The Third Monocular Depth Estimation Challenge

Jaime Spencer[1] Fabio Tosi[2] Matteo Poggi[2] Ripudaman Singh Arora[3] Chris Russell[4]
Simon Hadfield[5] Richard Bowden[5] GuangYuan Zhou[6] ZhengXin Li[7] Qiang Rao[6]
YiPing Bao[6] Xiao Liu[6] Dohyeong Kim[8] Jinseong Kim[8] Myunghyun Kim[8]
Mykola Lavreniuk[9] Rui Li[10] Qing Mao[10] Jiang Wu[10] Yu Zhu[10] Jinqiu Sun[10]
Yanning Zhang[10] Suraj Patni[11] Aradhye Agarwal[11] Chetan Arora[11] Pihai Sun[12]
Kui Jiang[12] Gang Wu[12] Jian Liu[12] Xianming Liu[12] Junjun Jiang[12] Xidan Zhang[13]
Jianing Wei[13] Fangjun Wang[13] Zhiming Tan[13] Jiabao Wang[14] Albert Luginov[15]
Muhammad Shahzad[15] Seyed Hosseini[16] Aleksander Trajcevski[16] James H. Elder[16]

Abstract

This paper discusses the results of the third edition of the Monocular Depth Estimation Challenge (MDEC). The challenge focuses on zero-shot generalization to the challenging SYNS-Patches dataset, featuring complex scenes in natural and indoor settings. As with the previous edition, methods can use any form of supervision, i.e. supervised or self-supervised. The challenge received a total of 19 submissions outperforming the baseline on the test set: 10 among them submitted a report describing their approach, highlighting a diffused use of foundational models such as Depth Anything at the core of their method. The challenge winners drastically improved 3D F-Score performance, from 17.51% to 23.72%.

1. Introduction

Monocular depth estimation (MDE) aims at predicting the distance from the camera to the points of the scene depicted by the pixels in the captured image. It is a highly ill-posed problem due to the absence of geometric priors usually available from multiple images. Nonetheless, deep learning has rapidly advanced this field and made it a reality, enabling results far beyond imagination.

[1]Independent [2]University of Bologna [3]Blue River Technology [4]Oxford Internet Institute [5]University of Surrey [6]ByteDance [7]University of Chinese Academy of Science [8]RGA Inc. [9]Space Research Institute NASU-SSAU, Kyiv, Ukraine [10]Northwestern Polytechnical University, Xi'an [11]Indian Institute of Technology, Delhi [12]Harbin Institute of Technology [13]Fujitsu [14]GuangXi University [15]University of Reading [16]York University

For years, most proposed approaches have been tailored to training and testing in a single, defined domain – e.g., automotive environments [33] or indoor settings [64] – often ignoring their ability to generalize to unseen environments. Purposely, the Monocular Depth Estimation Challenge (MDEC) in the last years has encouraged the community to delve into this aspect, by proposing a new benchmark for evaluating MDE models on a set of complex environments, comprising natural, agricultural, urban, and indoor settings. The dataset comes with a validation and a testing split, without any possibility of training/fine-tuning over it thus forcing the models to generalize.

While the first edition of MDEC [90] focused on benchmarking self-supervised approaches, the second [91] additionally opened the doors to supervised methods. During the former, the participants outperformed the baseline [30, 92] in all image-based metrics (AbsRel, MAE, RMSE), but could not improve pointcloud reconstructions [65] (F-Score). The latter, instead, brought new methods capable of outperforming the baseline on both aspects, establishing a new State-of-the-Art (SotA). The third edition of MDEC, detailed in this paper, ran in conjunction with CVPR2024, following the successes of the second one by allowing submissions of methods exploiting any form of supervision, *e.g.* supervised, self-supervised, or multi-task.

Following previous editions, the challenge was built around SYNS-Patches [1, 92]. This dataset was chosen because of the variegated diversity of environments it contains, including urban, residential, industrial, agricultural, natural, and indoor scenes. Furthermore, SYNS-Patches contains dense high-quality LiDAR ground-truth, which is very challenging to obtain in outdoor settings. This allows

for a benchmark that accurately reflects the real capabilities of each model, potentially free from biases.

While the second edition counted 8 teams outperforming the SotA baseline in either pointcloud- or image-based metrics, this year 19 submissions achieved this goal. Among these, 10 submitted a report introducing their approach, 7 of whose outperformed the winning team of the second edition. This demonstrates the increasing interest – and efforts – in MDEC.

In the remainder of the paper, we will provide an overview of each submission, analyze their results on SYNS-Patches, and discuss potential future developments.

