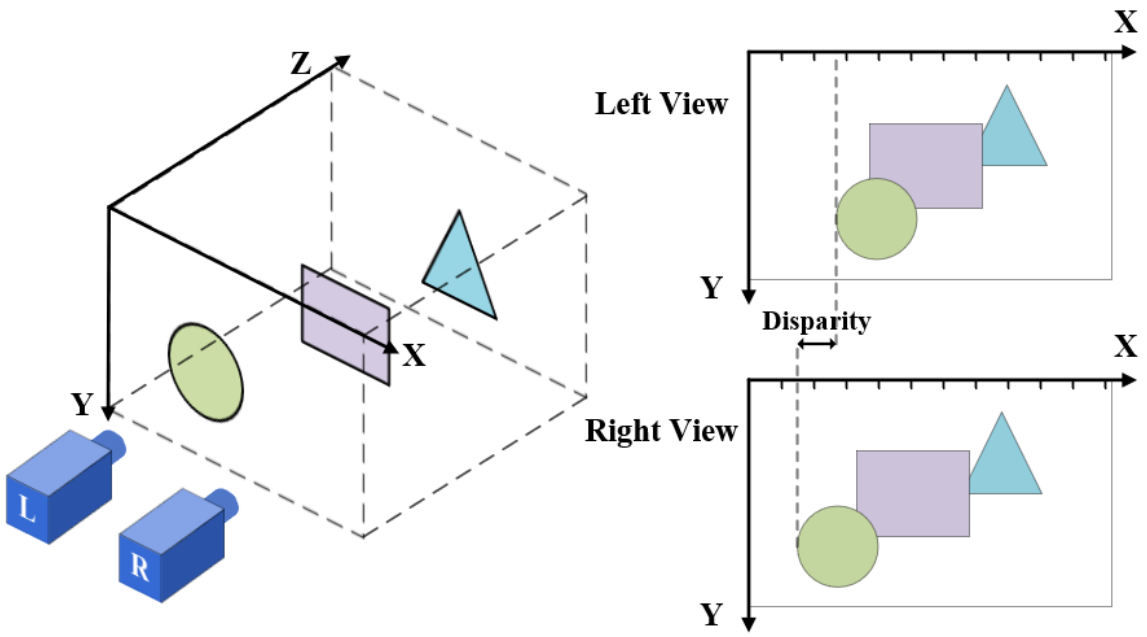
Diving into the Fusion of Monocular Priors for Generalized Stereo Matching

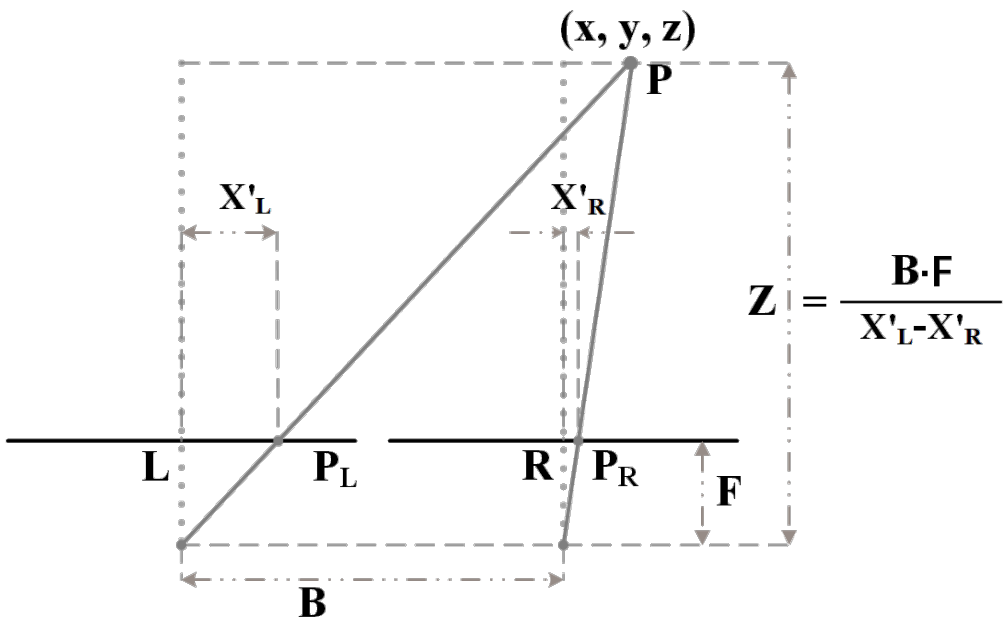
Chengtang Yao[1,2], Lidong Yu[3,4], Zhidan Liu[1,2], Jiaxi Zeng[1,2],
Yuwei Wu[1,2], Yunde Jia[2,1]

[1]Beijing Institute of Technology, China, [2]Shenzhen MSU-BIT University, China, [3]NVIDIA, [4]NEOLIX

Introduction



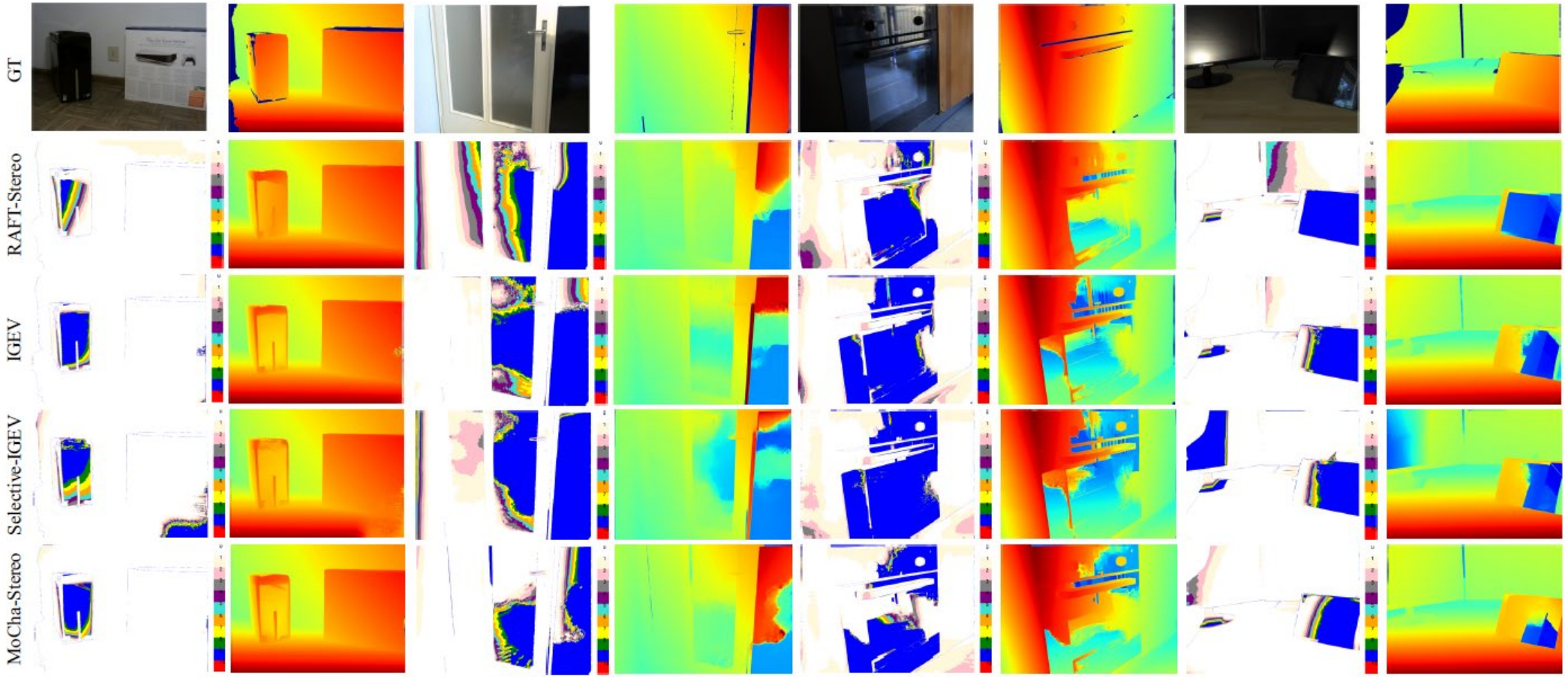
Disparity Estimation



Depth Computation

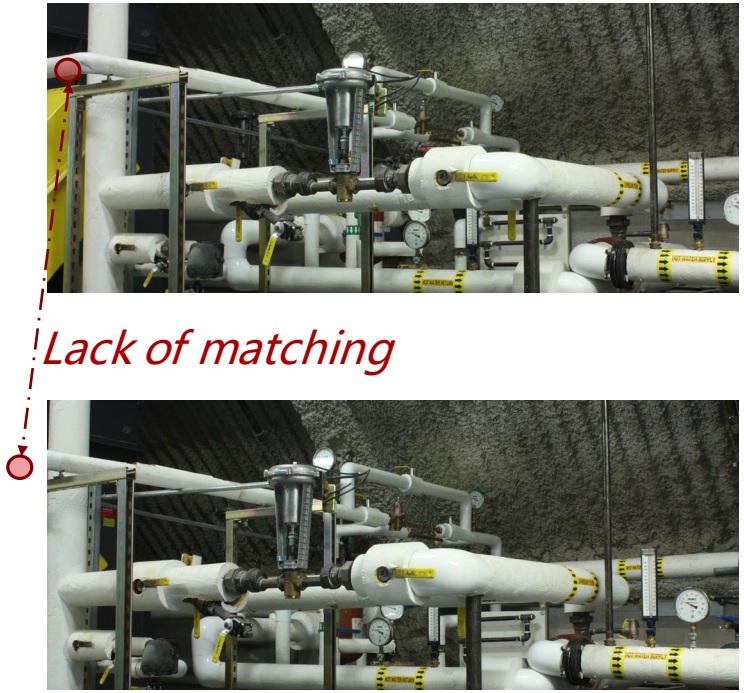
Problem

Challenges on ill-posed regions like occlusions and non-Lambertian surfaces.

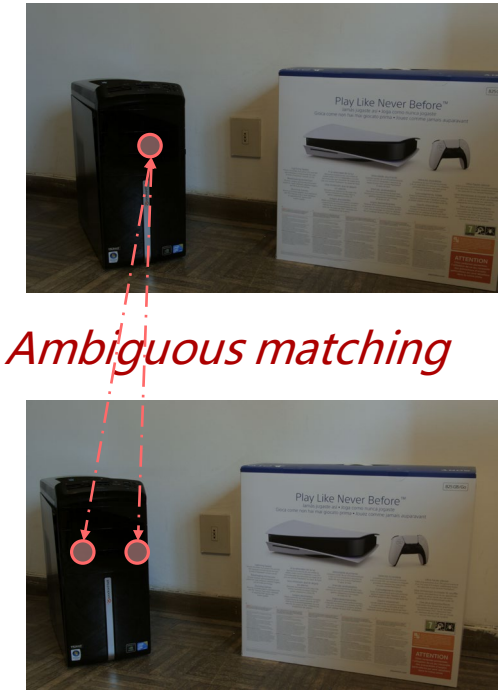


Problem

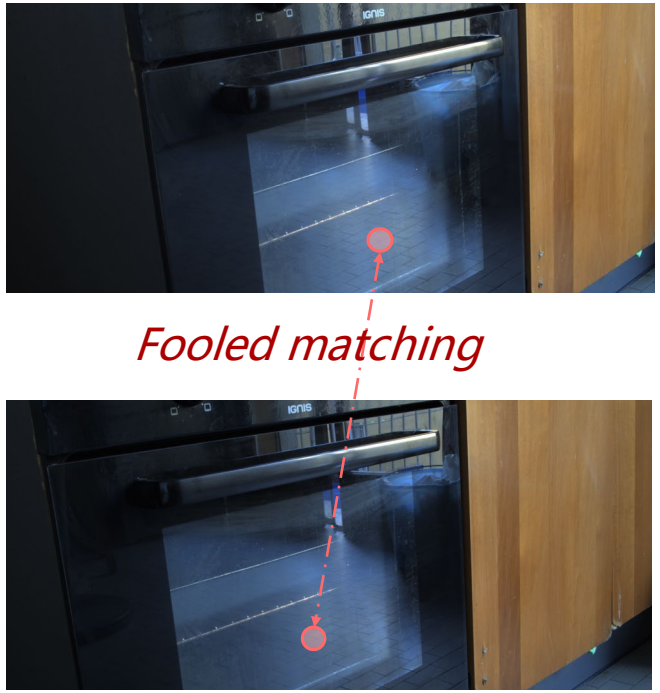
The matching formulation makes it naturally hard for the stereo matching to handle ill-posed regions like occlusions and non-Lambertian surfaces.



Occlusion



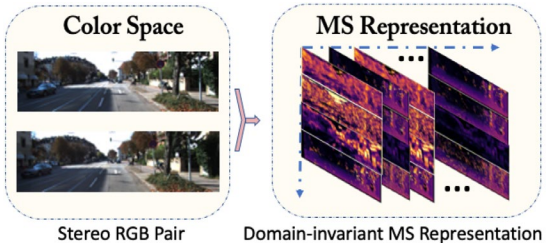
Texture-less region



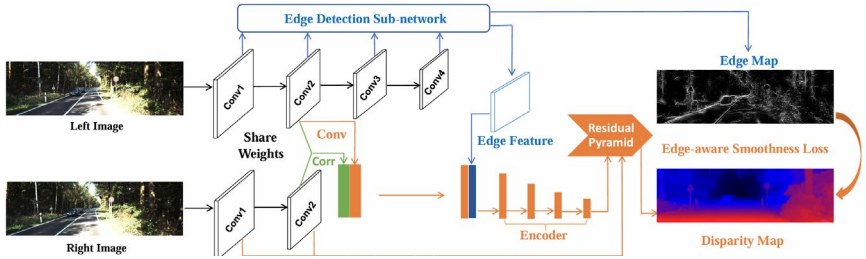
Glass

Motivation

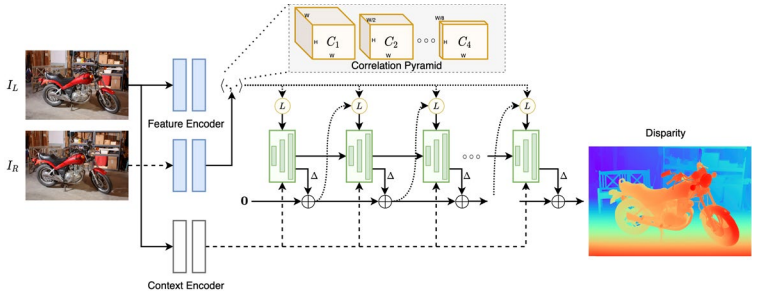
Biased monocular priors from limited data



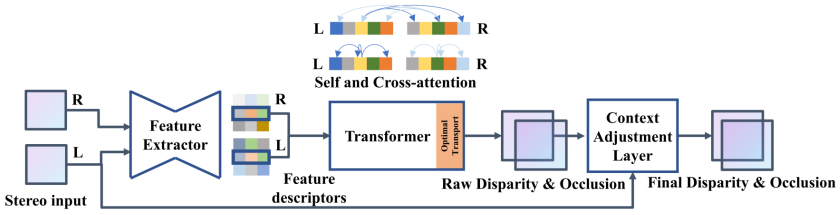
Domain-invariant Features [1]



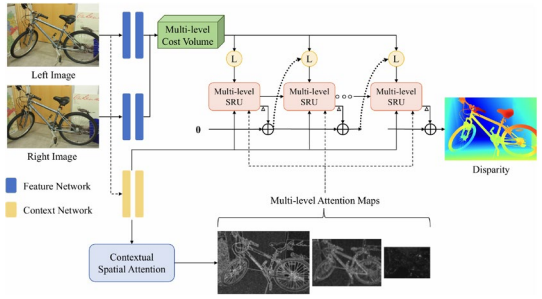
Edge Features [2]



Context Features [3]



Receptive Field [4]

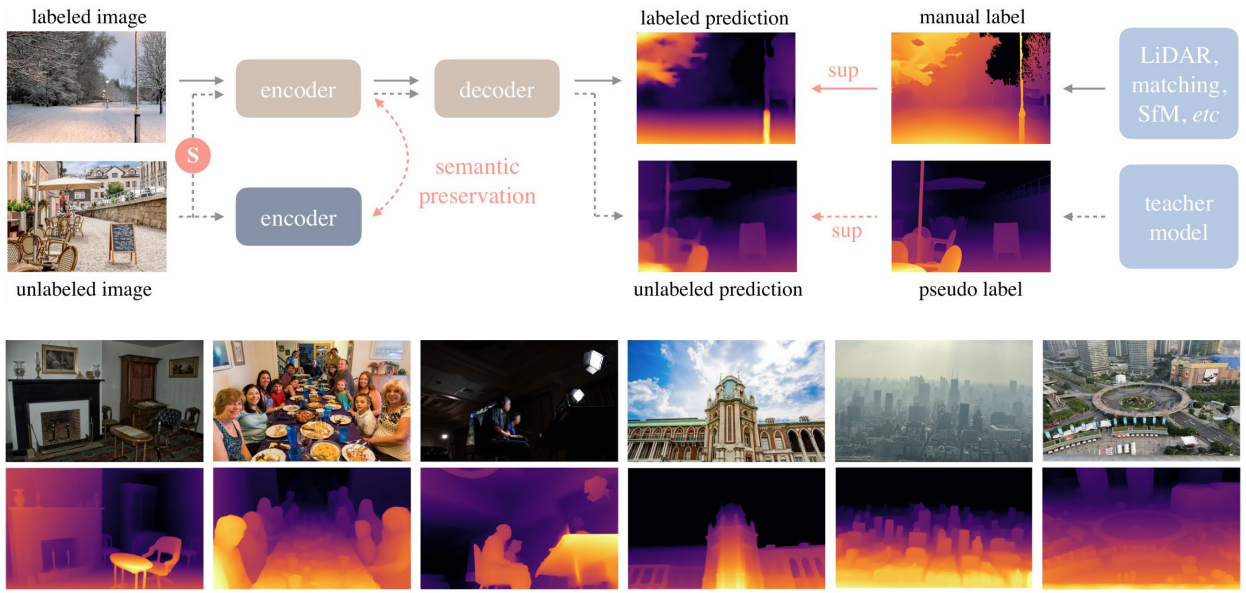


Frequency Information [5]

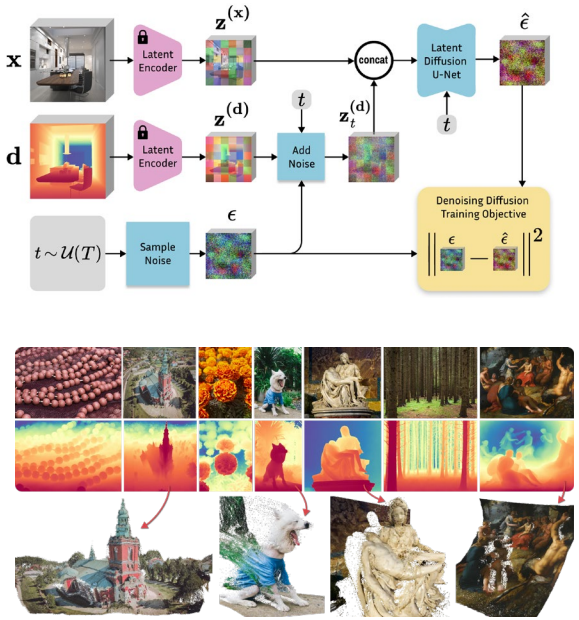
[1] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization[C]. 3DV, 2020.
[2] Song X, Zhao X, Fang L, et al. Edgestereo: An effective multi-task learning network for stereo matching and edge detection[J]. International Journal of Computer Vision (IJCV), 2020.
[3] Lipson L, Teed Z, Deng J. Raft-stereo: Multilevel recurrent field transforms for stereo matching[C]. 3DV, 2021.
[4] 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[C]. ICCV, 2021.
[5] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching[C]. CVPR, 2024.

Motivation

Unbiased monocular priors from large foundation model



DepthAnything [1]

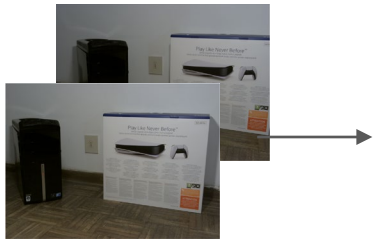


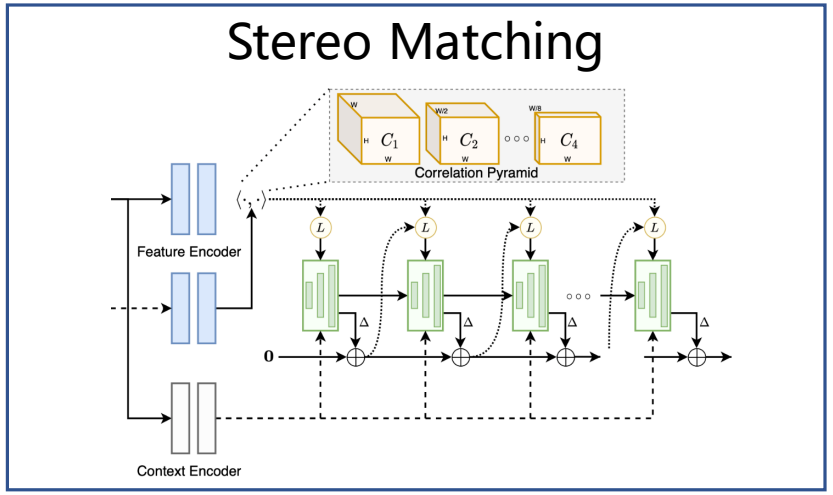
Marigold [2]

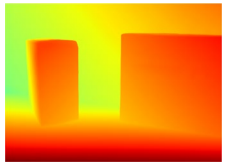
[1] Yang L, Kang B, Huang Z, et al. Depth anything: Unleashing the power of large-scale unlabeled data[C]. CVPR, 2024.
[2] Ke B, Obukhov A, Huang S, et al. Repurposing diffusion-based image generators for monocular depth estimation[C]. CVPR, 2024.

Motivation

Fusing unbiased monocular priors from large foundation model for generalized stereo matching



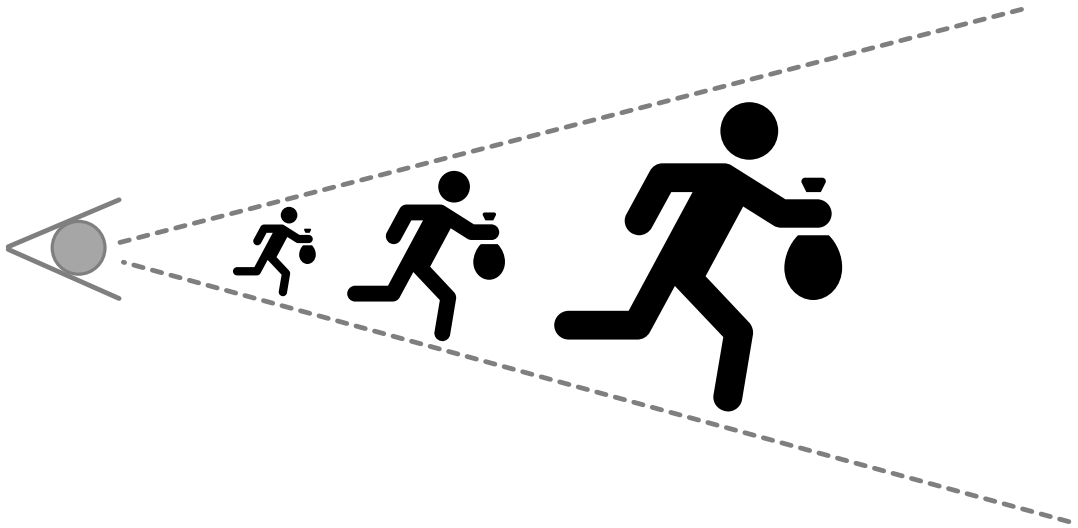




Challenges

The fusion is nontrivial:

- The gap between the affine-invariant relative depth from monocular depth and absolute depth from disparity;

