NTIRE 2023 Challenge on HR Depth From Images of Specular and Transparent Surfaces

Published Version:
NTIRE 2023 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, Jun Shi, Dafeng Zhang, Yong A, Yixiang Jin, Dingzhe Li, Chao Li, Zhiwen Liu, Qi Zhang, Yixing Wang, Shi Yin. - ELETTRONICO. - (2023), pp. 1384-1395. (Intervento presentato al convegno IEEE/CVF Conference on Computer Vision and Pattern Recognition tenutosi a Vancouver, Canada nel 17-24 June 2023) [10.1109/CVPRW59228.2023.00143].

Availability:
This version is available at: https://hdl.handle.net/11585/955889 since: 2024-02-06

Published:
DOI: http://doi.org/10.1109/CVPRW59228.2023.00143

NTIRE 2023 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
Jun Shi Dafeng Zhang Yong A Yixiang Jin Dingzhe Li
Chao Li Zhiwen Liu Qi Zhang Yixing Wang Shi Yin

Abstract

This paper reports about the NTIRE 2023 challenge on HR Depth From images of Specular and Transparent surfaces, held in conjunction with the New Trends in Image Restoration and Enhancement workshop (NTIRE) workshop at CVPR 2023. This challenge is held to boost the research on depth estimation, mainly to deal with two of the open issues in the field: high-resolution images and non-Lambertian surfaces characterizing specular and transparent materials. The challenge is divided into two tracks: a stereo track focusing on disparity estimation from rectified pairs and a mono track dealing with single-image depth estimation. The challenge attracted about 100 registered participants for the two tracks. In the final testing stage, 5 participating teams submitted their models and fact sheets, 2 and 3 for the Stereo and Mono tracks, respectively.

1. Introduction

Since the advent of computer vision, estimating depth from images has always been the object of study for a large part of the research community. Indeed, recovering depth represents the first pivotal step to pave the way to several downstream applications, ranging from augmented reality, robotics, autonomous navigation, and more. Depth can be measured either by means of dedicated, active sensors – LiDARs, ToFs, Radars, etc. – or through standard imaging sensors by developing algorithms / deep neural networks. Although depth sensing technologies grew fast in the last decade and proved a mature reality, some challenges still preclude their unbound deployment.

Among them, one is resolution. Indeed, on the one hand, active depth sensors usually provide sparse depth measurements, rarely reaching 1 Megapixel (Mpx); on the other hand, although standard cameras feature resolutions up to dozens of Mpx, processing them with deep neural networks requires significant computational efforts.

Another one is represented by *non-Lambertian* materials, which are, again, challenging for active sensors and image-based techniques. Indeed, they often break the assumptions behind the working principles of most depth sensing techniques, both in the case of active sensors – e.g., the refraction of a light beam emitted by a LiDAR, or its projection on an object behind a transparent surface – and image-based approaches – e.g., stereo algorithms would fail to estimate depth for a transparent object, since matches would be found for the content behind it. Nonetheless, in several practical applications, it is crucial to properly estimate the *correct* depth for these materials too – e.g., a grasping arm dealing with transparent objects would fail into manipulating them if not equipped with a depth perception technologies being not appropriate to deal with them.

This NTIRE 2023 Challenge on HR Depth from Images of Specular and Transparent Surfaces aims at pushing forward the development of state-of-the-art solutions for depth estimation that can effectively deal with the aforementioned challenges. Purposely, we employ the Booster dataset [76, 78] in this challenge, which is the only benchmark implementing proving grounds for both, featuring 12Mpx images with several transparent and reflective materials. The challenge is organized into two tracks: one focusing on *Stereo* approaches, estimating depth as the *disparity* between pixels into two, rectified stereo images, and the other aimed at assessing the accuracy of single-image depth estimation techniques (*Mono*). The challenge has 49 and 51 registered participants for two tracks, respectively. Among them, 2 and 3 participating teams submitted their models and fact sheets during the final testing stage, respectively. Some adopt off-the-shelf, existing solutions, while others combine different methodologies and exploit their synergy

*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 2023 HR Depth from Images of Specular and Transparent Surfaces challenge organizers. The other authors participated in the challenge.
Appendices B and C contain the authors' team names and affiliations.
The NTIRE website: https://cvlai.net/ntire/2023/

to obtain better results. The outcome of this challenge is discussed in detail in Section 4.

2. Related Work

This section introduces the literature relevant to stereo and monocular depth estimation.

