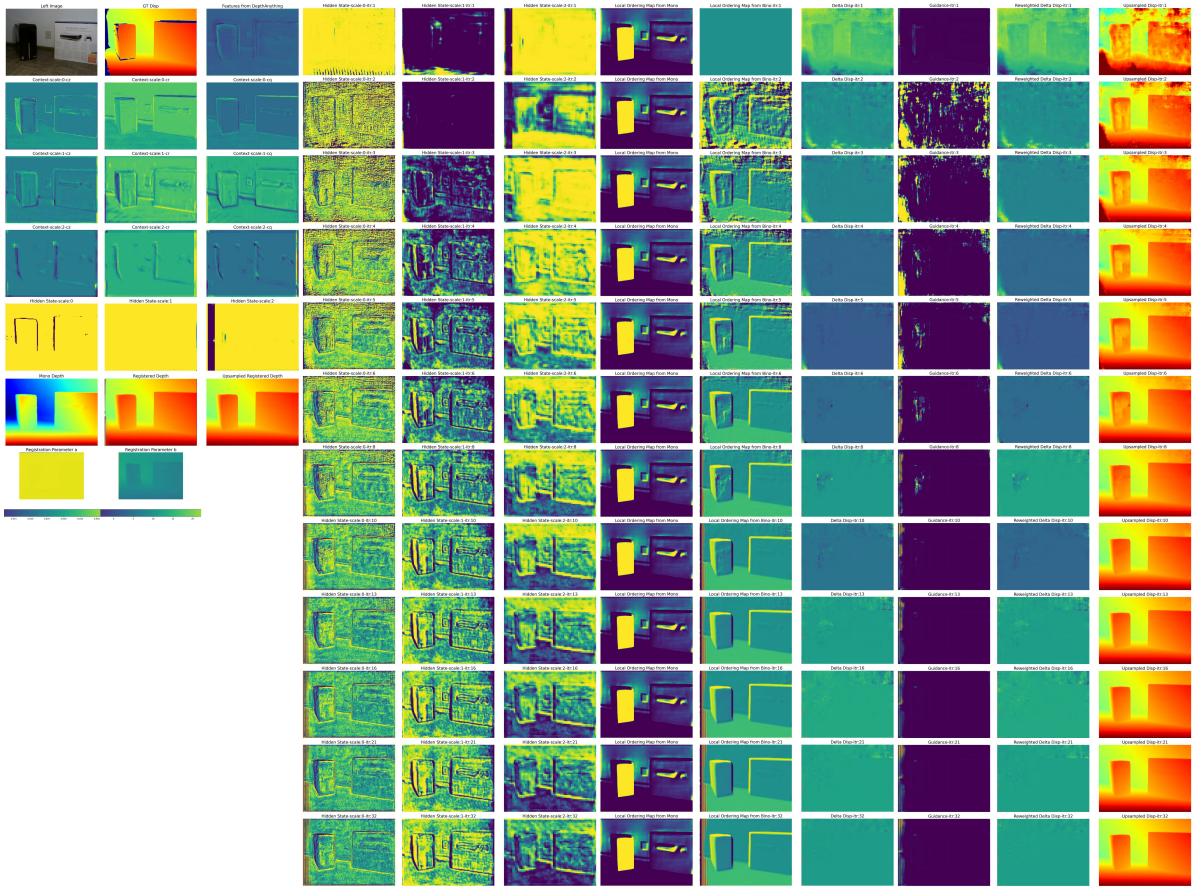

Figure 13. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

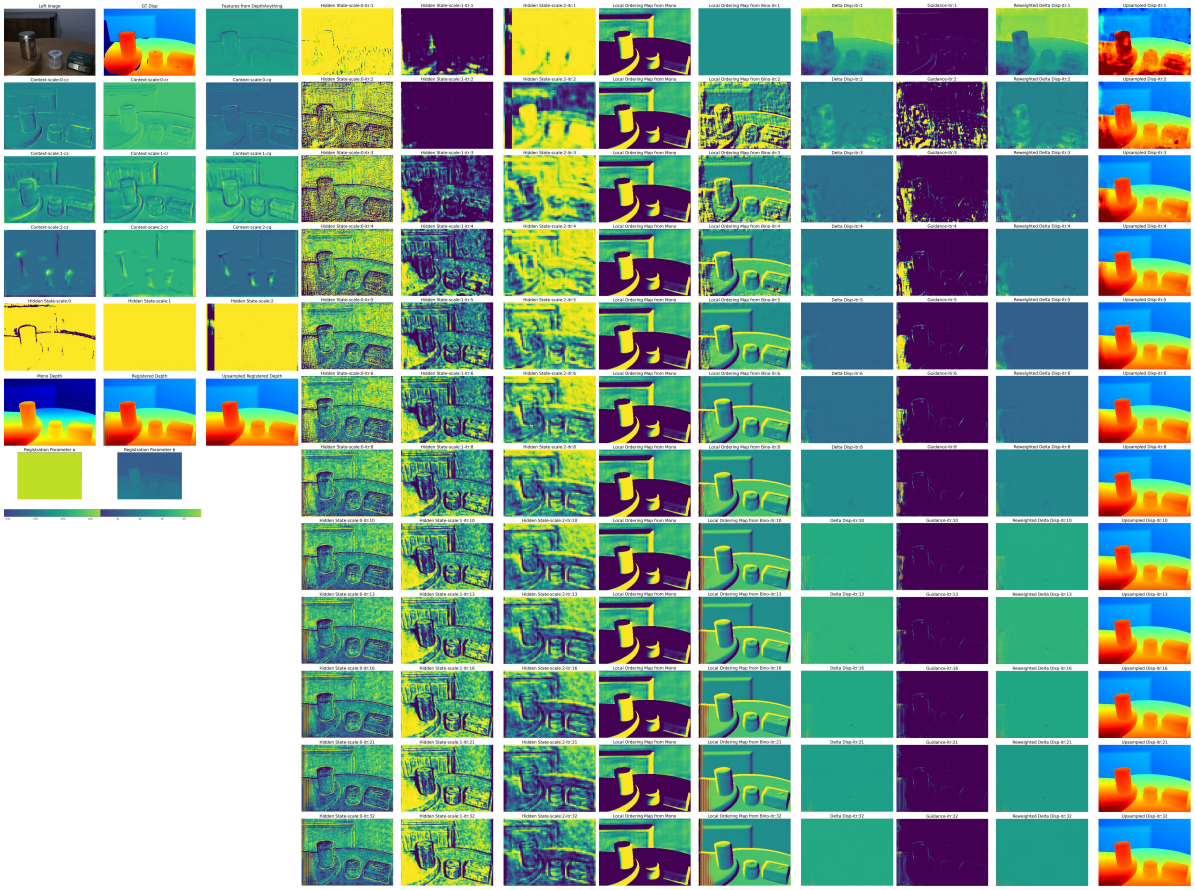