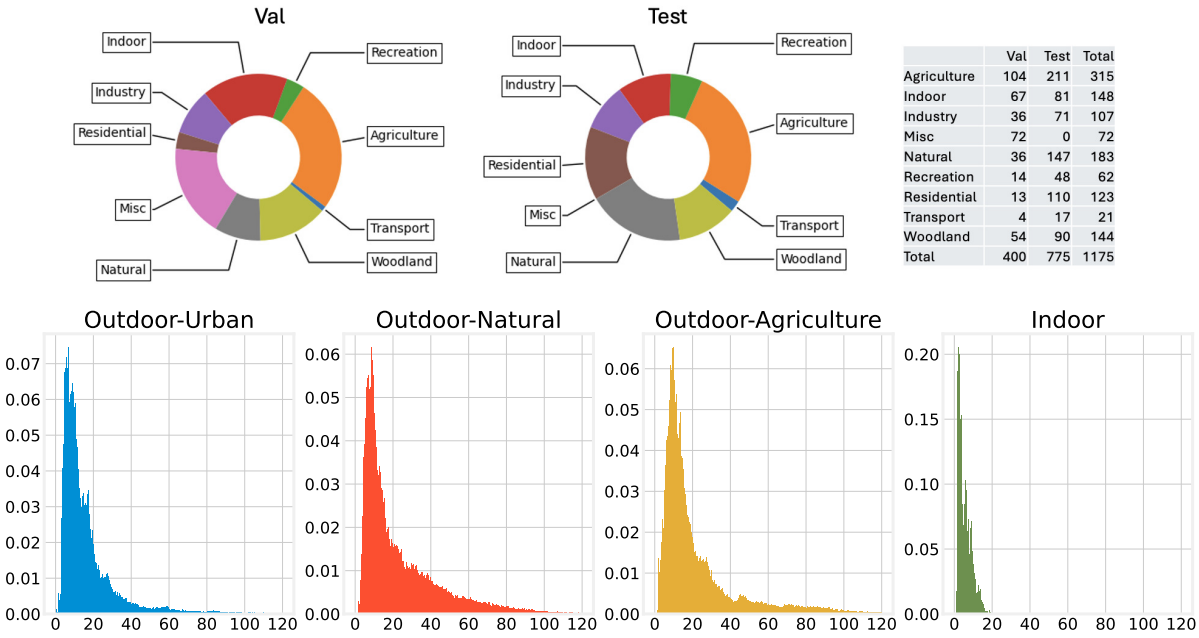
2. Related Work

Supervised MDE. Early monocular depth estimation (MDE) efforts utilized supervised learning, leveraging ground truth depth labels. Eigen et al. [26] proposed a pioneering end-to-end convolutional neural network (CNN) for MDE, featuring a scale-invariant loss and a coarse-to-fine architecture. Subsequent advancements incorporated structured prediction models such as Conditional Random Fields (CRFs) [54, 120] and regression forests [82]. Deeper network architectures [80, 109], multi-scale fusion [63], and transformer-based encoders [8, 16, 79] further enhanced performance. Alternatively, certain methods framed depth estimation as a classification problem [6, 7, 28, 51]. Novel loss functions were also introduced, including gradient-based regression [53, 104], the berHu loss [50], an ordinal relationship loss [14], and scale/shift invariance [80].

Self-Supervised MDE. To overcome the dependence on costly ground truth annotations, self-supervised methods were developed. Garg et al. [30], for the first time, proposed an algorithm based on view synthesis and photometric consistency across stereo image pairs, the importance of which for was extensively analyzed by Poggi et al. [74]. Godard et al. [34] introduced Monodepth, which incorporated differentiable bilinear interpolation [44], virtual stereo prediction, and a SSIM+L_1 reconstruction loss. Zhou et al. [130] presented SfM-Learner, which required only monocular video supervision by replacing the known stereo transform with a pose estimation network. Following the groundwork laid by these frameworks, subsequent efforts focused on refining the depth estimation accuracy by integrating feature-based reconstructions [89, 119, 124], semantic segmentation [122], adversarial losses [3], proxy-depth representations [5, 18, 48, 70, 83, 97, 107], trinocular supervision [75] and other constraints [9, 61, 103]. Other works focused on improving depth estimates at object boundaries [96, 99]. Moreover, attention has also been given to challenging cases involving dynamic scenarios during the training phase, which pose difficulties in providing accurate supervision signals for such networks. This has been addressed, for example, by incorporating uncertainty estimates [48, 73, 112], motion masks [11, 22, 37, 98], optical flow [59, 81, 118], or via the minimum reconstruction loss [35]. Finally, several architectural innovations, including 3D (un)packing blocks [38], position encoding [36], transformer-based encoders [2, 127], sub-pixel convolutions [71], progressive skip connections [60], and self-attention decoders [46, 110, 129], allowed further improvements. Among them, lightweight models tailored for real-time applications with memory and runtime constraints have also been developed [4, 19, 43, 68, 69, 72, 108].

Generalization and "In-the-Wild" MDE. Estimating depth in the wild refers to the challenging task of developing methods that can generalize to a wide range of unknown settings [14, 15]. Early works in this area focused on predicting relative (ordinal) depth [14, 15]. Nonetheless, the limited suitability of relative depth in many downstream contexts has driven researchers to explore affine-invariant depth estimation [53, 113]. In the affine-invariant setting, depth is estimated up to an unknown global offset and scale, offering a compromise between ordinal and metric representations. Researchers have employed various strategies to achieve generalization, including leveraging annotations from large datasets to train monocular depth models [79, 80, 111], including internet photo collections [53, 113], as well as from automotive LiDAR [33, 38, 42], RGB-D/Kinect sensors [17, 64, 95], structure-from-motion reconstructions [52, 53], optical flow/disparity estimation [80, 109], and crowd-sourced annotations [14]. However, the varying accuracy of these annotations may have impacted model performance, and acquiring new data sources remains a challenge, motivating the exploration of self-supervised approaches [116, 125]. For instance, KBR(++) [93, 94] leverage large-scale self-supervision from curated internet videos. The transition from CNNs to vision transformers has further boosted performance in this domain, as demonstrated by DPT (MiDaS v3) [79] and Omnidata [25]. Furthermore, a few works like Metric3D [114] and ZeroDepth [40] revisited the depth estimation by explicitly feeding camera intrinsics as additional input. A notable recent trend involves training generative models, especially diffusion models [29, 41, 88] for monocular depth estimation [24, 45, 47, 84, 85].

