DEFOM-Stereo: Depth Foundation Model Based Stereo Matching

Supplementary Material

A. Evaluation of Depth Anything V2

In this section, we present the evaluation of Depth Anything V2 [15] on some common stereo datasets. To achieve cross-scene robustness, the relative depth estimation models are usually trained with a scale and shift (affine) invariant loss [9] on the inverse depth space, thus predicting an affine disparity with unknown scale and shift. Given an inverse depth map predicted by Depth Anything V2 is $\mathbf{z}$ and its corresponding ground truth disparity map $\mathbf{d}_{gt}$, they must satisfy the following affine transformation,

$$\mathbf{d}_{gt} = s\mathbf{z} + t, \tag{1}$$

where s is the scale and t is the shift. For each image example, we can use the least square to find the solution for the scale and shift, $\hat{s}$ and $\hat{t}$. The aligned disparity map from the depth estimate is computed as,

$$\hat{\mathbf{d}} = \hat{s}\mathbf{z} + \hat{t}. \tag{2}$$

The quality of the depth estimate of Depth Anything V2 can be accessed with the end-point error (EPE) between $\mathbf{d}_{gt}$ and $\hat{\mathbf{d}}$. To evaluate the scale consistency, we further compute a ratio map between the ground truth disparity $\mathbf{d}_{gt}$ and the aligned disparity map $\hat{\mathbf{d}}$,

$$\mathbf{r} = \frac{\mathbf{d}_{gt}}{\hat{\mathbf{d}}}, \tag{3}$$

where $\hat{\mathbf{d}}$ is clamped with a minimum (set to 0.05) in advance to avoid meaningless division.

The distribution of the values in $\mathbf{r}$ reveals the scale consistency. If the depth estimate were scale-consistent, most values in $\mathbf{r}$ approximate to 1, otherwise there must be many ratios that deviate to 1. Therefore, we compute the standard deviation of $\mathbf{r}$ to assess the scale consistency.

We perform evaluation on three realistic datasets, KITTI 2015 [8], Middlebury (half resolution) [10] and ETH3D [11], and two synthetic datasets, Scene Flow [7] and CREStereo [4]. For the realistic datasets, we evaluate on their entire trainsets, *i.e.*, 200 examples for KITTI 2015, 15 samples for Middlebury, and 27 samples for ETH3D. For Scene Flow and CREStereo, we evaluate on 200 random samples from their trainsets.

Tab. 1 presents the quantitative results. Even given the unknown scale and shift, disparity errors are large, especially for the synthetic Scene Flow [7] and CREStereo [4] datasets, whose STD is very high too, indicating the scale inconsistency within the image is serious. This is because the synthetic stereo dataset is about unnatural scenes. As the stereo model is usually pre-trained on the synthetic stereo dataset, the scale inconsistency of DEFOM poses challenges for recovering disparity from its depth estimate. In contrast, Depth Anything V2 presents better results on realistic datasets. Although the EPE on Middlebury-half is up to 5, it is mainly due to its high resolution. In contrast, for two synthetic datasets, both the EPE and STD are larger, indicating Depth Anything V2 does not predict depth maps with good scale consistency.

Sceneflow		CREStereo		KITTI-2015		Middlebury-half		ETH3D	
EPE	STD	EPE	STD	EPE	STD	EPE	STD	EPE	STD
8.04	3.31	5.10	3.30	2.08	0.74	5.00	0.11	0.65	0.40

Table 1. Examination of Depth Anything V2 on typical stereo datasets via least-square affine alignment.

Fig. 1 visualizes examples from Scene Flow, KITTI, Middlebury, and ETH3D. The last column is the ratio map with a color bar. These visualizations further highlight the scale inconsistency issue, especially in the Scene Flow dataset, though some scale inconsistency is also evident in the real datasets, albeit to a lesser extent. Despite this, synthetic datasets, like Scene Flow, help to train the scale update module, as they pose more challenges to scale recovery.