Deep Stereo Matching. Deep stereo-matching networks that can perform end-to-end processing have emerged as the most popular and effective solution for estimating disparity. These networks can be classified into two categories: 2D and 3D architectures. The first category is promoted by DispNet [39], which has inspired more advanced deep architectures [35, 42, 49, 56, 60, 62, 73, 75]. On the other hand, GC-Net [27] pioneered the use of an explicit 3D feature cost volume that employs feature concatenation or difference. More recent networks have been developed based on this approach [6, 8, 9, 13, 22, 28, 54, 68, 72, 80]. Recently, novel deep stereo networks have taken inspiration from the state-of-the-art optical flow network RAFT [61] to design architectures that can iteratively refine their outputs for the stereo matching task [31, 36]. Alternatively, some networks employ Transformers [20, 34] to capture long-range contextual information that can help improve disparity predictions in challenging regions. Despite their success, deep learning-based stereo methods rely heavily on expensive and hard-to-source ground-truth depth labels for training. These methods perform at their best when a large amount of annotated data is available. Indeed, the availability of various benchmarks for training and evaluation facilitates the rapid evolution of stereo algorithms. In the beginning, datasets were restricted to controlled and indoor environments, and they were composed of only a few dozen samples. However, in the last decade, more comprehensive stereo datasets have emerged, such as KITTI 2012 [16] and 2015 [40], Middlebury 2014 [52], and ETH3D [53]. The high accuracy of state-of-the-art stereo networks on these datasets suggests that most of the challenges they present are nearly addressed. Nevertheless, the latest stereo datasets do not specifically focus on the most arduous open challenges for stereo matching, which are found in the Booster [78] dataset. In this challenge, we rely on this dataset that emphasizes several specular and transparent surfaces, the primary causes of failure in state-of-the-art stereo networks.

Monocular Depth Estimation. The monocular depth estimation task was initially accomplished using handcrafted features that encode perceptual cues such as texture gradient, object size, and linear perspective, which are vital for determining depth. These cues were the basis of early research in the field [51]. However, the development of deep learning has led to significant advancements in this area, allowing for the direct learning of depth-related priors from annotated data [7, 14, 30, 44, 67]. This research trend has been able to progress rapidly due to the availability of large-scale datasets with associated ground-truth depth labels [7, 14, 30, 44, 67], as well as the implementation of self-supervised strategies [17–19, 21, 24, 25, 43, 63, 64, 70, 82] to address the lack of annotations. These latter strategies exploit either stereo pairs or monocular videos, and the predicted depth is combined with known or estimated camera pose, respectively, to establish correspondences between adjacent images. Other approaches, such as AdaBins [2], DPT [45], and MiDaS [47] use adaptive bins and vision transformers for depth regression and leverage large-scale depth training by mixing multiple datasets. Nonetheless, the projection of depth maps into 3D space results in deformed point clouds, which has been effectively addressed by Yin *et al*. [74]. Furthermore, restoring high-frequency details in estimated depth maps for high-resolution images continues to be a challenge. To address this issue, Miangoleh *et al*. [41] have developed a framework that modifies the input of a pre-trained monocular network and merges multiple estimations.

However, in the monocular depth estimation literature little attention has been given to single-view depth estimation networks that can handle transparent and reflective surfaces due to the scarcity of datasets specifically suited for this task. Only recently, Booster [76] has been introduced, which features some very challenging yet accurately annotated non-Lambertian objects and images at much higher resolutions. Finally, few works have faced non-Lambertian depth estimation but using depth completion approaches and sparse depth measurements from active sensors [10, 50].

Competitions and Challenges on Depth Estimation. Finally, it is worth mentioning some past challenges focusing on depth perception from stereo or monocular images. Among them, the Robust Vision Challenge (ROB) [79] embracing both tasks, the Dense Depth for Autonomous Driving challenge (DDAD) [15], the Fast and Accurate Single-Image Depth Estimation on Mobile Devices Challenge (MAI) [23], the Argoverse Stereo Challenge [29] and the Monocular Depth Estimation Challenge (MDEC) [57, 58]. Despite the interest in this task, ours is the first challenge focusing on specular and transparent surfaces.

NTIRE 2023 Challenges. Our challenge is one of the NTIRE 2023 Workshop [1] series of challenges on: night photography rendering [55], HR depth from images of specular and transparent surfaces [77], image denoising [33], video colorization [26], shadow removal [65], quality assessment of video enhancement [37], stereo super-resolution [66], light field image super-resolution [69], image super-resolution (×4) [81], 360° omnidirectional image and video super-resolution [5], lens-to-lens bokeh effect transformation [11], real-time 4K super-resolution [12], HR nonhomogenous dehazing [1], efficient super-resolution [32].

[1] https://cvlai.net/ntire/2023/

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

We host the NTIRE 2023 Challenge on HR Depth from Images of Specular and Transparent Surfaces to boost the accuracy of state-of-the-art solutions for depth perception and make them capable of handling high-resolution images, as well as dealing with challenging, non-Lambertian surfaces such as mirrors, glasses and so on. We now report the main details of the challenge.