Figure 14. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

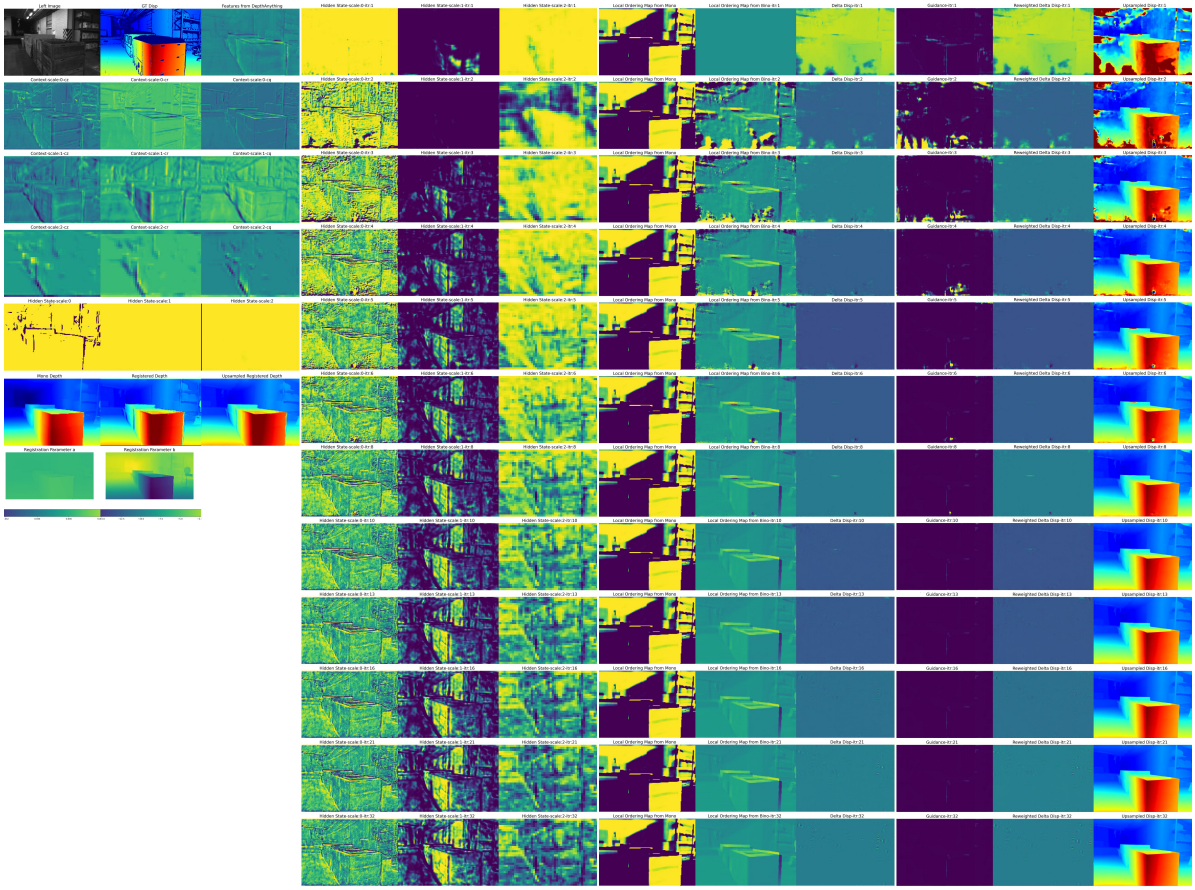

Figure 15. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