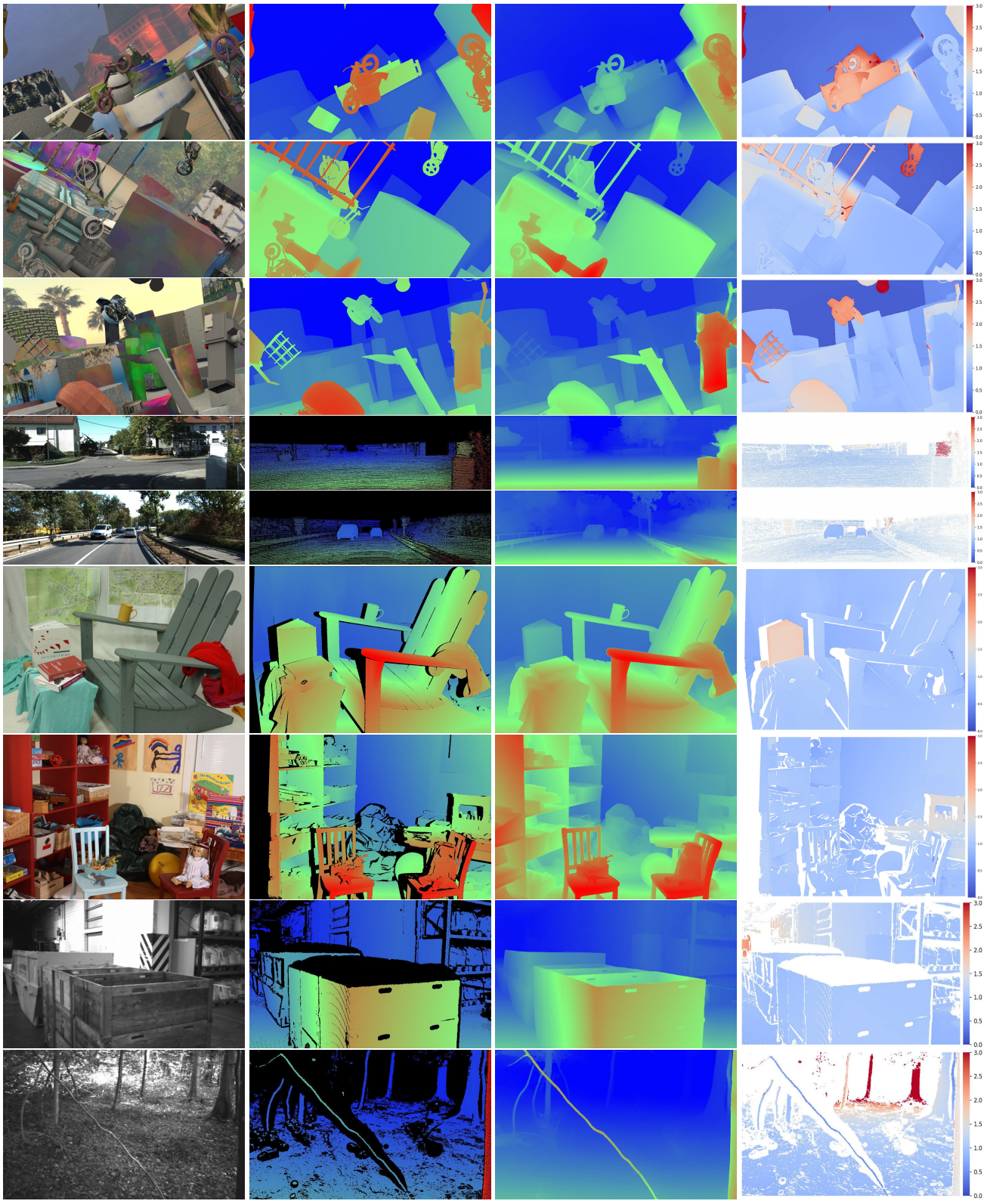


Figure 1. Visualization of the aligned depth estimate of Depth Anything V2 on some examples of the stereo datasets. **Row 1-3:** Scene Flow. **Row 4:** KITTI 2012. **Row 5:** KITTI 2015. **Row 6-7:** Middlebury. **Row 8-9:** ETH3D. Best viewed in color and by zooming in.

Models	Proposed Modules				Scene Flow		KITTI 2012	KITTI 2015	Middlebury-half	ETH3D	Params. (M)	Time (s)
	CCE	CFE	DI	SU	EPE	Bad 1.0	Bad 3.0	Bad 3.0	Bad 2.0	Bad 1.0		
Baseline					0.56	6.66	4.65	5.57	10.67	3.45	11.11	0.222†
+CCE	✓				0.49	6.08	4.40	5.84	8.42	2.82	12.10	0.242
+CFE		✓			0.50	6.17	4.13	5.53	10.45	2.83	13.89	0.243
+CCE+CFE	✓	✓			0.49	5.95	4.02	5.75	8.31	2.53	14.11	0.246
+DI			✓		0.57	6.74	4.57	5.63	12.40	2.77	11.11	0.242
+DI+SU			✓	✓	0.50	6.03	4.15	5.12	8.15	2.67	15.51	0.244
Full Model (ViT-S)	✓	✓	✓	✓	0.46	5.57	4.29	5.29	6.76	2.61	18.51	0.255
Full Model (ViT-L)	✓	✓	✓	✓	0.42	5.10	3.76	4.99	5.91	2.35	47.30	0.316

Table 2. **Ablation study of proposed networks on the Scene Flow test set and zero-shot generation.** The baseline is RAFT-Stereo with two levels of correlation pyramids. The parameters counted here are the trainable ones. The time is the inference time for 960×540 inputs. †*We found that pre-defining the neighbor sampling indexes within the search radius can significantly accelerate the inference instead of repeatedly defining them in every lookup as RAFT-Stereo's implementation. We also apply this trick to the baseline, otherwise, its inference time would be 0.329s.*

B. Ablation Study

B.1. Main Ablation

In this section, we verify the effectiveness of the proposed components via a main ablation study. The supplementary material will present a detailed ablation study of the combined encoders' design choices and the proposed scale update module's number of steps. In the main ablation study, we mainly use the ViT-S as the ViT backbone for our model and train all the variants on Scene Flow for 200k steps. For accuracy comparison, both the in-domain test and zero-shot generalization evaluation are presented. We also list trainable parameters and the inference time of the model variants for comparing computational complexities. The results are shown in Tab. 2.

Effecitveness of combined encoders. Compared with the baseline, the combined context encoder (CCE) and the combined feature encoder (CFE) achieves about 10% improvement on Scene Flow, while CCE performs slightly better than CFE. Their combination does not give much additional gain. Clear improvement appears on KITTI 2012, Middlebury, and ETH3D, but KITTI 2015.