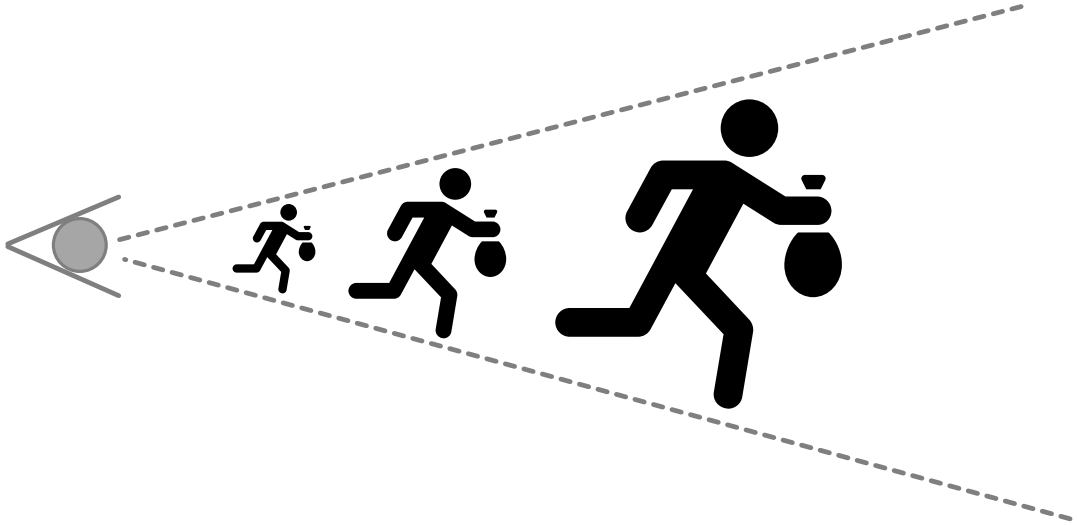


Scale Ambiguity in Monocular Depth

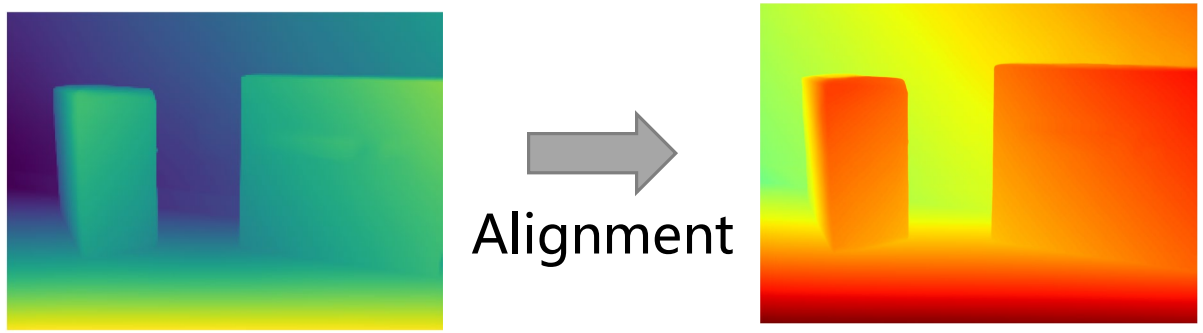
Challenges

The fusion is nontrivial:

- The gap between the affine-invariant relative depth from monocular depth and absolute depth from disparity;



Scale Ambiguity in Monocular Depth



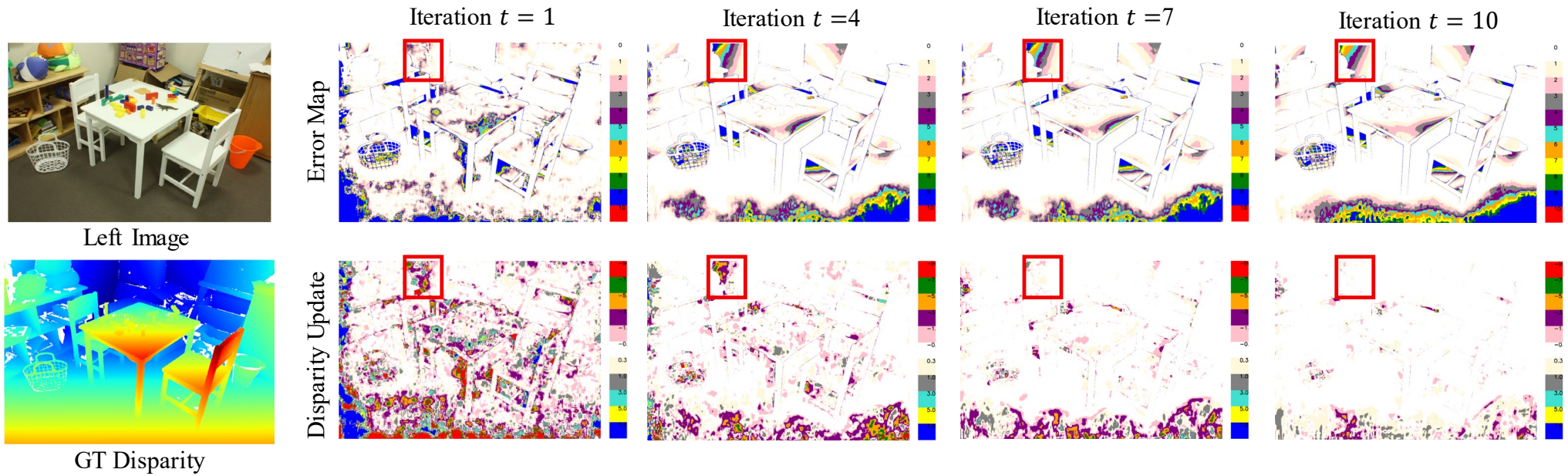
Relative Depth

Absolute Depth

Challenges

The alignment is nontrivial:

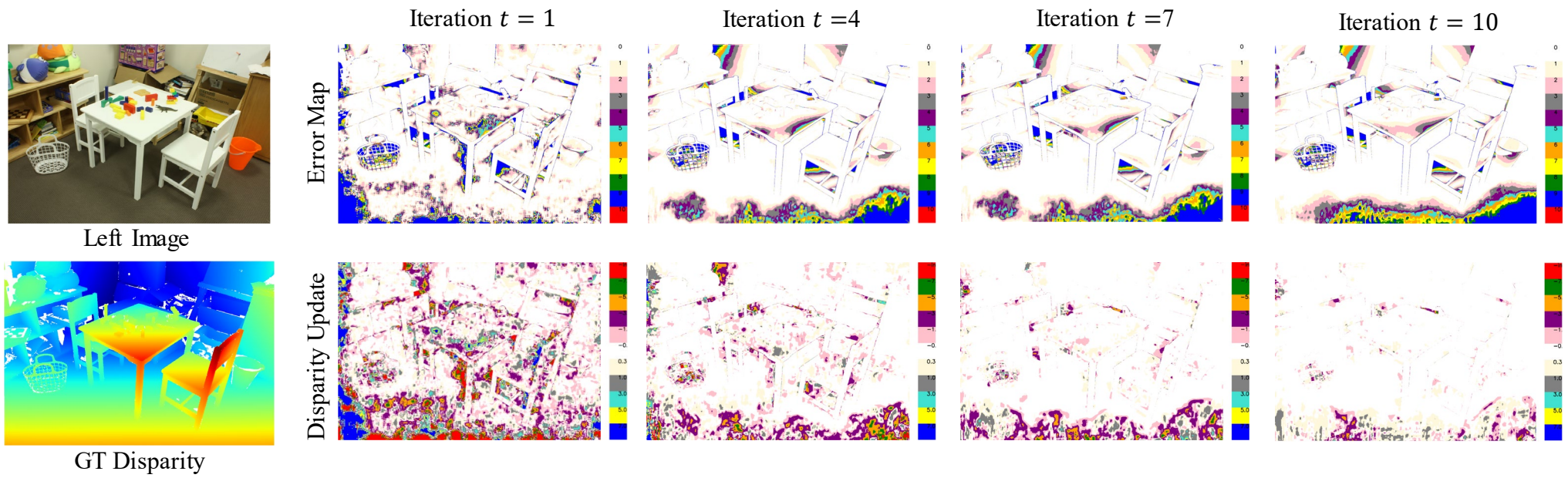
- Over-confidence in the disparity update will lead to local optima;



Challenges

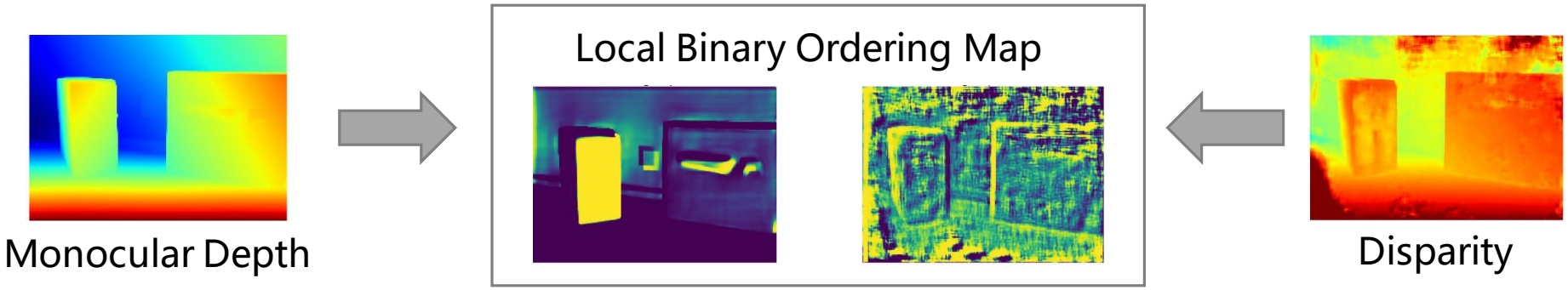
The alignment is nontrivial:

- Over-confidence in the disparity update will lead to local optima;
- Noisy disparity results computed at the first several iterations will misguide the fusion and slow down good fusion;



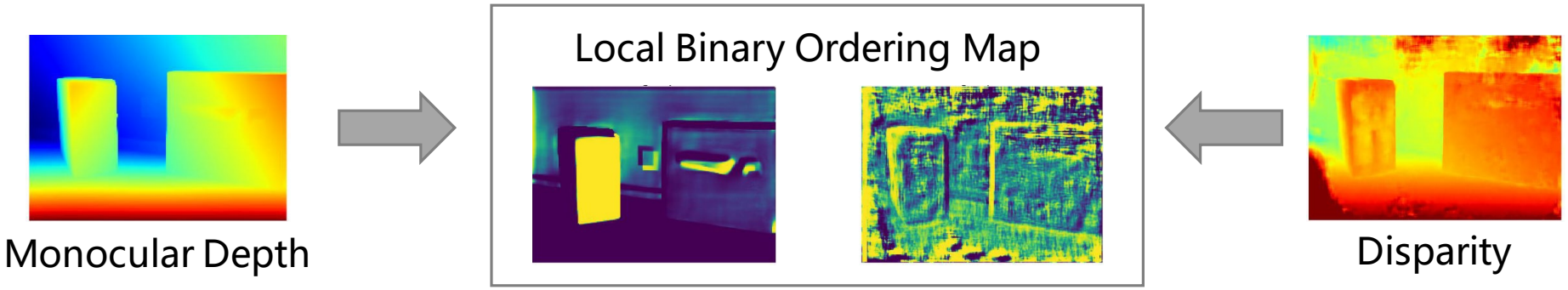
Contributions

- Local Binary Ordering Map: indicating whether two pixels are farther or closer;

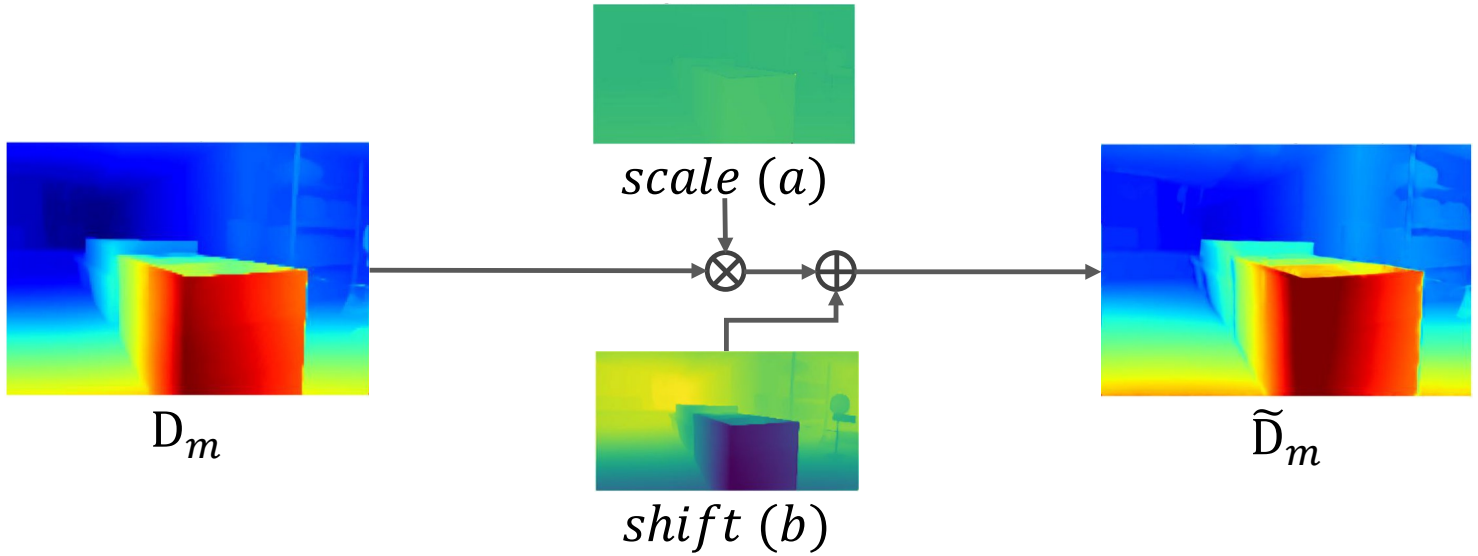


Contributions

- Local Binary Ordering Map: indicating whether two pixels are farther or closer;