Tracks. We include two tracks: *Stereo*, dealing with disparity estimation from rectified image pairs, and *Mono*, focusing on single-image depth estimation architectures.

- **Track 1: Stereo.** The goal of this track consists of obtaining high-quality, high-resolution disparity maps from 12Mpx stereo pairs. The main difficulties are the image resolution, prohibitive for most state-of-the-art existing stereo networks, and the presence of non-Lambertian objects, making the correspondence matching problem challenging.?
- **Track 2: Mono.** The goal of this track consists of estimating a depth map out of a single 12Mpx image. This task is more challenging than stereo depth estimation because of the inherent ill-posed nature of the problem. Moreover, the presence of several transparent objects and mirrors – being out-of-distribution elements in most depth estimation datasets – makes it even more challenging.

Datasets. The challenge is built over the Booster dataset [76, 78]. It consists of 419 high-resolution balanced and unbalanced stereo pairs, featuring 64 different scenes and respectively divided into 228 and 191 pairs for training and testing purposes – dividing the total number of scenes into 38 and 26. Booster has been recently extended [76] by the release of a second testing split devoted to the evaluation of monocular depth estimation methods and made of 187 single frames collected from 21 new scenes.

For this challenge, we adopt the original 228 training stereo pair as the *training split*, shared among the two tracks. Then, 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, resulting in 15 validation samples for each track, out of the total 26 and 28 available from the selected scenes. A visualization of the validation split is shown in Fig. 1. The remaining images of the two original testing splits are then retained as official stereo and mono *testing splits* for this challenge, resulting in 169 and 159 samples, respectively.

Evaluation Protocol. According to the specific track, Stereo or Mono, we select the official metrics used by

Figure 1. Validation scenes. Three scenes were used to validate methods for each track. Five different illuminations were available for each scene.

the Booster benchmark [76, 78]. For the Stereo track, we compute the percentage of pixels having disparity errors larger than a threshold τ (bad-τ, with $\tau \in [2, 4, 6, 8]$), as well as we measure the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). For the Mono Track, we compute the absolute error relative to the ground value (Abs Rel.), and the percentage of pixels having the maximum between the prediction/ground-truth and ground-truth/prediction ratios lower than a threshold (δ_i, with i being 1.05, 1.15, and 1.25). Also in this case, we estimate the mean absolute error (MAE), and Root Mean Squared Error (RMSE). For both tracks, any metric is computed on any valid pixel (*All*), or in the alternative, on pixels belonging to a specific material class i (*Class* i), to evaluate the impact of non-Lambertian objects. To rank submissions, we use only MAE and Abs. Rel – respectively for Stereo and Mono tracks – averaged over all pixels, highlighted in **red** in the tables. However, monocular networks estimate depth up to an unknown scale and shift factors. Thus, given a monocular depth prediction, $\hat{d}$, before computing metrics, we modulate it as $\alpha\hat{d} + \beta$, with α, β being a scale and shift factor. Following [48], α, β are estimated with Least Square Estimation (LSE) regression over the ground truth depth map d:

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

where p are the pixel locations of the depth maps.

4. Challenge Results

For the two tracks, 2 and 3 teams participated in the final testing phase respectively. Tables 1 and 2 report the main results and important information for these teams. The methods for stereo and mono tracks are briefly described in Section 5.1 and Section 5.2, while the team members are listed in Appendix B and Appendix C for the two tracks, respectively.

		All						Class 0		Class 1		Class 2		Class 3	
Rank	Team	MAE	RMSE	bad-2	bad-4	bad-6	bad-8	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
#2	RAFT-Stereo (ft) [78]	7.08	16.09	38.89	23.53	17.88	14.74	4.64	10.80	5.45	12.23	15.27	21.05	11.13	17.18
#3	Chengzhi-Group	21.21	42.03	38.64	28.96	25.51	23.37	7.86	17.39	11.43	21.98	52.88	61.87	42.01	54.07
#1	SRC-B	6.07	14.38	32.43	20.82	16.30	13.89	4.99	11.25	3.75	9.64	4.67	8.25	10.58	15.43

Table 1. **Stereo Track: Evaluation on the Challenge Test Set.** Predictions were evaluated at full resolution (4112×3008), on All pixels and on pixels belonging to classes from 0 to 3. Classes are ordered in an increasing level of difficulty, e.g., class 3 pixels belong to transparent and mirror surfaces. In **gold**, **silver**, and **bronze** we show first, second, and third-rank approaches, respectively.