Effecitveness of depth initialization. Initializing the disparity map with the modulated depth from DEFOM instead of zeros does not improve the model's performance on in-domain fitting. Nevertheless, DI achieves better generalization results on KITTI 2012, Middlebury, and ETH3D.

Effecitveness of scale update. When incorporating the scale update with depth initialization (+DI+SU), there is around 10% improvement on Scene Flow. Besides, apparent progress is also obtained in the generationalization evaluation of the four realistic datasets. Noteworthy, the Bad 2.0 of Middlebury is reduced by over 50%.

Effecitveness of all components. By integrating all the proposed modules into our complete model, we observe further improvements on the Scene Flow and Middlebury datasets. However, some slight performance drops are observed when compared to individual components on other datasets. For instance, on KITTI 2015, the full model (ViT-S) slightly underperforms the combination of depth initialization and scale update. Nevertheless, the full model demonstrates better overall performance. Additionally, using a larger ViT backbone, ViT-L, further enhances performance.

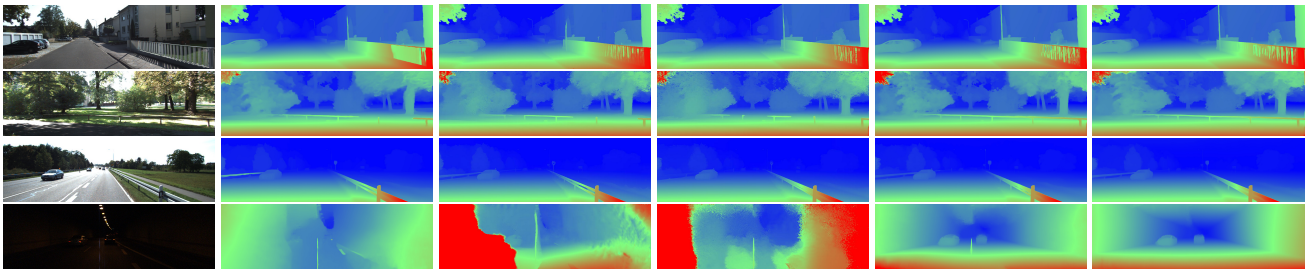
Trainbale Parameters. As we fixed the DEFOM, the new trainable parameters mainly come from the new DPT for CCE and CFE, and the ConvGRU for SU. using CCE and CFE meanwhile increases 2M parameters (+18%), using SU increases 5.4M parameters (+49%), and the full model (ViT-S) increases 7.4M parameters (+67%). The full model (ViT-L) has quadrupled the parameters, as the channels of the new DPT for CCE and CFE are defined to be proportional to those of the ViT backbone, following the fixed DPT of DEFOM for predicting depth.

Inference times. The increase in inference time is not as significant as the growth in model parameters, as the majority of the inference time is spent on recurrent update iterations. The total number of iterations in our model is set to match that of RAFT-Stereo. The proposed components contribute to a modest 10% increase in inference time individually, primarily due to the operation of DEFOM. The full model (ViT-S) results in a 15% increase in inference time, while the larger full model (ViT-L) sees a 42% increase.

B.2. Combined Encoders

In this section, we present an ablation study about combined encoders's design choices. Tab. 3 shows the results. For both the combined feature encoder and context encoder, we simultaneously experiment with other design choices for them, including using the DPT only to construct the encoders without CNNs and using the original fixed DPT instead of a new trainable DPT.

Can we simply abandon CNNs? The answer is **No**. We first ablate the CNNs in the encoders and use the feature maps from the new DPT head only as the matching feature maps and context maps. The results are listed in the first row of Tab. 3. The ablation would result in a significant performance drop on both in-domain test and zero-shot generation. For example, The EPE increases by over 35% on Scene Flow,