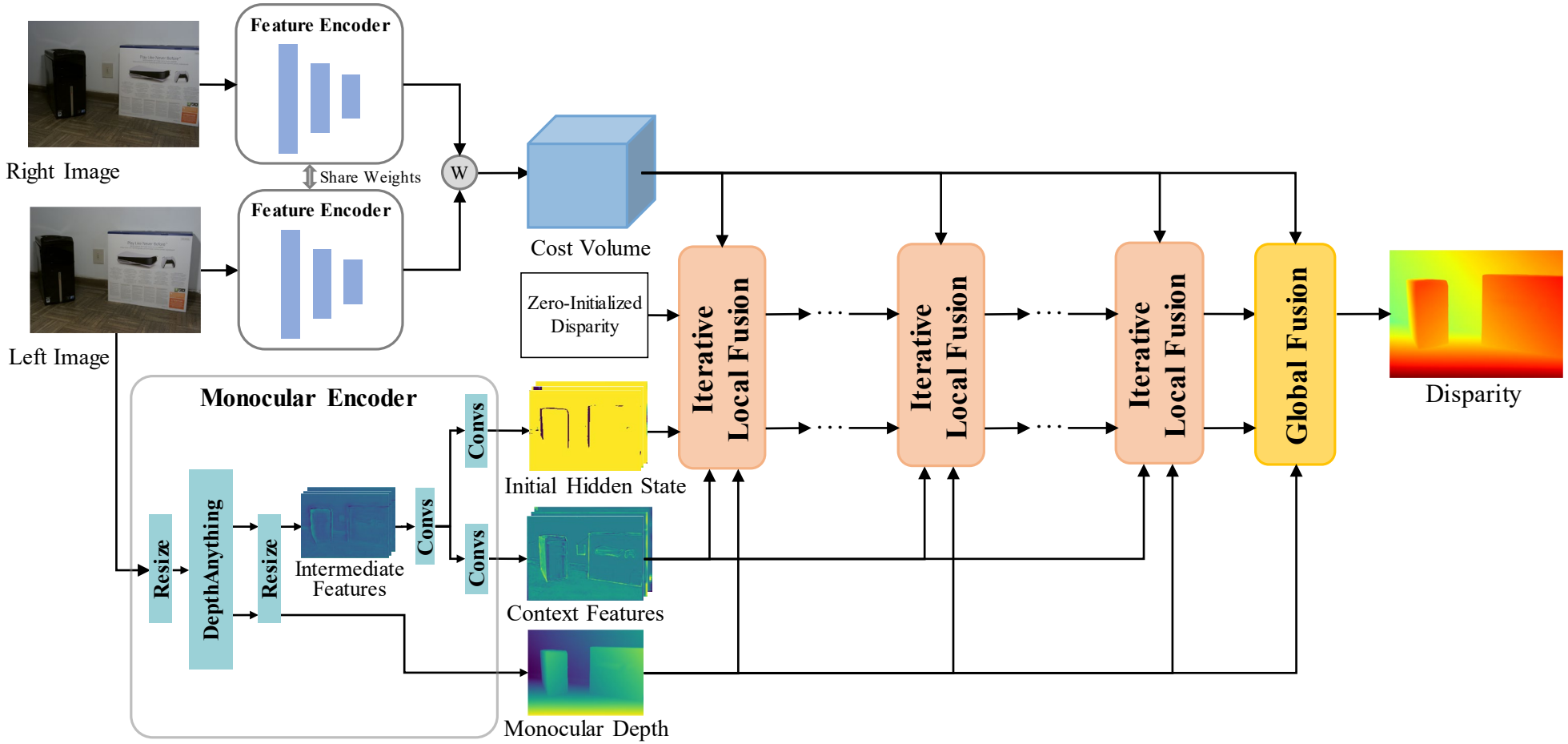


- A noisy linear regression problem about the registration parameters;