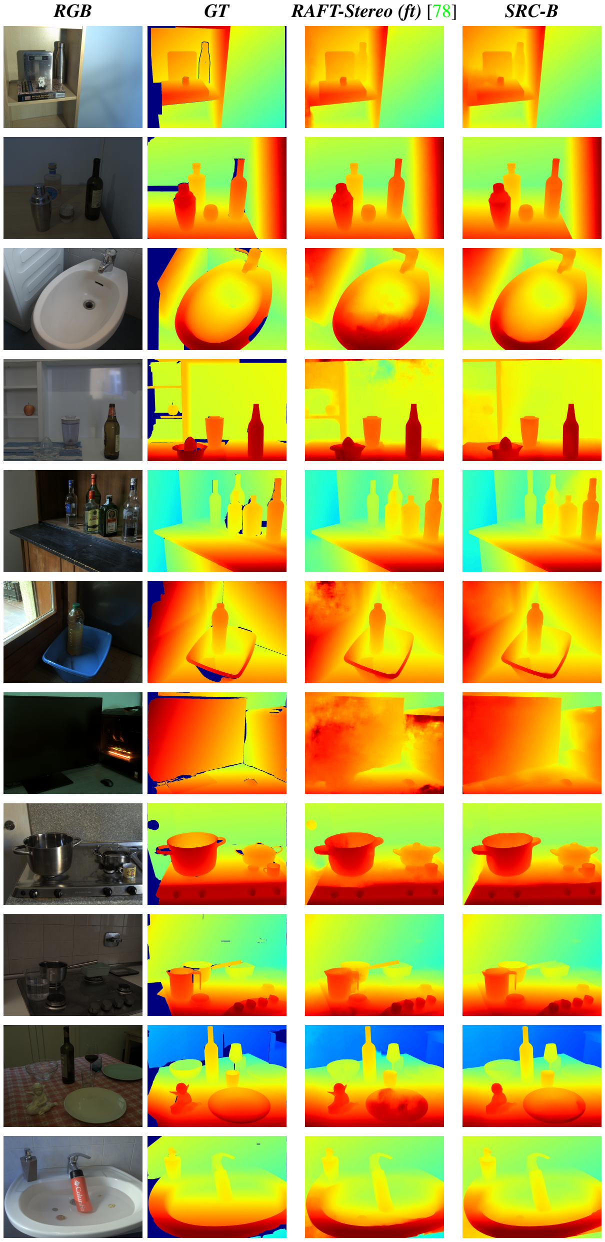

Figure 2. **Qualitative results – Stereo track.** From left to right: RGB reference image, ground-truth disparity, predictions by RAFT-Stereo (ft) [78] and the network proposed by ***SRC-B*** group.

4.1. Track 1: Stereo

Table 1 collects the results for this first track. In the first entry on top, we report the baseline method – i.e., the very same RAFT-Stereo [36] model fine-tuned on the Booster training split and reported in [78]. For the sake of space, we report bad metrics for All pixels only, while MAE and RMSE are shown for All pixels, as well as for the single classes of materials from 0 to 3. We can notice that one of the two methods failed to beat the baseline and achieved regularly worse results on any metric.

On the contrary, the other participant group was able to consistently outperform the RAFT-Stereo model fine-tuned on the Booster training set – thus winning this track of the challenge – by dropping overall MAE and RMSE by 1 and 4 points and bad metrics by about 6, 3, 1.5, and 1% respectively. More specifically, we can notice that the improvements come at the price of slightly higher MAE and RMSE for class 0 regions, which paves the way to a significant boost in class 1 (1.5 and 2.6), a dramatic improvement in class 2 (10.6 and 12.8) and a moderate boost in class 3 too (0.6 and 1.7). Fig. 2 shows some qualitative results taken from the stereo testing set: we can appreciate how the baseline (third column) sometimes generates noisy disparities, as shown in rows 4, 6, and 7, whereas the winning method provides smoother results (fourth column). Nonetheless, we highlight how some very challenging cases remain unsolved, as in the case of the water surface on the bottom-most row.

4.2. Track 2: Mono

Table 2 shows the results for the second track. The first entry on top reports the results by the baseline method – i.e., the DPT [46] model, fine-tuned on the Booster training split as detailed in [76]. For the sake of space, we report RMSE and δ metrics for All pixels only, while Abs. Rel and MAE are shown for All pixels, as well as for the single classes of materials from 0 to 3. Again, one of the methods failed to beat the baseline, not reaching its performance on any of the considered metrics.