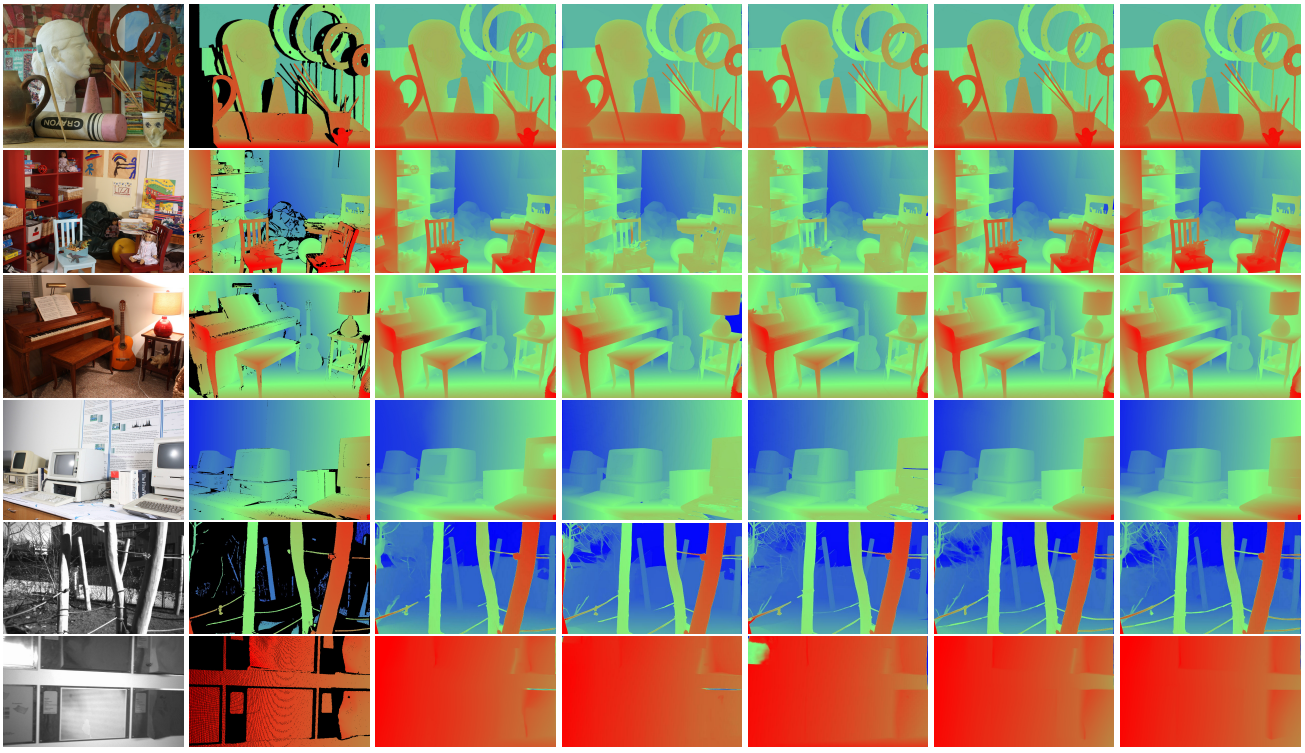


Figure 2. Zero-Shot qualitative comparison with RAFT-Stereo [6] Mocha-Stereo [2] and Selective-IGEV [13] on the four common realistic stereo datasets. **Row 1-2:** KITTI 2012. **Row 3-4:** KITTI 2015. **Row 5-6:** Middlebury-full. **Row 7-8:** Middlebury-half. **Row 9-10:** ETH3D. Best viewed in color and by zooming in.

Models	Scene Flow		KITTI 2015	Midd.-half	ETH3D
	EPE	Bad 1.0	Bad 3.0	Bad 2.0	Bad 1.0
without CNNs	0.619	7.078	5.998	6.655	3.872
without new DPT	0.473	5.815	5.450	5.265	2.278
Full Model	0.458	5.571	5.289	4.287	2.614

Table 3. **Ablation study on the design choices combined encoders.** Midd.-half represents Middlebury (half resolution).

and Bad 2.0 increases by over 35% on Middlebury (half resolution). The results indicate that the CNN feature is still necessary for the proposed model.

Is a new DPT beneficial? The answer is **overall Yes**. The second row of Tab. 3 shows the result of the model that used the fixed DPT of Depth Anything V2. There are about $3-5\%$ error increases on Scene Flow and KITTI 15 and a 23% rise on Middlebury (half resolution), while a 13% drop on ETH3D. We thus hypothesize that the fixed DPT is more favorable to data with a small disparity range (< 64), like ETH3D and a new trainable DPT is more helpful to the large disparity. As a new trainable DPT is generally better, we include it in our final model.

B.3. Iterations of Scale Update

In this section, we investigate the effect of the number of iterations of the scale update module in Tab. 4. Likewise, we fixed the total number of iterations as 18 in training and 32 in evaluation, and the number of SU iterations is set the same in training and evaluation. We perform training

on Scene Flow for 50k steps with a batch size of 4 and evaluation on the Scene Flow test set. The number of scale update iterations is increased from 0 to 10 and the 0 scale update iteration represents that the scale update module is not used. When the scale update iteration is 0, the model has the highest error metrics, and increasing it to 1 results in over 12% error reduction. The EPE continues to decrease until the scale update iteration reaches 8 and the performance for the 7-9 scale update iteration is closed. When the scale update iteration exceeds 9, the performance starts to drop obviously. Therefore, we select 8 as the number of scale update iterations in our final model.

SU Iter	0	1	3	5	7	8	9	10
EPE	0.752	0.668	0.651	0.650	0.640	0.636	0.637	0.660
Bad 1.0	9.018	8.030	7.804	7.709	7.683	7.697	7.660	8.243

Table 4. **Ablation study on scale update iterations.**

C. Zero-Shot Qualitative Comparison

In this section, we provide more visual comparison with RAFT-Stereo [6] , Mocha-Stereo [2] and Selective-IGEV [13] Therefore, we provide a visual comparison among RAFT-Stereo [6] with our DEFOM-Stereo (VIT-S), and DEFOM-Stereo (VIT-L). Fig. 2 presents the comparison on four common stereo datasets. We also provide the qualitative comparison on a more diverse stereo image dataset Flickr1024 [14] in Fig. 3 and Fig. 4. All the models are pretrained on the Scene FLow dataset only. The clear advantages of our models can be seen in the visual comparison.

