TRICKY 2024 Challenge on Monocular Depth from Images of Specular and Transparent Surfaces

TRICKY 2024 Challenge on Monocular Depth from Images of Specular and Transparent Surfaces

Pierluigi Zama Ramirez*, Alex Costanzino*, Fabio Tosi*, Matteo Poggi*, Luigi Di Stefano*, Jean-Baptiste Weibel*, Dominik Bauer*, Doris Antensteiner*, Markus Vincze*,
Jiaqi Li, Yachuan Huang, Junrui Zhang, Yiran Wang, Jinghong Zheng, Liao Shen, Zhiguo Cao, Ziyang Song, Zerong Wang, Ruijie Zhu, Hao Zhang, Rui Li, Jiang Wu, Xian Li, Yu Zhu, Jinqiu Sun, Yanning Zhang, Pihai Sun, Yuanqi Yao, Wenbo Zhao, Kui Jiang, Junjun Jiang, Mykola Lavreniuk, Pengzhi Li, and Jui-Lin Wang

Abstract. This paper reports on the TRICKY 2024 challenge on Monocular Depth From images of Specular and Transparent surfaces, held in conjunction with the Transparent & Reflective objects In the wild Challenges (TRICKY) workshop at ECCV 2024. The goal of this challenge is to foster research on depth estimation and address one of the challenges that remained open in the field: the perception of non-Lambertian surfaces. The challenge focuses on the popular single-image depth estimation task, attracting more than 50 registered participants. In the final testing stage, 6 participating teams submitted their models and fact sheets.

1 Introduction

Recovering the 3D geometry of a capture scene is one of the longest-standing research problems in computer vision. Starting from images, the very first step to tackle this task is estimating depth, which is also a milestone in higher-level applications such as autonomous driving, robotics, and more. To this purpose, the market proposes a variety of specialized sensors exploiting *active* technologies, ranging from LiDARs, Radars, Sonars, and Time-of-Flight (ToF). In the last decade, the rising impact of deep learning on the research community made depth estimation from images an appealing alternative, thanks to the lower costs of standard color cameras and the higher resolution they feature.

Among the many techniques relying on images, one that has always been considered the holy grail in 3D computer vision has become the most popular

*Pierluigi Zama Ramirez (pierluigi.zama@unibo.it), Alex Costanzino, Fabio Tosi, Matteo Poggi, Luigi Di Stefano, Jean-Baptiste Weibel, Dominik Bauer, Doris Antensteiner and Markus Vincze are the TRICKY 2024 workshop and challenge organizers. The other authors participated in the challenge. The last section contains the authors' team names and affiliations. The TRICKY 2024 website: https://sites.google.com/view/eccv24-tricky-workshop/

lately: single-image depth estimation. Although an ill-posed problem, the rising availability of training data and the advancements in architectural design yielded a number of models predicting depth – either relative [56,83] or metric [86] – with unprecedented accuracy. Nevertheless, as often happens when web-scale datasets become the fuel for progress, these models still struggle to deal with specific, challenging conditions falling in the long tail of the training data distribution.

The presence of objects made of *non-Lambertian* materials represents one of those challenges, both because belonging to such a long tail, as well as for their physical properties. Indeed, these materials are particularly difficult to perceive also for active sensors, as they violate the assumptions upon which their sensing technology is developed – e.g., light beams shot by LiDARs are refracted or surpass transparent surfaces. For what concerns single-image depth estimation, the ambiguities of context framed over reflective surfaces or the complete transparency of some surfaces make them even more ill-posed conditions inside an ill-posed problem itself, driving for instance a neural network to predict the depth of the content behind a transparent surface. Although this latter example may not be considered a real failure, we argue that, from a practical point of view, it is: indeed, depending on the application, it may be necessary to perceive the presence of a glassy door by properly estimating its depth for a robot navigating indoor; in such a case, failing to identify the door as an obstacle would not avoid colliding with it.

This TRICKY 2024 Challenge on Monocular Depth from Images of Specular and Transparent Surfaces pursues the advancement of state-of-the-art single-image depth estimation, by encouraging the development of solutions that can properly deal with non-Lambertian materials. Purposely, we employ the Booster dataset [90,93] as the proving ground of this challenge, a recent benchmark featuring high-resolution images with several non-Lambertian surfaces. This challenge follows the success of the NTIRE 2023 and 2024 challenges on HR Depth from Images of Specular and Transparent Surfaces [55, 92], which took place once per year at CVPR and focused on depth from either stereo or single images. However, the unprecedented speed at which the state-of-the-art advanced in the last months on the latter track [13,32,83,84] motivated us to propose this new challenge tailored to it. This initiative welcomed 54 registered participants. Among them, 6 teams submitted their models and fact sheets during the final phase, with some participants deploying off-the-shelf, existing solutions, whereas others combined different strategies and combined them to obtain better results. The final results of this challenge are summarized and discussed in Section 4.

2 Related Work

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

2.1 Monocular Depth Estimation.

Early approaches to single-image depth estimation relied on hand-crafted features to capture visual cues such as texture gradients and object proportions [63]. The advent of deep learning revolutionized the field, enabling direct learning from data and leading to significant improvements [4,12,36,54,80]. This progress has been fueled by the availability of extensive datasets with ground-truth depth annotations [4, 12, 36, 54, 80], coupled with the development of self-supervised techniques [1,18–20,23,30,31,45,50,51,53,76,77,81,95–98]. These self-supervised methods recast depth estimation as an image reconstruction task during training, using either stereo pairs or monocular video sequences. Building on these baselines, researchers have explored multi-task approaches that incorporate complementary data such as optical flow [60, 78, 89, 99] and semantic segmentation [21, 34, 91]. In addition, efforts have been made to predict depth uncertainty [25, 52], further improving the reliability and applicability of monocular depth estimation systems. In other cases, specialized self-supervised frameworks have been developed to address the challenges posed by dynamic objects in monocular video sequences [34, 38, 39, 48, 74].

In recent years, affine-invariant models for monocular depth estimation have emerged [41, 56, 59, 85], addressing the challenge of generalizing to unknown settings [5, 6]. This approach estimates depth up to an unknown global offset and scale, providing a compromise between ordinal and metric representations. MiDaS [59] pioneered this approach by training on diverse datasets to achieve cross-domain generalization, paving the way for subsequent works such as DPT [56], Omnidata [11], and Depth Anything [83,84]. Researchers have used various strategies to achieve generalization, including leveraging annotations from large datasets [56, 83], Internet photo collections [41, 70, 71, 85], automotive LiDAR [16,17,27], RGB-D sensors [7,49,73], teacher-student paradigms [2], and crowd-sourced annotations [5]. Recent works such as Metric3D [26, 87] and ZeroDepth [22] have revisited depth estimation by explicitly incorporating camera intrinsic properties. Other efforts have focused on improving the accuracy of depth maps, including point cloud shape recovery techniques [88] and high-frequency detail restoration [42, 47]. The field has continued to evolve, with emerging trends including the application of generative models, particularly diffusion models [24, 67], to depth estimation tasks [10, 13, 29, 32, 64–66].

Despite these advances, the challenge of accurately estimating depth for transparent and reflective surfaces remains largely unaddressed. This gap is primarily due to the lack of suitable datasets, with Booster [90] being a notable exception, providing high-resolution images with precisely annotated non-Lambertian objects. To address this problem, Costanzino *et al.* [9] proposed a method to generate pseudo-annotations for transparent and mirror (ToM) surfaces using existing depth models and material segmentation. Building on this direction, Tosi *et al.* [79] introduced an approach that uses conditioned diffusion models to generate challenging, out-of-distribution scenes with associated depth information, addressing both adverse weather conditions [15] and non-

Lambertian surfaces. Alternative approaches have explored depth completion techniques [8, 62] to address the complexity of non-Lambertian surfaces.

2.2 Depth Estimation Competitions and Challenges.

The field of depth estimation has been promoted by past – and concurrent – competitions and challenges addressing both stereo and monocular approaches. Notable events include the Robust Vision Challenge (ROB) [94], which covers both methods; the Dense Depth for Autonomous Driving (DDAD) Challenge [14]; and the Fast and Accurate Single-Image Depth Estimation on Mobile Devices Challenge (MAI) [28]. Stereo-specific challenges such as the Argoverse Stereo Challenge [35] have also made significant contributions. In the monocular domain, the Monocular Depth Estimation Challenge (MDEC) [68, 69, 72] has been particularly influential. Of note, the first and second editions of this challenge [55, 92], which were part of the NTIRE workshop at CVPR 2023 and CVPR 2024, respectively, continued to drive further exploration in this area.

3 TRICKY Challenge on Monocular Depth from Images of Specular and Transparent Surfaces

We host the TRICKY 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 non-Lambertian surfaces – such as mirrors, glasses, and more.

3.1 Datasets.

Our challenge takes place over the Booster dataset [90, 93]. It consists of 606 12Mpx high-resolution stereo pairs collected in 85 different scenes under various illuminations. Each scene is paired with an accurate high-resolution depth ground truth and material segmentation map. The scenes are divided into 38 for training purposes and 47 for testing, for a total of 228 training and 378 test samples. Images and ground truths are downsampled to 1028×752 in this challenge.

As in [55, 92], we adopt the original 228 training samples as the *training split*. We identify a *validation split* by sampling images with different illuminations from 6 scenes of the testing splits *Microwave, Mirror1, Pots, Desk, Mirror3, Sanitaries*, yielding 30 validation samples. The validation scenes are shown in Fig. 1. The remaining frames of the original testing split are the official *test split* for this challenge, resulting in 328 samples.

3.2 Evaluation Protocol.

We select the official metrics used by the Booster benchmark for the monocular benchmark [90, 93]. Specifically, we measure the percentage of pixels having the

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

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). Following the second edition of the NTIRE challenge [92], we compute metrics on three different sets of pixels, following [9]: *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 the most precise $\delta < 1.05$ metric, highlighted in **red** in the tables. We define two rankings based on performance on *ToM* and *All* regions, respectively. Finally, 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 [59], 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

A total of 6 teams participated in the final testing phase. Every method proposed by the teams is briefly described in Section 5, while the respective team members are listed in Section 8.

Table 1 reports the results achieved by these teams, as well as the baseline we set, for *Tom*, *All*, and *Other* pixels. At the very bottom of each subtable, 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. 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 87%, as well as to reduce the Abs Rel. below 3%. The improvements are consistent on *Other* pixels as well – and, consequently, on *All*.

Table 1: Evaluation on the Challenge Test Set. Predictions evaluated at resolution 1028×752 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. ZoeDepth results were reported as a baseline.

		ToM					
Team	Rank	$\boldsymbol{\delta < 1.05}$	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	RMSE
Smartlab	**1**	**90.75**	97.58	98.41	0.03	0.03	0.03
SixSeven	**2**	**90.61**	99.23	99.83	0.02	0.02	0.03
3DCreators	**3**	**87.68**	99.79	99.89	0.03	0.03	0.03
HIT-AIIA	4	83.01	99.72	99.94	0.03	0.03	0.03
Lavreniuk	5	82.29	98.38	99.58	0.03	0.03	0.04
THU-LW	6	60.92	91.50	98.26	0.05	0.06	0.07
ZoeDepth	7	48.03	84.22	93.96	0.08	0.08	0.09

		All					
Team	Rank	$\boldsymbol{\delta < 1.05}$	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	RMSE
Smartlab	**2**	**89.59**	98.33	99.05	0.03	0.02	0.04
SixSeven	**3**	**87.53**	97.54	98.72	0.03	0.03	0.04
3DCreators	**1**	**93.31**	99.87	99.94	0.02	0.02	0.03
HIT-AIIA	4	84.91	99.05	99.65	0.03	0.03	0.04
Lavreniuk	5	81.87	97.17	98.69	0.04	0.03	0.05
THU-LW	6	68.38	92.97	97.61	0.06	0.05	0.08
ZoeDepth	7	62.76	91.00	96.30	0.06	0.06	0.08

	Other					
Team	$\delta < 1.05$	$\delta < 1.15$	$\delta < 1.25$	Abs Rel.	MAE	RMSE
Smartlab	89.35	98.38	99.11	0.03	0.03	0.04
SixSeven	86.84	97.38	98.64	0.03	0.03	0.04
3DCreators	93.38	99.86	99.94	0.02	0.02	0.03
HIT-AIIA	84.57	98.92	99.63	0.03	0.03	0.04
Lavreniuk	81.44	97.01	98.59	0.04	0.03	0.06
THU-LW	67.58	93.39	97.74	0.06	0.05	0.08
ZoeDepth	62.62	91.14	96.13	0.07	0.05	0.09

Finally, we can appreciate the substantial improvement achieved by the two absolute winners, SmartLab and 3DCreators, respectively, according to *ToM* and *All* rankings. In Fig. 2 we report some qualitative examples: we can appreciate how, in some cases, any of the submitted models can properly handle *ToM* regions – as for the shower in the first column. However, we can still observe failure cases in most of them in the presence of water surfaces (second column).

5 Challenge Methods

5.1 Baseline - ZoeDepth [3]

We utilize the ZoeDepth model as the baseline approach, a cutting-edge network for monocular depth estimation. This model is built upon DPT [57], an encoder-decoder architecture that employs a vision transformer (ViT) as the core of its encoder. Additionally, it features a metric bins module aimed at learning a metric depth representation. We utilize the pre-trained weights made available by the authors.

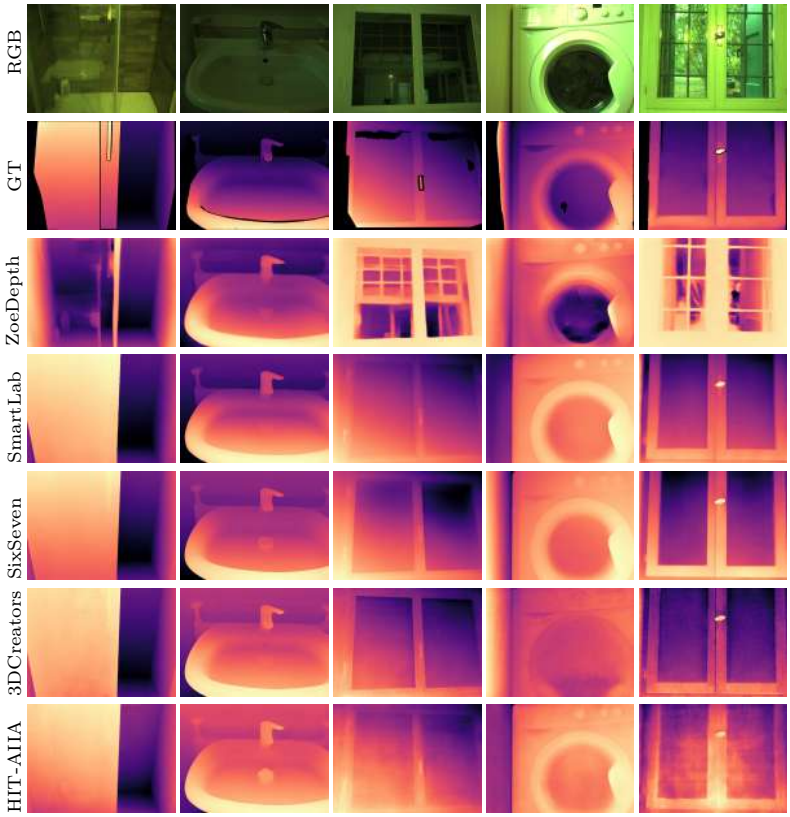

Fig. 2: Qualitative results. From top to bottom: RGB reference image, ground-truth disparity, predictions by ZoeDepth [3] and four among the participant methods.

5.2 Team 1 - SmartLab

The team proposes a depth estimation framework based on the recent Depth Anything v2 [84] framework due to its outstanding generalization performance. The team proposes a customized data augmentation strategy and a regional loss function to enhance its performance in the specular and transparent areas of the Booster [90, 93] datasets. A pipeline overview is shown in Fig. 3.

Initially, the model is fine-tuned using the Hypersim [61] and the training set of the Booster datasets. The global scale-shift-invariant loss function is the same as MiDaS [58]. Given the diversity of RGB exposure and lighting conditions in the Booster dataset, a random tone-mapping is applied to the Hypersim dataset during the training phase to enhance robustness to varying illuminations.

In addition to the global scale-shift-invariant loss, a local gradient alignment loss is applied for non-Lambertian surfaces. Let D and $D*$ denote the predicted and the ground-truth (GT) depth maps, respectively. Let ∇ represent the gradi-

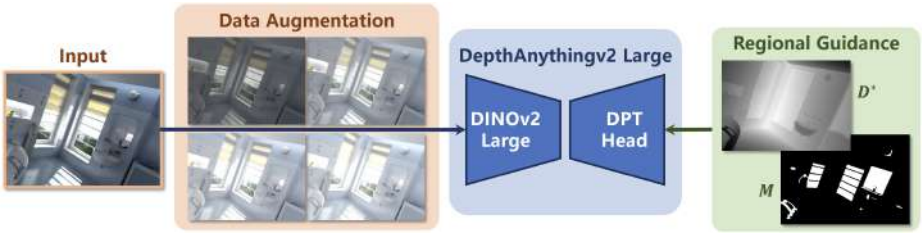

Fig. 3: Team 1 - Smartlab

ent magnitude. Given the non-Lambertian surface mask M, the regional gradient matching loss can be expressed as:

$$\mathcal{L}_{region} = \frac{1}{|M|} \sum_{i}^{M} ||s\nabla D_i + t - \nabla D_i^*|| \tag{2}$$

$$(s,t) = \arg\min_{s,t} ||(s\nabla D + t) - \nabla D^*||_2^2 \tag{3}$$

where s and t are the least squares coefficients. The Booster training set provides annotated masks M, while M for the Hypersim dataset is obtained by thresholding and binarizing the reflectance coefficient map.

These improvements enhance the depth estimation model's performance for non-Lambertian surfaces while keeping the basic framework unchanged. The inference time on an A6000 GPU is 0.248 seconds.

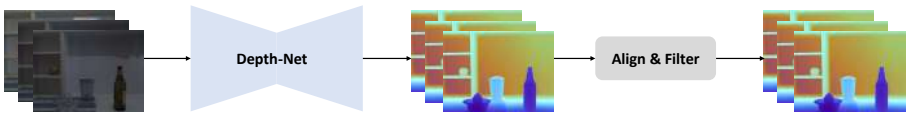

Fig. 4: Team 2 - SixSeven

5.3 Team 2 - SixSeven

Due to its excellent generalization ability on unseen scenes, the team builds its framework upon Depth Anything V2 [84]. The overall pipeline is as shown in Fig. 4.

Depth Anything V2 is first fine-tuned on the Booster [93] training dataset. For better convergence, the last activation function is replaced with a ReLU. Scale s and shift t are computed to align the depth prediction d with ground-truth d_{gt}. Then an MSE loss is applied to the aligned depth prediction $\hat{d}$ and

ground depth.

$$Loss = \|\hat{d} - d_{gt}\|^2, \qquad \text{with} \qquad \hat{d} = sd + t. \tag{4}$$

All the images are resized to 748*748 for training. The model is trained with Adam optimizer and a learning rate of 5e-7 for 100 epochs. A single training takes approximately 1 day on a single A6000 GPU card.

In the Booster dataset, for each scene, there are images acquired under different lighting conditions, whose depth maps are consistent. To fully exploit these images, the framework follows Marigold [32] to ensemble the predictions under different lighting conditions. Specifically, for the depth predictions of a scene under N types of lighting conditions $d_1, \cdots, d_N$, they jointly align them to a canonical scale and range with optimized scale $\hat{s}_i$ and $\hat{t}_i$ in an iterative way. The objective minimizes the distances between each pair of scaled and shifted predictions $(\hat{d}_i, \hat{d}_j)$ as follows.

$$L_{post} = \min_{\substack{s_1,\ldots,s_N \\ t_1,\ldots,t_N}} \sqrt{\frac{1}{b} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\| \hat{\mathbf{d}}_i - \hat{\mathbf{d}}_j \right\|_2^2}, \tag{5}$$

where $b = \binom{N}{2}$. Besides, in each optimization step, the merged depth map m is calculated by taking the pixel-wise median. An extra regularization term $\mathbf{R} = |\min(m)| + |1 - \max(m)|$ is added to prevent to the trivial solution. They take the merged depth m as the final ensembled prediction for all images of the same scene under different lighting conditions.

Finally, to further enhance the model's accuracy, median filtering is applied to the output depth maps. The kernel size for the median filter is 5, which, on the one hand, does not overly smooth the image and thus avoids losing edge information and, on the other hand, effectively filters out noises.

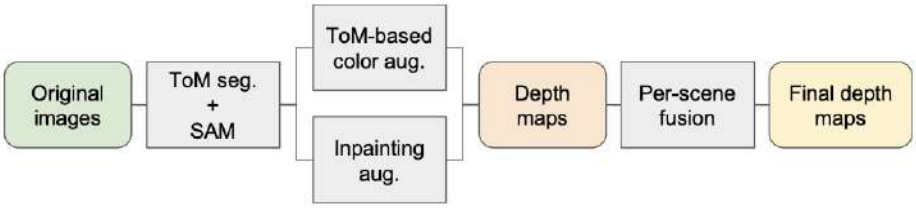

Fig. 5: Team 3 - 3DCreators

5.4 Team 3 - 3DCreators

The team adopts a training-free approach, as illustrated in Fig. 5. The Metric3D-v2-giant model [26] is used as the baseline model as it is trained on a large dataset and is capable of high generalization.

Given the test images, transparent or reflective areas (ToM) are segmented using multiple methods [44,46,82], and the segmented pixels are used as prompts for refinement with the Segment Anything Model (SAM) [33]. A parallel approach is employed to handle different degrees of reflection. For images with large reflective areas (e.g., mirrors, windows), an inpainting method [75] is used to complete the images from which ToM areas have been removed. For other cases, the uniform coloring strategy of [9] is employed for coloring ToM surfaces. Both color jittering and horizontal flips are applied during this process.

Finally, multiple estimations are fused using the median operation. After obtaining these fused estimations, depth maps from varying illuminations of the same scene are fused into one final depth map, again using the median operation.

5.5 Team 4 - HIT-AIIA

The team employed the Depth-Anything-V2-Large model [84] as their starting network, known for its strong generalization ability due to training on large-scale synthetic and unlabeled real-world datasets. This model has prior knowledge of handling transparent and reflective surfaces. Their fine-tuning strategy follows the ZoeDepth pipeline but uses the pre-trained encoder from Depth-Anything-V2-Large instead of the MiDas encoder [3].

Initially, the team attempted to fine-tune the model using the competition's training set. However, this strategy significantly degraded the model's ability to predict depth details, likely due to noise and incomplete labels in the training set. Conversely, fine-tuning exclusively on the synthetic Hypersim dataset [61] led to inconsistent depth predictions for objects with varying transmissive properties.

These observations prompted the team to use a mixed dataset for fine-tuning. The mixed dataset consists of the synthetic datasets Hypersim and MIDepth [43], along with the real-world Booster dataset. For Hypersim, a subset with a maximum depth of up to 10 meters is selected from the training split. For MIDepth, which includes multi-illumination and transparent/reflective scenes, a subset with a maximum depth of up to 5 meters is selected from the full training split. The model was trained using the Adam optimizer with learning rates set to 5e-6 for the encoder and 5e-5 for the decoder. Training was conducted on two RTX 3090 GPUs with a batch size of 8. The minimum and maximum depths were set to 0.001 meters and 10 meters, respectively, and the training spanned a total of 10 epochs.

5.6 Team 5 - Lavreniuk

The proposed DeepBlend method, shown in Fig. 6, utilizes the Depth Anything V2 Large model [84], which was initially fine-tuned on the MSD and Trans10K datasets, followed by the Booster dataset. For dataset preparation, the team employed multiple pipelines. For Trans10K, pseudo-labeled depth masks are created using a blending technique that combines the original image and the transparent object mask, improving upon the method in [9]. This technique preserves object form and enhances depth accuracy. For MSD, depth maps were averaged from

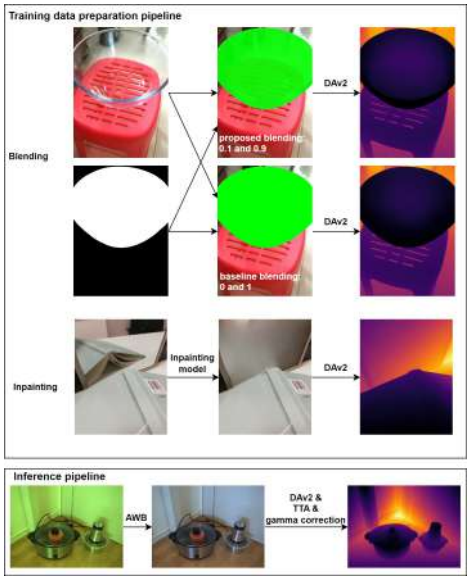

Fig. 6: Team 5 - Lavreniuk

two sources: one from the blended approach and another from a restored image produced by an inpainting model based on fast Fourier convolutions.

During inference, test-time augmentations were applied, including flipping and color jittering, similar to [37]. To refine the depth map and mitigate the effects of varying lighting conditions, an automatic white balance algorithm was employed, along with gamma correction applied to the final output.

5.7 Team 6 - THU-LW

The team utilized Marigold [32] as the baseline method for depth estimation and developed a Depth Consistency Module (DCM) to integrate 512×512 and 768×768 depth maps, thereby enhancing depth estimation accuracy. Drawing inspiration from techniques for consistent depth estimation [40], the team designed a similar architecture for both the depth consistency module and the dataset used. The DCM employs a straightforward encoder-decoder architecture, with the encoder comprising two down-sampling convolutional layers followed by five residual blocks. The decoder includes two transposed convolutional layers, incorporating skip connections from the encoder to the decoder. The model was implemented using PyTorch and trained for 50,000 iterations with a learning rate of 1e-4. For training, a batch size of 4 and randomly cropping to 192×192 are employed. By optimizing the depth maps obtained at both 512 and 768 resolutions through this network, the performance in complex scenes is improved without requiring fine-tuning on specific datasets. Moreover, the lightweight nature of this network minimizes computational overhead.

6 Conclusion

The TRICKY 2024 challenge highlighted how the very latest advancements in single-image depth estimation can push the bar very high for what concerns perceiving transparent and mirroring objects, despite their challenging appearance. Indeed, state-of-the-art solutions such as Marigold [32] and Depth Anything v2 [84] can properly deal with most of these elements, mostly thanks to the high-quality data they have processed during training [61]. Nonetheless, some failure cases still occur when dealing with some very rare situations – i.e., the water surface depicted in Fig. 2, second column – suggesting that further investigation is necessary to design solutions that are reliable also in such circumstances.

We hope the TRICKY 2024 challenge will inspire more practitioners and researchers to push the boundaries of depth estimation on specular and transparent surfaces further.

7 TRICKY 2024 Challenge Organizers

Title:
TRICKY 2024 Challenge on Monocular 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], Luigi Di Stefano[1], Jean-Baptiste Weibel[2], Dominik Bauer[3], Doris Antensteiner[4], Markus Vincze[2]

Affiliations:
[1] University of Bologna, Italy
[2] TU Wien
[3] Columbia University
[4] AIT - Austrian Institute of Technology

8 Teams and Affiliations

8.1 SmartLab

Members:
Jiaqi Li [1] (lijiaqi_mail@hust.edu.cn), Yachuan Huang[1] (yachuan@hust.edu.cn), Junrui Zhang[2] (jr_z@hust.edu.cn), Yiran Wang[1] (wangyiran@hust.edu.cn), Jinghong Zheng[1] (deepzheng@hust.edu.cn), Liao Shen[1] (leoshen@hust.edu.cn), Zhiguo Cao[1] (zgcao@hust.edu.cn).

Affiliations:
[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China.
[2] School of Future Technology, Huazhong University of Science and Technology, China.

8.2 SixSeven

Members:
Ziyang Song[1,2] (songziyang@mail.ustc.edu.cn), Zerong Wang[2] (wangzerong@live.com), Ruijie Zhu[1] (ruijiezhu@mail.ustc.edu.cn), Hao Zhang[2] (haozhang@vivo.com).

Affiliations:
[1] University of Science and Technology of China.
[2] Vivo Mobile Communication Co., Ltd.

8.3 3DCreators

Members:
Rui Li (lirui.david@gmail.com), Jiang Wu (18392713997@163.com), Xian Li (llxx@mail.nwpu.edu.cn), Yu Zhu (yuzhu@nwpu.edu.cn), Jinqiu Sun* (sunjinqiu@nwpu.edu.cn), Yanning Zhang (ynzhang@nwpu.edu.cn).

Affiliations:
[1] Northwestern Polytechnical University.

8.4 HIT-AIIA

Members:
Pihai Sun[1] (pihaisun@stu.hit.edu.cn), Yuanqi Yao[1] (yuanqiyao@stu.hit.edu.cn), Wenbo Zhao[1] (wbzhao@hit.edu.cn), Kui Jiang[1] (jiangkui@hit.edu.cn), Junjun Jiang[1] (jiangjunjun@hit.edu.cn).

Affiliations:
[1] Harbin Institute of Technology

8.5 Lavreniuk

Members:
Mykola Lavreniuk[1] (nick_93@ukr.net)

Affiliations:
[1] Space Research Institute NASU-SSAU, Kyiv, Ukraine

8.6 THU-LW

Members:
Pengzhi Li[1] (lpz21@mails.tsinghua.edu.cn), Jui-Lin Wang[1] (931949915@qq.com)

Affiliations:
[1] Tsinghua University

References

1. Aleotti, F., Tosi, F., Poggi, M., Mattoccia, S.: Generative adversarial networks for unsupervised monocular depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 3
2. Aleotti, F., Zaccaroni, G., Bartolomei, L., Poggi, M., Tosi, F., Mattoccia, S.: Real-time single image depth perception in the wild with handheld devices. Sensors **21**(1), 15 (2020) 3
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023) 5, 6, 7, 10
4. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Proc. NeurIPS (2016) 3
5. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. Advances in neural information processing systems **29** (2016) 3
6. Chen, W., Qian, S., Fan, D., Kojima, N., Hamilton, M., Deng, J.: OASIS: A large-scale dataset for single image 3d in the wild. In: CVPR (2020) 3
7. Cho, J., Min, D., Kim, Y., Sohn, K.: Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. arXiv preprint arXiv:2110.11590 (2021) 3
8. Choi, J., Jung, D., Lee, Y., Kim, D., Manocha, D., Lee, D.: Selfdeco: Self-supervised monocular depth completion in challenging indoor environments. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 467–474. IEEE (2021) 4
9. Costanzino, A., Zama Ramirez, P., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Learning depth estimation for transparent and mirror surfaces. In: The IEEE International Conference on Computer Vision (2023), iCCV 3, 5, 10
10. Duan, Y., Guo, X., Zhu, Z.: DiffusionDepth: Diffusion denoising approach for monocular depth estimation. arXiv preprint arXiv:2303.05021 (2023) 3
11. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: ICCV. pp. 10786–10796 (2021) 3
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Proc. NeurIPS (2014) 3
13. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. arXiv preprint arXiv:2403.12013 (2024) 2, 3
14. Gaidon, A., Shakhnarovich, G., Ambrus, R., Guizilini, V., Vasiljevic, I., Walter, M., Pillai, S., Kolkin, N.: Dense depth for autonomous driving (DDAD) challenge (https://sites.google.com/view/mono3d-workshop) (2021) 4
15. Gasperini, S., Morbitzer, N., Jung, H., Navab, N., Tombari, F.: Robust monocular depth estimation under challenging conditions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 3
16. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal of Robotics Research **32**(11), 1231–1237 (2013). https://doi.org/10.1177/0278364913491297 3
17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 3
18. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proc. CVPR (2017) 3

19. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proc. ICCV (2019) 3
20. GonzalezBello, J.L., Kim, M.: Forget about the lidar: Self-supervised depth estimators with med probability volumes. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12626–12637. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/951124d4a093eeae83d9726a20295498-Paper.pdf 3
21. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. arXiv preprint arXiv:2002.12319 (2020) 3
22. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: ICCV (2023) 3
23. Guo, X., Li, H., Yi, S., Ren, J., Wang, X.: Learning monocular depth by distilling cross-domain stereo networks. In: Proc. ECCV (2018) 3
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020) 3
25. Hornauer, J., Belagiannis, V.: Gradient-based uncertainty for monocular depth estimation. In: European Conference on Computer Vision. pp. 613–630. Springer (2022) 3
26. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. arXiv preprint arXiv:2404.15506 (2024) 3, 9
27. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2702–2719 (2019) 3
28. Ignatov, A., Malivenko, G., Plowman, D., Shukla, S., Timofte, R.: 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. pp. 2545–2557 (June 2021) 4
29. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: DDP: Diffusion model for dense visual prediction. In: ICCV (2023) 3
30. Jiang, H., Larsson, G., Maire Greg Shakhnarovich, M., Learned-Miller, E.: Self-supervised relative depth learning for urban scene understanding. In: Proc. ECCV (2018) 3
31. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proc. CVPR (2020) 3
32. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 3, 9, 11, 12
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 10
34. Klingner, M., Termöhlen, J.A., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 582–600. Springer (2020) 3

35. Kretzschmar, H., Liniger, A., Alvarez, J.M., Wang, Y., Casser, V., Yu, F., Pavone, M., Li, B., Geiger, A., Ondruska, P., Li, L.E., Angelov, D., Leonard, J., Van Gool, L.: Argoverse stereo competition (https://cvpr2022.wad.vision/) (2021, 2022) 4
36. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016) 3
37. Lavreniuk, M., Bhat, S.F., Müller, M., Wonka, P.: Evp: Enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment (2023), https://arxiv.org/abs/2312.08548 11
38. Li, H., Gordon, A., Zhao, H., Casser, V., Angelova, A.: Unsupervised monocular depth learning in dynamic scenes. In: Kober, J., Ramos, F., Tomlin, C. (eds.) Proceedings of the 2020 Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 155, pp. 1908–1917. PMLR (16–18 Nov 2021), https://proceedings.mlr.press/v155/li21a.html 3
39. Li, H., Poggi, M., Tosi, F., Mattoccia, S., et al.: On-site adaptation for monocular depth estimation with a static camera. In: BMVC. pp. 901–907 (2023) 3
40. Li, P., Ding, Y., Wang, H., Tang, C., Li, Z.: The devil is in the edges: Monocular depth estimation with edge-aware consistency fusion. arXiv preprint arXiv:2404.00373 (2024) 11
41. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018) 3
42. Li, Z., Bhat, S.F., Wonka, P.: 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) 3
43. Liang, Y., Zhang, Z., Xian, C., He, S.: Delving into multi-illumination monocular depth estimation: A new dataset and method. IEEE Transactions on Multimedia pp. 1–15 (2024). https://doi.org/10.1109/TMM.2024.3353544 10
44. Lin, J., Wang, G., Lau, R.W.: Progressive mirror detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3697–3705 (2020) 10
45. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5667–5675 (2018) 3
46. Mei, H., Yang, X., Wang, Y., Liu, Y., He, S., Zhang, Q., Wei, X., Lau, R.W.: Don't hit me! glass detection in real-world scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3687–3696 (2020) 10
47. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: 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. pp. 9685–9694 (2021) 3
48. Moon, J., Bello, J.L.G., Kwon, B., Kim, M.: From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior. arXiv preprint arXiv:2312.10118 (2023) 3
49. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 3

50. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: Towards real-time unsupervised monocular depth estimation on cpu. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 5848–5854. IEEE (2018) 3
51. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proc. CVPR (2020) 3
52. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020) 3
53. Poggi, M., Tosi, F., Mattoccia, S.: Learning monocular depth estimation with unsupervised trinocular assumptions. In: 2018 International conference on 3d vision (3DV). pp. 324–333. IEEE (2018) 3
54. Ramamonjisoa, M., Du, Y., Lepetit, V.: Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In: Proc. CVPR (2020) 3
55. Ramirez, P.Z., Tosi, F., Di Stefano, L., Timofte, R., Costanzino, A., Poggi, M., Salti, S., Mattoccia, S., Shi, J., Zhang, D., 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. pp. 1384–1395 (2023) 2, 4
56. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021) 2, 3
57. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (October 2021) 6
58. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. (2020) 7
59. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022) 3, 5
60. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12240–12249 (2019) 3
61. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021) 7, 10, 12
62. Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., Song, S.: Clear grasp: 3d shape estimation of transparent objects for manipulation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3634–3642. IEEE (2020) 4
63. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Depth perception from a single still image. In: Proc. AAAI (2008) 3
64. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. arXiv preprint arXiv:2306.01923 (2023) 3
65. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023) 3

66. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors. arXiv preprint arXiv:2406.01493 (2024) 3
67. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 3
68. Spencer, J., Qian, C.S., Russell, C., Hadfield, S., Graf, E., Adams, W., Schofield, A.J., Elder, J.H., Bowden, R., Cong, H., Mattoccia, S., Poggi, M., Suri, Z.K., Tang, Y., Tosi, F., Wang, H., Zhang, Y., Zhang, Y., Zhao, C.: The monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 623–632 (January 2023) 4
69. Spencer, J., Qian, C.S., Trescakova, M., Russell, C., Hadfield, S., Graf, E., Adams, W., Schofield, A.J., Elder, J., Bowden, R., Anwar, A., Chen, H., Chen, X., Cheng, K., Dai, Y., Hoa, H.T., Hossain, S., Huang, J., Jing, M., Li, B., Li, C., Li, B., Liu, Z., Mattoccia, S., Mercelis, S., Nam, M., Poggi, M., Qi, X., Ren, J., Tang, Y., Tosi, F., Trinh, L., Uddin, S.M.N., Umair, K.M., Wang, K., Wang, Y., Wang, Y., Xiang, M., Xu, G., Yin, W., Yu, J., Zhang, Q., Zhao, C.: The second monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2023) 4
70. Spencer, J., Russell, C., Hadfield, S., Bowden, R.: Kick back & relax: Learning to reconstruct the world by watching slowtv. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15768–15779 (October 2023) 3
71. Spencer, J., Russell, C., Hadfield, S., Bowden, R.: Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. arXiv preprint arXiv:2403.01569 (2024) 3
72. Spencer, J., Tosi, F., Poggi, M., Arora, R.S., Russell, C., Hadfield, S., Bowden, R., Zhou, G., Li, Z., Rao, Q., Bao, Y., Liu, X., Kim, D., Kim, J., Kim, M., Lavreniuk, M., Li, R., Mao, Q., Wu, J., Zhu, Y., Sun, J., Zhang, Y., Patni, S., Agarwal, A., Arora, C., Sun, P., Jiang, K., Wu, G., Liu, J., Liu, X., Jiang, J., Zhang, X., Wei, J., Wang, F., Tan, Z., Wang, J., Luginov, A., Shahzad, M., Hosseini, S., Trajcevski, A., Elder, J.H.: The third monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2024) 4
73. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: Proc. of the International Conference on Intelligent Robot Systems (IROS) (Oct 2012) 3
74. Sun, Y., Hariharan, B.: Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 3
75. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2149–2159 (2022) 10
76. Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 3
77. Tosi, F., Aleotti, F., Ramirez, P.Z., Poggi, M., Salti, S., Stefano, L.D., Mattoccia, S.: Distilled semantics for comprehensive scene understanding from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 3

78. Tosi, F., Aleotti, F., Ramirez, P.Z., Poggi, M., Salti, S., Stefano, L.D., Mattoccia, S.: Distilled semantics for comprehensive scene understanding from videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4654–4665 (2020) 3
79. Tosi, F., Zama Ramirez, P., Poggi, M.: Diffusion models for monocular depth estimation: Overcoming challenging conditions. In: European Conference on Computer Vision (ECCV) (2024) 3
80. Wang, L., Zhang, J., Wang, Y., Lu, H., Ruan, X.: CLIFFNet for monocular depth estimation with hierarchical embedding loss. In: Proc. ECCV (2020) 3
81. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: Proc. ICCV (2019) 3
82. Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., Luo, P.: Segmenting transparent objects in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 696–711. Springer (2020) 10
83. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024) 2, 3
84. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024) 2, 3, 7, 8, 10, 12
85. Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., Renyin, D.: Diversedepth: Affine-invariant depth prediction using diverse data. arXiv preprint arXiv:2002.00569 (2020) 3
86. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9043–9053 (2023) 2
87. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3D: Towards zero-shot metric 3d prediction from a single image. In: ICCV (2023) 3
88. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 204–213 (2021) 3
89. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018) 3
90. Zama Ramirez, P., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Di Stefano, L., Mattoccia, S.: Booster: a benchmark for depth from images of specular and transparent surfaces. arXiv preprint arXiv:2301.08245 (2023) 2, 3, 4, 7
91. Zama Ramirez, P., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Geometry meets semantics for semi-supervised monocular depth estimation. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 298–313. Springer (2019) 3
92. Zama Ramirez, P., Tosi, F., Di Stefano, L., Timofte, R., Costanzino, A., Poggi, M., Salti, S., Mattoccia, S., Zhang, Y., Wu, C., He, Z., Yin, S., Dong, J., Liu, Y., Jiang, H., Shi, J., A, Y., Jin, Y., Li, D., Ke, B., Obukhov, A., Wang, T., Metzger, N., Huang, S., Schindler, K., Huang, Y., Li, J., Zhang, J., Wang, Y., Huang, Z., Liu, T., Cao, Z., Li, P., Wang, J.L., Zhu, W., Geng, H., Zhang, Y., Lan, L., Xu, K., Sun, T., Xu, Q., Saini, S., Gupta, A., Mistry, S.K., Shukla, A., Jakhetiya, V., Jaiswal, S., Sun, Y., Zheng, Z., Ning, Y., Cheng, J.H., Liu, H.I., Huang, H.W.,

Yang, C.Y., Jiang, Z., Peng, Y.H., Huang, A., Hwang, J.N.: 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 Workshops (2024) 2, 4, 5

93. Zama Ramirez, P., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L.: Open challenges in deep stereo: The booster dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21168–21178 (June 2022) 2, 4, 7, 8
94. Zendel, O., Dai, A., Puig Fernandez, X., Geiger, A., Koltun, V., Kontschieder, P., Kortylewski, A., Lin, T.Y., Sattler, T., Scharstein, D., Schilling, H., Uhrig, J., Wulff, J.: The robust vision challenge (http://www.robustvision.net/) (2018, 2020, 2022) 4
95. Zhao, C., Poggi, M., Tosi, F., Zhou, L., Sun, Q., Tang, Y., Mattoccia, S.: Gasmono: Geometry-aided self-supervised monocular depth estimation for indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16209–16220 (2023) 3
96. Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: Monovit: Self-supervised monocular depth estimation with a vision transformer. In: 2022 international conference on 3D vision (3DV). pp. 668–678. IEEE (2022) 3
97. Zhou, C., Zhang, H., Shen, X., Jia, J.: Unsupervised learning of stereo matching. In: The IEEE International Conference on Computer Vision (ICCV). IEEE (October 2017) 3
98. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017) 3
99. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 36–53 (2018) 3