As for the remaining methods, both were able to beat the DPT model consistently. For what concerns the top #2 method, it manages to reduce the error metrics on All pixels by about 0.3, 0.28, and 0.03, respectively on Abs. Rel, MAE, and RMSE, with average increases on the δ met-

		All						Class 0		Class 1		Class 2		Class 3	
Rank	Team	**Abs. Rel**	MAE	RMSE	$\delta < 1.05$	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	Abs Rel.	MAE	Abs Rel.	MAE	Abs Rel.	MAE
#3	DPT (ft) [76]	**0.1458**	**0.1596**	**0.2075**	**29.52**	**60.38**	**77.97**	**0.1557**	**0.1696**	**0.1834**	**0.1756**	**0.1741**	**0.1554**	**0.1850**	**0.1660**
#4	lillian	**0.1607**	**0.1787**	**0.2251**	**29.20**	**57.12**	**73.84**	**0.1696**	**0.1820**	**0.2027**	**0.1891**	**0.1786**	**0.1560**	**0.1901**	**0.1689**
#2	yshk	**0.1162**	**0.1319**	**0.1752**	**37.05**	**70.24**	**84.73**	**0.1185**	**0.1363**	**0.1193**	**0.1177**	**0.1533**	**0.1362**	**0.1423**	**0.1279**
#1	cv_challenge	**0.0738**	**0.0858**	**0.1187**	**52.23**	**85.44**	**93.58**	**0.0730**	**0.0840**	**0.0812**	**0.0898**	**0.0798**	**0.0693**	**0.1018**	**0.0926**

Table 2. Mono Track: Evaluation on the Challenge Test Set. Predictions were evaluated at full resolution (4112×3008), on All pixels, and on pixels belonging to classes from 0 to 3. Classes are ordered in an increasing level of difficulty, e.g., class 3 pixels belong to transparent and mirror surfaces. In **gold**, **silver**, and **bronze** we show first, second, and third-rank approaches, respectively.

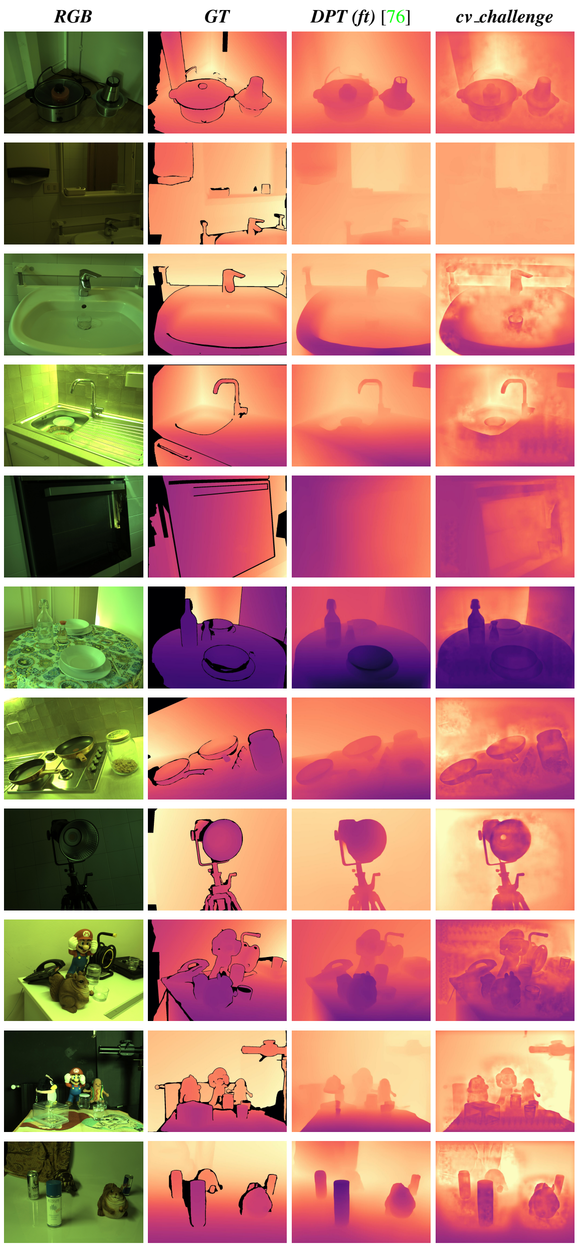

Figure 3. **Qualitative results – Mono track.** From left to right: RGB reference image, ground-truth disparity, predictions by DPT (ft) [76] and the network proposed by ***cv_challenge*** group.

rics of 18, 10, and 7 points. The drops in the error metrics are consistent on any of the 4 classes, with reductions in Abs Rel./MAE of about 0.04/0.03, 0.04/0.03, 0.02/0.02, and 0.04/0.04 for classes from 0 to 3.

Finally, the winning method achieves a substantial improvement over the baseline by reducing any error metric to half in most cases. Fig. 3 shows some qualitative examples from the mono testing set: although the baseline apparently produces smoother depth maps, its accuracy results to be, on average, inferior to the one of the winning method.

5. Challenge Methods

We now describe each submitted solution in detail.

