Universiteit Antwerpen

This item is the archived peer-reviewed author-version of:

The second monocular depth estimation challenge

Reference:
Spencer Jaime, Qian C. Stella, Trescakova Michaela, Russell Chris, Hadfield Simon, Graf Erich W., Adams Wendy J., Schofield Andrew J., Elder James, Bowden Richard, ...- The second monocular depth estimation challenge
2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 17-24 June, 2023, Vancouver, BC, Canada- ISSN 2160-7516 - IEEE, 2023, p. 3064-3076
Full text (Publisher's DOI): https://doi.org/10.1109/CVPRW59228.2023.00308
To cite this reference: https://hdl.handle.net/10067/2022880151162165141

uantwerpen.be
Institutional repository IRUA

This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

The Second Monocular Depth Estimation Challenge

Jaime Spencer[1] C. Stella Qian[2] Michaela Trescakova[3] Chris Russell[4] Simon Hadfield[1]
Erich W. Graf[3] Wendy J. Adams[3] Andrew J. Schofield[2] James Elder[5] Richard Bowden[1]
Ali Anwar[6] Hao Chen[7] Xiaozhi Chen[8] Kai Cheng[9] Yuchao Dai[10] Huynh Thai Hoa[11]
Sadat Hossain[11] Jianmian Huang[12] Mohan Jing[9] Bo Li[10] Chao Li[13] Baojun Li[12]
Zhiwen Liu[13] Stefano Mattoccia[14] Siegfried Mercelis[6] Myungwoo Nam[11] Matteo Poggi[14]
Xiaohua Qi[9] Jiahui Ren[10] Yang Tang[15] Fabio Tosi[14] Linh Trinh[6] S. M. Nadim Uddin[11]
Khan Muhammad Umair[11] Kaixuan Wang[8] Yufei Wang[10] Yixing Wang[13] Mochu Xiang[10]
Guangkai Xu[9] Wei Yin[8] Jun Yu[9] Qi Zhang[13] Chaoqiang Zhao[15]

Abstract

This paper discusses the results for the second edition of the Monocular Depth Estimation Challenge (MDEC). This edition was open to methods using any form of supervision, including fully-supervised, self-supervised, multi-task or proxy depth. The challenge was based around the SYNS-Patches dataset, which features a wide diversity of environments with high-quality dense ground-truth. This includes complex natural environments, e.g. forests or fields, which are greatly underrepresented in current benchmarks.

The challenge received eight unique submissions that outperformed the provided SotA baseline on any of the pointcloud- or image-based metrics. The top supervised submission improved relative F-Score by 27.62%, while the top self-supervised improved it by 16.61%. Supervised submissions generally leveraged large collections of datasets to improve data diversity. Self-supervised submissions instead updated the network architecture and pretrained backbones. These results represent a significant progress in the field, while highlighting avenues for future research, such as reducing interpolation artifacts at depth boundaries, improving self-supervised indoor performance and overall natural image accuracy.

[1]University of Surrey [2]Aston University [3]University of Southampton [4]Amazon [5]York University [6]University of Antwerp [7]Zhejiang University [8]DJI Technology [9]University of Science and Technology of China [10]Northwestern Polytechnical University [11]DeltaX [12]Independent [13]VIVO [14]University of Bologna [15]East China University of Science and Technology

1. Introduction

Monocular depth estimation (MDE) refers to the task of predicting the distance from the camera to each image pixel. Unlike traditional geometric correspondence and triangulation techniques, this requires only a single image. Despite the ill-posed nature of the problem, deep learning has shown rapid improvements in this field.

Unfortunately, many existing approaches have focused solely on training and evaluating in an automotive urban setting. This puts into question their ability to adapt to previously unseen environments. The proposed Monocular Depth Estimation Challenge (MDEC) aims to mitigate this by evaluating models on a complex dataset consisting of natural, agricultural, urban and indoor scenes. Furthermore, this is done in a zero-shot fashion, meaning that the models must be capable of generalizing.

The first edition of MDEC [77] focused on benchmarking self-supervised approaches. The submissions outperformed the baseline [25,78] in all image-based metrics (AbsRel, MAE, RMSE), but provided slightly inferior pointcloud reconstructions [62] (F-Score). The second edition of MDEC, detailed in this paper, ran in conjunction with CVPR2023. This edition was open to any form of supervision, *e.g.* supervised, self-supervised or multi-task. The aim was to evaluate the state of the field as a whole and determine the gap between different supervision strategies.

The challenge was once again centered around SYNS-Patches [1, 78]. This dataset was chosen due its diversity, which includes urban, residential, industrial, agricultural, natural and indoor scenes. Furthermore, SYNS-Patches contains dense high-quality LiDAR ground-

truth, which is exceedingly rare in outdoor environments. This ensures that the evaluations accurately reflect the capabilities of each model.

Eight teams out of the 28 final submissions outperformed the State-of-the-Art (SotA) baseline in either pointcloud- or image-based metrics. Half of these submission were supervised using ground-truth depths, while the remaining half were self-supervised with the photometric reconstruction loss [25,28]. As expected, supervised submissions typically outperformed self-supervised ones. However, the novel self-supervised techniques generally outperformed the provided baseline, even in pointcloud reconstructions. The remainder of the paper will provide the technical details of each submission, analyze their results on SYNS-Patches and discuss potential directions for future research.

2. Related Work

Supervised. Eigen *et al*. [22] introduced the first end-to-end CNN for MDE, which made use of a scale-invariant loss and a coarse-to-fine network. Further improvements to the network architecture included the use of CRFs [53, 100], regression forests [72], deeper architectures [67, 88], multi-scale prediction fusion [60] and transformer-based encoders [9,15,66]. Alternatively, depth estimation was formulated as a discrete classification problem [7, 8, 24, 49]. In parallel, novel losses were proposed in the form of gradient-based regression [51, 84], the berHu loss [47], an ordinal relationship loss [14] and scale/shift invariance [67].