Adverse Weather and Transparent/Specular Surfaces. Existing monocular depth estimation networks have struggled under adverse weather conditions. Approaches have addressed low visibility [89], employed day-night branches using GANs [100, 126], utilized additional sensors [31], or faced trade-offs [101]. Recently, md4all [32] enabled robust performance across conditions without compromising ideal setting performance. Furthermore, estimating depth for transparent or mirror (ToM) surfaces posed a unique challenge [121, 123]. Costanzino et al. [21] is the only work dedicated to this, introducing novel datasets [77, 78]. Their



	Val	Test	Total
Agriculture	104	211	315
Indoor	67	81	148
Industry	36	71	107
Misc	72	0	72
Natural	36	147	183
Recreation	14	48	62
Residential	13	110	123
Transport	4	17	21
Woodland	54	90	144
Total	400	775	1175

Figure 1. SYNS-Patches Properties. Top: Distribution of images per category in the validation split and the test split respectively. Bottom: 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).

approach relied on segmentation maps or pre-trained networks, generating pseudo-labels by inpainting ToM objects and processing them with a pre-trained depth model [80], enabling fine-tuning of existing networks to handle ToM surfaces.

3. The Monocular Depth Estimation Challenge

The third edition of the Monocular Depth Estimation Challenge[1] was organized on CodaLab [67] as part of a CVPR2024 workshop. The development phase lasted four weeks, using the SYNS-Patches validation split. During this phase, the leaderboard was public but the usernames of the participants were anonymized. Each participant could see the results achieved by their own submission.

The final phase of the challenge was open for three weeks. At this stage, the leaderboard was completely private, disallowing participants to see their own scores. This choice was made to encourage the evaluation on the validation split rather than the test split and, together with the fact that all ground-truth depths were withheld, severely avoiding any possibility of overfitting over the test set by conducting repeated evaluations on it.

Following the second edition [91], any form of supervision was allowed, in order to provide a more comprehensive overview of the monocular depth estimation field as a whole. This makes it possible to better study the gap between different techniques and identify possible, future research directions. In this paper, we report results only for submissions that outperformed the baseline in any pointcloud-/image-based metric on the Overall dataset.

Dataset. The challenge takes place based on the SYNS-Patches dataset [1, 92], chosen due to the diversity of scenes and environments. A breakdown of images per category and some representative examples are shown in Figure 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 such dense ground-truth, depth boundaries were obtained using Canny edge-detection on the log-depth maps, allowing us to compute additional fine-grained metrics for these challenging regions. As outlined in [91, 92], the images were manually checked to remove dynamic object artifacts.

Evaluation. Participants were asked to provide the up-to-scale disparity prediction for each dataset image. The evaluation server bilinearly upsampled the predictions to the target resolution and inverted them into depth maps. Although self-supervised methods trained with stereo pairs and supervised methods using LiDAR or RGB-D data should be capable of predicting metric depth, in order to ensure comparisons are as fair as possible, the evaluation aligned any predictions with the ground-truth using the median depth. We set a maximum depth threshold of 100 meters.

Metrics. Following the first and second editions of the challenge [90, 91], we use a mixture of image-/pointcloud-/edge-based metrics. Image-based metrics are the most common (MAE, RMSE, AbsRel) and are computed using

[1] https://codalab.lisn.upsaclay.fr/competitions/17161

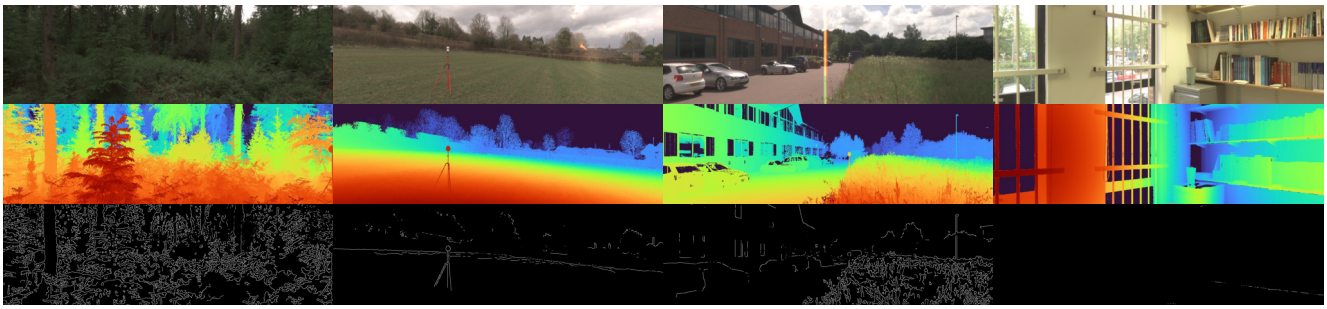


Figure 2. SYNS-Patches Dataset. We show samples from diverse scenes, including complex urban, natural, and indoor spaces. High-quality ground-truth depth covers about 78.20% of the image, from which depth boundaries are computed as Canny edges in log space.

pixel-wise comparisons between the predicted and ground-truth depth map. Pointcloud-based metrics [65] (F-Score, IoU, Chamfer distance) instead bring the evaluation in the 3D domain, evaluating the reconstructed pointclouds as a whole. Among these, we select 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 [49].