5.1. Track 1: Stereo

5.1.1 Baseline - RAFT-Stereo (ft) [78]

Our baseline for the Stereo track is the state-of-the-art RAFT-Stereo architecture [36], a recent method for two-view stereo based on the original RAFT optical flow framework [61]. Specifically, RAFT-Stereo first extracts features from the left and right input images and then builds a 3D cost volume by computing the similarity between pixels of the same height in the images. The architecture then uses multi-level GRU units to update the disparity field and improve its global consistency iteratively. In our experiments, we use the available model trained by the authors and fine-tune it on the Booster training set augmented with additional images from the Middlebury 2014 dataset. Specifically, following the training protocol described in [76, 78], we run 100 training epochs on image crops of size 884456 randomly extracted from images resized to half or quarter of the original resolution. This strategy allows the network, referred to as RAFT-Stereo (ft), to compensate for most errors due to non-Lambertian surfaces and better handle specular and transparent objects in the scene.

5.1.2 Team 1 - SRC-B)

The team Samsung Research China - Beijing (SRC-B) (CodaLab: xiaozhazha) proposed an architecture to address the challenges of accurate depth estimation in high-resolution images with non-Lambertian surfaces consisting of two main stages, which are shown in Fig 4.

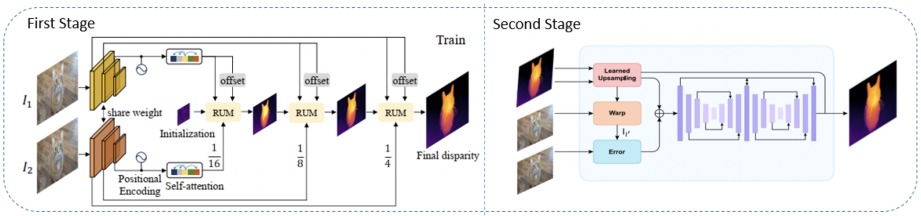

Figure 4. **Network Architecture – Team *Samsung Research China - Beijing (SRC-B)*.**

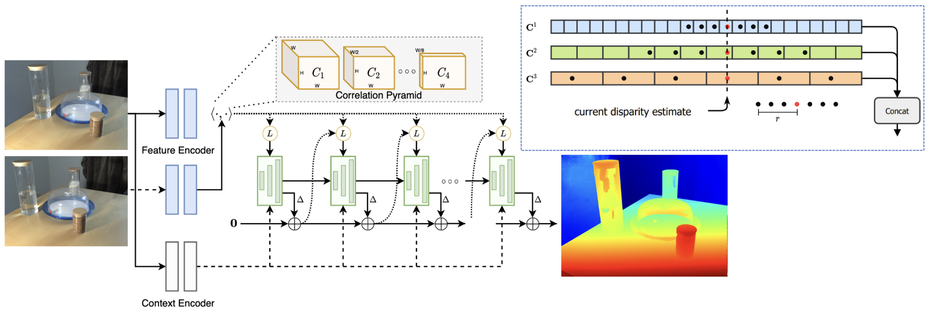

Figure 5. **Network Architecture – Team *Chengzhi-Group*.**

In the first stage of the network, they adopt the CREStereo [31] approach, which uses a hierarchical network to predict disparities in a coarse-to-fine manner. This approach employs several techniques, such as an adaptive group local correlation layer that uses cross and self-attention 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 to improve performance.

In the second stage, they employ an error-aware refinement module based on left-right warping. This module leverages high-frequency information from the original left image and error maps to correct estimation errors caused by the smooth prediction in the first stage.

The proposed network is implemented using the Pytorch framework and trained on 2 v100 GPUs with a batch size of 8. They use Adam optimizer with a standard learning rate of 0.0004. In the first stage, the training process is set to 102,600 iterations. They fine-tune the CREStereo module on the Booster training dataset with a pre-trained model obtained from [31]. Following this, they fix the weights of the CREStereo module and fine-tuned the error-aware module for an additional 57,000 iterations in the second stage.

During the training phase, they apply several augmentation techniques, including random scaling, cropping, chromatic augmentation, and random occlusions, to the training samples. These techniques help to improve the robustness and generalization of the proposed method.

During the inference phase, they use a stacked cascaded architecture to handle high-resolution image inputs. They first downsample the image pair to $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ to construct an image pyramid that is then fed into the network. This helps to capture both the fine and coarse details of the input images and improves the accuracy of the depth estimation.

5.1.3 Team 2 – Chengzhi-Group

The team Chengzhi-Group (CodaLab: chengzhi) uses an off-the-shelf stereo network to participate in the challenge. Specifically, they deploy RAFT-Stereo [36], whose architecture is sketched in Fig. 5, with the weights officially released by the authors on github (*raft-middlebury.pth* model). According to [36], the model has been trained on synthetic data for 200k steps, with a batch size of 8 360×720 crops, and with 22 updates of the disparity estimates, by using a one-cycle learning rate schedule with a minimum learning rate of 1e-4. As data augmentation, the image saturation was adjusted between 0 and 1.4, the right image was shifted vertically to simulate imperfect rectification that is common in datasets such as ETH3D and Middlebury, and image/disparities have been stretched by random factors in $[2^{-0.2}, 2^{-0.4}]$ to simulate a range of possible disparity distributions. After training on synthetic data, the model has been fine-tuned on 384×1000 random crops of the 23 Middlebury 2014 training images for 4000 steps, with a batch size of 2 and 22 update iterations. Inference is performed at half-resolution, using 32 update iterations.

