NTIRE 2024 Challenge on HR Depth from Images of Specular and Transparent Surfaces

Pierluigi Zama Ramirez Fabio Tosi Luigi Di Stefano Radu Timofte
Alex Costanzino Matteo Poggi Samuele Salti Stefano Mattoccia Yangyang Zhang
Cailin Wu Zhuangda He Shuangshuang Yin Jiaxu Dong Yangchenxu Liu
Hao Jiang Jun Shi Yong A Yixiang Jin Dingzhe Li Bingxin Ke
Anton Obukhov Tinafu Wang Nando Metzger Shengyu Huang Konrad Schindler
Yachuan Huang Jiaqi Li Junrui Zhang Yiran Wang Zihao Huang Tianqi Liu
Zhiguo Cao Pengzhi Li Jui-Lin Wang Wenjie Zhu Hui Geng Yuxin Zhang
Long Lan Kele Xu Tao Sun Qisheng Xu Sourav Saini Aashray Gupta
Sahaj K. Mistry Aryan Shukla Vinit Jakhetiya Sunil Jaiswal Yuejin Sun
Zhuofan Zheng Yi Ning Jen-Hao Cheng Hou-I Liu Hsiang-Wei Huang
Cheng-Yen Yang Zhongyu Jiang Yi-Hao Peng Aishi Huang Jenq-Neng Hwang

Abstract

This paper reports on the NTIRE 2024 challenge on HR Depth From images of Specular and Transparent surfaces, held in conjunction with the New Trends in Image Restoration and Enhancement (NTIRE) workshop at CVPR 2024. This challenge aims to advance the research on depth estimation, specifically to address two of the main open issues in the field: high-resolution and non-Lambertian surfaces. The challenge proposes two tracks on stereo and single-image depth estimation, attracting about 120 registered participants. In the final testing stage, 2 and 8 participating teams submitted their models and fact sheets for the two tracks.

1. Introduction

Recovering the 3D structure of a scene directly from images has been one of the most studied topics in computer vision. Depth estimation represents the first step for this purpose and a cornerstone for higher-level applications such as augmented reality, autonomous or assisted driving, robotics, and more. Although a variety of custom, *active* sensors exists for this task – LiDARs, Radars, Time-of-Flight (ToF), just to name a few – approaches estimating depth from one or multiple color images have gained higher and higher popularity with the advent of deep learning. Despite the steady improvements we witnessed in the last decade, estimating depth in certain conditions remains an open challenge. In particular, we identify two as the main sources of trouble.

The first is spatial resolution. Specifically, any depth sensor mentioned before provides depth maps at a relatively low resolution, usually not higher than 1 Megapixel (Mpx). On the contrary, color cameras nowadays reach a resolution of one or two orders of magnitude higher yet introduce significant computational complexity for processing.

The second is caused by *non-Lambertian* surfaces, resulting in a hard challenge for both active depth sensors and image-based approaches. Materials featuring this property violate the assumptions upon which active sensors are developed – e.g., light beams emitted by LiDARs are refracted or surpass transparent surfaces. Image-based techniques are also affected, with stereo matching algorithms or monocular depth estimation models, for instance, failing to estimate the real distance of a transparent surface in favor of the distance of objects behind it. Although one may feel that this latter example might not represent a real failure, we argue it is: indeed, in several real applications, it might be crucial to properly perceive the real depth for transparent objects as well – for instance, when willing to grasp some glassy objects or when navigating and willing to avoid a glass door.

This NTIRE 2024 Challenge on HR Depth from Images of Specular and Transparent Surfaces aims to encourage the development of state-of-the-art methodologies for estimat-

*Pierluigi Zama Ramirez (pierluigi.zama@unibo.it), Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, Luigi Di Stefano and Radu Timofte are the NTIRE 2024 HR Depth from Images of Specular and Transparent Surfaces challenge organizers. The other authors participated in the challenge.
Appendix contains the authors' team names and affiliations.
The NTIRE website: https://cvlai.net/ntire/2024/

ing depth from single images that are robust and effective at dealing with the aforementioned challenges. For this purpose, we employ the Booster dataset [98, 100] in this challenge, a recent benchmark that represents a proving ground for what concerns high-resolution and non-Lambertian surfaces, thanks to its 12Mpx images and the abundant presence of transparent and reflective objects. Following the format of the first edition, the challenge is organized into two tracks: one focusing on *Stereo* approaches, recovering depth through triangulation from the *disparity* estimated between pixels into two rectified frames, and the other limiting the input to a single image (*Mono*). The challenge has 120 registered participants. Among them, 8 and 2 teams for the monocular and stereo tracks submitted their models and fact sheets during the final phase. Some adopt off-the-shelf, existing solutions, while others combine different methodologies and exploit their synergy to obtain better results. The outcome of this edition of our challenge is reported and discussed in detail in Section 4.

2. Related Work

We review the literature relevant to stereo and monocular depth estimation, which is the object of our challenge.

Deep Stereo Matching. Deep networks estimating dense disparity maps in end-to-end manner have emerged as the preferred paradigm to tackle stereo matching [53, 55]. This revolution ignited with DispNet [45], a 2D CNN followed by more and more advanced architectures [39, 50, 54, 62, 68, 75, 77, 92, 97]. An alternative family of model emerged with GC-Net [29], that builds an explicit *cost volume* and then processes it with 3D convolutions, an approach followed, again, by several following-up works [5, 8, 9, 13, 22, 30, 67, 80, 85, 91, 103]. In the last three years, two further trends emerged with Transformers [20, 37, 43] and optimization-inspired architectures [76]. The latter in particular, starting with RAFT-Stereo [76], has conquered the main stage lately [26, 34, 81, 88, 90, 102]. The steady advances in the design of deep stereo models brought, through the years, to a saturation of the most popular benchmarks, starting with KITTI 2012 [16] and 2015 [47], then proceeding with ETH3D [66] and, only lately, Middlebury 2014 [65]. Nevertheless, these benchmarks do not specifically focus on the most arduous open challenges for stereo matching, which are the main objects of study in the Booster [100] dataset. Accordingly, in this challenge, we rely on the latter.

Monocular Depth Estimation. To estimate depth out of a single image, hand-crafted features at first were used to encode perceptual cues such as texture gradient, object size, and linear perspective – the cornerstones of early research in the field [64]. The advent of deep learning made it possible to tackle this task and to achieve unprecedented results by directly learning from data [6, 14, 33, 56, 84]. The increasing availability of large-scale datasets annotated with ground-truth depth labels [6, 14, 33, 56, 84] played a crucial role in the quick escalation of this field, side by side with the introduction of self-supervised paradigms [17–19, 21, 25, 27, 52, 78, 79, 87, 105, 106] to address the lack of annotations – specifically, by casting the depth estimation task as an image reconstruction problem during training, by exploiting either stereo pairs or monocular videos. In the last four years, the development of affine-invariant monocular depth estimation models [58, 60] has gained popularity. MiDaS [60] represents the pivotal work in this direction, training a CNN on a mixture of several datasets to achieve cross-domain generalization – followed by DPT [58], Depth Anything [93], and Marigold [28]. Other works focused on recovering the real point cloud shapes from the deformed ones obtained from monocular depth maps [96] or restoring high-frequency details [36, 48].