4. Challenge Submissions

We now highlight the technical details for each submission, as provided by the authors themselves. Each submission is labeled based on the supervision used, including ground-truth (**D**), proxy ground-truth (**D***), DepthAnything [111] pretraining (†) and monocular (**M**) or stereo (**S**) photometric support frames. Teams are numbered according to rankings.

Baseline – S

J. Spencer *j.spencermartin@surrey.ac.uk*
C. Russell *chris.russell@oii.ox.ac.uk*
S. Hadfield *s.hadfield@surrey.ac.uk*
R. Bowden *r.bowden@surrey.ac.uk*

Challenge organizers' submission from the first edition.

Network. ConvNeXt-B encoder [56] with a base Monodepth decoder [34, 62] from [92].

Supervision. Self-supervised with a stereo photometric loss [30] and edge-aware disparity smoothness [34].

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

Team 1: PICO-MR – †D*

G. Zhou *zhouguangyuan@bytedance.com*
Z. Li *lizhengxin17@mails.ucas.ac.cn*
Q. Rao *raoqiang@bytedance.com*
Y. Bao *baoyiping@bytedance.com*
X. Liu *liuxiao@foxmail.com*

Network. Based on Depth-Anything [111] with a BEiT384-L backbone, starting from the authors' weights pre-trained on 1.5M labeled images and 62M+ unlabeled images.

Supervision. The model is fine-tuned in a supervised manner, with proxy labels derived from stereo images. The final loss function integrates the SILog loss, SSIL loss, Gradient loss, and Random Proposal Normalization (RPNL) loss.

Training. The network was fine-tuned on the CityScapes dataset [20], resizing the input to 384×768 resolution, while keeping proxy labels at $1024 \times \times2048$ resolution. Random flipping is used to augment data, the batch size is set to 16 and the learning rate to 0.000161. The fine-tuning is carried out to predict metric depth and early stops at 4 epochs, a strategic choice to prevent overfitting and ensure the model's robustness to new data.

Team 3: RGA-Robot – †S

D. Kim *figure317@rgarobot.com*
J. Kim *jsk24@rgarobot.com*
M. Kim *wiseman218@rgarobot.com*

Network. It uses the Depth Anything [111] pre-trained model to estimate relative depth, accompanied by an auxiliary network to convert it into metric depth. This latter is NAFNet [13], processing the final feature maps and relative depth map predicted by the former model together with the input image.

Supervision. Self-supervised loss with two main terms: image reconstruction loss and smoothness loss. The former integrates perceptual loss with photometric loss as used in monodepth2 [35], with the former using a pre-trained VGG19 backbone [87], following a similar approach as in ESRGAN [106].

Training. The train is carried out on Kitti Eigen-Zhou with batch size 8 and learning rate $1e{-}4$ for 4 epochs. Only NAFNet is trained, while the Depth Anything model remains frozen.

Team 4: EVP++ – †D

M. Lavreniuk *nick_93@ukr.net*

Network. The architecture is based on Depth Anything [111], incorporating a VIT-L encoder [23] for feature extraction and the ZoeDepth metric bins module [8] as a de-

coder. This module computes per-pixel depth bin centers, which are linearly combined to produce metric depth.

Supervision. The models were trained in a supervised manner using ground-truth depth information obtained from various datasets, employing the SILog loss function.

Training. The models were trained on both indoor and outdoor data, respectively on the NYUv2 dataset [64] with an image size of 392×518, and on KITTI [33], Virtual KITTI 2 [10], and DIODE outdoor [102] with an image size of 518×1078. The batch size was set to 16, the learning rate to 0.000161, and the maximum depth to 10 for indoor scenes. For outdoor scenes, the batch size was set to 1, the learning rate to 0.00002, and the maximum depth to 80. Both models were trained for 5 epochs.

Team 6: 3DCreators – †D

R. Li *lirui.david@gmail.com*
Q. Mao *maoqing@mail.nwpu.edu.cn*
J. Wu *18392713997@mail.nwpu.edu.cn*
Y. Zhu *yuzhu@nwpu.edu.cn*
J. Sun *sunjinqiu@nwpu.edu.cn*
Y. Zhang *ynzhang@nwpu.edu.cn*

Network. An architecture made of two sub-networks. The first model consists of a pre-trained ViT-large backbone [23] from Depth Anything [111] and a ZoeDepth decoder [8]. The second is Metric3D [115], which uses ConvNext-Large [57] backbone and a LeRes decoder [117].