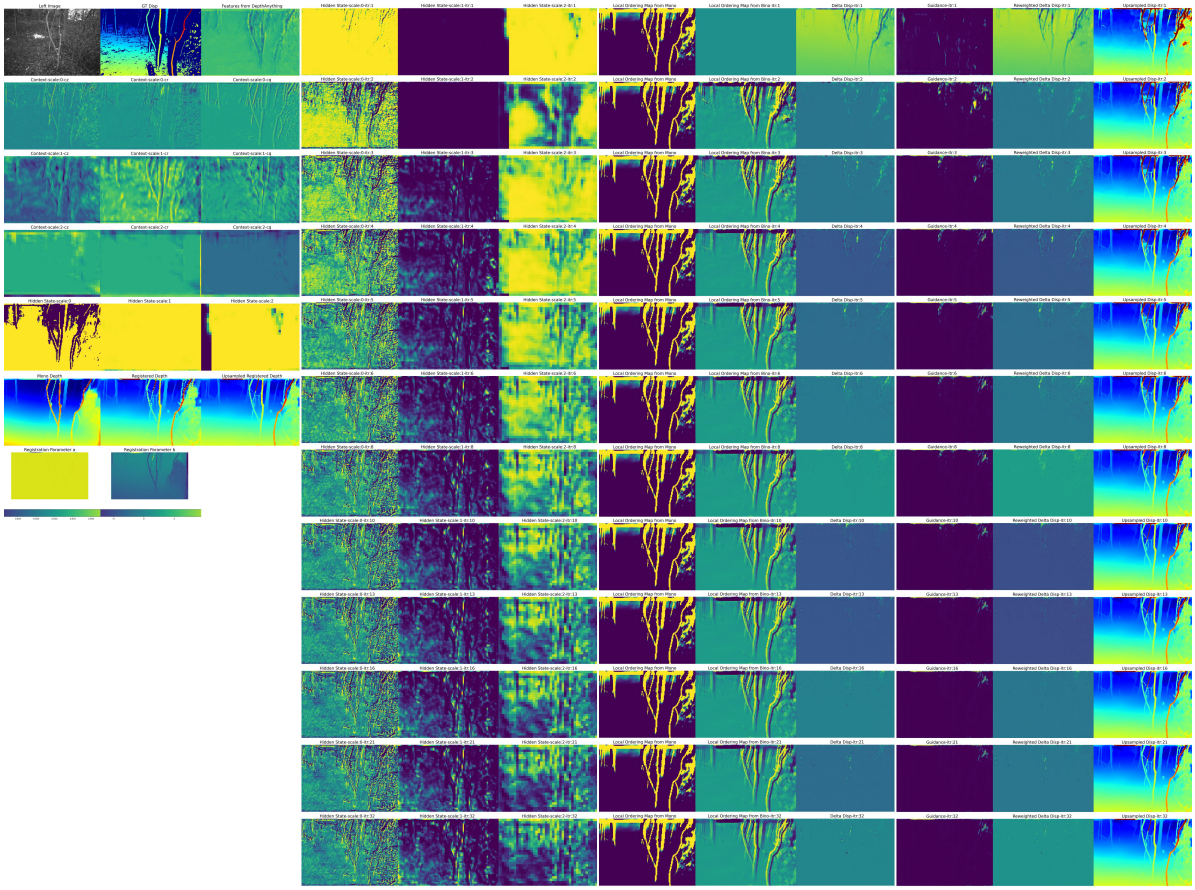

Figure 16. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

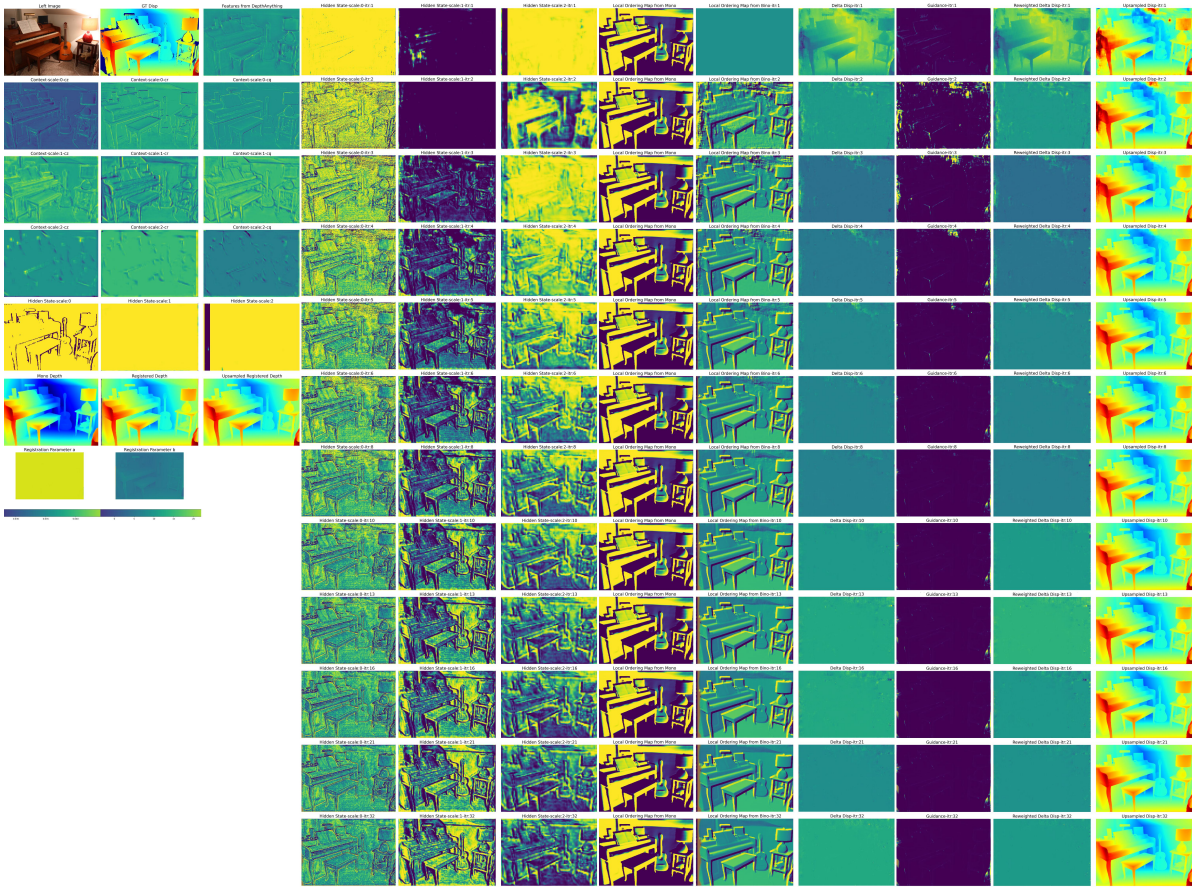