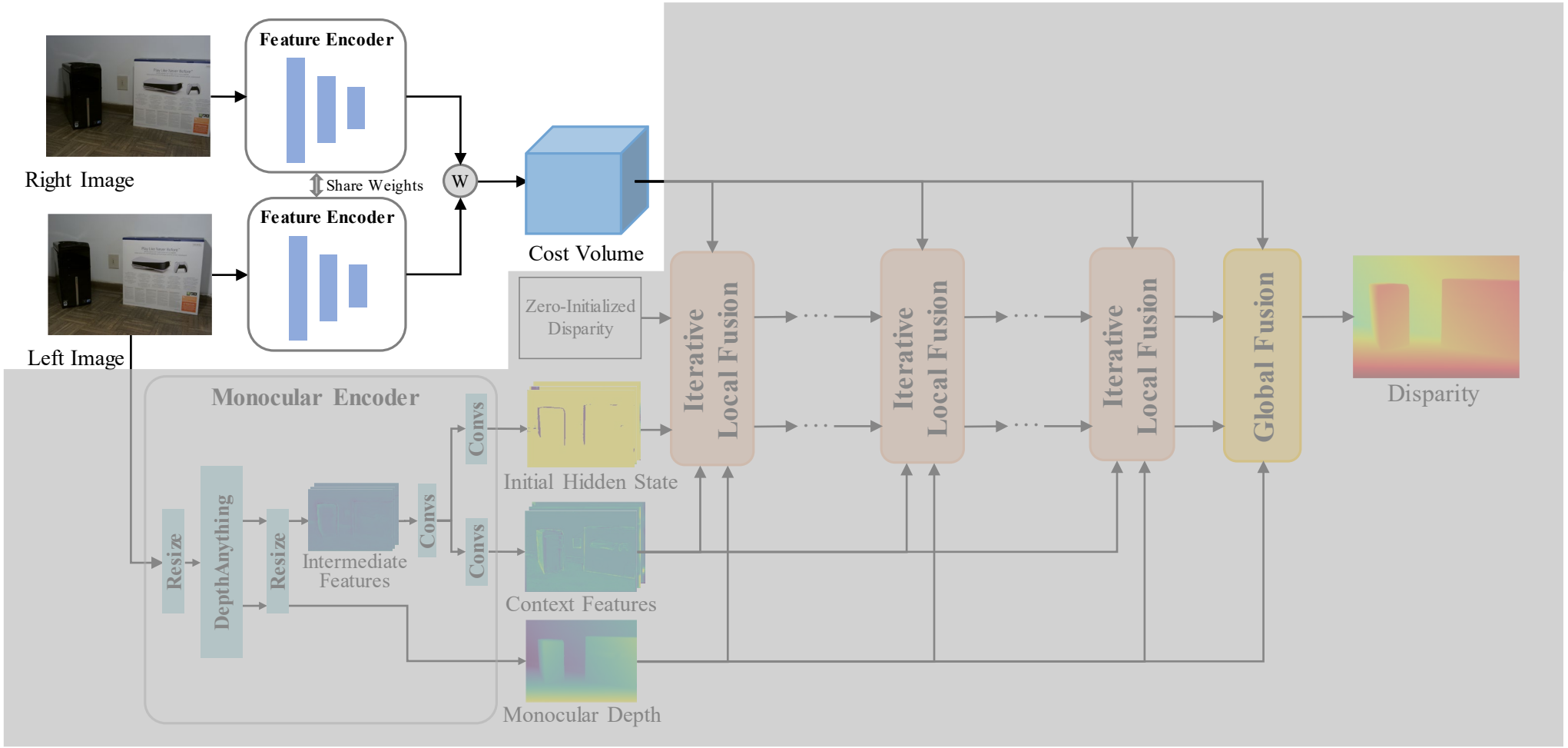
Model

Model Pipeline



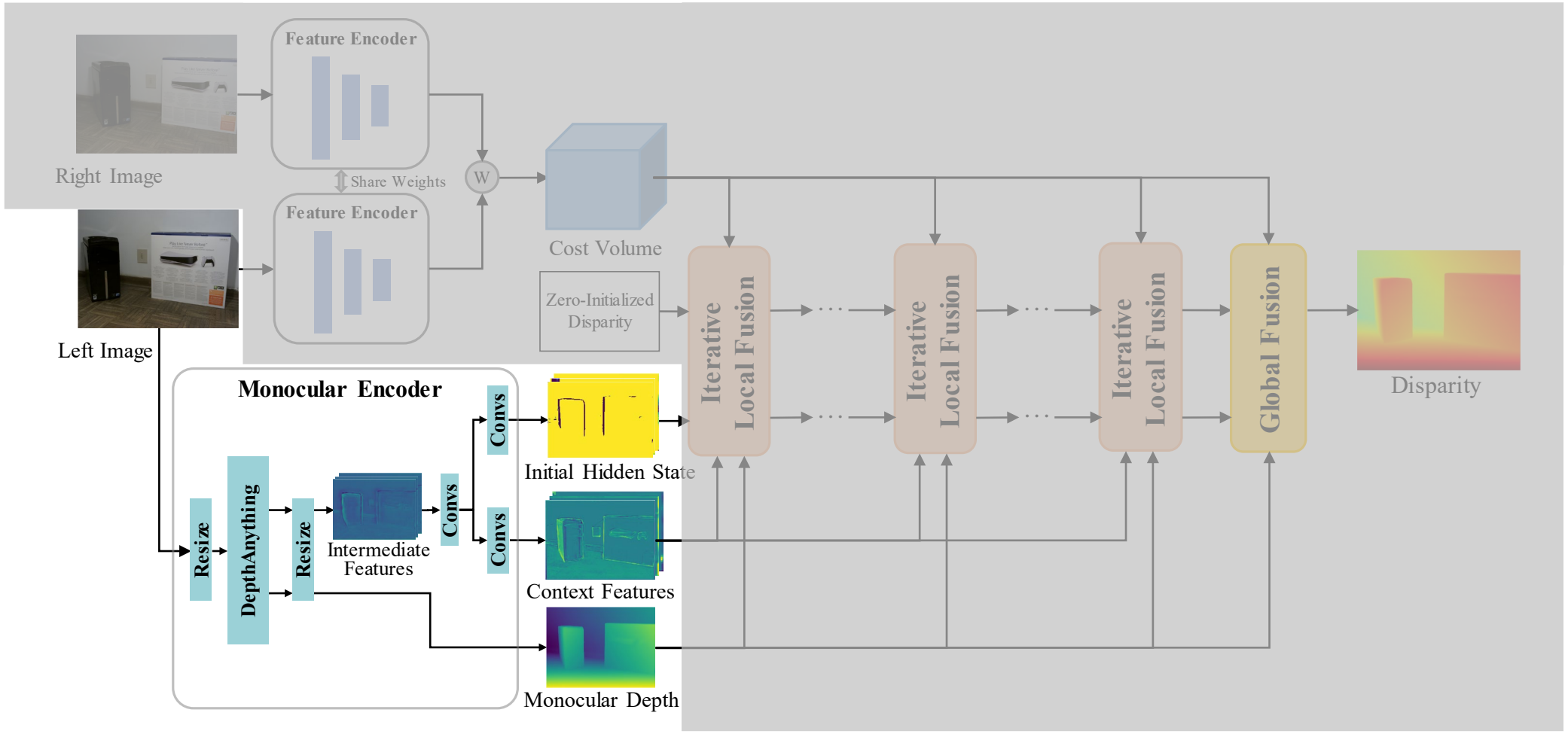
Model

➢ Feature Encoder



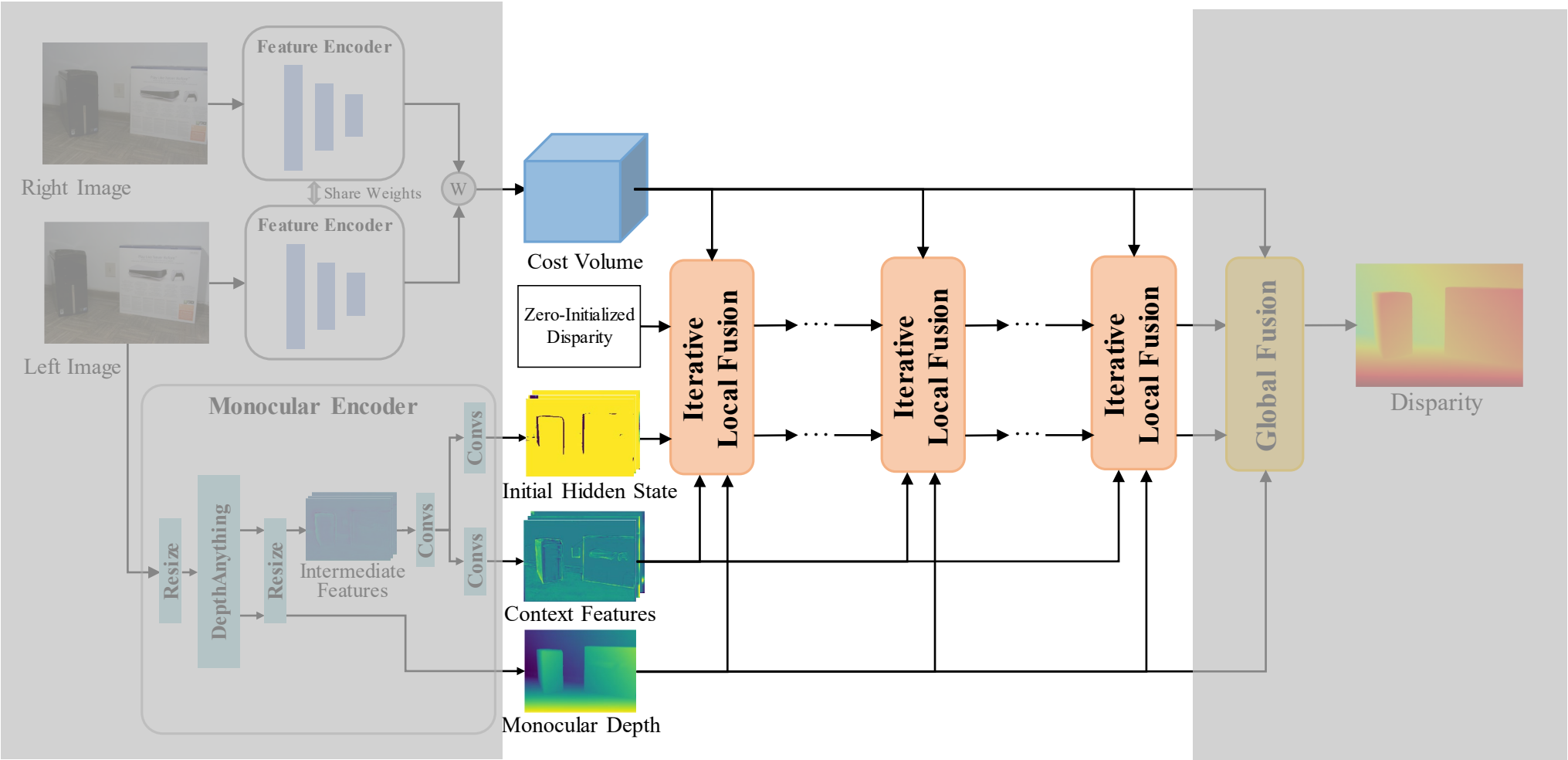
Model

- Monocular Encoder



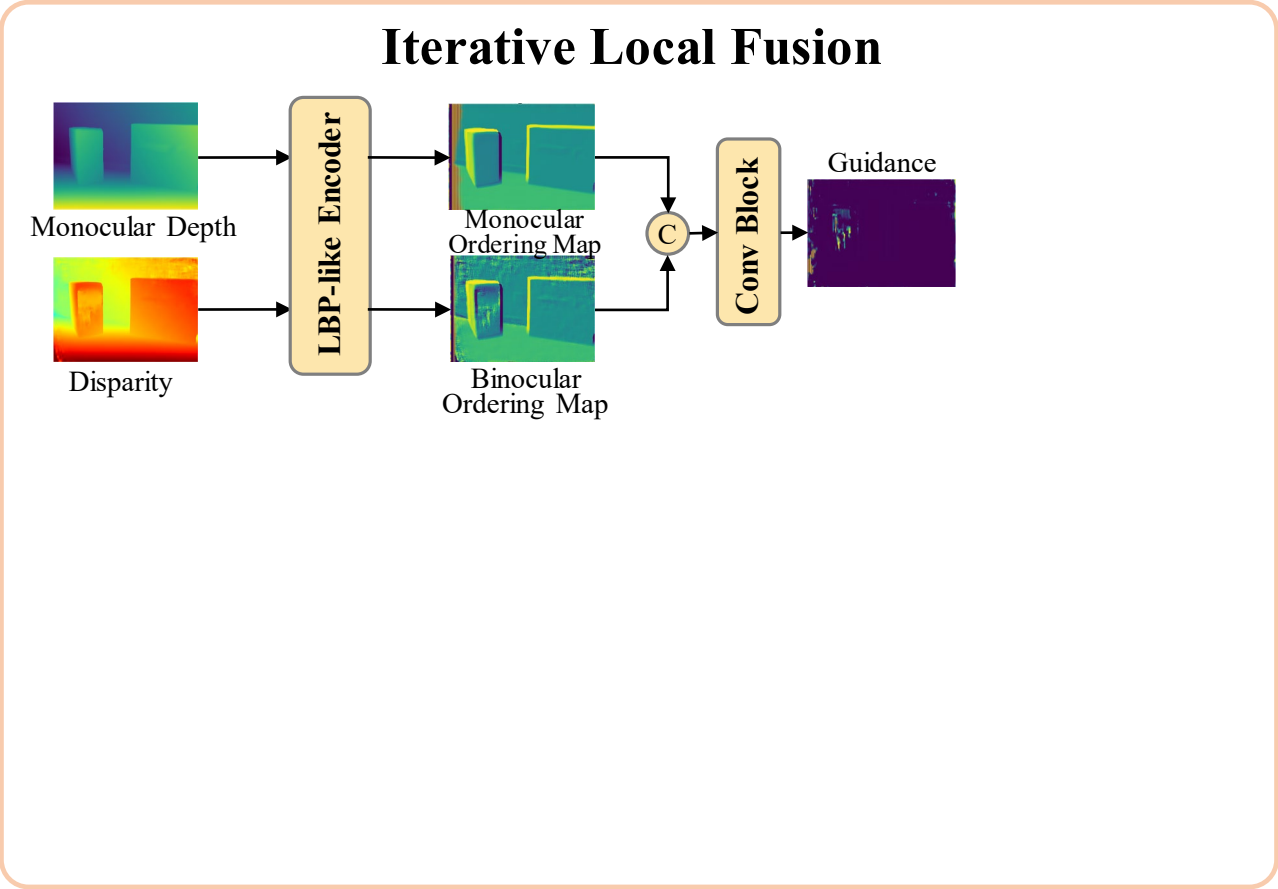
Model

➢ Iterative Local Fusion



Model