Supervision. The first network is fine-tuned with the KITTI dataset using SILog loss. The second network uses the released pre-trained weights trained by a diverse collection of datasets as detailed in [115].

Training. The first network is fine-tuned using batch size 16 for 5 epochs. At inference, test-time augmentation – i.e., color jittering and horizontal flipping – is used to combine the predictions by the two models: the same image is augmented 10 times and processed by the two models, then the predictions are averaged.

Team 7: visioniitd – D

S. Patni *suraj.patni@cse.iitd.ac.in*
A. Agarwal *aradhye.agarwal.cs520@cse.iitd.ac.in*
C. Arora *chetan@cse.iitd.ac.in*

Network. The model is ECoDepth [66], which provides effective conditioning for the MDE task to diffusion methods like stable diffusion. It is based on a Comprehensive Image Detail Embedding (CIDE) module which utilizes ViT embeddings of the image and subsequently transforms them to yield a semantic context vector. These embeddings are used to condition the pre-trained UNet backbone in Stable Diffusion, which produces hierarchical feature maps from its decoder. These are resized to a common dimension and passed to the Upsampling decoder and depth regressor to produce the final depth.

Supervision. Supervised training using the ground truth depth with SILog loss as the loss function with variance focus (λ) 0.85. Ground-truth depth is transformed as $\frac{1}{(1+x)}$.

Training. Trained on NYUv2 [64], KITTI [33], virtual KITTI v2 [10] for 25 epochs, with one-cycle learning rate (min: $3e{-}5$, max: $5e{-}4$) and batch size 32 on $8\times$ A100 GPUs.

Team 9: HIT-AIIA – †D

P. Sun *23s136164@stu.hit.edu.cn*
K. Jiang *jiangkui@hit.edu.cn*
G. Wu *gwu@hit.edu.cn*
J. Liu *hitcslj@hit.edu.cn*
X. Liu *csxm@hit.edu.cn*
J. Jiang *jiangjunjun@hit.edu.cn*

Network. It involves the pre-trained Depth Anything encoder and pre-trained CLIP model. The latter is introduced to calculate the similarity between the keywords 'indoor' or 'outdoor' and features extracted from the input image to route it to two, different instances of Depth Anything specialized on indoor or outdoor scenarios.

Supervision. Two instances of Depth Anything are fine-tuned on ground-truth labels, respectively from NYUv2 and KITTI for indoor and outdoor environments.

Training. The training resolution is 392×518 on NYUv2 and 384×768 on KITTI. The batch size is 16 and both instances are trained for 5 epochs.

Team 10: FRDC-SH – †D

X. Zhang *zhangxidan@fujitsu.com*
J. Wei *weijianing@fujitsu.com*
F. Wang *wangfangjun@fujitsu.com*
Z. Tan *zhmtan@fujistu.com*

Network. The depth network is the Depth Anything [111] pre-trained model – based on ZoeDepth [8] with a DPT_BEiT_L384 – and further fine-tuned.

Supervision. Trained on ground-truth depth, with SILog and Hyperbolic Chamfer Distance losses.

Training. The model is fine-tuned on NYU-v2 [64], 7Scenes [86], SUNRGBD [128], DIODE [102], KITTI [33], DDAD [39], and Argoverse [12] – without any resizing of the image resolution – for 20 epochs with batch size 32, a learning rate set to 1.61e-04, and a 0.01 weight decay.

Team 15: hyc123 – D

J. Wang *601533944@qq.com*

Network. Swin encoder [55] with skip connections and a decoder with channel-wise self-attention modules.

Supervision. Trained with ground truth depths, using a loss consisting of a combination of two L1 losses and an SSIM loss, weighted accordingly.

Training. The model was trained on Kitti Eigen-Zhou split using images of size 370×1224 for 100 epochs.

Team 16: ReadingLS – †MD*

A. Luginov *a.luginov@pgr.reading.ac.uk*
M. Shahzad *m.shahzad2@reading.ac.uk*

Network. The depth network is SwiftDepth [58], a compact model with only 6.4M parameters.

Supervision. Self-supervised monocular training with the minimum reconstruction loss [35], enhanced by offline knowledge distillation from a large MDE model [111].

Training. The model is trained in parallel on Kitti Eigen-Zhou and a selection of outdoor YouTube videos, similarly to KBR [93]. Both training and prediction are performed with the input resolution of 192×640. The teacher model [111] is not trained on either these datasets or SYNS-Patches.

Team 19: Elder Lab – D

S. Hosseini *smhh@yorku.ca*
A. Trajcevski *atrajcev@yorku.ca*
J. H. Elder *jelder@yorku.ca*

Network. An off-the-shelf semantic segmentation model [105] is used at first to segment the image. Then, the depth of pixels on the ground plane is estimated by predicting the camera angle from the height of the highest pixel on the ground. Then, depth is propagated vertically for pixels above the ground, while the Manhattan frame is estimated with [76] to identify both Manhattan and non-Manhattan segments in the image and propagate depth along them in 3D space. Finally, the depth map is completed according to heat equations [27], with pixels for which depth has been already estimated imposing forcing conditions, while semantic boundaries and the image frame impose reflection boundary conditions.