Figure 17. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

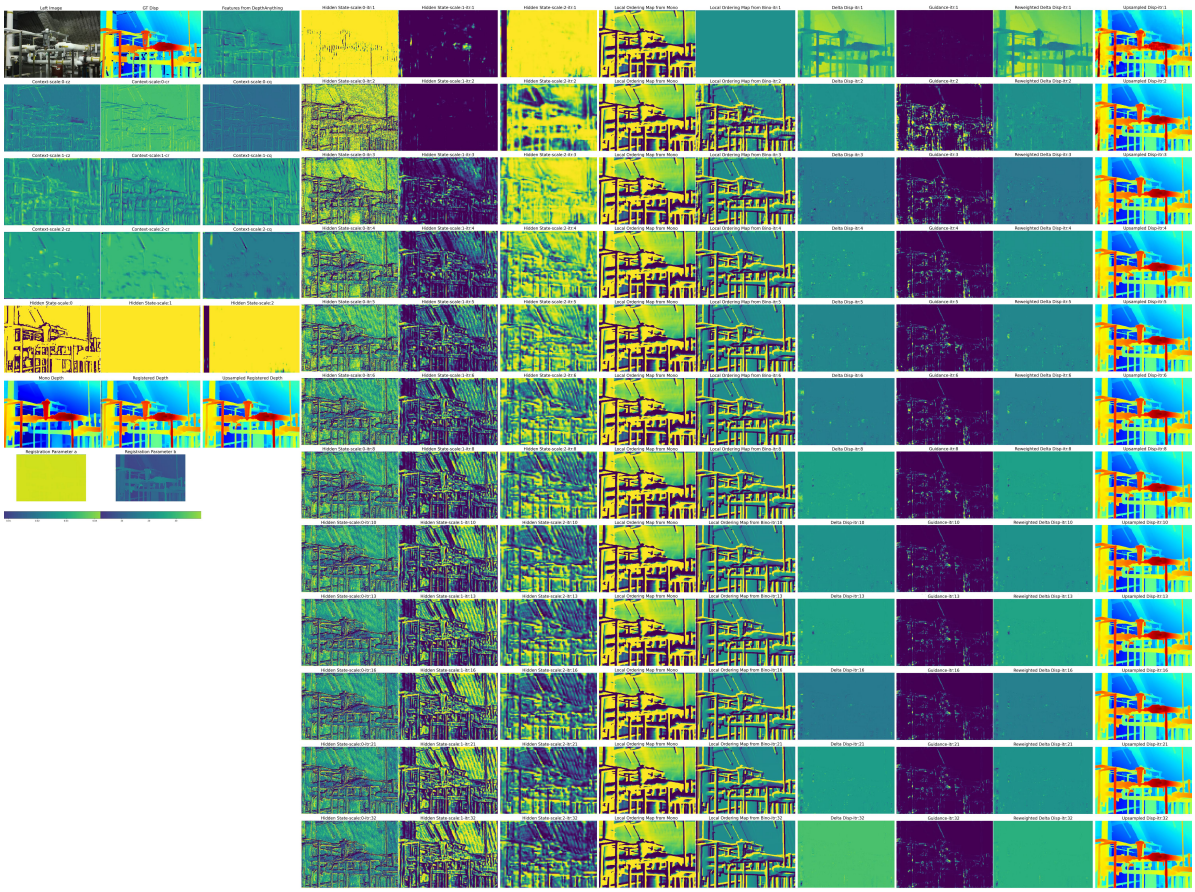

Figure 18. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