➢ Iterative Local Fusion

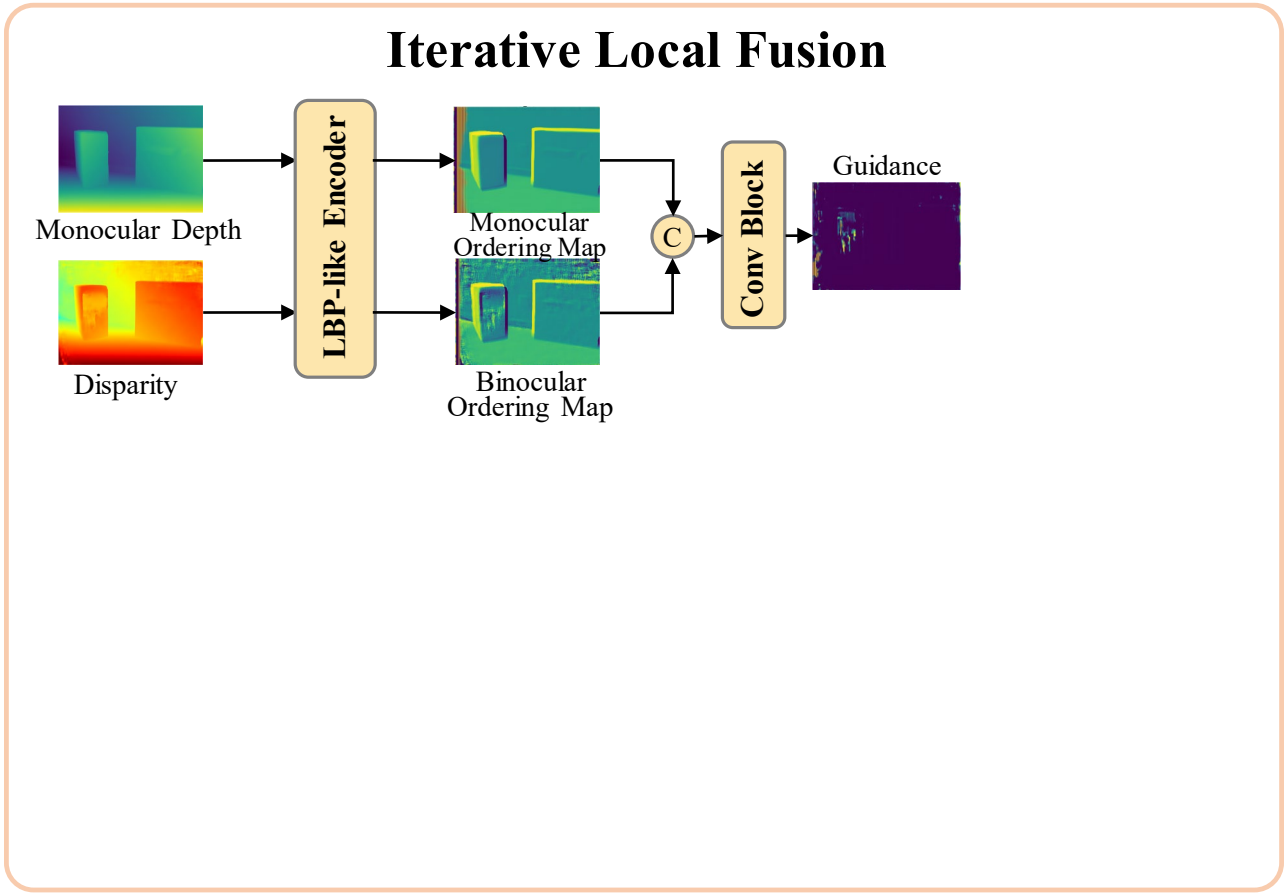


$$M(u,v) = \{\sigma(D(u',v') - D(u,v))\}$$

$$(u',v') \in \mathcal{N}_{(u,v)}$$

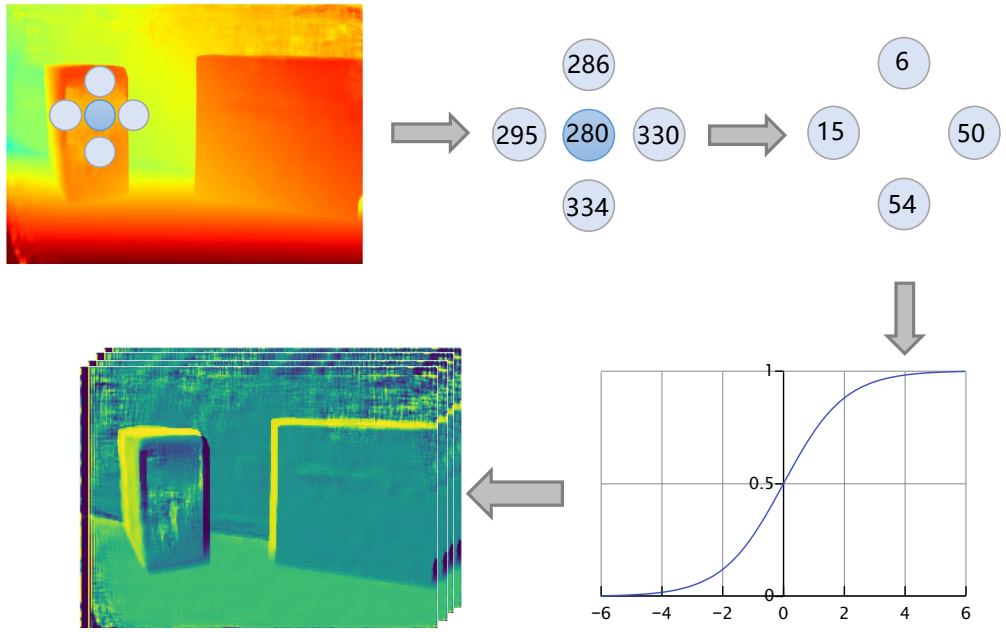
Model

➢ Iterative Local Fusion



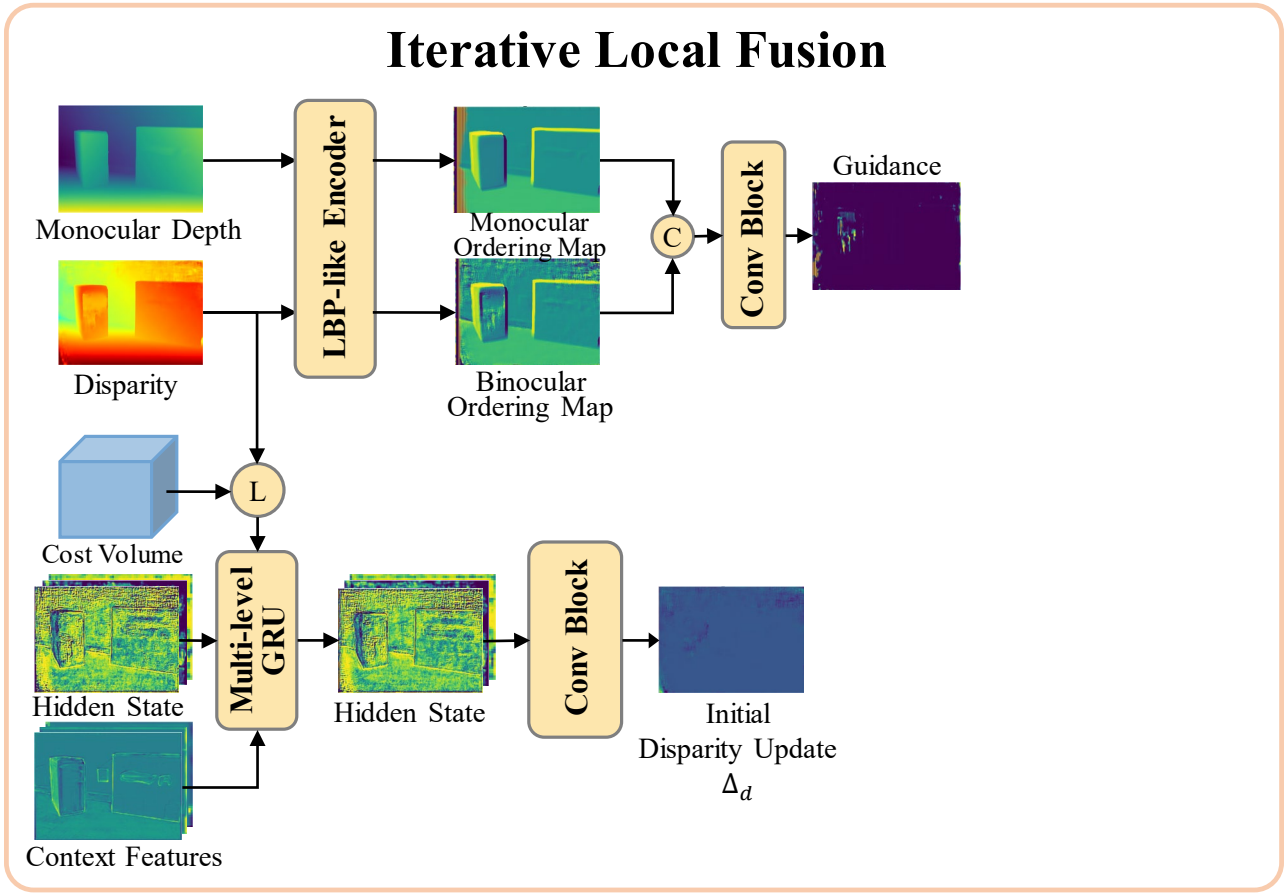
$$M(u,v) = \{\sigma(D(u',v') - D(u,v))\}$$