Supervision. Ground-truth depth is used for training three kernel regression models.

Training. Three simple statistical models are trained on CityScapes [20] and NYUv2 [64]: 1) A kernel regression model to estimate ground elevation angle from the vertical image coordinate of the highest observed ground pixel. The ground truth elevation angle is computed by fitting a plane (constrained to have zero roll) to the ground truth ground plane coordinates; 2) A kernel regression model to estimate the depth of ground pixels from their vertical coordinate, conditioned on semantic class; 3) median depth of non-ground pixels in columns directly abutting the bottom of the image frame, conditioned on semantic class.

5. Results

Submitted methods were evaluated on the testing split of SYNS-Patches [1, 92]. Participants were allowed to submit methods without any restriction on the supervision or the predictions by the model, which can be either relative or metric. Accordingly, to ensure a fair comparison among the methods, the submitted predictions are aligned to ground-truth depths according to median depth scaling.

5.1. Quantitative Results

Table 1 highlights the results of this third edition of the challenge, with the top-performing techniques, ordered using F-Score performance, achieving notable improvements over the baseline method. A first, noteworthy observation is the widespread adoption of the Depth Anything model [111], pre-trained on 62M of images, as the backbone architecture by the leading teams, including PICO-MR, RGA-Robot, EVP++, 3DCreators, HIT-AIIA, FRDC-SH, and ReadingLS, demonstrating its effectiveness and versatility.

Specifically, Team PICO-MR, which secured the top position on the leaderboard, achieved an F-score of 23.72, outperforming the baseline method by a remarkable 72.9%. This represents a significant improvement over the previous state-of-the-art method, DJI&ZJU, which achieved an F-score of 17.51 in the "The Second Monocular Depth Estimation Challenge" [91]. In particular, Team PICO-MR's result shows a 35.5% increase in performance compared to DJI&ZJU, highlighting the rapid progress made in monocular depth estimation within a relatively short period. This improvement can be also clearly observed in the other metrics considered, both accuracy and error – notably, achieving the second absolute results on F-Edges, MAE, and RMSE. Their success can be attributed to the fine-tuning of the Depth Anything model on the Cityscapes dataset using a combination of SILog, SSIL, Gradient, and Random Proposal Normalization losses, as well as their strategic choice of fine-tuning for a few epochs to prevent overfitting and ensure robustness to unseen data.

Team RGA-Robot, in the third place, achieved an F-score of 22.79, outperforming the baseline by 66.1%. Their novel approach of augmenting the Depth Anything model, maintained frozen, with an auxiliary network, NAFNet, to convert relative depth predictions into metric depth, combined with self-supervised loss terms, shows the effectiveness of this approach in enhancing depth accuracy. In terms of the F-Edges metric, this method achieves the best result.

Team EVP++, ranking fourth, achieved an F-score of 20.87, surpassing the baseline by 52.1%. Their approach involved training the Depth Anything model on both indoor and outdoor datasets, adapting image sizes, batch sizes, and learning rates to each scenario, and highlighting the importance of tailoring model parameters to the specific characteristics of the target environment. This strategy notably improves the results in terms of standard 2D error metrics, yielding the lowest MAE, RMSE, and AbsRel.

Several other teams also surpassed both the baseline method and the previous state-of-the-art from the second edition of the challenge. Team 3DCreators achieved an

Table 1. SYNS-Patches Results. We provide metrics across the whole test split of the dataset. Top-performing entries generally leverage the pre-trained Depth Anything [111] model. Only a few methods use self-supervised losses or proxy depth labels.

	Train	Rank	F↑	F-Edges↑	MAE↓	RMSE↓	AbsRel↓	Acc-Edges↓	Comp-Edges↓
PICO-MR	†D*	**1**	**23.72**	11.01	3.78	6.61	21.24	3.90	4.45
Anonymous	?	2	23.25	10.78	3.87	6.70	21.70	3.59	9.86
RGA-Robot	†S	3	22.79	**11.52**	5.21	9.23	28.86	4.15	0.90
EVP++	†D	4	20.87	10.92	**3.71**	**6.53**	**19.02**	2.88	6.77
Anonymous	?	5	20.77	9.96	4.33	7.83	27.80	3.45	13.25
3DCreators	†D	6	20.42	10.19	4.41	7.89	23.94	3.61	5.80
visioniitd	D	7	19.07	9.92	4.53	7.96	23.27	3.26	8.00
Anonymous	?	8	18.60	9.43	3.92	7.16	20.12	2.89	15.65
HIT-AIIA	†D	9	17.83	9.14	4.11	7.73	21.23	2.95	17.81
FRDC-SH	†D	10	17.81	9.75	5.04	8.92	24.01	3.16	14.16
Anonymous	?	11	17.57	9.13	4.28	8.36	23.35	3.18	20.66
Anonymous	?	12	16.91	9.07	4.14	7.35	22.05	3.24	18.52
Anonymous	?	13	16.71	9.25	5.48	11.05	34.20	**2.57**	18.04
Anonymous	?	14	16.45	8.89	5.29	10.53	33.67	2.60	18.73
hyc123	D	15	15.92	9.17	8.25	13.88	43.88	4.11	**0.74**
ReadingLS	†MD*	16	14.81	8.14	5.01	8.94	29.39	3.28	30.28
Baseline	S	17	13.72	7.76	5.56	9.72	32.04	3.97	21.63
Anonymous	?	18	13.71	7.55	5.49	9.44	30.74	3.61	18.36
Anonymous	?	19	11.90	8.08	6.33	10.89	30.46	2.99	33.63
Elder Lab	D	20	11.04	7.09	8.76	15.86	63.32	3.22	40.61