5.2. Track 2: Mono

5.2.1 Baseline - DPT (ft) [76]

For the Mono track, we adopt the DPT architecture as the baseline, which represents the state-of-the-art network for the monocular depth estimation task. Specifically, the DPT architecture relies on an encoder-decoder structure that

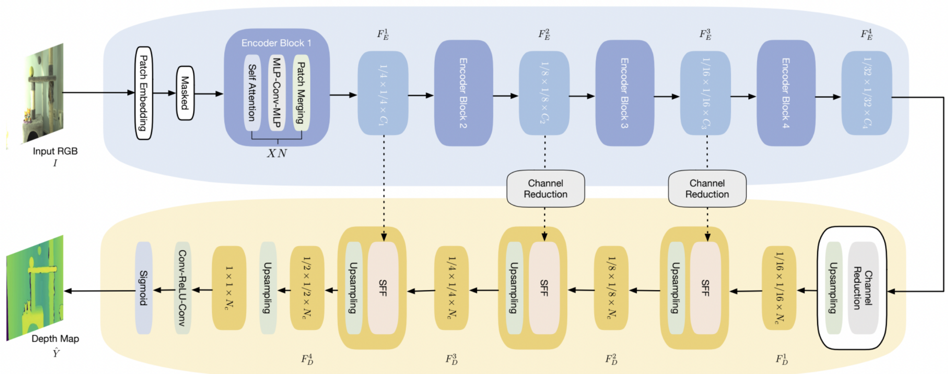

Figure 6. **Network Architecture – Team *lillian*.**

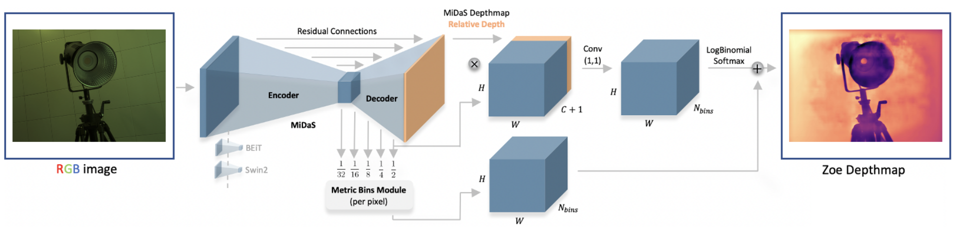

Figure 7. **Network Architecture – Team *cv_challenge*.**

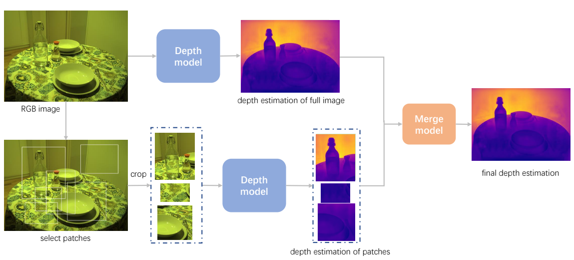

Figure 8. **Boosting Strategy – Team *cv_challenge*.**

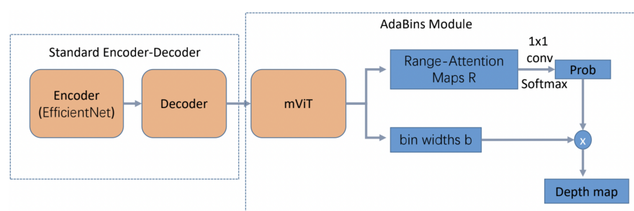

Figure 9. **Network Architecture – Team *yshk*.**

leverages a vision transformer (ViT) as a building block for the encoder. This allows the network to avoid explicit downsampling operations, which are typical of standard fully-convolutional networks and ensures a representation with constant dimensionality throughout all processing stages, as well as maintaining a global receptive field. Similar to the Stereo track, we use the available weights provided by the authors and fine-tune the network on the training images of the Booster dataset. Following [76], the fine-tuning process involves running 50 epochs on batches of random 2878×2105 crops, which are further resized to network resolution (384×384) and extracted from randomly horizontally flipped and color-jittered images.