Despite these steady advances, little attention has been given to single-view depth estimation networks capable of handling transparent and reflective surfaces effectively. This is mostly because of the scarcity of datasets specifically suited for this task – except for Booster [98], featuring some very challenging yet accurately annotated non-Lambertian objects in high-resolution images. On this track, Costanzino *et al*. [12] developed a strategy for retrieving pseudo-annotation for non-Lambertian objects by using monocular depth estimation models and material segmentation masks while others have faced non-Lambertian depth estimation through depth completion approaches [10, 63].

Competitions/Challenges on Depth Estimation. We mention some past – and concurrent – challenges built around the depth estimation task, both from stereo and monocular images. Among them, the Robust Vision Challenge (ROB) [101] covering both, the Dense Depth for Autonomous Driving challenge (DDAD)[15], the Fast and Accurate Single-Image Depth Estimation on Mobile Devices Challenge (MAI) [24], the Argoverse Stereo Challenge [32] and the Monocular Depth Estimation Challenge (MDEC) [70–72]. Finally, we recall the first edition of this challenge [57], part of the NTIRE workshop at CVPR 2023.

NTIRE 2024 Challenges. This challenge is one of the NTIRE 2024 Workshop [1] associated challenges on: dense and non-homogeneous dehazing [1], night photography rendering [2], blind compressed image enhancement [94], shadow removal [82], efficient super resolution [61], image super resolution ($\times 4$) [7], light field image super-resolution [86], stereo image super-resolution [83], HR depth from images of specular and transparent surfaces [99], bracketing image restoration and enhancement [104], portrait quality assessment [4], quality assessment for AI-generated content [41], restore any image model (RAIM) in the wild [38], RAW image super-

[1] https://cvlai.net/ntire/2024/

resolution [11], short-form UGC video quality assessment [35], low light enhancement [42], and RAW burst alignment and ISP challenge.

3. NTIRE Challenge on HR Depth from Images of Specular and Transparent Surfaces

We host the NTIRE 2024 Challenge on HR Depth from Images of Specular and Transparent Surfaces to encourage the community to develop state-of-the-art solutions capable of dealing with high-resolution images and non-Lambertian surfaces – such as mirrors, glasses, and more. We now introduce the main details of the challenge.

Tracks. Our challenge is organized into two tracks: *Stereo*, focusing on estimating the disparity between pairs of rectified images, and *Mono*, which instead requires estimating depth from a single input image.

- **Track 1: Stereo.** In this track, the participants are asked to obtain high-quality, high-resolution disparity maps from 12Mpx stereo pairs. The main difficulties are represented by the resolution itself, which is prohibitive for most state-of-the-art existing stereo networks, and the presence of non-Lambertian objects, violating the basic assumptions allowing to retrieve depth out of correspondences.
- **Track 2: Mono.** Conversely, this track consists of estimating depth out of a single 12Mpx image. This problem is more challenging than the former because of the inherent ill-posed nature of the problem. Furthermore, the presence of several transparent objects and mirrors – rarely appearing in most depth estimation datasets – makes it even more complex.

Datasets. Our challenge takes place over the Booster dataset [98, 100], consisting of 419 high-resolution balanced and unbalanced stereo pairs, collected in 64 different scenes and respectively divided into 228 and 191 pairs for training and testing purposes – with 38 and 26 for the two sets respectively. An extended version of Booster [98] releases a second testing split, dedicated to the evaluation of monocular depth estimation approaches and made of 187 single frames, collected from 21 new environments.

As in the first edition [57], we adopt the original 228 training stereo pair as the *training split* for both tracks. We identify two distinct *validation splits* by sampling images with different illuminations from 3 scenes of the stereo and monocular testing splits – respectively *Microwave, Mirror1, Pots* for the Stereo track, and *Desk, Mirror3, Sanitaries* for the Mono track, yielding 15 validation samples for each track, out of the total 26 and 28 available from the selected scenes. The remaining frames of the two original testing splits become the official stereo and mono *testing splits* for this challenge, resulting in 169 and 159 samples.

Evaluation Protocol. Depending on the specific track, Stereo or Mono, we select the official metrics used by the Booster benchmark [98, 100]. For the former track, we measure the percentage of pixels with disparity errors larger than a threshold τ (bad-τ, with $\tau \in [2, 4, 6, 8]$), as well as the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). For the latter Track, we measure the percentage of pixels having the maximum between the prediction/ground-truth and ground-truth/prediction ratios lower than a threshold ($\delta < i$, with i being 1.05, 1.15, and 1.25) and the absolute error relative to the ground truth value (Abs Rel.), as well as the mean absolute error (MAE), and Root Mean Squared Error (RMSE). Differently from the previous edition [57], we compute metrics on three different sets of pixels, following [12]: *ToM* regions – i.e., those belonging to non-Lambertian surfaces – *All* pixels and *Others* – i.e., the difference between *All* and *ToM* sets. To rank submissions, we use bad-2 and $\delta < 1.05$ – respectively for Stereo and Mono tracks – averaged over all pixels, highlighted in **red** in the tables. We define two rankings based on performance on *ToM* and *All* regions respectively[2]. Finally, since most monocular networks estimate depth up to an unknown scale and shift factors, before computing metrics we recover metric depth from predicted maps $\hat{d}$ as $\alpha\hat{d}+\beta$, with α, β being a scale and shift factor. According to [60], α, β are estimated with Least Square Estimation (LSE) regression over the ground truth depth map d:

$$(\alpha, \beta) = \arg\min_{\alpha,\beta} \sum_p \left(\alpha\hat{d}(p) + \beta - d(p)\right)^2 \tag{1}$$

where p are the pixel locations where both predictions and ground truth depths are defined.

4. Challenge Results

For the two distinct tracks, 2 and 8 teams participated respectively in the final testing phase. We are now discussing the outcome of both in Sections 4.1 and 4.2. Each method for stereo and mono tracks is briefly described in Section 5.1 and Section 5.2, with team members listed in the appendix.

4.1. Track 1: Stereo

Table 1 collects the results for this first track. At the bottom, we report the baseline method – i.e., the CREStereo [34] model using the weights publicly available. From left to right, we report bad-τ metrics, MAE, and RMSE metrics for *Tom*, *All*, and *Other* pixels respectively. On the right of the team's name, we report their overall rank, computed according to bad-2 errors on *ToM* and *All* regions.