D. Qualitative Comparison of RVC Models

This section presents a visual comparison among robust vision challenge models. We compare our RVC model with the previous best-performing model of individual benchmarks, *i.e.*, UCFNet_RVC [12] on KITTI 2015, CREStereo++_RVC [3] on Middlebury and LoS_RVC [5] on ETH3D. And our model demonstrates more accurate results simultaneously.

E. Evaluation on Ill-Pose Regions

To further show the detailed improvement in occluded and textureless areas, we evaluate the models on Middlebury, as the indoor scene contains sufficient occluded and textureless regions. We follow LoS to use SSIM to extract textureless regions from the image. Tab. 5 shows the results, where the proportions of different regions are also counted. There is obvious improvement in these ill-posed areas. Fig. 6 visualizes some examples for the evaluation.

Methods	All(100%)	Non-occluded (87.92%)	Occluded(12.08%)	Textureless(59.60%)
Our Baseline	13.44	10.64	30.33	13.01
Mocha-Stereo	11.49	9.11	25.79	12.25
Ours (ViT-S)	6.76	4.29	20.83	7.05
Ours (ViT-L)	5.91	3.26	20.64	6.04

Table 5. Zero-shot evaluation (Bad 2.0) on different areas of Midd.-half.

F. Transparent or Mirror Surfaces

We follow the reviewer's suggestion to evaluate our model on the dataset about transparent or mirror (ToM) surfaces. We examine the models on the Booster datasett [16] which features reflective and glass surfaces, and some examples are shown in Fig. 7. We find that DAv2 performs usually well for these reflective and transparent materials and our model also works if these ill-posed factors are not too serious. When there is a large mirror, our model cannot work, and DAv2(ViT-S) also slightly fails. For readers who require a very robot model for ToM, we refer them to Stereo Anywhere [1], which is specifically designed for this problem.

References

[1] 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. 5

[2] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 27768–27777, 2024. 4, 5, 7, 8, 10

[3] 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/CVF International Conference on Computer Vision*, pages 3318–3327, 2023. 5, 9

[4] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 1

[5] 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. 5, 9

[6] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 4, 5, 7, 8

[7] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 1

[8] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 1

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1

[10] 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. Springer, 2014. 1

[11] 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. 1

[12] Zhelun Shen, Xibin Song, Yuchao Dai, Dingfu Zhou, Zhibo Rao, and Liangjun Zhang. Digging into uncertainty-based pseudo-label for robust stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5, 9