5.2.2 Team 1 - lillian)

Team lillian (CodaLab: lillian) employs SimMIM [71] as framework with SwinV2-B [38] as backbone, as shown in Fig. 6. They rescale the depth range of Booster to the one of NYU dataset and perform several data augmentations, such as horizontal and vertical flip, random crop, random brightness, random contrast, random gamma, random hue saturation, RGB shift, random sun flare, Gaussian noise and Gaussian Blur. By utilizing these techniques, the performance of the model on the Booster dataset improved significantly, resulting in converging to a lower absolute relative error. Additionally, they use the sigmoid function that results in a wider range of depth values, which in turn can facilitate the convergence of the depth estimation model. The input of the sigmoid is scaled by a constant factor to obtain faster convergence.

5.2.3 Team 2 - cv_challenge)

The cv_challenge team (CodaLab: cv_challenge) takes advantage of the ZoeDepth model [4] (shown in Fig. 7), employing the NYU Depth v2 dataset and part of indoor images in DIODE dataset to train the model. To improve the detail of the inferred depth maps, they combine the

ZoeDepth with a content-adaptive multi-resolution merging algorithms [41], selecting patches from the input image and feeding them to the model using resolutions adaptive to the local depth cue density. Such patch-based estimates are then merged into the full-image estimation, making the depth prediction contain more high-frequency details, as depicted in Fig. 8. However, unlike in [41], they do not apply multi-resolution for each full-image estimation or each patch estimation, improving the efficiency of the model significantly. In fact, since the ZoeDepth employed is transformer-based and can draw information from the whole image, capturing long-range dependencies effectively, the preliminary experiments performed by the team let them conclude that directly omitting the multi-resolution merging does not hurt the depth result distinctly.

5.2.4 Team 3 - yshk)

Team yshk (CodaLab: wyx0821) takes advantage of AdaBins [3] to estimate the depth values. The overall architecture is sketched in Fig. 9 and mainly contains two components: a standard Encoder-Decoder block and the AdaBins Module. The encoder is based on a pre-trained EfficientNet B5 [59] model and the decoder is feature upsampling. The Adabins Module takes the output of the decoder as input and produces the depth image. The most important part of Adabins Module is the mViT block, which outputs the bin widths b and the Range-Attention Maps R. The former is estimated adaptively for each image and defines how the depth interval is divided, the latter is obtained as the dot product between pixel-wise features and transformer output embeddings. Finally, R and b are combined to calculate the depth map. R is handled by a 1×1 Conv to obtain N-channels, that are projected into probabilities over N classes by a Softmax operation. For each pixel, its depth value is obtained by the linear combination of Softmax scores and the depth-bin-centers. The training is performed on the NYU v2 dataset.

Acknowledgments

We thank the NTIRE 2023 sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).
We thank the NTIRE 2023 Challenge on HR Depth from Images of Specular and Transparent Surfaces sponsor:
Eyecan.ai (https://www.eyecan.ai/)

A. NTIRE 2023 Organizers

Title:
NTIRE 2023 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[2,3]
Affiliations:
[1] University of Bologna, Italy
[2] Computer Vision Lab, University of Würzburg, Germany
[3] Computer Vision Lab, ETH Zürich, Switzerland

B. Track 1: Teams and Affiliations

Chengzhi-Group

Members:
Chengzhi Cao[1] (chengzhicao@mail.ustc.edu.cn), Fanrui Zhang[1], Qiang Zhan[1], Kunyu Wang[1]
Affiliations:
[1] University of Science and Technology of China

Samsung Research China - Beijing (SRC-B)

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

C. Track 2: Teams and Affiliations

cv_challenge

Members:
Chao Li[1] (lichao@vivo.com), Zhiwen Liu[1], Qi Zhang[1], Yixing Wang[1]
Affiliations:
[1] VIVO

lillian

Members:
Liangyan Li[1] (lil61@mcmaster.ca), Runchen Liang[1], Yangyi Liu[1], Huan Liu[1], Siyu Song[1], Jun Chen[1]
Affiliations:
[1] McMaster University

yshk

Members:
Shi Yin[1] (yinshi2021@njtech.edu.cn)
Affiliations:
[1] Nanjing Tech University

References

[1] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, June 2021.

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

[5] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[6] 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.

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

[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, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[12] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[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] 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.

[24] 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.

[25] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proc. CVPR*, 2020.

[26] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[33] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[34] 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.

[35] 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.

[36] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *arXiv preprint arXiv:2109.07547*, 2021.

[37] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[38] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022.

[39] 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.

[40] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] 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.

[42] 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.

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

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

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

[46] 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.

[47] 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 Trans. Pattern Anal. Mach. Intell.*, 2020.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

[55] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

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

[57] 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.

[58] 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, Mohan 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.

[59] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[66] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

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

[68] 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.

[69] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

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

[71] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling, 2022.

[72] 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.

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

[74] 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.

[75] 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.

[76] 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.

[77] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, 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 Workshops*, 2023.

[78] 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.

[79] 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.

[80] 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.

[81] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[82] 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.