Both methods participating in this track outperformed the baseline, with MiMcAlgo [Stereo] consistently achieving lower error rates than SRC-B [Stereo] on *ToM* and *All*

[2] we will observe that the two coincide on the Stereo track

		ToM						All						Other					
Team	Rank	**bad-2**	bad-4	bad-6	bad-8	MAE	RMSE	**bad-2**	bad-4	bad-6	bad-8	MAE	RMSE	bad-2	bad-4	bad-6	bad-8	MAE	RMSE
MiMcAlgo [Stereo]	**#1**	**52.46**	**33.56**	**23.38**	**18.75**	**6.70**	**11.51**	**32.56**	**16.32**	**11.01**	**8.61**	**3.50**	**8.31**	**29.18**	**12.70**	**8.31**	**6.31**	**2.85**	**7.09**
SRC-B [Stereo]	**#2**	**59.27**	**38.24**	**31.08**	**27.04**	**8.81**	**13.21**	**32.79**	**19.39**	**14.59**	**11.86**	**4.19**	**9.22**	**28.51**	**14.60**	**10.02**	**7.66**	**2.94**	**6.92**
CREStereo [baseline]	**#3**	**59.64**	**47.26**	**40.27**	**35.41**	**24.69**	**42.28**	**35.75**	**23.51**	**18.98**	**16.42**	**12.13**	**28.46**	**28.34**	**13.93**	**7.91**	**4.64**	**2.95**	**7.59**

Table 1. **Stereo Track: Evaluation on the Challenge Test Set.** Predictions evaluated at full resolution (4112×3008) on All pixels and pixels belonging to ToM (Transparent or Mirror) or Other materials. In **gold**, **silver**, and **bronze**, we show first, second, and third-rank approaches, respectively. We rank methods on the **bad-2** metric.

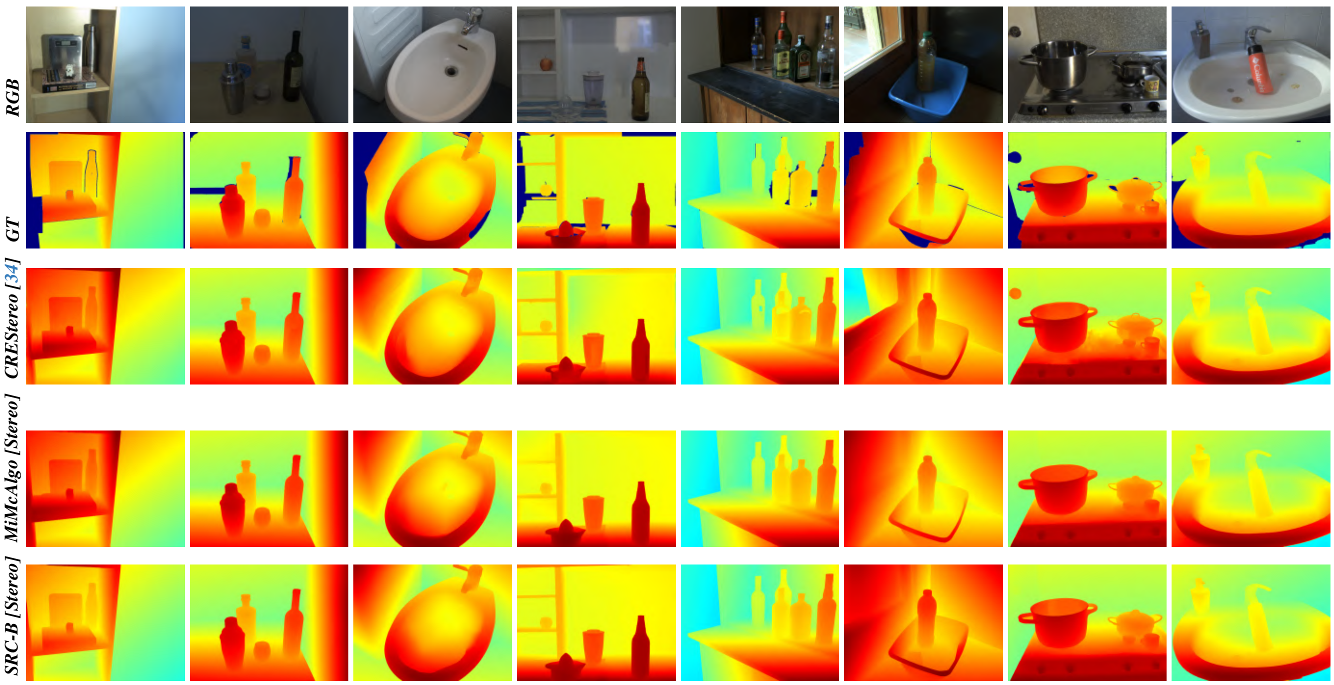

Figure 1. **Qualitative results – Stereo track.** From top to bottom: RGB reference image, ground-truth disparity, predictions by CREStereo [34], MiMcAlgo [Stereo], and SRC-B [Stereo].

pixels, with very few exceptions on *Other* regions – i.e., on bad-2 and RMSE. Interestingly, we can notice how the two achieve very close bad-2 rates on *All* pixels, as a compromise between the much more accurate results achieved by MiMcAlgo [Stereo] on *ToM* regions – i.e., about 7% lower error – and the slightly lower errors on *Other* pixels yielded by SRC-B [Stereo] – that is 0.7%, yet represents the majority of the pixels in the images. Fig. 1 shows some results from the stereo testing set. We can notice how both submitted methods learn to deal with some specific challenges – the bottles in column 5 and the window in column 6 – while they still struggle at properly dealing with very challenging elements, such as the water surface in the rightmost column.

4.2. Track 2: Mono

Table 2 shows the results for the second track. At the very bottom, we report the results achieved by the baseline method – i.e., the ZoeDepth [3] model using the weights provided by the authors. From left to right, we report deltas, Abs Rel., MAE, and RMSE metrics for *Tom*, *All*, and *Other* pixels respectively. We report two different rankings, according to the performance observed on the reference metrics computed over *ToM* and *All* pixels respectively.

All of the submitted methods consistently outperformed the ZoeDepth baseline. For what concerns *ToM* regions, the top #3 methods manage to push the strictest accuracy metric – $\delta < 1.05$ – beyond 70%, as well as to reduce the Abs Rel. below 4%. The improvements are consistent on *Other* pixels as well – and, consequently, on *All*. There, the gain over the baseline is minor compared to what was observed on *ToM* regions, yet consistent.

Finally, we can appreciate the substantial improvement achieved by the two absolute winners, MiMcAlgo [Mono] and SmartLab, respectively, according to *ToM* and *All* rankings. Fig. 2 shows some qualitative examples from the mono testing set: we can appreciate how, in some cases, any of the submitted models can properly handle *ToM* regions – as for the oven in the third row. However, we can still observe failure cases in most of them in the presence of mirrors (first row) or water surfaces (second row).