M**=Monocular – **S**=Stereo – **D=Proxy Depth – **D**=Ground-truth Depth – †=Pre-trained Depth Anything model*

F-score of 20.42, outperforming the baseline by 48.8% by fine-tuning and combining predictions from the Depth Anything model and Metric3D. Team visioniitd follows surpassing the baseline using ECoDepth, which conditions Stable Diffusion's UNet backbone with Comprehensive Image Detail Embeddings. Team HIT-AIIA and FRDC-SH also achieved notable improvements, with F-score of 17.83 and 17.81, respectively, using specialized model instances and fine-tuning on diverse datasets.

Finally, the remaining teams outperformed the baseline either on the F-score or any of the other metrics, yet not surpassing the winner of the previous edition. Team hyc123, with an F-score of 15.92, outperformed the baseline by 16.0% using a Swin encoder with skip connections and a decoder with channel-wise self-attention modules, while Team ReadingLS outperforms the baseline by distilling knowledge from Depth Anything to a lightweight network based on SwiftDepth, further improved using minimal reconstruction loss during training. Finally, Team Elder Lab employed an off-the-shelf semantic segmentation model and estimated depth using techniques such as predicting camera angle, propagating depth along Manhattan and non-Manhattan segments, and completing the depth map using heat equations. They achieved an F-score of 11.04, 19.5% lower than the baseline score of 13.72, yet they obtained 3.22 Acc-Edge, beating the baseline.

5.2. Qualitative Results

Figure 3 provides qualitative results for the depth predictions of each submission. A notable trend among the top-performing teams, such as PICO-MR, RGA-Robot, EVP++, and 3DCreators, is the adoption of the Depth Anything model as a backbone architecture. While Depth Anything represents the current state-of-the-art in monocular depth estimation, the qualitative results highlight that there are still significant challenges in accurately estimating depth, particularly for thin structures in complex outdoor scenes. This is evident in columns 2, 4, 5, and 6 of Figure 3, where objects like trees and branches are not well-recovered, despite the impressive quantitative performance of these methods as shown in Table 1. Interestingly, Team visioniitd, which employs a novel approach called ECoDepth to condition Stable Diffusion's UNet backbone with Comprehensive Image Detail Embeddings, demonstrates a remarkable ability to estimate depth for thin structures. Yet, they are outperformed quantitatively by other methodologies, suggesting that estimating depth in smooth regions may be more challenging than in thin structures.

The qualitative results also reveal some method-specific anomalies. For instance, hyc123 exhibits salt-and-pepper noise artifacts, while Elder Lab's method, which ranks last, generates overly smooth depth maps that lose important scene objects. These anomalies highlight the importance of developing robust techniques that can handle diverse scene characteristics. Grid-like artifacts are observed in the predictions of top-performers PICO-MR and RGA-Robot, particularly in regions where the network seems uncertain about depth estimates. This suggests that further improvements in network architecture and training strategies may be necessary to mitigate these artifacts.