$$(u',v') \in \mathcal{N}_{(u,v)}$$



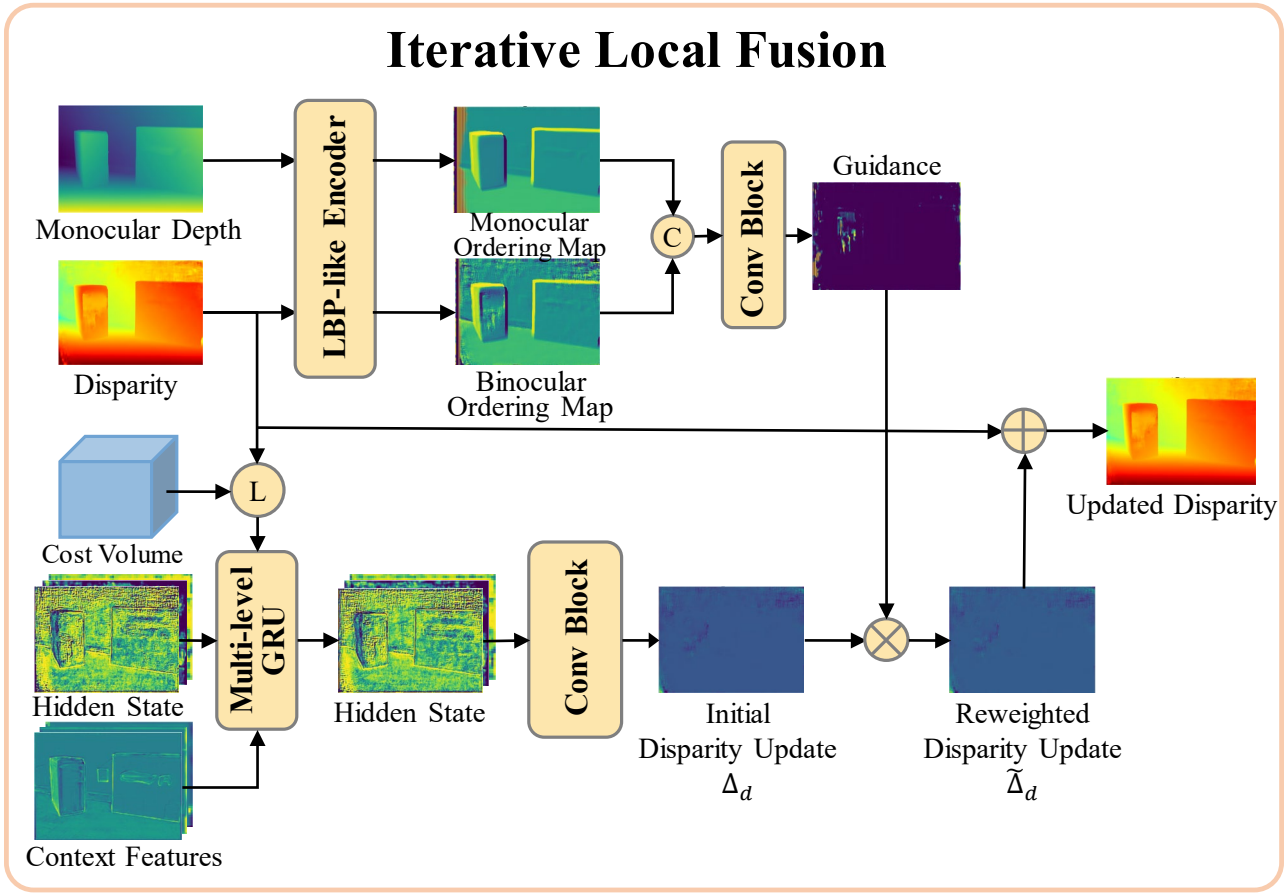
Model

➢ Iterative Local Fusion



Model

➢ Iterative Local Fusion

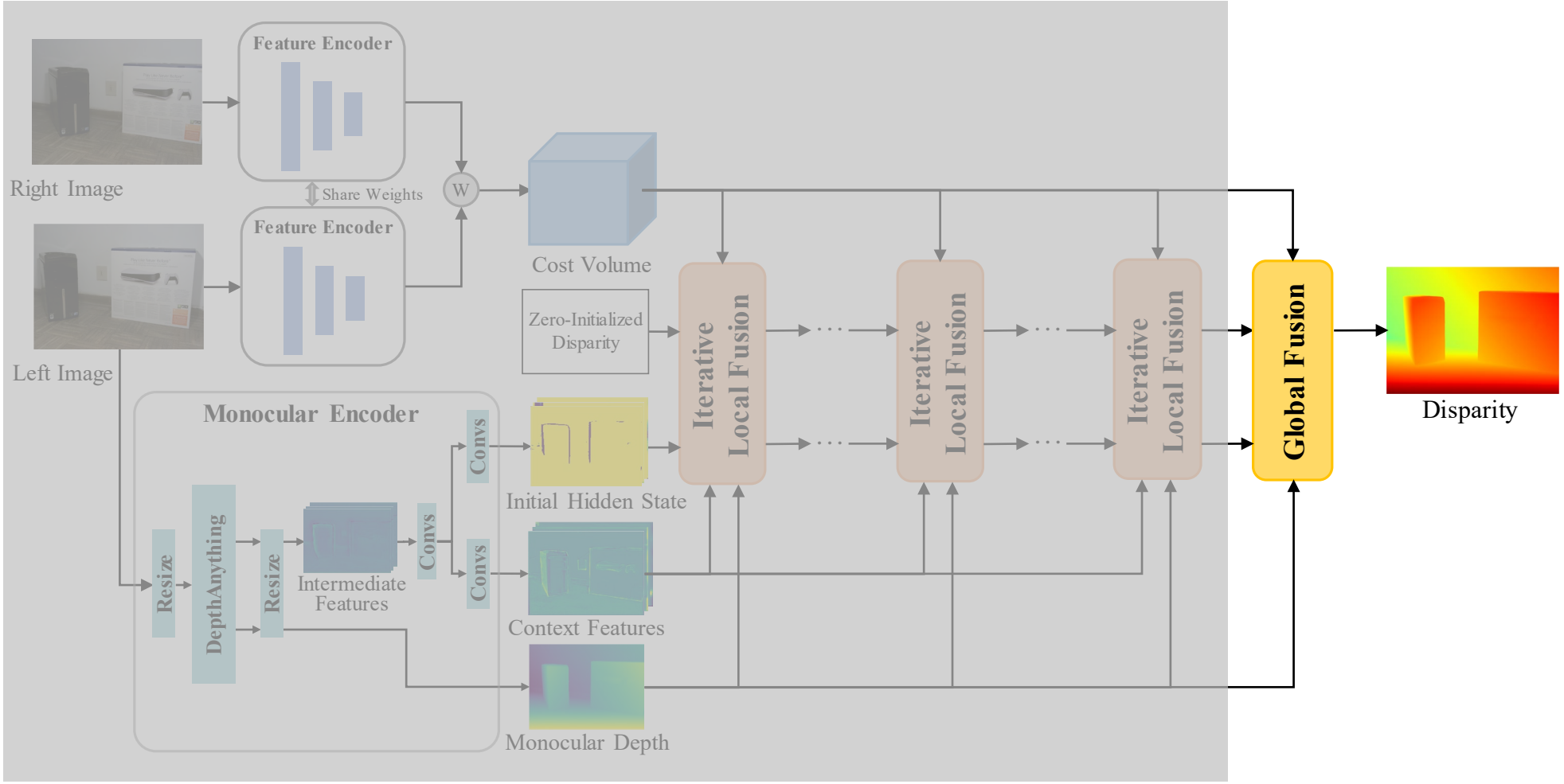


$$\tilde{\Delta}_d = \Delta_d(1 + G \cdot r \cdot t/T)$$

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

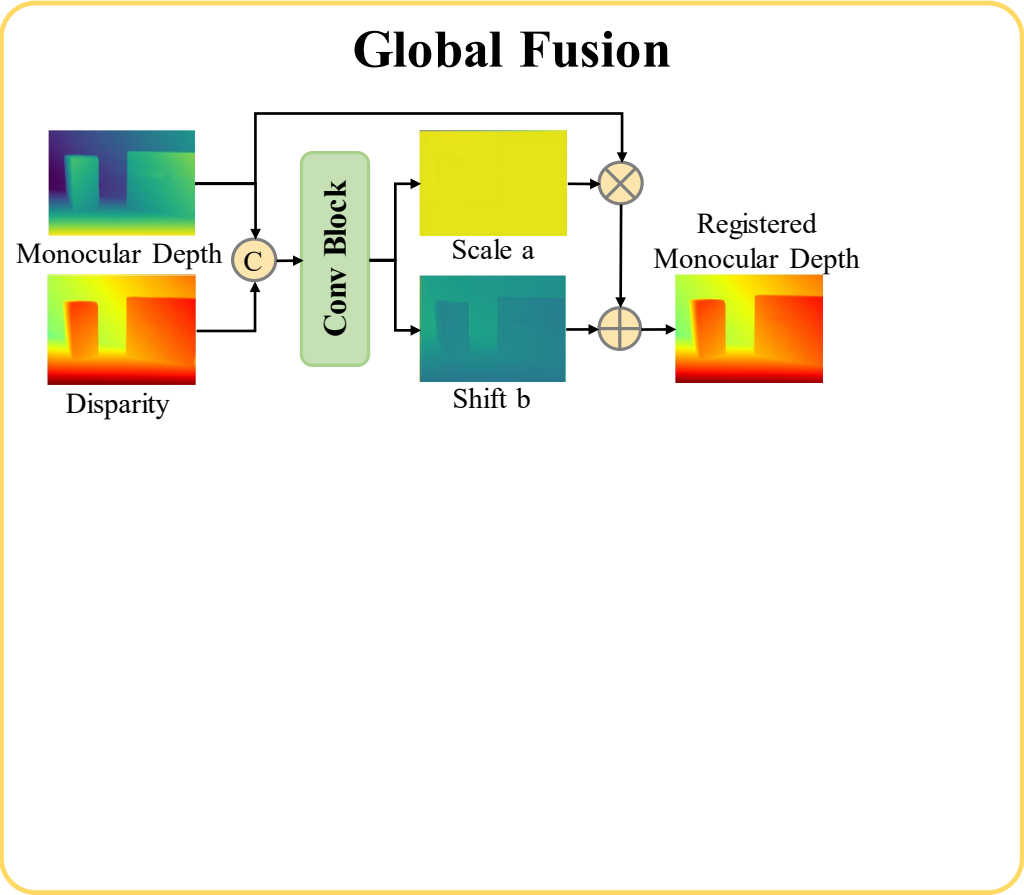
Model

- Global Fusion



Model

- Global Fusion

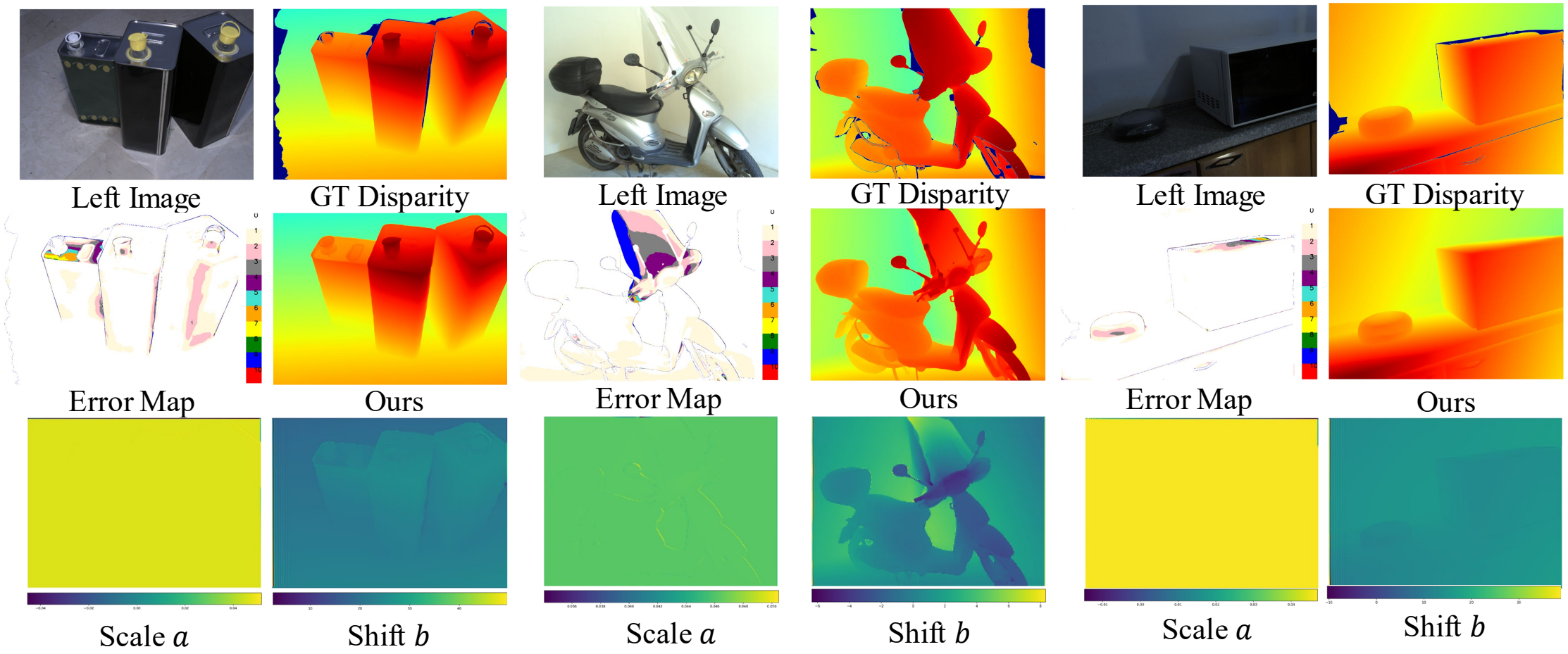


$$\widetilde{D}_m = a \cdot D_m + b$$

$$a, b = \mathcal{F}(D_m, D_d^T)$$

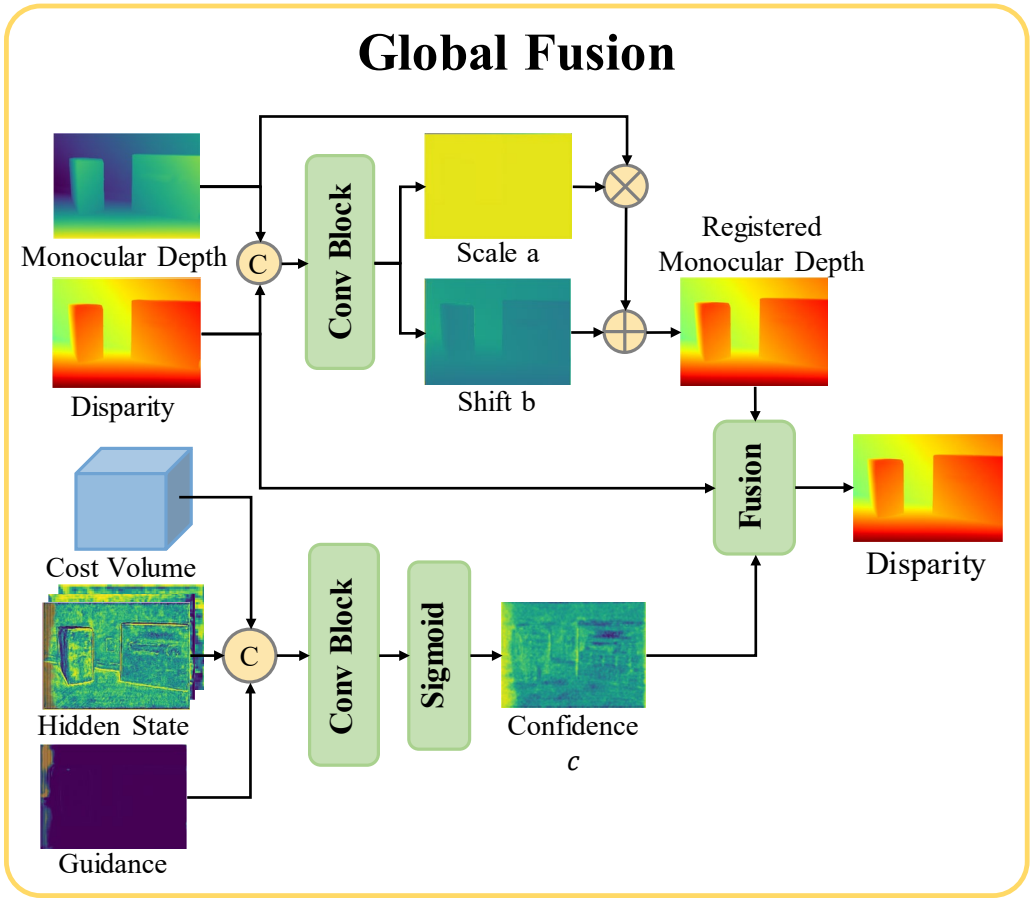
Model

- Global Fusion



Model

- Global Fusion



$$\widetilde{D}_m = a \cdot D_m + b$$

$$a, b = \mathcal{F}(D_m, D_d^T)$$

$$D_f = c \cdot D_d^T + (1 - c) \cdot \widetilde{D}_m$$

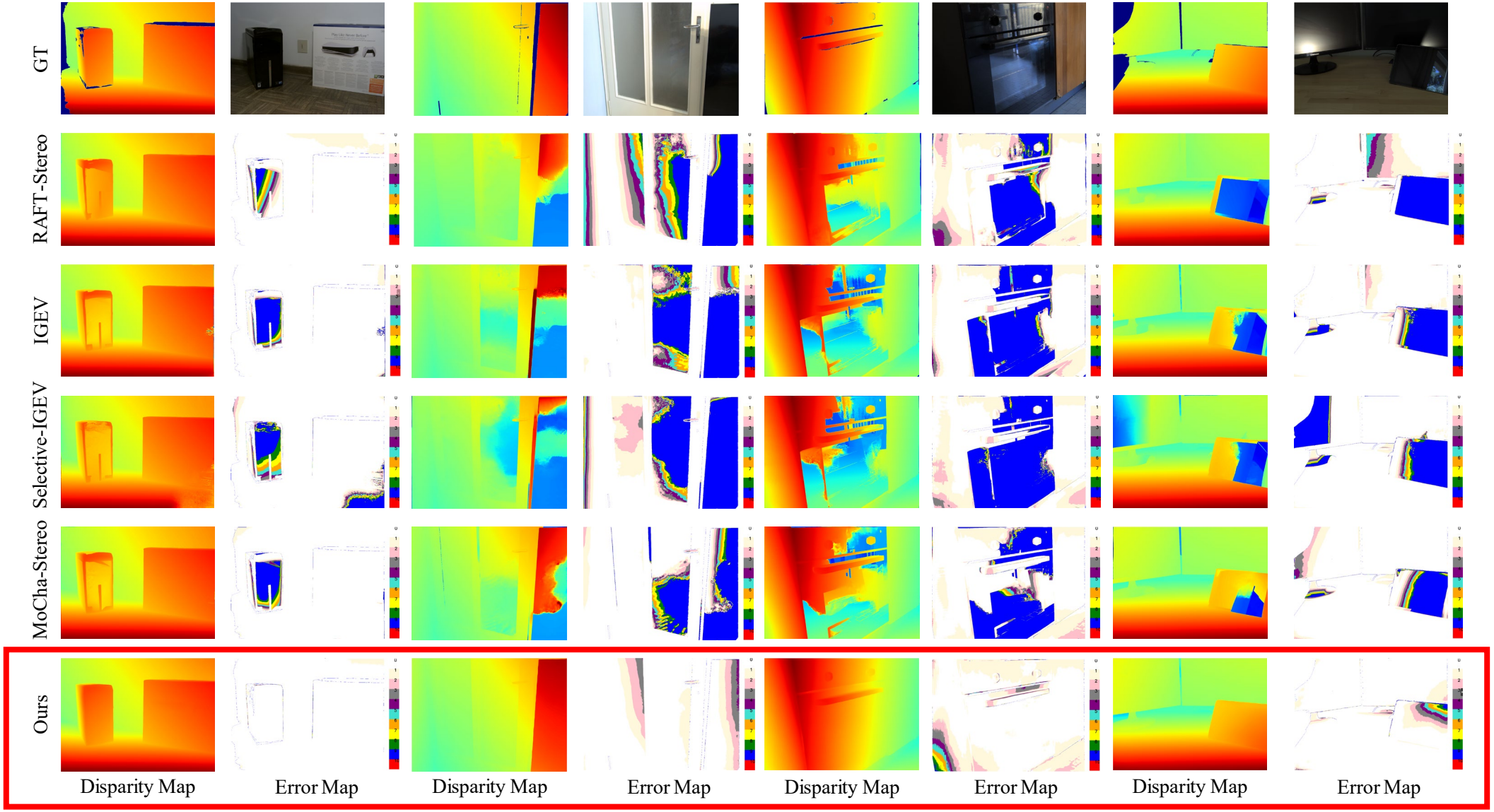
Experiments

- Zero-Shot Generalization on Non-Lambertian Surfaces

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 [25]		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 [46]		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 [46]		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 [51]		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 [43]	✓	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 [22]		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**