	ToM							All							Other					
Team	Rank	δ <1.05	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	RMSE	Rank	δ <1.05	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	RMSE	$\delta < 1.05$	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	RMSE
MiMcAlgo [Mono]	**#1**	**77.11**	**98.09**	**99.66**	**3.38**	**3.44**	**4.60**	**#2**	**71.64**	**94.86**	**98.46**	**5.03**	**4.39**	**8.09**	**69.93**	**93.99**	**98.06**	**5.54**	**4.68**	**9.04**
SmartLab	**#2**	**75.78**	**99.08**	**99.84**	**3.40**	**3.44**	**4.65**	**#1**	**79.97**	**97.95**	**99.53**	**3.77**	**3.25**	**6.32**	**79.59**	**97.33**	**99.32**	**4.09**	**3.38**	**7.13**
PD&HPC	**#3**	**70.04**	**96.79**	99.56	**3.98**	**4.16**	**5.06**	#6	65.43	93.27	96.35	6.20	5.37	8.67	63.68	92.33	95.90	6.77	5.70	9.76
UW IPL	#4	63.48	96.47	**99.64**	4.21	4.42	5.29	**#3**	**68.08**	**95.52**	**98.80**	**5.19**	**4.48**	**8.00**	**67.25**	**94.34**	**98.48**	**5.64**	**4.71**	**8.89**
Marigold-LCM	#5	62.88	92.59	98.25	5.40	5.57	7.21	#4	66.27	89.62	96.10	6.87	5.68	10.38	63.96	88.14	96.09	7.50	5.95	11.35
THU-808	#6	59.72	90.00	97.92	5.71	5.93	7.23	#5	65.81	90.14	96.69	6.61	5.55	10.08	64.44	89.46	97.08	7.06	5.66	10.92
SRC-B [Mono]	#7	57.01	95.31	96.88	5.73	5.83	6.93	#7	63.61	90.62	95.13	7.16	5.88	10.07	61.65	88.28	94.83	7.94	6.07	11.24
DVision	#8	56.59	90.28	97.23	5.83	6.08	7.12	#8	61.95	91.01	96.09	6.72	5.86	9.48	61.50	91.23	96.58	7.02	5.81	10.17
ZoeDepth [Baseline]	#9	45.21	82.27	93.06	8.04	8.71	9.57	#9	61.31	87.97	94.38	7.60	6.38	10.88	60.23	87.43	93.71	8.34	6.31	12.18

Table 2. **Mono Track: Evaluation on the Challenge Test Set.** Predictions evaluated at full resolution (4112×3008) on All pixels and pixels belonging to ToM (Transparent or Mirror) or Other materials. In **gold**, **silver**, and **bronze**, we show first, second, and third-rank approaches, respectively. We rank methods on two metrics, $\boldsymbol{\delta}$ **<1.05** computed on either ToM or All pixels.

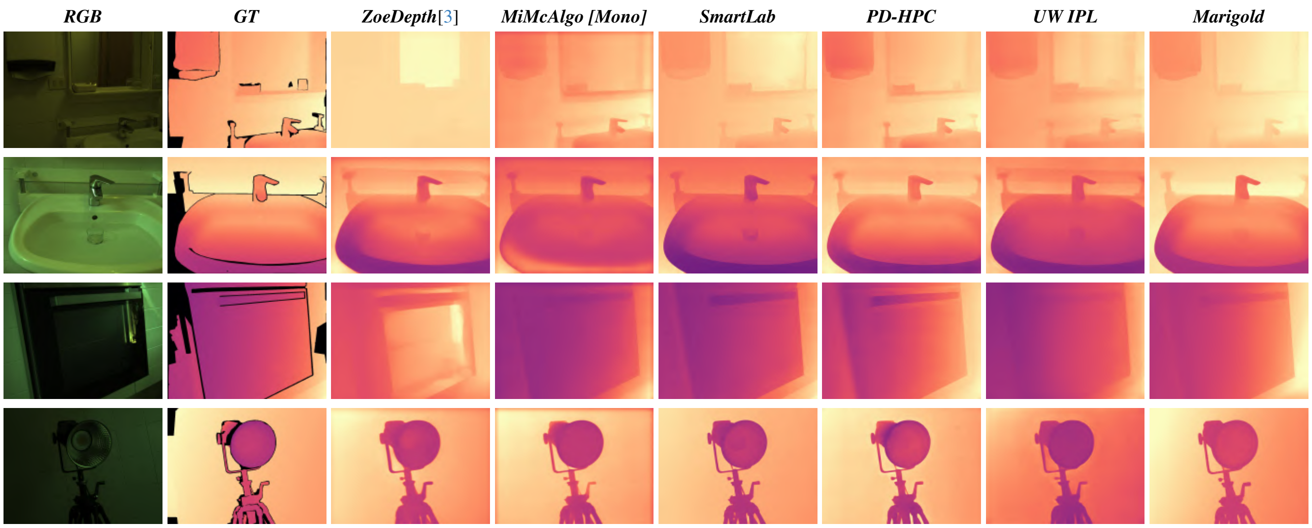

Figure 2. **Qualitative results – Mono track.** From left to right: RGB reference image, ground-truth disparity, predictions by ZoeDepth [3] and five among the participant methods.

5. Challenge Methods

5.1. Track 1: Stereo

5.1.1 Baseline - CREStereo [34]

For the Stereo track, we select the state-of-the-art CREStereo architecture [34] as the baseline. It consists of a hierarchical network with recurrent refinement, designed to update disparities in a coarse-to-fine manner. At its core, an adaptive group correlation layer (AGCL) is designed to mitigate the impact of non-ideal rectification, where an alternate 2D-1D local search strategy with deformable windows is employed for robust matching. Conversely to the all-pairs correlation module in RAFT-Stereo [40], AGCL computes correlations only in local search windows, reducing memory and computation requirements. We process images at quarter resolution and upsample predicted disparity maps to the original resolution.

5.1.2 Team 1 - MiMcAlgo [Stereo]

The team MiMcAlgo [Stereo] (CodaLab: MiDualCam) proposed a teacher-student framework for learning to handle non-Lambertian surfaces.

Specifically, IGEV-Stereo [89] is adopted as the baseline architecture for both the teacher and student, as it relies on the strong semantical and context information extracted from the left image to refine disparity predictions with convGRUs. Furthermore, the team observed that training the model with low-resolution images is more likely to predict the correct disparity for ToM objects, whereas a network trained with high-resolution images is more prone to errors in these regions, which may be related to the fact that mirror objects require larger receptive fields and high-level semantic information to be recognized [23].