[13] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 4, 5, 7, 8

[14] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5, 7, 8

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

[16] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022. CVPR. 5

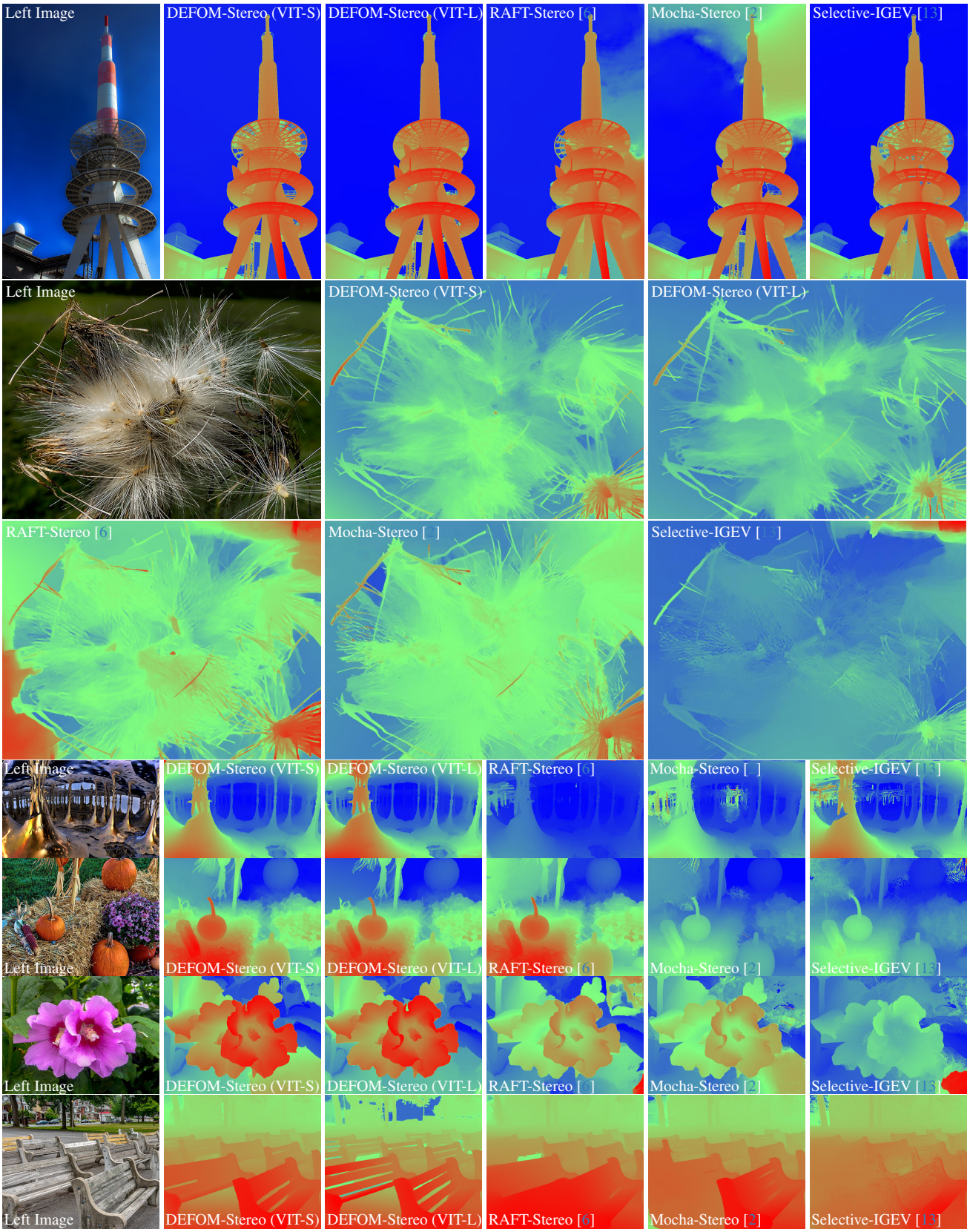


Figure 3. Zero-Shot qualitative comparison with RAFT-Stereo [6], Mocha-Stereo [2] and Selective-IGEV [13] on Flickr1024 [14]. Best viewed in color and by zooming in.

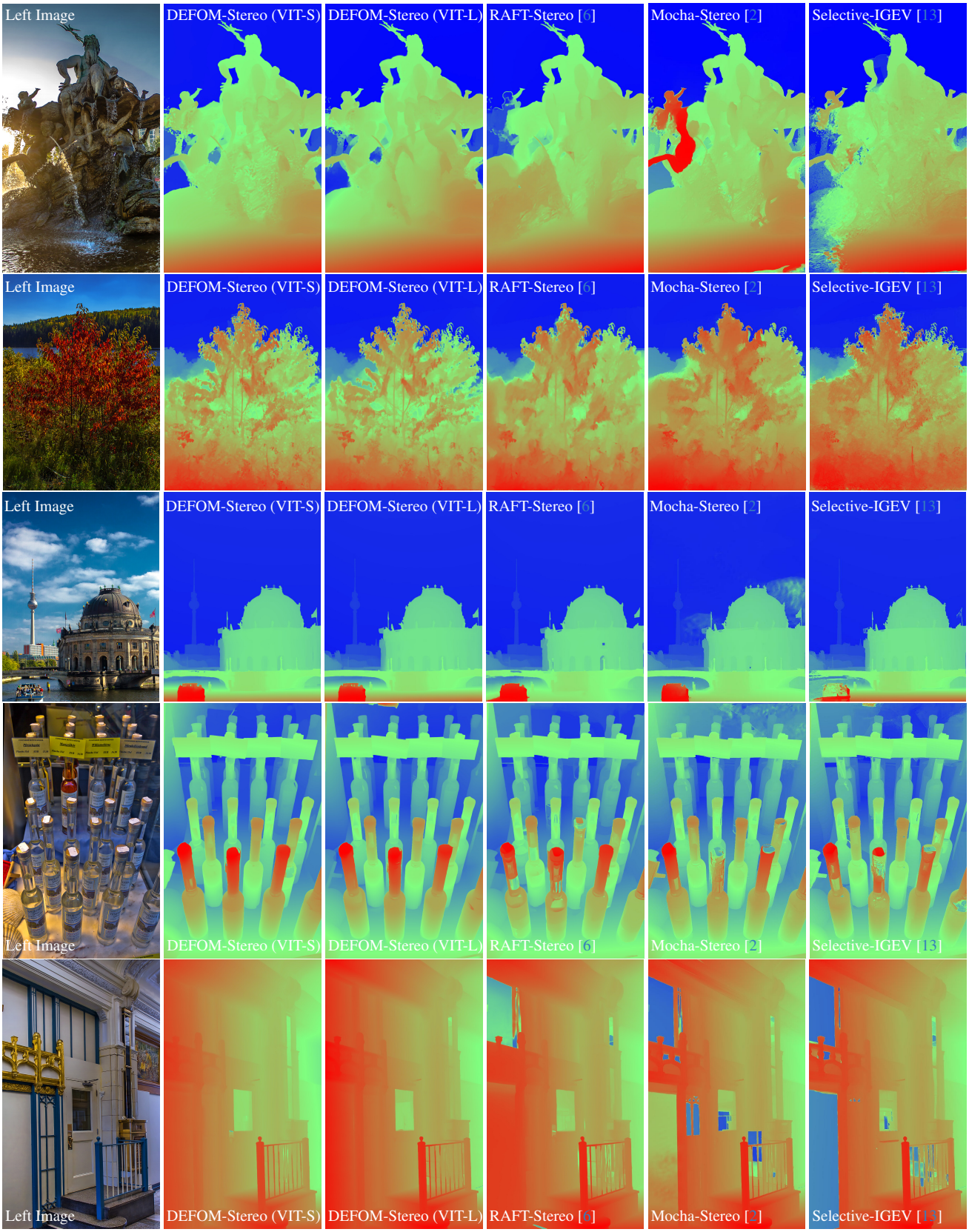


Figure 4. Zero-Shot qualitative comparison with RAFT-Stereo [6], Mocha-Stereo [2] and Selective-IGEV [13] on Flickr1024 [14]. Best viewed in color and by zooming in.