Recent approaches focused on the generalization capabilities of MDE by training with collections of datasets [7, 23, 66, 67, 69, 82]. This relied on the availability of ground-truth annotations, including automotive data LiDAR [27, 32, 38], RGB-D/Kinect [16, 61, 79], SfM reconstructions [50,51], optical flow/disparity estimation [67,88] or crowd-sourced annotations [14]. These annotations varied in accuracy, which may have impacted the final model's performance. Furthermore, this increased the requirements for acquiring data from new sources, making it challenging to scale to larger amounts of data.

Self-Supervised. Instead of relying on costly annotations, Garg *et al*. [25] proposed an algorithm based on view synthesis and the photometric consistency across stereo pairs. Monodepth [28] incorporated differentiable bilinear interpolation [42], virtual stereo prediction and a SSIM+L_1 reconstruction loss. SfM-Learner [108] required only monocular video supervision by replacing the known stereo transform with a pose estimation network.

Artifacts due to dynamic objects were reduced by incorporating uncertainty [45, 65, 93], motion masks [12, 20, 31], optical flow [57, 68, 98] or the minimum reconstruction loss [29]. Meanwhile, robustness to unreliable photometric appearance was improved via feature-based reconstructions [76, 99, 105] and proxy-depth supervision [45, 73, 86]. Developments in network architecture design included 3D (un-)packing blocks [32], positional encoding [30], transformer-based encoders [2, 106], sub-pixel convolutions [64], progressive skip connections [58] and self-attention decoders [43, 91, 107].

Challenges & Benchmarks. The majority of MDE approaches have been centered around automotive data. This includes popular benchmarks such as Kitti [27, 81] or the Dense Depth for Autonomous Driving Challenge [32]. The Robust Vision Challenge series [104], while generalization across multiple datasets, has so far consisted only of automotive [27] and synthetic datasets [10, 70].

More recently, Ignatov *et al*. introduced the Mobile AI Challenge [40], investigating efficient MDE on mobile devices in urban settings. Finally, the NTIRE2023 [102] challenge, concurrent to ours, targeted high-resolution images of specular and non-lambertian surfaces.

The Monocular Depth Estimation Challenge series [77]—the focus of this paper—is based on the MonoDepth Benchmark [78], which provided fair evaluations and implementations of recent SotA self-supervised MDE algorithms. Our focus lies on zero-shot generalization to a wide diversity of scenes. This includes common automotive and indoor scenes, but complements it with complex natural, industrial and agricultural environments.

3. The Monocular Depth Estimation Challenge

The second edition of the Monocular Depth Estimation Challenge[1] was organized on CodaLab [63] as part of a CVPR2023 workshop. The initial development phase lasted four weeks, using the SYNS-Patches validation split. The leaderboard for this phase was anonymous, where all method scores were publicly available, but usernames remained hidden. Each participant could see the metrics for their own submission.

The final challenge stage was open for two weeks. In this case, the leaderboard was completely private and participants were unable to see their own scores. This encouraged evaluation on the validation split rather than the test split. Combined with the fact that all ground-truth depths were withheld, the possibility of overfitting due to repeated evaluations was severely limited.

This edition of the challenge was extended to any form of supervision, with the objective of providing a more comprehensive overview of the field as a whole. This allowed us to determine the gap between different techniques and identify avenues for future research. We report results only for submissions that outperformed the baseline in any pointcloud-/image-based metric on the Overall dataset.

Dataset. The challenge is based on the SYNS-Patches dataset [1, 78], chosen due to the diversity of scenes and

[1] https://codalab.lisn.upsaclay.fr/competitions/10031

Table 1. SYNS-Patches. Distribution of images per category in the val/test splits.

	Agriculture	Indoor	Industry	Misc	Natural	Recreation	Residential	Transport	Woodland	Total
Val	104	67	36	72	36	14	13	4	54	400
Test	211	81	71	0	147	48	110	17	90	775
Total	315	148	107	72	183	62	123	21	144	1,175

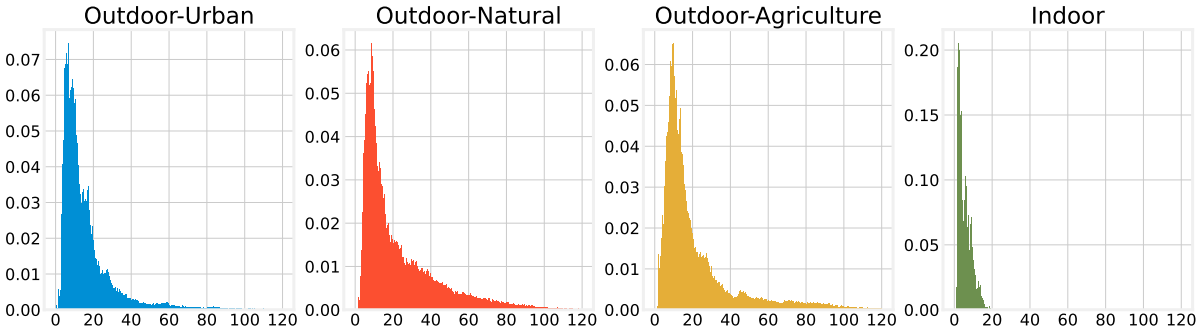

Figure 1. Depth Distribution Per Scene Type. Indoor scenes are limited to 20m, while outdoor scenes reach up to 120m. Natural and Agriculture scenes contain a larger percentage of long-range depths (20-80m), while urban scenes focus on the mid-range (20-40m).