Experiments

- Zero-Shot Generalization on Non-Lambertian Surfaces



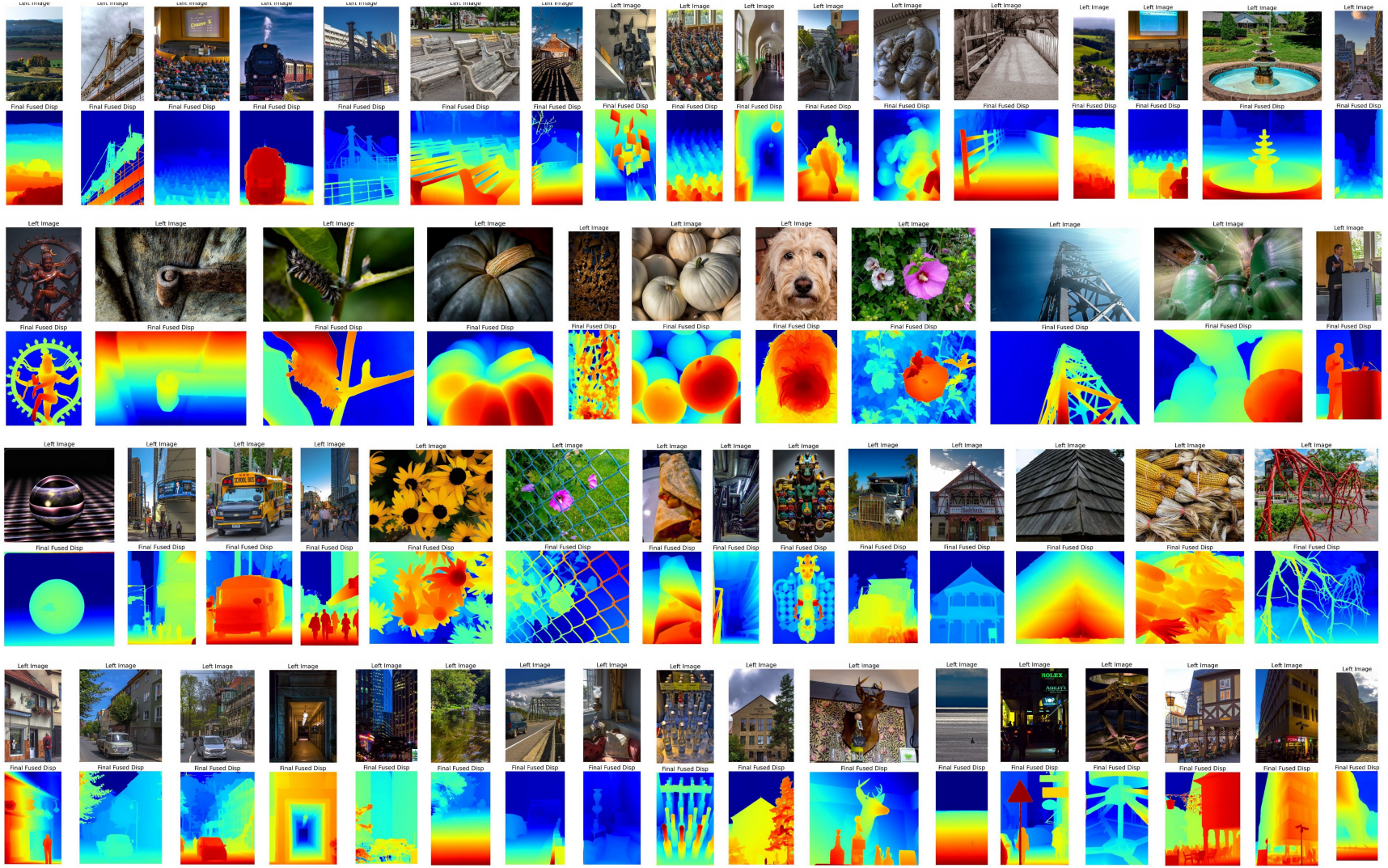
Experiments

➢ Zero-Shot Generalization in General Scenes

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 [58]	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 [23]	2022		1.32	5.30	1.09	5	2.34	10.9	-	-	-	-	1.16	10.7
Mask-CFNet [35]	2023		-	5.80	-	4.8	-	13.7	-	-	-	-	-	5.7
STTR* [21]	2021		2.14	9.5	2.51	9.62	9.13	21.76	5.03	13.49	35.98	78.84	-	-
PCWNet [40]	2022		-	5.60	-	4.2	-	15.8	-	15.8	-	-	3.8	14.4
RAFTStereo* [22]	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* [51]	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* [25]	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* [46]	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* [46]	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* [43]	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**

Experiments

- Zero-Shot Generalization in the Wild



Experiments

- Effectiveness and Robustness of Modules under Zero-Shot Testing

Exp	Middlebury (H)	
	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

- Comparison of Monocular Depth Estimation Methods

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

Experiments

- Model Scalability with Increasing Data from the TranScene Dataset

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**

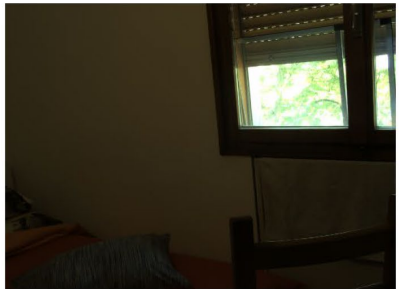
Discussion

- Transparent scenes -> both the transparent surfaces and the behind scene are significant

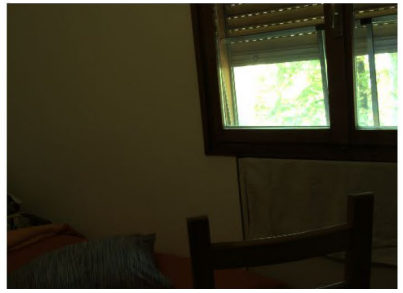
Left Image

Right Image

Left Image

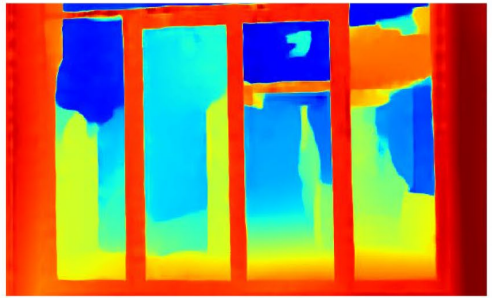


Right Image

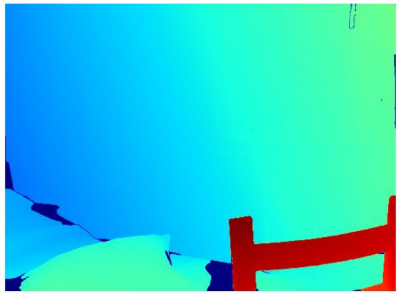


GT

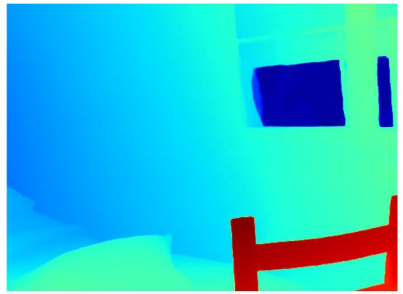
Prediction



GT



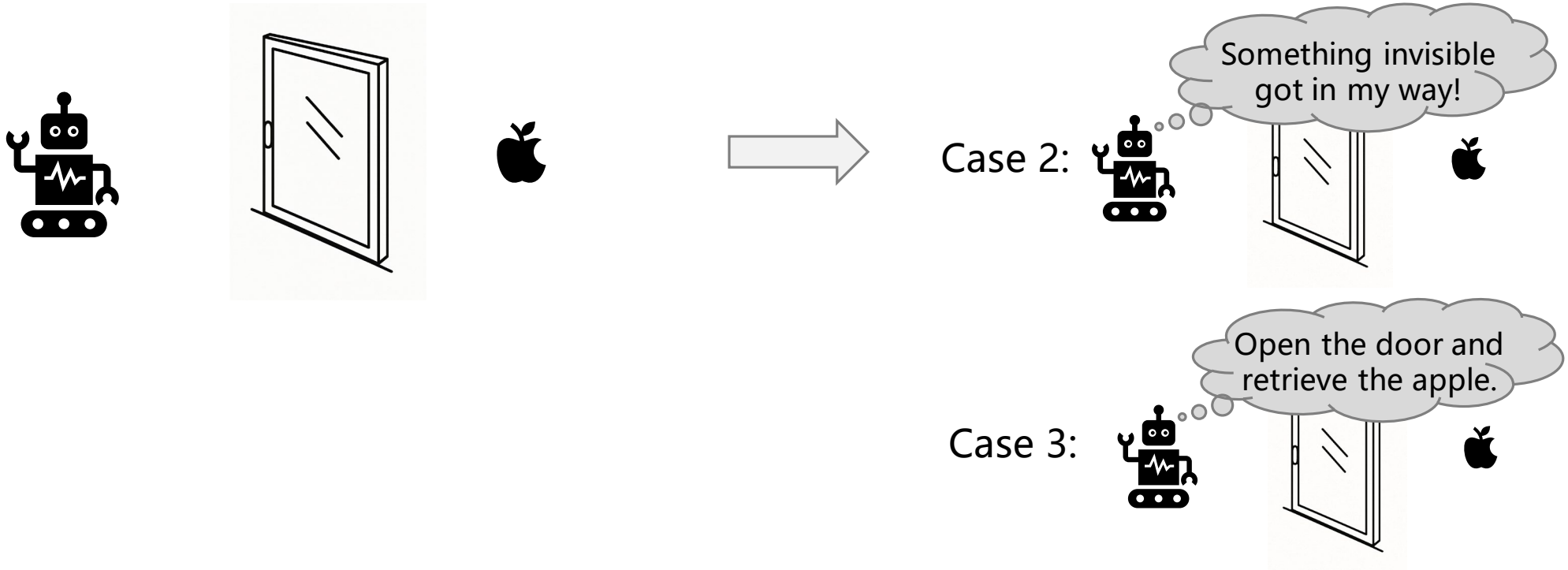
Prediction



Discussion

- Transparent scenes -> both the transparent surfaces and the behind scene are significant

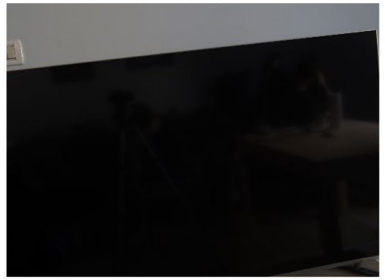
Task: Retrieve the apple behind the glass door.



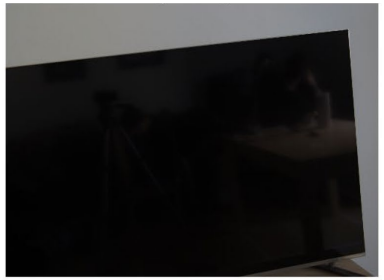
Discussion

- Close black screen and dark tunnel -> information from video streams and segmentation

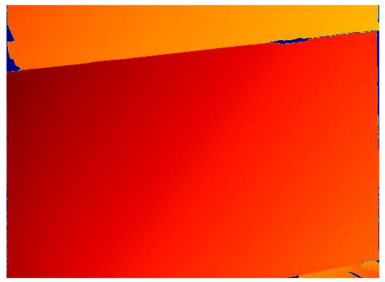
Left Image



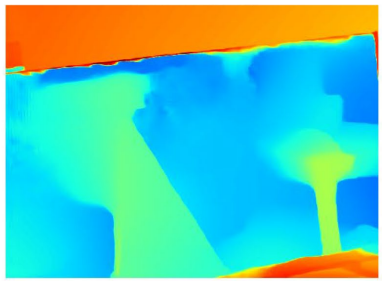
Right Image



GT



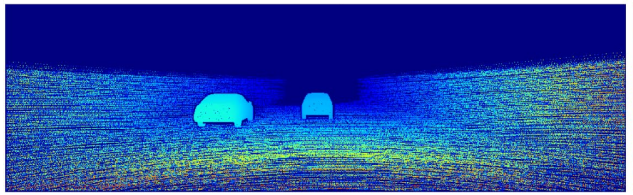
Prediction



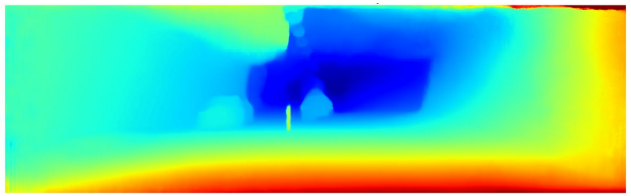
Left Image

Right Image

GT



Prediction



Thanks for Your Attention

Paper

Project

Demo Live