Accordingly, the Booster training set was downsampled to $\frac{1}{5}$ and $\frac{1}{8}$ of the input resolution to train two different teacher networks. Then, N different weak augmentations (mainly on color and brightness) were applied to the unlabeled training images, downsampled by the aforementioned factors, and sent to the two teacher networks for inference to produce 2×N predictions. These are upsampled to the original resolution and, among them, the best are selected as pseudo-labels and used to train the student network on $\frac{1}{8}$, specifically by alternating between labeled and unlabeled data to learn more robust feature representations (see Fig. 3

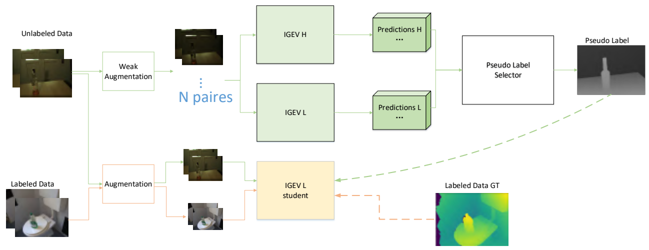

Figure 3. **Network Architecture – Team *MiMcAlgo [Stereo]*.**

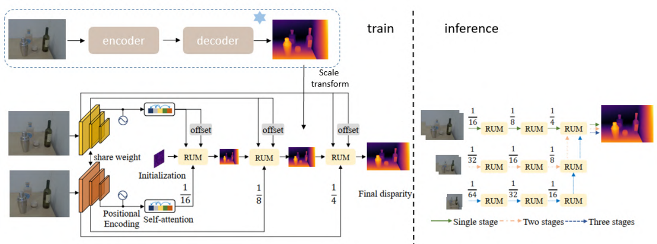

Figure 4. **Network Architecture – Team *Samsung R&D Institute China-Beijing (SRC-B [Stereo])*.**

for an overview). Pseudo-labels for training are obtained by computing the per-pixel average across the N predictions at $\frac{1}{8}$ resolution and selecting those closest to the average as *GT1* labels. Then, among the N predictions at $\frac{1}{5}$, those closest to *GT1* are selected as *GT2*. Finally, the bad-4 error between the two is computed: if it exceeds a threshold (15%), *GT1* are selected as final pseudo-labels – assuming that *GT2* fails on ToM objects because of its lower receptive field, otherwise *GT2* are used for training the student.

5.1.3 Team 2 – SRC-B [Stereo]

Samsung R&D Institute China-Beijing (SRC-B [Stereo]) (CodaLab: pixinsight) proposed a two-branch architecture combining the power of stereo and mono, shown in Fig 4.

In the first branch, it adopts Depth-Anything [93], which maximizes the preservation of the DinoV2's semantic features and utilizes both labeled and unlabeled images to facilitate better monocular depth estimation. The second branch implements CREStereo [34], a hierarchical network to predict disparities in a coarse-to-fine manner. This approach employs an adaptive group local correlation layer that uses cross and self attention [73] to aggregate global context information, a 2D-1D alternate local strategy to handle imperfect epipolar images, a deformable search window to reduce matching ambiguity, and feature map grouping [22] to improve performance. The relative disparity predicted by Depth-Anything [93] is aligned with the metric disparity predicted by the stereo network using least squares to calculate a global translation and scaling. Then, the aligned monocular disparity replaces the prediction of the stereo network and is used to carry out subsequent iterative optimization. Overview in Fig. 4.

The framework is implemented using Pytorch [51] and trained on 4× 3090 GPUs. In the first stage, the disparity maps of the Booster training dataset are aligned to the Depth-Anything prediction range and used to fine-tune the depth head of Depth-Anything itself for 100 iterations with an L1 loss, starting from the pre-trained model from [93]. Then, in the second stage, the monocular module is frozen, and CREStereo is fine-tuned for an additional 1000 epochs. During the training phase, they apply several augmentation techniques to the training samples, including random scaling, cropping, chromatic augmentation, and random occlusions. During the inference phase, images are downsampled to $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ to construct an image pyramid that is then fed into the network following [34].

5.2. Track 2: Mono

5.2.1 Baseline - ZoeDepth [3]

For the Mono track, we adopt the ZoeDepth model as the baseline, a state-of-the-art network for the monocular depth estimation task. It relies on DPT [59] as its main backbone, an encoder-decoder model that leverages a vision transformer (ViT) as a building block for the encoder, enriched by a metric bins module designed for learning a metric depth representation. Similar to the Stereo track, we use the available weights provided by the authors.

5.2.2 Team 1 - Marigold-LCM

The Marigold-LCM team (CodaLab: *anton*) combines the recently proposed Marigold depth estimator [28] with Latent Consistency Models (LCM) [44, 69] to achieve efficient inference while maintaining high-quality depth predictions. Marigold leverages the pre-trained Stable Diffusion model for conditional depth generation, with only the latent U-Net component being fine-tuned during training on a dataset of 73K synthetic samples from Hypersim and Virtual KITTI. To enhance inference efficiency, the team employs LCM by distilling knowledge from Marigold into a student model, which is trained to produce outputs identical to Marigold's through a self-consistency function. During testing, Marigold's U-Net is replaced with the trained student model, the DDIM scheduler is replaced with the LCM scheduler, and only 3 denoising steps are run, along with test-time ensembling of 10 samples, ultimately achieving high-quality depth predictions with significantly fewer denoising steps. The input images are downsampled to the resolution of 374 × 512 for inference and then upsampled to the original resolution. Notably, the Booster dataset is not seen during the original fine-tuning of Marigold or distillation of Marigold-LCM. Overview in Fig. 5.