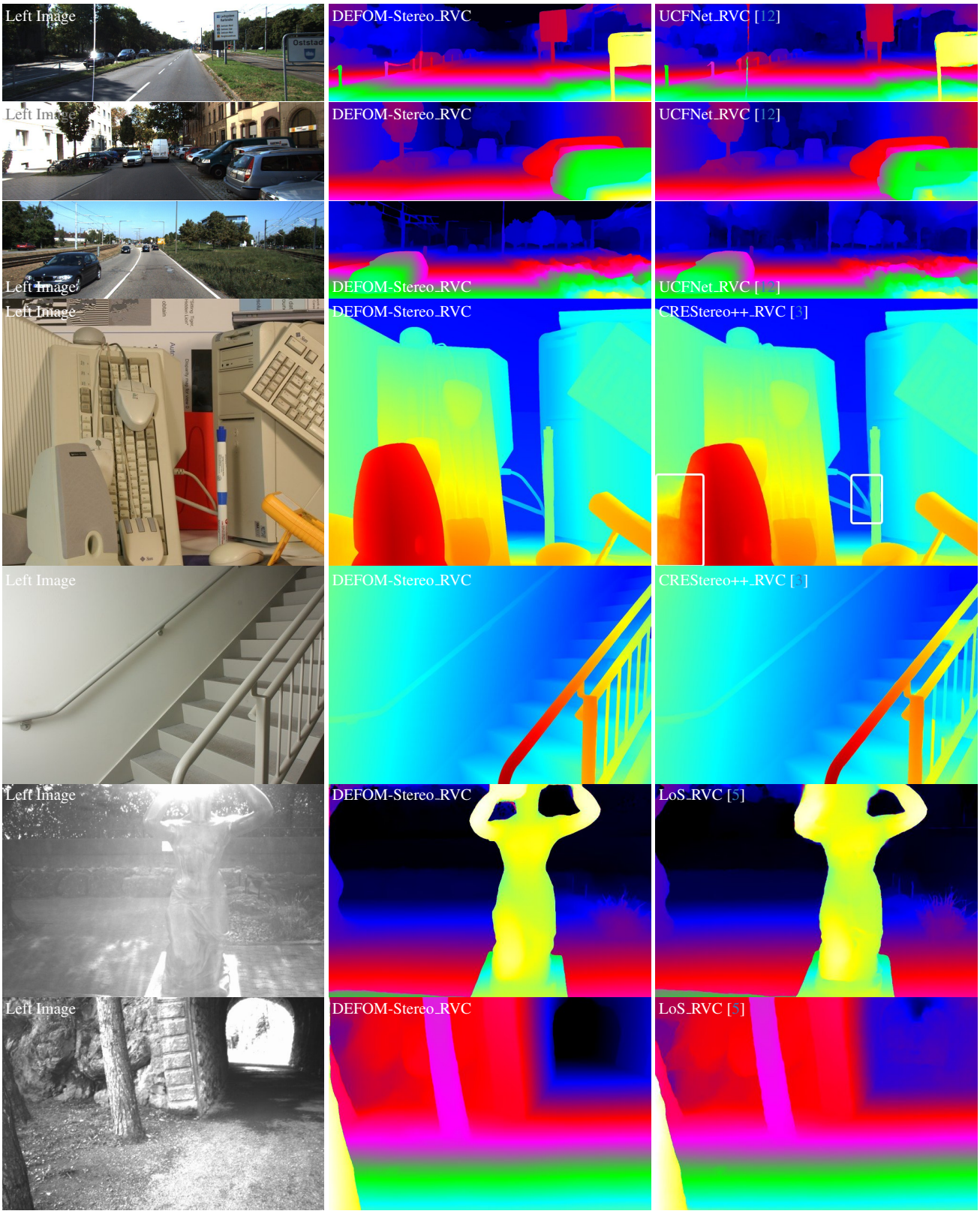


Figure 5. Qualitative Comparison among top-performing RVC models, including UCFNet_RVC [12], CREStereo++_RVC [3], LoS_RVC [5] and our model. **Row 1-3:** KITTI 2015. **Row 4-5:** Middlebury. **Row 5-6:** ETH3D. Best viewed in color and by zooming in.