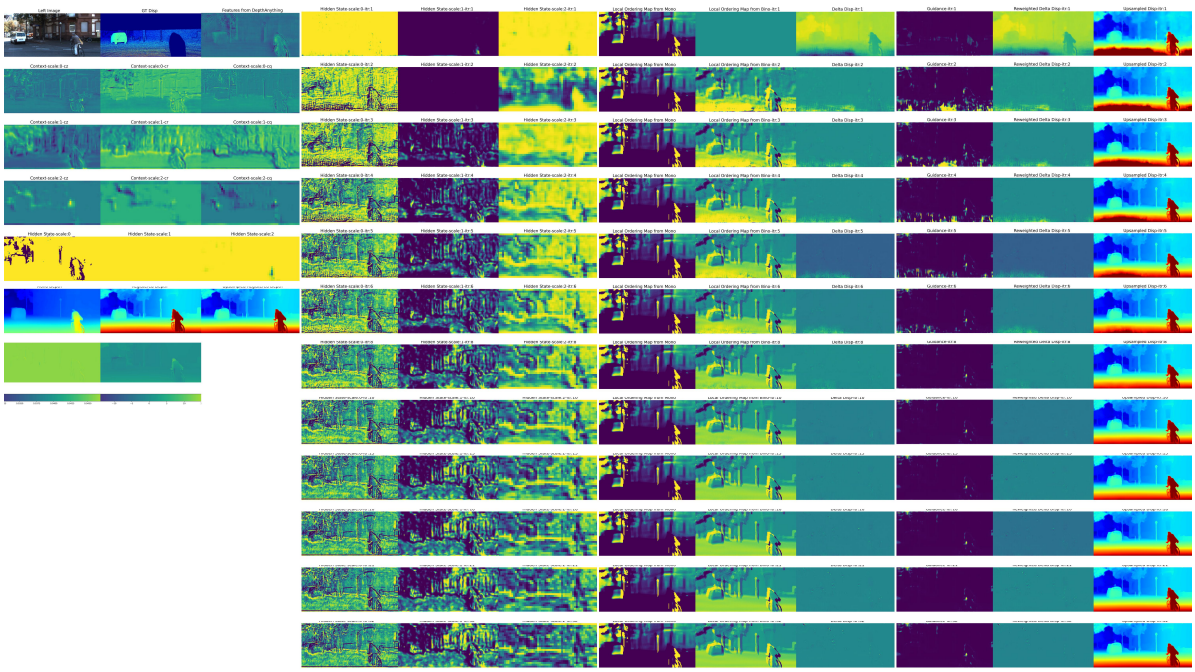

Figure 19. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

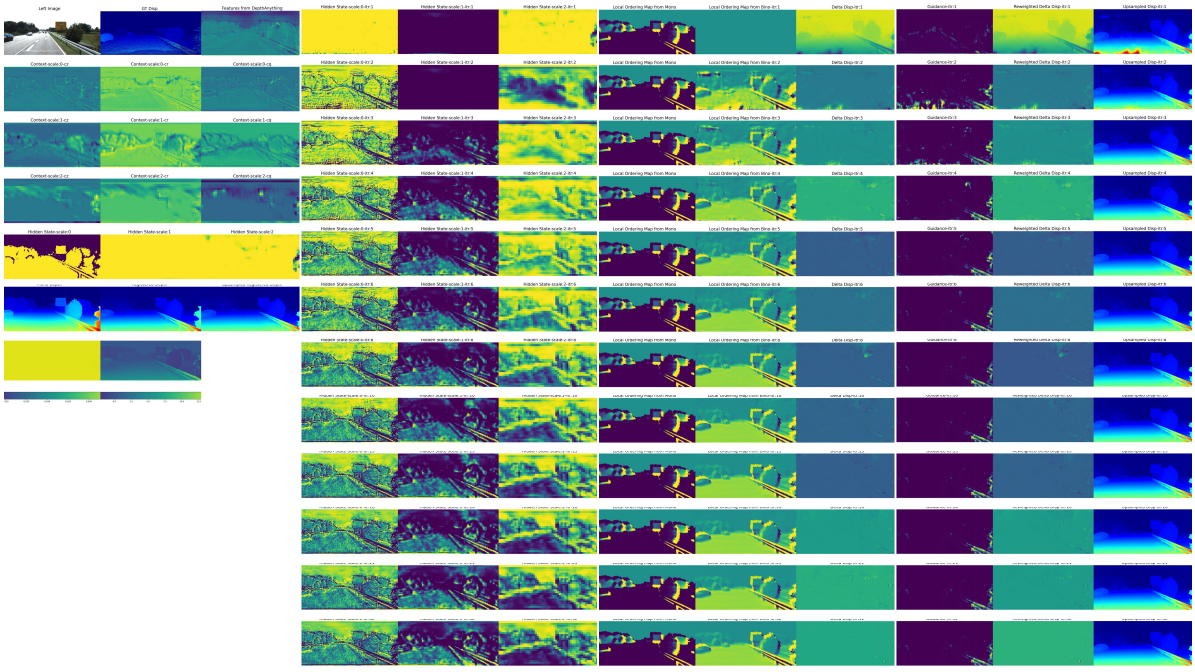

Figure 20. The visualization of intermediate results. itr: the current iteration. cz, cr, cq: context used in GRU. $scale$: scale $0 \sim 2$ represents resolution from high to low.