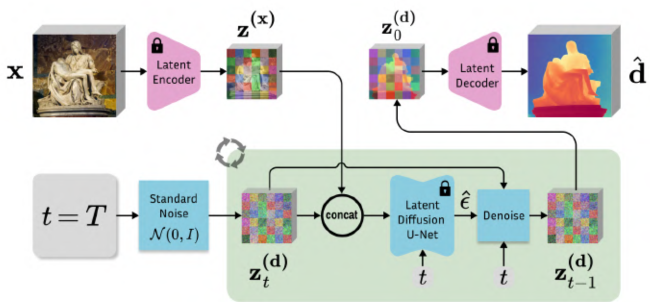

Figure 5. **Network Architecture – Team *Marigold-LCM*.**

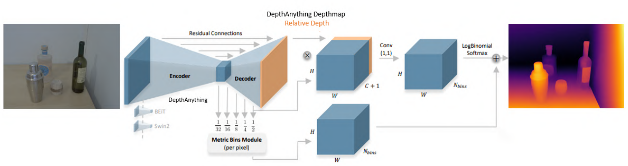

Figure 6. **Network Architecture – Team *SRC-B [Mono]*.**

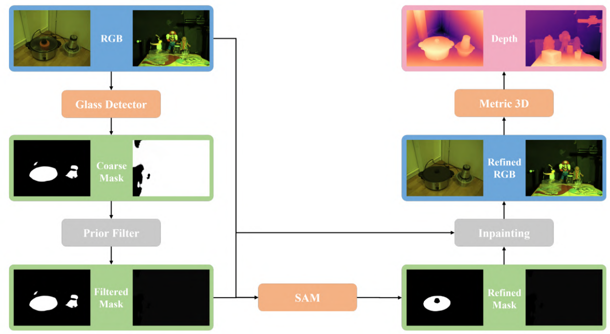

Figure 7. **Network Architecture – Team *SmartLab*.**

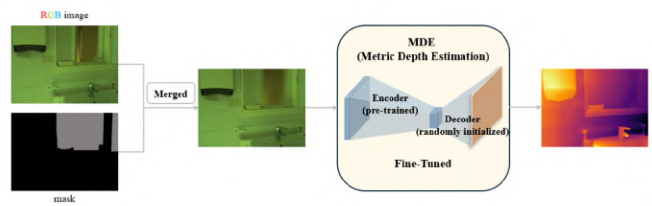

Figure 8. **Network Architecture – Team *PD&HPC*.**

5.2.3 Team 2 - SRC-B [Mono]

The SRC-B [Mono] team exploits the Depth Anything [93] base model, which follows the training strategy of MiDaS [60] by using a mixed training set and extending it to 62M unlabeled data. The Depth Anything model enhances the preservation of semantic features from DINOv2 [49] while using a teacher-student framework to train on unlabeled data. Specifically, the method employs an affine-invariant loss, introduces strong color and spatial distortions, and integrates the Depth Anything encoder into the ZoeDepth [3] framework to convert relative disparity to metric depth. For fine-tuning on the Booster dataset, the team adjusts the input image dimensions to 770×770 and conducts fine-tuning over 100 epochs. Overview in Fig. 6.

5.2.4 Team 3 - SmartLab

The "SmartLab" team presents a training-free approach for estimating depth in scenes with transparent and mirror surfaces. The method employs a coarse-to-fine strategy, first using a glass detection model, GDNet [46], to generate a coarse mask of potential transparent and mirror surfaces. The refined masks are used to sample points within the masked regions, which are fed into the Segment Anything Model (SAM) [31] to obtain more precise masks enriched with semantic information. The masked regions of the input image are inpainted using a strategy of filling with the most frequently occurring color similarly to [12]. Finally, the Metric3D [95] depth predictor is used to estimate the depth of the transparent and mirror surfaces based on the inpainted image. Overview in Fig. 7.

5.2.5 Team 4 - THU-808

The THU-808 team's "FuseDepth" method is inspired by the Boosting Monocular Depth (BMD) [48] work, which suggests that different resolution inputs yield depth maps with varying levels of detail. The approach builds upon the Marigold [28] diffusion-based depth estimation method. The team devises a depth fusion approach tailored to inputs of different resolutions, specifically 512 and 1024. Instead of training a depth fusion network that could disrupt the original depth distribution, they employ guided filtering to fuse the depth maps. The threshold and radius for the guided filtering are set to 64 and $1e^{-8}$, respectively. Although simple, this method effectively enhances depth accuracy by leveraging the varying levels of detail obtained from different resolution inputs.

5.2.6 Team 5 - PD&HPC

The PD&HPC team's "DepthBlur" approach begins by fine-tuning the Depth Anything model [93] on the Booster training set, modeling the training phase as in ZoeDepth [3]. This fine-tuning process significantly improves the model's performance in handling complex surfaces. The team investigates the effect of image preprocessing techniques on further enhancing the model's accuracy. They apply Gaussian blurring to the ToM portions of the images, which are identified using a segmentation model. This preprocessing step mimics the real-world light scattering effect on these surfaces and provides additional visual cues for depth estimation. However, the team observes that hardware limita-

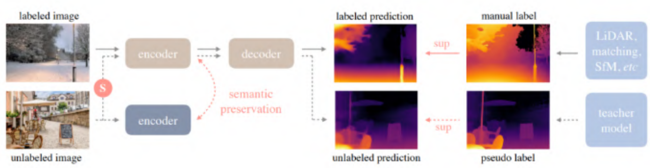

Figure 9. **Network Architecture – Team *DVision*.**

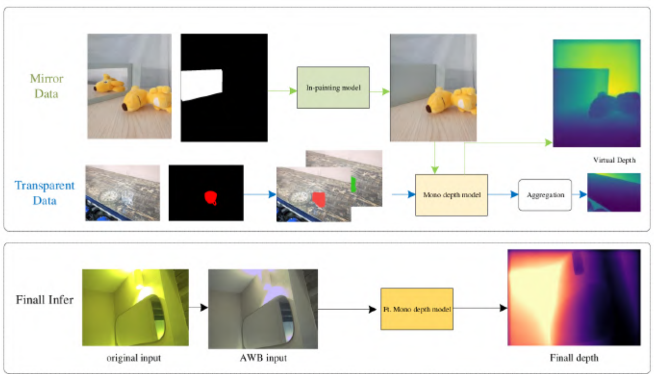

Figure 10. **Network Architecture – Team *MiMcAlgo [Mono]*.**

tions constrain the input image resolution during training, impacting the effectiveness of the results when upscaling to match the test set resolution. Overview in Fig. 8.

5.2.7 Team 6 - DVision

The DVision team's "Masked-Depth-Anything" method addresses the challenge of training the Depth Anything [93] model with the provided dataset, downsampling the high-resolution images to (3, 518, 714) for compatibility with the model. To minimize information loss from ToM surfaces during resizing, they divide the images into smaller sections of size (3, 1400, 1400) with a 20% overlap. The team focuses on retraining Depth Anything with the segments containing ToM surfaces and introduces a MaskLoss function to prioritize the model's attention on these surfaces. The MaskLoss function computes the mean squared error between the predicted and actual depth values for pixels corresponding to the ToM mask. However, using MaskLoss alone results in good predictions on ToM surfaces but poor performance on other surfaces. To address this issue, the team incorporates three additional loss functions with pre-assigned weights: SMLoss (Sobel filter-based edge loss), SSIMLoss (Structural Similarity Index Measure loss), and L1Loss (mean absolute error loss). The final loss function is a weighted combination of these four losses.

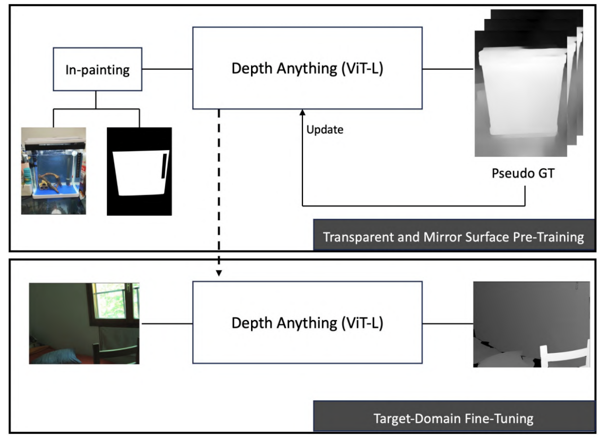

Figure 11. **Network Architecture – Team *UW IPL*.**

5.2.8 Team 7- MiMcAlgo [Mono]

The MiMcAlgo [Mono] team's (CodaLab: *Sunyj*) method uses the state-of-the-art Depth Anything-L [93] depth network as the base model. They adapt the fine-tuning method from [12] with additional improvements to address the challenging task. Here, mirror data is inpainted using the LaMa [74] model, while transparent data follows the default processing flow from [12]. The model is fine-tuned with the official weights on the MSD and Trans 10K test sets using specific hyperparameters and data augmentation techniques. A gray-world algorithm, horizontal flipping, and averaging are applied during inference to improve performance. A gamma coefficient of 0.5 is used to adapt the results to the test set's depth range. Overview in Fig. 10.