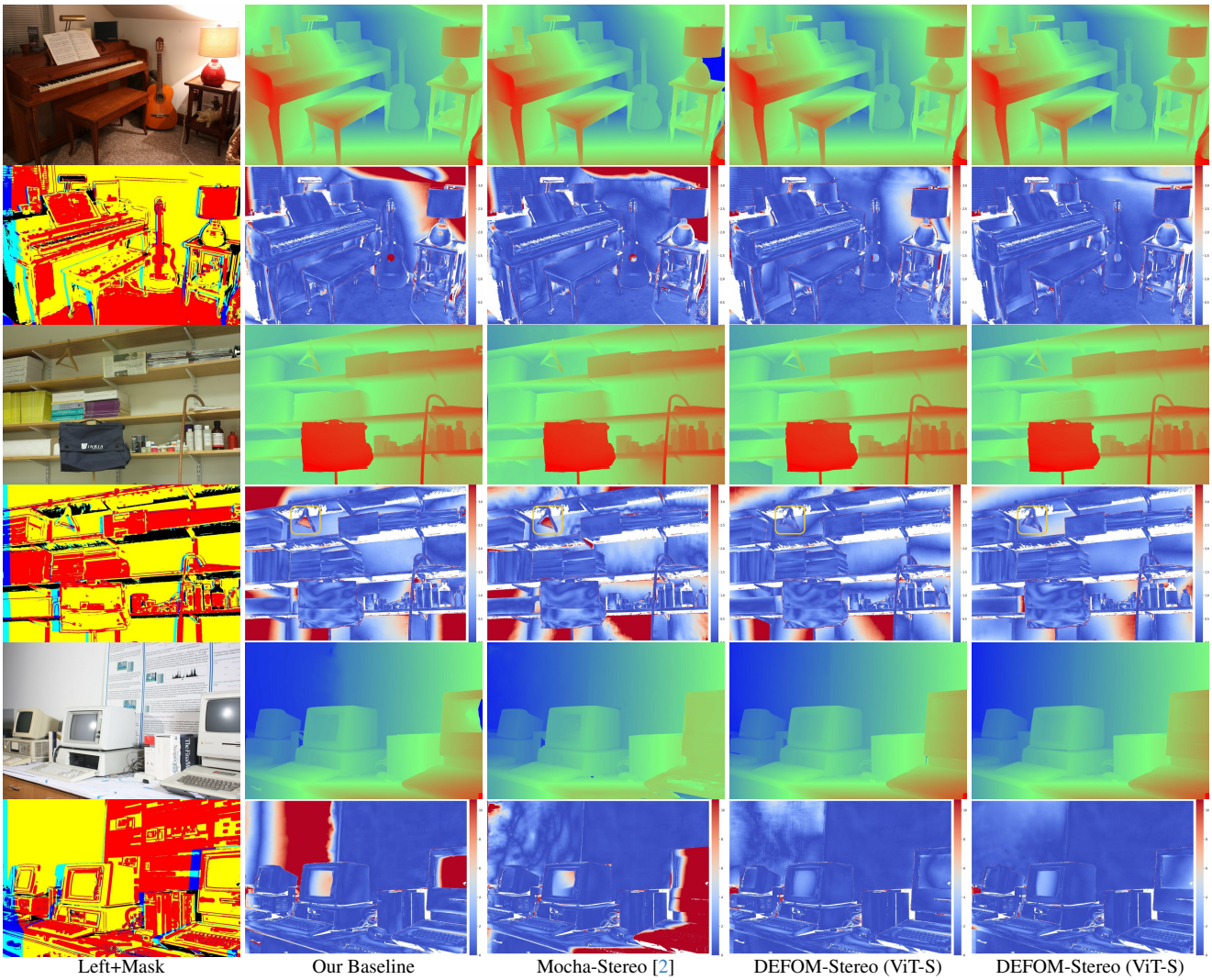


Figure 6. Visual comparison on ill-posed areas. **Odd Rows:** Left Image and Disparity Maps. **Even Rows:** Region Masks and Error Maps. Non-occluded and textured regions are in red. Non-occluded and textureless regions are in yellow. Occluded and textured regions are in blue. Occluded and textureless regions are in cyan. Best viewed in color and by zooming in.

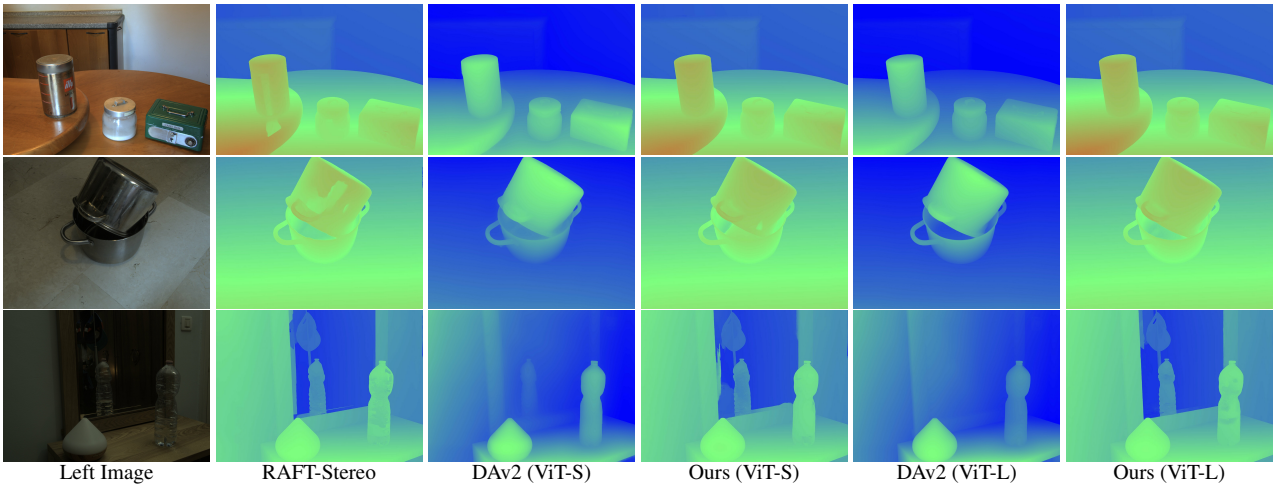


Figure 7. Examination of the depth models on the Booster dataset. Best viewed in color and by zooming in.