This is the final peer-reviewed author’s accepted manuscript (postprint) of the following publication:

**Published Version:**

**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

**Terms of use:**
Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher’s website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/). When citing, please refer to the published version.

(Article begins on next page)
NTIRE 2023 Challenge on
HR Depth from Images of Specular and Transparent Surfaces

Pierluigi Zama Ramirez
Alex Costanzino
Jun Shi
Chao Li

Fabio Tosi
Matteo Poggi
Dafeng Zhang
Zhiwen Liu

Luigi Di Stefano
Samuele Salti
Yong A
Qi Zhang

Radu Timofte
Stefano Mattoccia
Yixiang Jin
Yixing Wang

Shi Yin

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.
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 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 hand-crafted 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 the Booster benchmark [76, 78]. For the Stereo track, we compute the percentage of pixels having disparity errors larger than a threshold $\tau$ (bad-$\tau$, 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 ($\delta_i$, 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, $d$, before computing metrics, we modulate it as $\alpha d + \beta$, with $\alpha, \beta$ being a scale and shift factor. Following [48], $\alpha, \beta$ are estimated with Least Square Estimation (LSE) regression over the ground truth depth map $d$:

$$
\begin{align*}
(\alpha, \beta) &= \arg \min_{\alpha, \beta} \sum_{p} \left( \alpha \hat{d}(p) + \beta - d(p) \right)^2
\end{align*}
$$

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.
### 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.

### 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 \( \delta \) 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 \( \delta \) met-

---

**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.

| Rank | Team               | All MAE | All RMSE | bad-2 MAE | bad-2 RMSE | bad-4 MAE | bad-4 RMSE | bad-6 MAE | bad-6 RMSE | bad-8 MAE | bad-8 RMSE | Class 0 MAE | Class 0 RMSE | Class 1 MAE | Class 1 RMSE | Class 2 MAE | Class 2 RMSE | Class 3 MAE | Class 3 RMSE |
|------|-------------------|---------|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 2    | RAFT-Stereo (ft)  | 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     | 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     |          |          |

*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.
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.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Abs. Rel</th>
<th>MAE</th>
<th>RMSE</th>
<th>δ &lt; 1.05</th>
<th>δ &lt; 1.15</th>
<th>δ &lt; 1.25</th>
<th>Abs. Rel</th>
<th>MAE</th>
<th>RMSE</th>
<th>δ &lt; 1.05</th>
<th>δ &lt; 1.15</th>
<th>δ &lt; 1.25</th>
<th>Abs. Rel</th>
<th>MAE</th>
<th>RMSE</th>
<th>δ &lt; 1.05</th>
<th>δ &lt; 1.15</th>
<th>δ &lt; 1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>03</td>
<td>DPT (ft) [76]</td>
<td>0.1458</td>
<td>0.1596</td>
<td>0.2075</td>
<td>29.52</td>
<td>60.38</td>
<td>77.97</td>
<td>0.1557</td>
<td>0.1696</td>
<td>0.1834</td>
<td>0.1756</td>
<td>0.1741</td>
<td>0.1554</td>
<td>0.1850</td>
<td>0.1660</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>lillian</td>
<td>0.1607</td>
<td>0.1787</td>
<td>0.2251</td>
<td>29.20</td>
<td>57.12</td>
<td>73.84</td>
<td>0.1696</td>
<td>0.1820</td>
<td>0.2077</td>
<td>0.1891</td>
<td>0.1786</td>
<td>0.1560</td>
<td>0.1901</td>
<td>0.1689</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>yshk</td>
<td>0.1162</td>
<td>0.1319</td>
<td>0.1752</td>
<td>37.05</td>
<td>70.24</td>
<td>84.73</td>
<td>0.1185</td>
<td>0.1363</td>
<td>0.1193</td>
<td>0.1177</td>
<td>0.1533</td>
<td>0.1362</td>
<td>0.1423</td>
<td>0.1279</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01</td>
<td>cv_challenge</td>
<td>0.0736</td>
<td>0.0858</td>
<td>0.1187</td>
<td>22.23</td>
<td>85.44</td>
<td>93.58</td>
<td>0.0736</td>
<td>0.0840</td>
<td>0.0812</td>
<td>0.0898</td>
<td>0.0798</td>
<td>0.0693</td>
<td>0.1018</td>
<td>0.0926</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. 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.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Abs. Rel</th>
<th>MAE</th>
<th>RMSE</th>
<th>δ &lt; 1.05</th>
<th>δ &lt; 1.15</th>
<th>δ &lt; 1.25</th>
<th>Abs. Rel</th>
<th>MAE</th>
<th>RMSE</th>
<th>δ &lt; 1.05</th>
<th>δ &lt; 1.15</th>
<th>δ &lt; 1.25</th>
<th>Abs. Rel</th>
<th>MAE</th>
<th>RMSE</th>
<th>δ &lt; 1.05</th>
<th>δ &lt; 1.15</th>
<th>δ &lt; 1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>03</td>
<td>DPT (ft) [76]</td>
<td>0.1458</td>
<td>0.1596</td>
<td>0.2075</td>
<td>29.52</td>
<td>60.38</td>
<td>77.97</td>
<td>0.1557</td>
<td>0.1696</td>
<td>0.1834</td>
<td>0.1756</td>
<td>0.1741</td>
<td>0.1554</td>
<td>0.1850</td>
<td>0.1660</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>lillian</td>
<td>0.1607</td>
<td>0.1787</td>
<td>0.2251</td>
<td>29.20</td>
<td>57.12</td>
<td>73.84</td>
<td>0.1696</td>
<td>0.1820</td>
<td>0.2077</td>
<td>0.1891</td>
<td>0.1786</td>
<td>0.1560</td>
<td>0.1901</td>
<td>0.1689</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>yshk</td>
<td>0.1162</td>
<td>0.1319</td>
<td>0.1752</td>
<td>37.05</td>
<td>70.24</td>
<td>84.73</td>
<td>0.1185</td>
<td>0.1363</td>
<td>0.1193</td>
<td>0.1177</td>
<td>0.1533</td>
<td>0.1362</td>
<td>0.1423</td>
<td>0.1279</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01</td>
<td>cv_challenge</td>
<td>0.0736</td>
<td>0.0858</td>
<td>0.1187</td>
<td>22.23</td>
<td>85.44</td>
<td>93.58</td>
<td>0.0736</td>
<td>0.0840</td>
<td>0.0812</td>
<td>0.0898</td>
<td>0.0798</td>
<td>0.0693</td>
<td>0.1018</td>
<td>0.0926</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
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 $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ 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
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 $2878 \times 2105$ crops, which are further resized to network resolution ($384 \times 384$) 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 $1 \times 1$ 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


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


Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate