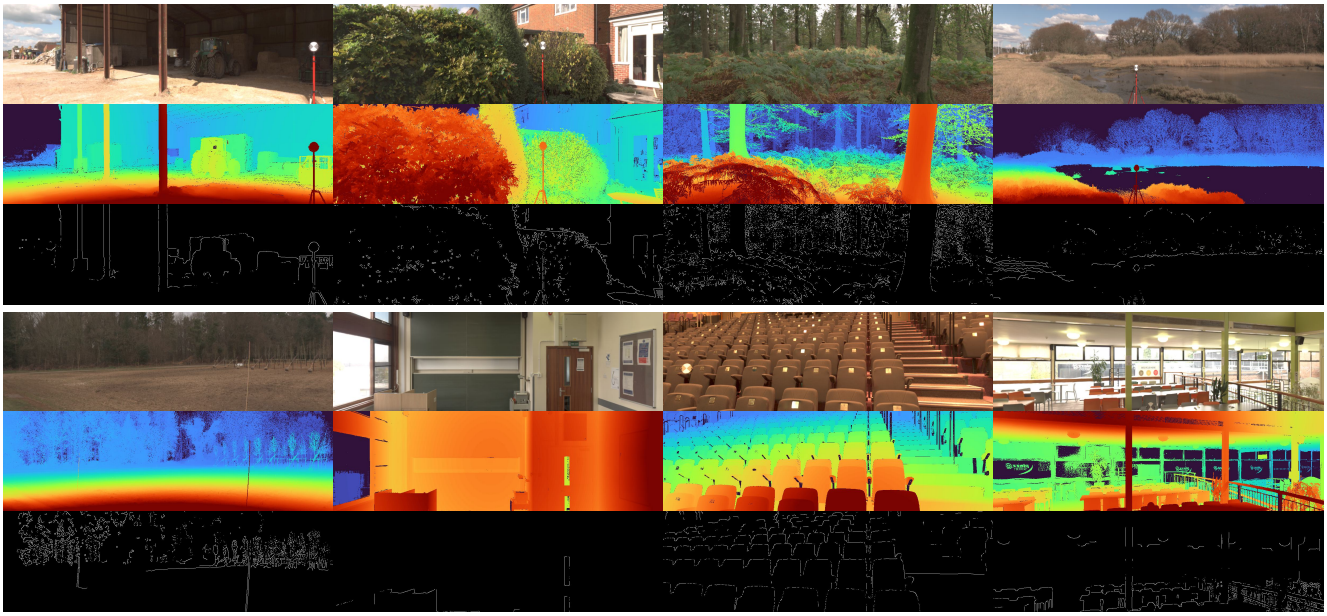

Figure 2. SYNS-Patches. Sample images from the diverse dataset scenes, including complex urban, natural and indoor settings. The dataset contains high-quality ground-truth with 78.20% coverage. Depth boundaries were computed as Canny edges in the log-depth maps.

environments. A breakdown of images per category and some representative examples are shown in Table 1 and Figure 2. SYNS-Patches also provides extremely high-quality dense ground-truth LiDAR, with an average coverage of 78.20% (including sky regions). Given the dense ground-truth, depth boundaries were obtained using Canny edge-detection on the log-depth maps. This allows us to compute additional fine-grained metrics for these challenging regions. As outlined in [78], the images are manually checked to remove dynamic object artifacts.

Evaluation. Participants provided the unscaled disparity prediction for each dataset image. The evaluation server bilinearly upsampled the predictions to the target resolution and inverted them into depth maps. Self-supervised methods trained with stereo pairs and supervised methods using LiDAR or RGB-D data should be capable of predicting met-

ric depth. Despite this, in order to ensure comparisons are as fair as possible, the evaluation aligned predictions with the ground-truth using the median depth. We set a maximum depth threshold of 100 meters.

Metrics. We follow the metrics used in the first edition of the challenge [77], categorized as image-/pointcloud-/edge-based. Image-based metrics represent the most common metrics (MAE, RMSE, AbsRel) computed using pixel-wise comparisons between the predicted and ground-truth depth map. Pointcloud-based metrics [62] (F-Score, IoU, Chamfer distance) instead evaluate the reconstructed pointclouds as a whole. In this challenge, we report reconstruction F-Score as the leaderboard ranking metric. Finally, edge-based metrics are computed only at depth boundary pixels. This includes image-/pointcloud-based metrics and edge accuracy/completion metrics from IBims-1 [46].

4. Challenge Submissions

We outline the technical details for each submission, as provided by the authors. Each submission is labeled based on the supervision used, including ground-truth (**D**), proxy ground-truth (**D***) and monocular (**M**) or stereo (**S**) photometric support frames. The first half represent supervised methods, while the remaining half are self-supervised.

Baseline – S

J. Spencer[1] *j.spencermartin@surrey.ac.uk*
C. Russell[3] *cmruss@amazon.de*
S. Hadfield[1] *s.hadfield@surrey.ac.uk*
R. Bowden[1] *r.bowden@surrey.ac.uk*

Challenge organizers submission from the first edition.

Network. ConvNeXt-B encoder [56] with a base Monodepth decoder [28, 59] from [78].

Supervision. Self-supervised with a stereo photometric loss [25] and edge-aware disparity smoothness [28].

Training. Trained for 30 epochs on Kitti Eigen-Zhou with an image resolution of 192×640.

Team 1: DJI&ZJU – D

W. Yin[8] *yvanwy@outlook.com*
K. Cheng[9] *chengkai21@mail.ustc.edu.cn*
G. Xu[9] *xugk@mail.ustc.edu.cn*
H. Chen[7] *haochen.cad@zju.edu.cn*
B. Li[10] *libo@nwpu.edu.cn*
K. Wang[8] *wkx1993@gmail.com*
X. Chen[8] *xiaozhi.chen@dji.com*

Network. ConvNeXt-Large [56] encoder, pretrained on ImageNet-22k [21], and a LeReS decoder [97] with skip connections and a depth range of $[0.3, 150]$ meters.

Supervision. Supervised using ground-truth depths from a collection of datasets [3, 6, 13, 16, 17, 26, 32, 36, 90, 92, 103]. The final loss is composed of the SILog loss [22], pairwise normal regression loss [97], virtual normal loss [95] and a random proposal normalization loss (RPNL). RPNL enhances the local contrast by randomly cropping patches from the predicted/ground-truth depth and applying median absolute deviation normalization [75].

Training. The network was trained using a resolution of 512×1088. In order to train on mixed datasets directly with metric depth, all ground-truth depths were rescaled as $\hat{y}' = \hat{y} f_c / f$, where f is the original focal length and f_c is an arbitrary focal length. This way, the network assumed all images were taken by the same pinhole camera, which improved convergence.

Team 2: Pokemon – D

M. Xiang[10] *xiangmochu@mail.nwpu.edu.cn*
J. Ren[10] *renjiahui@mail.nwpu.edu.cn*
Y. Wang[10] *wangyufei777@mail.nwpu.edu.cn*
Y. Dai[10] *daiyuchao@nwpu.edu.cn*

Network. Two-stage architecture. The first part was composed of a SwinV2 backbone [54] and a modified NeWCRFs decoder [100] with a larger attention window. The second stage used an EfficientNet [80] with 5 inputs (RGB, low-res depth and high-res depth) to refine the high-resolution depth.

Supervision. Supervised training using LiDAR/synthetic depth and stereo disparities from a collection of datasets [5, 6, 11, 16–18, 22, 34, 37–39, 61, 71, 83–85, 88, 89, 92, 94, 96]. Losses included the SILog loss [22] ($\lambda = 0.85$) for metric datasets, SILog ($\lambda = 1$) for scale-invariant training, the Huber disparity loss for Kitti disparities and an affine disparity loss [67] for datasets with affine ambiguities.

Training. The final combination of losses depended on the ground-truth available from each dataset, automatically mixed by learning an uncertainty weight for each dataset [44]. Since each dataset contained differently-sized images, they were resized to have a shorter side of 352 and cropped into square patches. Some datasets used smaller crops of size 96×352, such that the deepest feature map fell entirely into the self-attention window (11×11). A fusion process based on [60] merged low-/high-resolution predictions into a consistent high-resolution prediction.

Team 3: cv-challenge – D

C. Li[12] *lichao@vivo.com*
Q. Zhang[12] *zhangqi.aiyj@vivo.com*
Z. Liu[12] *zhiwen.liu@vivo.com*
Y. Wang[12] *wangyixing@vivo.com*

Network. Based on ZoeDepth [9] with a BEiT384-L backbone [4].

Supervision. Supervised with ground-truth depth from Kitti and NYUD-v2 [61] using the SILog loss.

Training. The original ZoeDepth [9] and DPT [66] were pretrained on a collection of 12 datasets. The models were then finetuned on Kitti (384×768) or NYUD-v2 (384×512)

for outdoor/indoor scenes, respectively. Different models were deployed on an automatic scene classifier. The fine-tuned models were combined with a content-adaptive multi-resolution merging method [60], where patches were combined based on the local depth cue density. Since the transformer-based backbone explicitly captured long-term structural information, the original double-estimation step was omitted.

Team 4: DepthSquad – D

M. Nam[11] *mwn0221@deltax.ai*
H. T. Hoa[11] *hoaht@deltax.ai*
K. M. Umair[11] *mumairkhan@deltax.ai*
S. Hossain[11] *sadat@deltax.ai*
S. M. N. Uddin[11] *sayednadim@deltax.ai*

Network. Based on the PixelFormer architecture [2] which used a Swin [55] encoder and self-attention decoder blocks with cross-attention skip connections. Disparity was predicted as a discrete volume [7], with the final depth map given as the weighted average using the bin probabilities.
Supervision. Supervised using the SILog loss w.r.t. the LiDAR ground-truth.
Training. The model was trained on the Kitti Eigen-Zhou (KEZ) split using images of size 370×1224 for 20 epochs. Additional augmentation was incorporated in the form of random cropping and rotation, left-right flipping and CutDepth [41]. When predicting on SYNS-Patches, images were zero-padded to 384×1248 to ensure the compatibility of the training resolution. These borders were remove prior to submission.

Team 5: imec-IDLab-UAntwerp – MS

L. Trinh[6] *khaclinh.trinh@student.uantwerpen.be*
A. Anwar[6] *ali.anwar@uantwerpen.be*
S. Mercelis[6] *siegfried.mercelis@uantwerpen.be*

Network. Pretrained ConvNeXt-v2-Huge [87] encoder with an HR-Depth decoder [58], modified with deformable convolutions [19]. The pose network instead used ResNet-18 [35].
Supervision. Self-supervised using the photometric loss [29] and edge-aware smoothness.
Training. Trained on the Kitti Eigen-Benchmark (KEB) split with images of size 192×640. The network was trained for a maximum of 30 epochs, with the encoder remaining frozen after 6 epochs.

Team 6: GMD – MS

B. Li[12] *1966431208@qq.com*
J. Huang[12] *huang176368745@gmail.com*

Network. ConvNeXt-XLarge [56] backbone and an HR-Depth [58] decoder.
Supervision. Self-supervised based on the photometric loss [29].
Training. Trained on KEZ using a resolution of 192×640.

Team 7: MonoViTeam – MSD*

C. Zhao[14] *zhaocq@mail.ecust.edu.cn*
M. Poggi[13] *m.poggi@unibo.it*
F. Tosi[13] *fabio.tosi5@unibo.it*
Y. Tang[14] *yangtang@ecust.edu.cn*
S. Mattoccia[13] *stefano.mattoccia@unibo.it*

Network. MonoViT [106] architecture, composed of MPViT [48] encoder blocks and a self-attention decoder.
Supervision. Self-supervised on Kitti Eigen (KE) using the photometric loss [29] (stereo and monocular support frames) and proxy depth regression. Regularized using edge-aware disparity smoothness [28] and depth gradient consistency w.r.t. the proxy labels.
Training. Proxy depths were obtained by training a self-supervised RAFT-Stereo network [52] on the trinocular Multiscopic [101] dataset. The stereo network was trained for 1000 epochs using 256×480 crops. The monocular network was trained on KE for 20 epochs using images of size 320×1024.

Team 8: USTC-IAT-United – MS

J. Yu[9] *harryjun@ustc.edu.cn*
M. Jing[9] *jing_mohan@mail.ustc.edu.cn*
X. Qi[9] *xiaohua000109@163.com*

Network. Predictions were obtained as a mixture of multiple networks: DiffNet [107], FeatDepth [74] and MonoDEVSNet [33]. DiffNet and FeatDepth used a ResNet backbone, while MonoDEVSNet used DenseNet [38].
Supervision. Self-supervised using the photometric loss [29].
Training. The three models were trained with different resolutions: 320×1024, 376×1242, 384×1248, respectively. All predictions were interpolated to 376×1242 prior to ensembling using a weighted average with coefficients $\{0.35, 0.3, 0.35\}$.

5. Results

Participant submissions were evaluated on SYNS-Patches [1, 78]. As previously mentioned, this paper only discusses submissions that outperformed the baseline in any pointcloud-/image-based metric across the Overall dataset. Since both challenge phases ran independently and participants were responsible for generating the predictions, we cannot guarantee that the testing/validation metrics used the same model. We therefore report results only for the test split. All methods were median aligned w.r.t. the ground-truth, regardless of the supervision used. This ensures that the evaluations are identical and comparisons are fair.

Table 2. SYNS-Patches Results. We provide metrics across the whole dataset and per scene-category. As expected, supervised methods generally outperform self-supervised ones. The largest gap can be found in Indoor scenes, self-supervised methods were trained exclusively on automotive data. Teams DepthSquad & imec-IDLab-UAntwerp outperformed the challenge baseline [78] by incorporating more advanced network architectures.

		Train	Rank	F↑	F-Edges↑	MAE↓	RMSE↓	AbsRel↓	Acc-Edges↓	Comp-Edges↓
	DJI&ZJU	D	**1**	**17.51**	8.80	**4.52**	8.72	**24.32**	3.22	21.65
	Pokemon	D	2	16.94	**9.63**	4.71	**8.00**	25.35	3.56	19.95
	cv-challenge	D	3	16.70	9.36	4.91	8.63	24.33	**3.02**	18.07
	imec-IDLab-UAntwerp	MS	4	16.00	8.49	5.08	8.96	28.46	3.74	**11.32**
Overall	GMD	MS	5	14.71	8.13	5.17	8.97	29.43	3.75	17.29
	Baseline	S	6	13.72	7.76	5.56	9.72	32.04	3.97	21.63
	DepthSquad	D	7	12.77	7.68	5.17	8.83	29.92	3.56	35.26
	MonoViTeam	MSD*	8	12.44	7.49	5.05	8.59	28.99	3.10	38.93
	USTC-IAT-United	MS	9	11.29	7.18	5.81	9.58	32.82	3.47	43.38
	DJI&ZJU	D	**1**	**16.41**	7.37	**3.81**	7.82	**21.85**	2.91	24.36
	imec-IDLab-UAntwerp	MS	4	16.28	7.27	4.49	7.98	26.18	3.67	**13.11**
	GMD	MS	5	15.21	6.80	4.60	8.00	27.55	3.73	16.26
	Pokemon	D	2	15.10	**8.48**	4.03	**6.90**	23.67	3.36	19.13
Outdoor-Urban	cv-challenge	D	3	15.01	7.79	4.26	7.70	22.88	2.87	15.73
	Baseline	S	6	14.09	6.48	4.77	8.43	29.10	3.89	22.75
	DepthSquad	D	7	12.90	5.92	4.49	7.80	27.44	3.26	35.36
	MonoViTeam	MSD*	8	12.52	5.89	4.37	7.62	26.46	**2.83**	40.33
	USTC-IAT-United	MS	9	11.31	5.73	5.14	8.69	30.64	3.13	40.15
	Pokemon	D	2	**14.90**	6.75	6.26	**10.47**	28.40	3.54	14.44
	cv-challenge	D	3	14.66	**6.79**	6.35	10.86	27.09	**3.08**	19.73
	imec-IDLab-UAntwerp	MS	4	14.43	6.02	6.51	11.43	30.57	3.59	**9.44**
	DJI&ZJU	D	**1**	14.31	6.07	**5.97**	10.81	**26.48**	3.45	17.75
Outdoor-Natural	GMD	MS	5	12.89	5.74	6.77	11.62	32.57	3.68	13.97
	Baseline	S	6	12.10	5.32	7.46	12.86	36.89	3.84	18.35
	DepthSquad	D	7	11.54	6.03	6.87	11.52	33.66	3.36	32.47
	MonoViTeam	MSD*	8	10.98	5.38	6.66	11.13	32.19	3.13	36.01
	USTC-IAT-United	MS	9	9.26	4.92	7.69	12.22	38.14	3.36	42.92
	DJI&ZJU	D	**1**	**16.36**	5.24	**5.17**	10.13	29.07	3.43	18.84
	Pokemon	D	2	15.58	**6.40**	5.25	**9.09**	27.45	3.64	18.30
	imec-IDLab-UAntwerp	MS	4	14.94	5.49	5.70	10.14	30.70	3.75	**10.27**
	cv-challenge	D	3	14.68	5.82	5.61	10.02	**25.90**	**3.17**	17.67
Outdoor-Agriculture	GMD	MS	5	14.03	5.06	5.65	9.98	30.40	3.80	15.94
	Baseline	S	6	12.26	4.76	6.10	10.84	33.58	4.00	18.73
	DepthSquad	D	7	11.56	4.55	5.61	9.79	31.16	3.60	35.30
	MonoViTeam	MSD*	8	11.15	4.52	5.62	9.61	31.43	3.17	39.06
	USTC-IAT-United	MS	9	10.27	3.97	6.34	10.76	33.61	3.40	38.73
	DJI&ZJU	D	**1**	**33.20**	33.12	**0.70**	1.63	**13.08**	**2.89**	33.52
	cv-challenge	D	3	33.08	**33.57**	0.87	1.35	16.52	2.90	21.76
	Pokemon	D	2	32.05	32.53	0.83	**1.26**	16.00	4.11	45.68
	imec-IDLab-UAntwerp	MS	4	22.49	29.53	1.06	1.59	23.38	4.38	**14.52**
Indoor	**Baseline**	S	6	21.11	28.96	1.04	1.51	22.77	4.60	37.09
	GMD	MS	5	20.25	29.52	1.03	1.48	23.37	3.96	35.68
	MonoViTeam	MSD*	8	19.82	28.62	0.97	1.42	20.91	3.63	43.46
	USTC-IAT-United	MS	9	19.81	28.93	1.02	1.49	21.83	5.19	69.50
	DepthSquad	D	7	19.18	28.34	1.09	1.61	23.24	5.14	44.02

M**=Monocular – **S**=Stereo – **D=Proxy Depth – **D**=Ground-truth Depth*

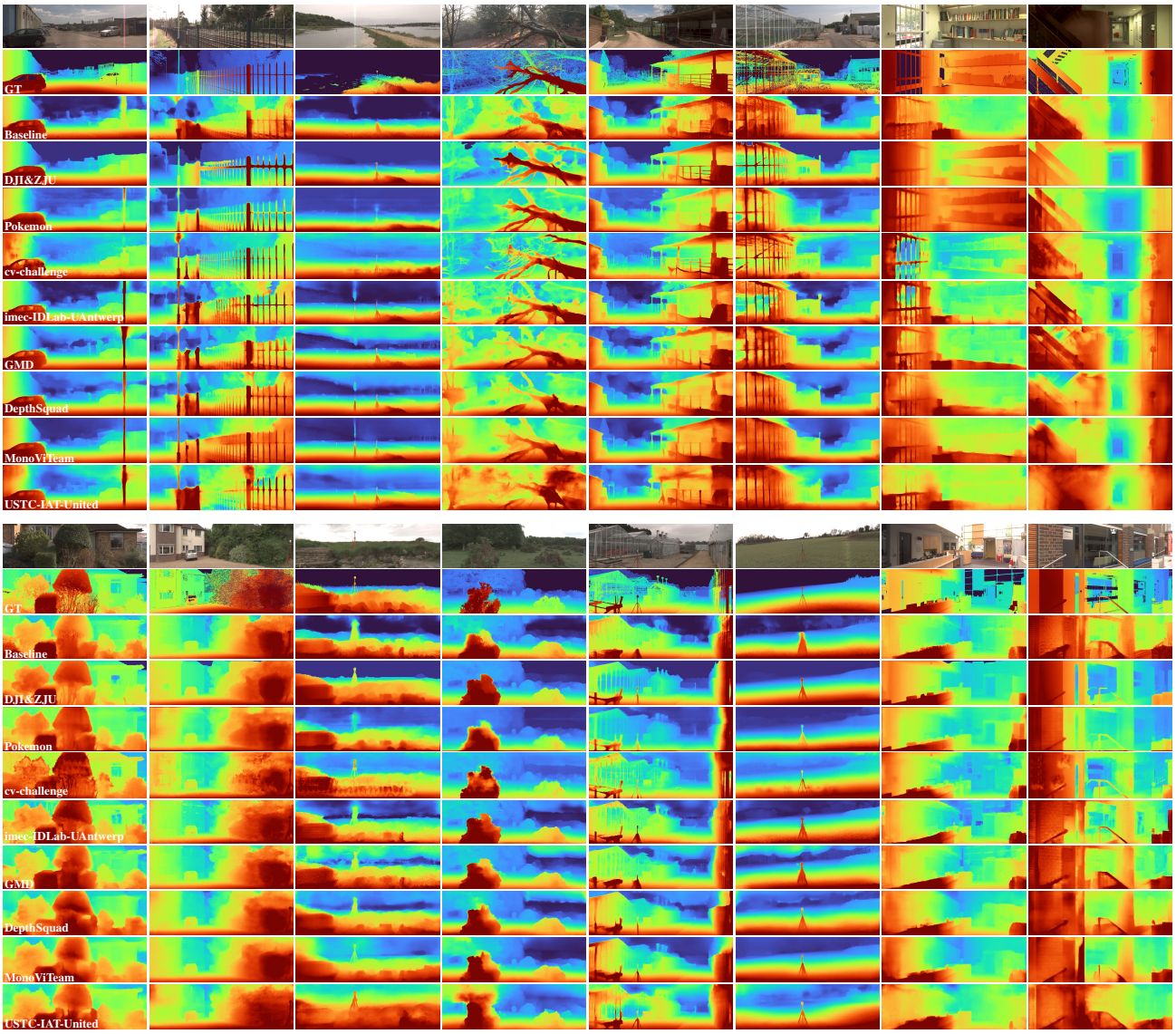

Figure 3. SYNS-Patches Depth Visualization. Best viewed in color and zoomed in. Most methods struggle with thin structures, such as branches and railings. Object boundaries are also characterized by "halos", caused by interpolation between foreground and background objects. Notable improvements can be seen in Natural and Agricultural scenes, where the top submissions provide much higher levels of detail than the baseline.

5.1. Quantitative Results

Table 2 shows the overall performance for each submission across the whole dataset, as well as each category. Each subset is ordered using F-Score performance. We additionally show the ranking order based on Overall F-Score for ease of comparison across categories.

The Overall top F-Score and AbsRel were obtained by Team DJI&ZJU, supervised using ground-truth depths from a collection of 10 datasets. This represents a relative improvement of 27.62% in F-Score (13.72% – Baseline) and 18% in AbsRel (29.66% – OPDAI) w.r.t. the first edition of the challenge [77]. The top-performing self-supervised method was Team imec-IDLab-UAntwerp, which leveraged improved pretrained encoders and deformable decoder convolutions. This submission provided relative improvements of 16.61% F-Score and 4.04% AbsRel over the first edition.

As expected, supervised approaches using ground-truth depth generally outperformed self-supervised approaches based on the photometric error. However, it is interesting to note that supervising a model with only automotive data (*e.g.* Team DepthSquad, trained on KEZ) was not sufficient to guarantee generalization to other scene

types. Meanwhile, as discussed in [78], improving the pretrained backbone (Teams imec-IDLab-UAntwerp & GMD) is one of the most reliable ways of increasing performance. Alternative contributions, such as training with proxy depths (MonoViTeam) or ensembling different architectures (USTC-IAT-United), can improve traditional image-based results but typically result in slightly inferior reconstructions.