5.2.9 Team 8- UW IPL

The UW IPL team's "DepthanyTM" method builds upon the Depth Anything [93] model. The pipeline initializes the ViT-L model with pre-trained Depth Anything weights, then in-paints selected data from Trans10K and MSD following the strategy proposed in [12]. The team fine-tunes the model on the in-painted data using pseudo-ground-truth depth and further fine-tunes it on the Booster training set. The encoder is initialized with Depth Anything weights pre-trained on NYUv2, while the decoder is randomly initialized. The ZoeDepth codebase is used to predict metric depth, and the model is fine-tuned for 5 epochs at different stages using default parameters. Overview in Fig. 11.

Acknowledgements

This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2024 sponsors: Meta Reality Labs, OPPO, KuaiShou, Huawei and University of Würzburg (Computer Vision Lab).

A. NTIRE 2024 Organizers

Title:
NTIRE 2024 Challenge on HR Depth from Images of Specular and Transparent Surfaces
Members:
Pierluigi Zama Ramirez[1] (pierluigi.zama@unibo.it), Alex Costanzino[1], Fabio Tosi[1], Matteo Poggi[1], Samuele Salti[1], Stefano Mattoccia[1], Luigi Di Stefano[1], Radu Timofte[2]
Affiliations:
[1] University of Bologna, Italy
[2] Computer Vision Lab, University of Würzburg, Germany

B. Track 1: Teams and Affiliations

MiMcAlgo [Stereo]

Members:
Yangyang Zhang[1] (zhangyangyang@xiaomi.com), Cailin Wu[1], Zhuangda He[1], Shuangshuang Yin[1], Jiaxu Dong[1], Yangchenxu Liu[1], Hao Jiang[1]
Affiliations:
[1] Xiaomi Inc., China

Samsung R&D Institute China-Beijing (SRC-B)

Members:
Jun Shi[1] (jun7.shi@samsung.com), Yong A[1], Yixiang Jin[1], Dingzhe Li[1]
Affiliations:
[1] Samsung R&D Institute China-Beijing (SRC-B)

C. Track 2: Teams and Affiliations

Marigold-LCM

Members:
Bingxin Ke[1] (bingxin.ke@geod.baug.ethz.ch), Anton Obukhov[1], Tianfu Wang[1], Nando Metzger[1], Shengyu Huang[1], Konrad Schindler[1]
Affiliations:
[1] Photogrammetry and Remote Sensing, ETH Zürich

Samsung R&D Institute China-Beijing (SRC-B)

Members:
Jun Shi[1] (jun7.shi@samsung.com), Yong A[1], Yixiang Jin[1], Dingzhe Li[1]
Affiliations:
[1] Samsung R&D Institute China-Beijing (SRC-B)

SmartLab

Members:
Yachuan Huang[1] (email: yachuan@hust.edu.cn), Jiaqi Li[1], Junrui Zhang[1], Yiran Wang[1], Zihao Huang[1], Tianqi Liu[1], Zhiguo Cao[1]
Affiliations:
National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China

THU-808

Members:
Pengzhi Li[1] (lpz21@mails.tsinghua.edu.cn), Jui-Lin Wang[1]
Affiliations:
[1]Tsinghua University, China

PD&HPC

Members:
Wenjie Zhu (zhuwenjie@nudt.edu.cn), Hui Geng, Yuxin Zhang, Long Lan, Kele Xu, Tao Sun, Qisheng Xu
Affiliations:
National University of Denfense Technology, Changsha, China

DVision

Members:
Sourav Saini (2021ucs0118@iitjammu.ac.in), Aashray Gupta, Sahaj K. Mistry, Aryan Shukla, Vinit Jakhetiya, Sunil Jaiswal
Affiliations:
Indian Institute of Technology Jammu, India
K-Lens GmbH

MiMcAlgo [Mono]

Members:
Yuejin Sun (sunyuejin@xiaomi.com), Zhuofan Zheng, Yi Ning, Hao Jiang
Affiliations:
Xiaomi Technology Co., Ltd

UW IPL

Members:
Jen-Hao Cheng (andyhci@uw.edu), Hou-I Liu, Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, Yi-Hao Peng,

Aishi Huang, Jenq-Neng Hwang
Affiliations:
University of Washington, National Yang Ming Chiao Tung University, Carnegie Mellon University, University of Illinois Urbana-Champaign

References

[1] Cosmin Ancuti, Codruta O Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[2] Nikola Banić, Egor Ershov, Artyom Panshin, Oleg Karasev, Sergey Korchagin, Shepelev Lev, Alexandr Startsev, Daniil Vladimirov, Ekaterina Zaychenkova, Dmitrii R Iarchuk, Maria Efimova, Radu Timofte, Arseniy Terekhin, et al. NTIRE 2024 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023.

[4] Nicolas Chahine, Marcos V. Conde, Sira Ferradans, Radu Timofte, et al. Deep portrait quality assessment. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018.

[6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Proc. NeurIPS*, 2016.

[7] Zheng Chen, Zongwei WU, Eduard Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. NTIRE 2024 challenge on image super-resolution (×4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[8] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2361–2379, 2019.

[9] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33, 2020.

[10] Jaehoon Choi, Dongki Jung, Yonghan Lee, Deokhwa Kim, Dinesh Manocha, and Donghwan Lee. Selfdeco: Self-supervised monocular depth completion in challenging indoor environments. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 467–474. IEEE, 2021.

[11] Marcos V. Conde, Florin-Alexandru Vasluianu, Radu Timofte, et al. Deep raw image super-resolution. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[12] Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *The IEEE International Conference on Computer Vision*, 2023. ICCV.

[13] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4384–4393, 2019.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proc. NeurIPS*, 2014.

[15] Adrien Gaidon, Greg Shakhnarovich, Rares Ambrus, Vitor Guizilini, Igor Vasiljevic, Matthew Walter, Sudeep Pillai, and Nick Kolkin. Dense depth for autonomous driving (DDAD) challenge (`https://sites.google.com/view/mono3d-workshop`), 2021.

[16] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010.

[17] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. CVPR*, 2017.

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proc. ICCV*, 2019.

[19] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12626–12637. Curran Associates, Inc., 2020.

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

[21] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proc. ECCV*, 2018.

[22] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019.

[23] Ruozhen He, Jiaying Lin, and Rynson WH Lau. Efficient mirror detection via multi-level heterogeneous learning. *arXiv preprint arXiv:2211.15644*, 2022.

[24] Andrey Ignatov, Grigory Malivenko, David Plowman, Samarth Shukla, and Radu Timofte. Fast and accurate single-image depth estimation on mobile devices, mobile ai 2021 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2545–2557, June 2021.

[25] Huaizu Jiang, Gustav Larsson, Michael Maire Greg Shakhnarovich, and Erik Learned-Miller. Self-

supervised relative depth learning for urban scene understanding. In *Proc. ECCV*, 2018.
[26] 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 (ICCV)*, pages 3318–3327, October 2023.
[27] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proc. CVPR*, 2020.
[28] 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 (CVPR)*, 2024.
[29] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
[30] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018.
[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
[32] Henrik Kretzschmar, Alex Liniger, Jose M. Alvarez, Yan Wang, Vincent Casser, Fisher Yu, Marco Pavone, Bo Li, Andreas Geiger, Peter Ondruska, Li Erran Li, Dragomir Angelov, John Leonard, and Luc Van Gool. Argoverse stereo competition (https://cvpr2022.wad.vision/), 2021, 2022.
[33] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.
[34] 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.
[35] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.
[36] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
[37] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021.
[38] Jie Liang, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Radu Timofte, Lei Zhang, et al. NTIRE 2024 restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.
[39] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
[40] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *arXiv preprint arXiv:2109.07547*, 2021.
[41] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.
[42] Xiaoning Liu, Zongwei WU, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.
[43] Jieming Lou, Weide Liu, Zhuo Chen, Fayao Liu, and Jun Cheng. Elfnet: Evidential local-global fusion for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17784–17793, 2023.
[44] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
[45] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
[46] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3687–3696, 2020.
[47] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[48] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain

Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9685–9694, 2021.

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[50] Jiahao Pang, Wenxiu Sun, Jimmy SJ. Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[52] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proc. CVPR*, 2020.

[53] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[54] Matteo Poggi and Fabio Tosi. Federated online adaptation for deep stereo. In *CVPR*, 2024.

[55] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[56] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *Proc. CVPR*, 2020.

[57] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, Samuele Salti, Stefano Mattoccia, Jun Shi, Dafeng Zhang, et al. Ntire 2023 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1384–1395, 2023.

[58] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021.

[59] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021.

[60] 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), 2022.

[61] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth NTIRE 2024 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[62] Tonmoy Saikia, Yassine Marrakchi, Arber Zela, Frank Hutter, and Thomas Brox. Autodispnet: Improving disparity estimation with automl. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1812–1823, 2019.

[63] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642. IEEE, 2020.

[64] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Depth perception from a single still image. In *Proc. AAAI*, 2008.

[65] 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.

[66] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269. IEEE, 2017.

[67] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, June 2021.

[68] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *ACCV*, 2018.

[69] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[70] Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, Andrew J. Schofield, James H. Elder, Richard Bowden, Heng Cong, Stefano Mattoccia, Matteo Poggi, Zeeshan Khan Suri, Yang Tang, Fabio Tosi, Hao Wang, Youmin Zhang, Yusheng Zhang, and Chaoqiang Zhao. The monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 623–632, January 2023.

[71] Jaime Spencer, C. Stella Qian, Michaela Trescakova, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, Andrew J. Schofield, James Elder, Richard Bowden, Ali Anwar, Hao Chen, Xiaozhi Chen, Kai Cheng, Yuchao Dai, Huynh Thai Hoa, Sadat Hossain, Jianmian Huang, Mo-

han Jing, Bo Li, Chao Li, Baojun Li, Zhiwen Liu, Stefano Mattoccia, Siegfried Mercelis, Myungwoo Nam, Matteo Poggi, Xiaohua Qi, Jiahui Ren, Yang Tang, Fabio Tosi, Linh Trinh, S M Nadim Uddin, Khan Muhammad Umair, Kaixuan Wang, Yufei Wang, Yixing Wang, Mochu Xiang, Guangkai Xu, Wei Yin, Jun Yu, Qi Zhang, and Chaoqiang Zhao. The second monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[72] Jaime Spencer, Fabio Tosi, Matteo Poggi, Ripudaman Singh Arora, Chris Russell, Simon Hadfield, Richard Bowden, GuangYuan Zhou, ZhengXin Li, Qiang Rao, YiPing Bao, Xiao Liu, Dohyeong Kim, Jinseong Kim, Myunghyun Kim, Mykola Lavreniuk, Rui Li, Qing Mao, Jiang Wu, Yu Zhu, Jinqiu Sun, Yanning Zhang, Suraj Patni, Aradhye Agarwal, Chetan Arora, Pihai Sun, Kui Jiang, Gang Wu, Jian Liu, Xianming Liu, Junjun Jiang, Xidan Zhang, Jianing Wei, Fangjun Wang, Zhiming Tan, Jiabao Wang, Albert Luginov, Muhammad Shahzad, Seyed Hosseini, Aleksander Trajcevski, and James H. Elder. The third monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[73] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021.

[74] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[75] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14362–14372, June 2021.

[76] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[77] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019.

[78] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[79] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[80] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[81] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 855–866, June 2023.

[82] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei WU, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[83] Longguang Wang, Yulan Guo, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, Radu Timofte, et al. NTIRE 2024 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[84] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. CLIFFNet for monocular depth estimation with hierarchical embedding loss. In *Proc. ECCV*, 2020.

[85] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900, 2019.

[86] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2024 challenge on light field image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[87] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proc. ICCV*, 2019.

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

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

[90] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[92] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, pages 636–651, 2018.

[93] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*,

2024.

[94] Ren Yang, Radu Timofte, et al. NTIRE 2024 challenge on blind enhancement of compressed image: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[95] 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.

[96] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

[97] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6044–6053, 2019.

[98] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Booster: a benchmark for depth from images of specular and transparent surfaces. *arXiv preprint arXiv:2301.08245*, 2023.

[99] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2024 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[100] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21168–21178, June 2022.

[101] Oliver Zendel, Angela Dai, Xavier Puig Fernandez, Andreas Geiger, Vladen Koltun, Peter Kontschieder, Adam Kortylewski, Tsung-Yi Lin, Torsten Sattler, Daniel Scharstein, Hendrik Schilling, Jonas Uhrig, and Jonas Wulff. The robust vision challenge (http://www.robustvision.net/), 2018, 2020, 2022.

[102] Jiaxi Zeng, Chengtang Yao, Lidong Yu, Yuwei Wu, and Yunde Jia. Parameterized cost volume for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18347–18357, October 2023.

[103] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[104] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. NTIRE 2024 challenge on bracketing image restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[105] Chaoqiang Zhao, Matteo Poggi, Fabio Tosi, Lei Zhou, Qiyu Sun, Yang Tang, and Stefano Mattoccia. Gasmono: Geometry-aided self-supervised monocular depth estimation for indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16209–16220, 2023.

[106] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.