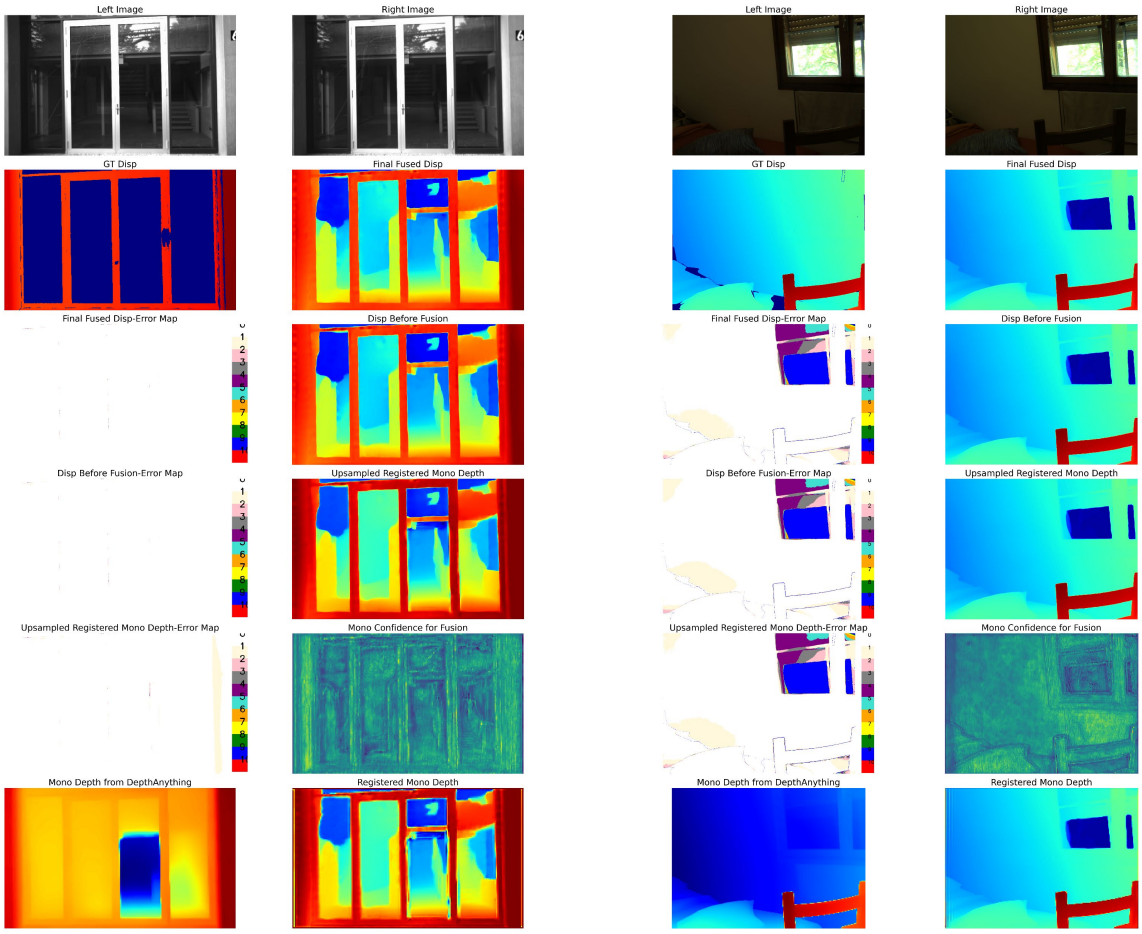

Figure 21. The visualization for failure case analysis.