The top submission (DJI&ZJU) consistently outperformed the other submissions across each scene category, demonstrating good generalization capabilities. However, Teams Pokemon & cv-challenge provided slightly better pointcloud reconstructions in Natural scenes. We theorize this might be due to the use of additional outdoor datasets, while DJI&ZJU primarily relies on automotive data. It is further interesting to note that self-supervised approaches such as Teams imec-IDLab-UAntwerp & GMD outperformed even some supervised methods in Urban reconstructions, despite training only on Kitti.

Finally, supervised methods provided the largest improvement in Indoor scenes, since self-supervised approaches were limited to urban driving datasets. DJI&ZJU relied on Taskonomy and DIML, Pokemon on ScanNet, SceneNet, NYUD-v2 and more and cv-challenge made use of ZoeDepth [9] pretrained on the DPT dataset collection [66]. This demonstrates the need for more varied training data in order to generalize across multiple scene types.

5.2. Qualitative Results

Figure 3 shows visualizations for each submission's predictions across varied scene categories. Generally, all approaches struggle with thin structures, such as the railings in images two and five or the branches in image four. Models vary between ignoring these thin objects (Baseline), treating them as solid objects (USTC-IAT-United) and producing inconsistent estimates (cv-challenge). Self-supervised methods are more sensitive to image artifacts (*e.g.* saturation or lens flare in images one and three) due to their reliance on the photometric loss. Meanwhile, supervised methods can be trained to be robust to the artifacts as long as the ground-truth is correct.

Object boundaries still present challenging regions, as demonstrated by the halos produced by most approaches. Even Team DJI&ZJU, while reducing the intensity of these halos, can sometimes produce over-pixelated boundaries. However, it is worth pointing out that many submissions significantly improve over the Baseline predictions [78]. In particular, Teams cv-challenge, imec-IDLab-UAntwerp & GMD show much greater levels of detail in Urban and Agricultural scenes, reflected by the improved Edge-Completion metric in Table 2. This is particularly impressive given the self-supervised nature of some of these submissions.

Unfortunately, self-supervised approaches show significantly inferior performance in Indoor settings, as they lack the data diversity to generalize. This can be seen by the fact that many self-supervised approaches produce incorrect scene geometry and instead predict ground-planes akin to outdoors scenes.

Images six, thirteen and sixteen highlight some interesting complications for monocular depth estimation. Transparent surfaces, such as the glass, are not captured when using LiDAR or photometric constraints. As such, most approaches ignore them and instead predict the depth for the objects behind them. However, as humans, we know that these represent solid surfaces and obstacles that cannot be traversed. It is unclear how an accurate supervision signal could be generated for these cases. This calls for more flexible depth estimation algorithms, perhaps relying on multi-modal distributions and discrete volumes.

6. Conclusions & Future Work

This paper has summarized the results for the second edition of MDEC. Most submissions provided significant improvements over the challenge baseline. Supervised submissions typically focused on increasing the data diversity during training, while self-supervised submissions improved the network architecture.

As expected, there is still a performance gap between these two styles of supervision. This is particularly the case in Indoor environments. This motivates the need for additional data sources to train self-supervised models, which are currently only trained on automotive data. Furthermore, accurate depth boundary prediction is still a highly challenging problem. Most methods frequently predicted "halos", representative of interpolation artifacts between the foreground and background.

Future challenge editions may introduce additional tracks for metric *vs*. relative depth prediction, as predicting metric depth is even more challenging. We hope this competition will continue to bring researchers into this field and strongly encourage any interested parties to participate in future editions of the challenge.

Acknowledgements

This work was partially funded by the EPSRC under grant agreements EP/S016317/1, EP/S016368/1, EP/S016260/1, EP/S035761/1.

References

[1] Wendy J Adams, James H Elder, Erich W Graf, Julian Leyland, Arthur J Lugtigheid, and Alexander Muryy. The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 6(1):35805, 2016. 1, 2, 5

[2] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip atten-

tion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023. 2, 5

[3] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 589–604. Springer, 2020. 4

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 4

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes–a diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 4

[6] Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escolano, and Miguel Cazorla. Uasol, a large-scale high-resolution outdoor stereo dataset. *Scientific data*, 6(1):162, 2019. 4

[7] 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. 2, 5

[8] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 480–496. Springer, 2022. 2

[9] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 4, 8

[10] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 2

[11] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 4

[12] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8001–8008, 2019. 2

[13] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[14] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 2

[15] Zeyu Cheng, Yi Zhang, and Chengkai Tang. Swindepth: Using transformers and multi-scale fusion for monocular-based depth estimation. *IEEE Sensors Journal*, 21(23):26912–26920, 2021. 2

[16] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. *arXiv preprint arXiv:2110.11590*, 2021. 2, 4

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4

[19] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 5

[20] Qi Dai, Vaishakh Patil, Simon Hecker, Dengxin Dai, Luc Van Gool, and Konrad Schindler. Self-supervised object motion and depth estimation from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[22] David Eigen and Rob Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In *International Conference on Computer Vision*, pages 2650–2658, 2015. 2, 4

[23] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11826–11835, 2019. 2

[24] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 2

[25] Ravi Garg, Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In *European Conference on Computer Vision*, pages 740–756, 2016. 1, 2, 4

[26] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 4

[27] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[28] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. *Conference on Computer Vision and Pattern Recognition*, pages 6602–6611, 2017. 2, 4, 5

[29] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. *International Conference on Computer Vision*, 2019-Octob:3827–3837, 2019. 2, 5

[30] Juan Luis Gonzalez Bello and Munchurl Kim. PLADE-Net: Towards Pixel-Level Accuracy for Self-Supervised Single-View Depth Estimation with Neural Positional Encoding and Distilled Matting Loss. In *Conference on Computer Vision and Pattern Recognition*, pages 6847–6856, 2021. 2

[31] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. 2

[32] Vitor Guizilini, Ambrus Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. *Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2020. 2, 4

[33] Akhil Gurram, Ahmet Faruk Tuna, Fengyi Shen, Onay Urfalioglu, and Antonio M López. Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12738–12751, 2021. 5

[34] Ankur Handa, Viorica Pătrăucean, Simon Stent, and Roberto Cipolla. Scenenet: An annotated model generator for indoor scene understanding. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5737–5743. IEEE, 2016. 4

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[36] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021. 4

[37] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. *arXiv preprint arXiv:2003.11172*, 2020. 4

[38] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 2, 4, 5

[39] Braden Hurl, Krzysztof Czarnecki, and Steven Waslander. Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2522–2529. IEEE, 2019. 4

[40] 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*, pages 2545–2557, 2021. 2

[41] Yasunori Ishii and Takayoshi Yamashita. Cutdepth: Edge-aware data augmentation in depth estimation. *arXiv preprint arXiv:2107.07684*, 2021. 5

[42] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. 2

[43] Adrian Johnston and Gustavo Carneiro. Self-Supervised Monocular Trained Depth Estimation Using Self-Attention and Discrete Disparity Volume. In *Conference on Computer Vision and Pattern Recognition*, pages 4755–4764, 2020. 2

[44] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. 4

[45] Maria Klodt and Andrea Vedaldi. Supervising the New with the Old: Learning SFM from SFM. In *European Conference on Computer Vision*, pages 713–728, 2018. 2

[46] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-Based Single-Image Depth Estimation Methods. In *European Conference on Computer Vision Workshops*, pages 331–348, 2018. 4

[47] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *International Conference on 3D Vision*, pages 239–248, 2016. 2

[48] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7286, 2022. 5

[49] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 663–678. Springer, 2019. 2

[50] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Mannequinchallenge: Learning the depths of moving people by watching frozen people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4229–4241, 2020. 2

[51] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

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

[53] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single

image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015. 2

[54] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 4

[55] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[56] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 4, 5

[57] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts ++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2624–2641, 2020. 2

[58] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation. *AAAI Conference on Artificial Intelligence*, 35(3):2294–2301, 2021. 2, 5

[59] 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. *Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 4

[60] 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. 2, 4, 5

[61] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 4

[62] Evin Pinar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2D to 3D: Re-thinking Benchmarking of Monocular Depth Prediction. *arXiv preprint*, 2022. 1, 4

[63] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022. 2

[64] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation. In *International Conference on Robotics and Automation*, pages 9250–9256, 2019. 2

[65] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the Uncertainty of Self-Supervised Monocular Depth Estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 3224–3234, 2020. 2

[66] 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. 2, 4, 8

[67] 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*, 2020. 2, 4

[68] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. 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*, pages 12240–12249, 2019. 2

[69] Haoyu Ren, Mostafa El-Khamy, and Jungwon Lee. Deep robust single image depth estimation neural network using scene understanding. In *CVPR Workshops*, volume 2, 2019. 2

[70] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017. 2

[71] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 4

[72] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5506–5514, 2016. 2

[73] Rui, Stückler Jörg, Cremers Daniel Yang Nan, and Wang. Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry. In *European Conference on Computer Vision*, pages 835–852, 2018. 2

[74] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-Metric Loss for Self-supervised Learning of Depth and Egomotion. In *European Conference on Computer Vision*, pages 572–588, 2020. 5

[75] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020. 4

[76] Jaime Spencer, Richard Bowden, and Simon Hadfield. DeFeat-Net: General monocular depth via simultaneous unsupervised representation learning. In *Conference on Computer Vision and Pattern Recognition*, pages 14390–14401, 2020. 2

[77] Jaime Spencer, C Stella Qian, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, Andrew J Schofield, James H Elder, Richard Bowden, Heng Cong, et al. The monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 623–632, 2023. 1, 2, 4, 7

[78] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *Transactions on Machine Learning Research*, 2022. Reproducibility Certification. 1, 2, 3, 4, 5, 6, 8

[79] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 2

[80] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, volume 97, pages 6105–6114, 2019. 4

[81] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity Invariant CNNs. *International Conference on 3D Vision*, pages 11–20, 2018. 2

[82] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 2

[83] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 4

[84] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *2019 International Conference on 3D Vision (3DV)*, pages 348–357. IEEE, 2019. 2, 4

[85] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 4

[86] Jamie Watson, Michael Firman, Gabriel Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. *International Conference on Computer Vision*, 2019-Octob:2162–2171, 2019. 2

[87] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 5

[88] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 2, 4

[89] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020. 4

[90] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021. 4

[91] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-Wise Attention-Based Network for Self-Supervised Monocular Depth Estimation. In *International Conference on 3D Vision*, pages 464–473, 2021. 2

[92] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 4

[93] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. In *Conference on Computer Vision and Pattern Recognition*, pages 1278–1289, 2020. 2

[94] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 4

[95] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2021. 4

[96] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 4

[97] 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. 4

[98] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[99] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2

[100] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3906–3915, 2022. 2, 4

[101] Weihao Yuan, Yazhan Zhang, Bingkun Wu, Siyu Zhu, Ping Tan, Michael Yu Wang, and Qifeng Chen. Stereo matching by self-supervision of multiscopic vision. In *2021*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5702–5709. IEEE, 2021. 5

[102] Pierluigi Zama Ramirez, Tosi Fabio, 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, and Shi Yin. 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. 2

[103] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 4

[104] Oliver Zendel, Angela Dai, Xavier Puig Fernandez, Andreas Geiger, Vladlen Koltun, Peter Kontschieder, Adam Kortylewski, Alina Kuznetsova, Tsung-Yi Lin, Torsten Sattler, Daniel Scharstein, Hendrik Schilling, Jonas Uhrig, and Jonas Wulff. Robust Vision Challenge 2022 — robustvision.net. http://robustvision.net/index.php. 2

[105] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian M. Reid. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. *Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018. 2

[106] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. *International Conference on 3D Vision*, 2022. 2, 5

[107] Hang Zhou, David Greenwood, and Sarah Taylor. Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. In *British Machine Vision Conference*, 2021. 2, 5

[108] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. *Conference on Computer Vision and Pattern Recognition*, pages 6612–6619, 2017. 2