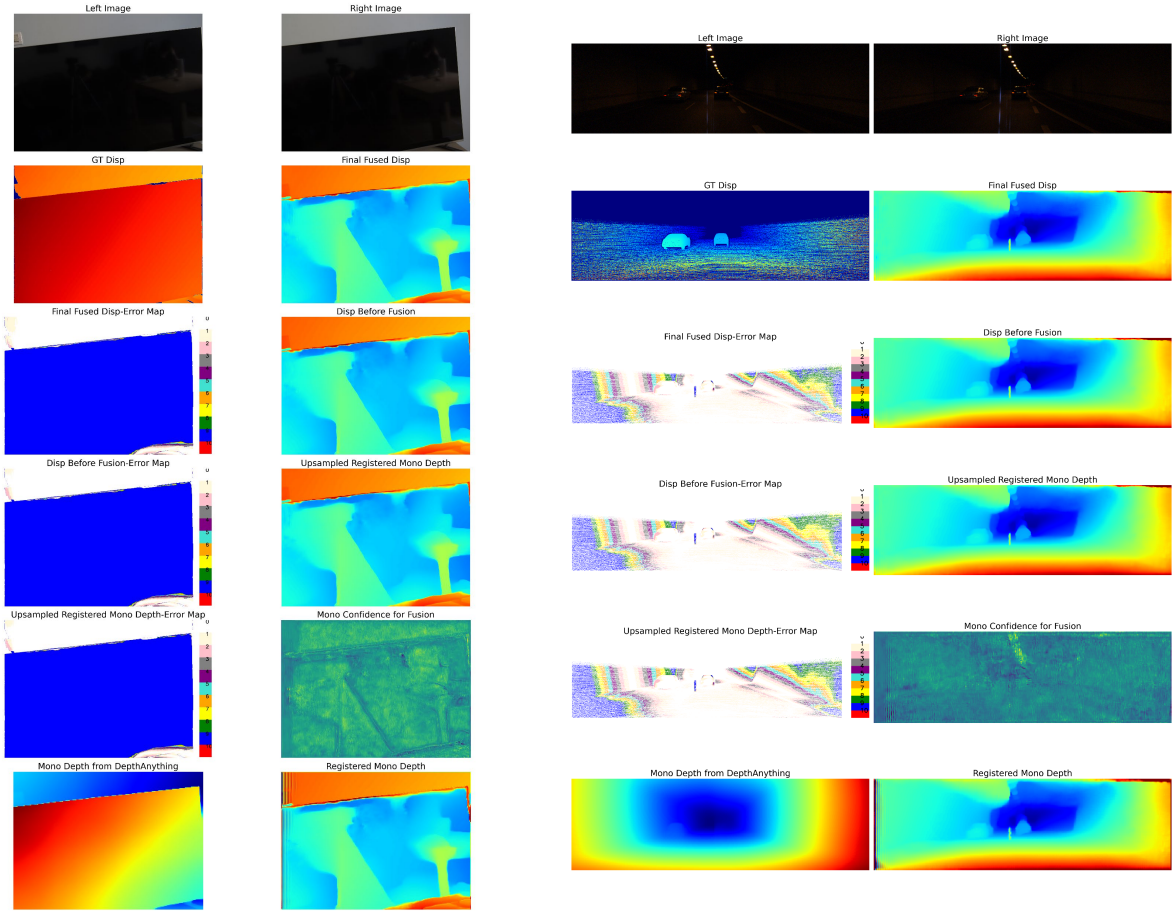

Figure 22. The visualization for failure case analysis.