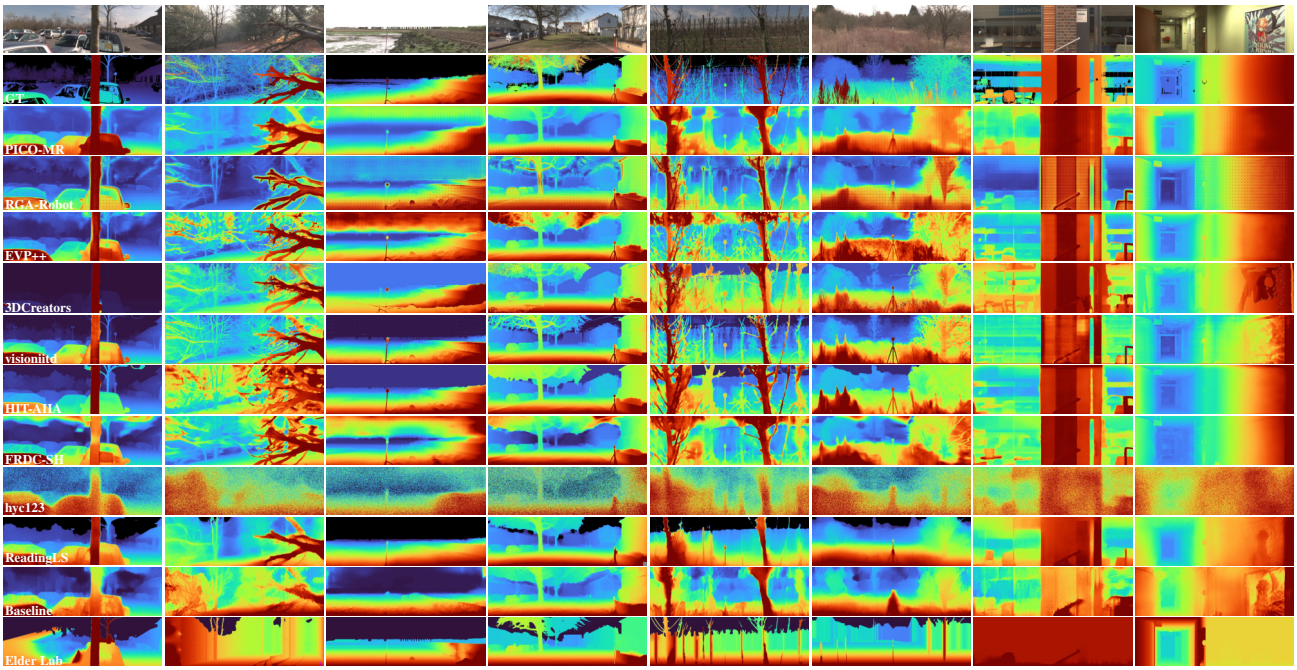


Figure 3. SYNS-Patches Depth Visualization. Best viewed in color and zoomed in. Methods are ranked based on their F-Score in Table 1. We can appreciate how thin structures still represent one of the hardest challenges to any method, such as branches and railings, for instance. Near depth discontinuities, most approaches tend to produce "halos", interpolating between foreground and background objects and thus failing to perceive sharp boundaries. Nonetheless, most methods expose higher level of detail compared to the baseline.

The indoor scenario in the last column shows the strong performance of methods like PICO-MR, EVP++, HIT-AIIA, and FRDC-SH in estimating scene structure. This can be attributed to their use of large-scale pre-training, fine-tuning on diverse datasets, and carefully designed loss functions that capture both global and local depth cues.

However, all methods still exhibit over-smoothing issues at depth discontinuities, manifesting as halo effects. While they outperform the baseline in this regard, likely due to their supervised training with ground truth or proxy labels, there remains significant room for improvement.

A notable limitation across all methods is the inability to effectively estimate depth for non-Lambertian surfaces, such as glass or transparent objects. This is evident in the penultimate right column and the first column, corresponding to the windshield. The primary reason for this limitation is the lack of accurate supervision for such surfaces in the training data, highlighting the need for novel techniques and datasets that explicitly address this challenge.

In conclusion, the qualitative results provide valuable insights into the current state of monocular depth estimation methods. While the adoption of large-scale pre-training and carefully designed architectures has led to significant improvements, challenges persist in accurately estimating depth for thin structures, smooth regions, and non-Lambertian surfaces. Addressing these limitations through novel techniques, improved training strategies, and diverse datasets will be crucial for further advancing this field.

6. Conclusions & Future Work

This paper has summarized the results for the third edition of MDEC. Over the various editions of the challenge, we have seen a drastic improvement in performance, showcasing MDE – in particular real-world generalization – as an exciting and active area of research.

With the advent of the first foundational models for MDE during the last months, we observed a diffused use of frameworks such as Depth Anything [111]. This ignited a major boost to the results submitted by the participants, with a much higher impact compared to the specific kind of supervision chosen for the challenge. Nonetheless, as we can appreciate from the qualitative results, any methods still struggle to accurately predict fine structures and discontinuities, hinting that there is still room for improvement despite the massive amount of data used to train Depth Anything.

We hope MDE will continue to attract new researchers and practitioners to this field and renew our invitation to participate in future editions of the challenge.

Acknowledgments. 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.

[2] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023.

[3] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[4] Filippo Aleotti, Giulio Zaccaroni, Luca Bartolomei, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Real-time single image depth perception in the wild with handheld devices. *Sensors*, 21(1):15, 2020.

[5] Lorenzo Andraghetti, Panteleimon Myriokefalitakis, Pier Luigi Dovesi, Belen Luque, Matteo Poggi, Alessandro Pieropan, and Stefano Mattoccia. Enhancing self-supervised monocular depth estimation with traditional visual odometry. In *2019 International Conference on 3D Vision (3DV)*, pages 424–433. IEEE, 2019.

[6] 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.

[7] 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.

[8] 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.

[9] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[10] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

[11] 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.

[12] Ming-Fang Chang, John W 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022.

[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.

[15] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. OASIS: A large-scale dataset for single image 3d in the wild. In *CVPR*, 2020.

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

[17] 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.

[18] Hyesong Choi, Hunsang Lee, Sunkyung Kim, Sunok Kim, Seungryong Kim, Kwanghoon Sohn, and Dongbo Min. Adaptive confidence thresholding for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12808–12818, 2021.

[19] Antonio Cipolletta, Valentino Peluso, Andrea Calimera, Matteo Poggi, Fabio Tosi, Filippo Aleotti, and Stefano Mattoccia. Energy-quality scalable monocular depth estimation on low-power cpus. *IEEE Internet of Things Journal*, 9(1):25–36, 2021.

[20] 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.

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

[22] 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.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[24] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023.

[25] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. pages 10786–10796, 2021.

[26] 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.

[27] James H Elder. Are edges incomplete? *International Journal of Computer Vision*, 34(2):97–122, 1999.

[28] 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.

[29] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024.

[30] 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.

[31] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 751–760. IEEE, 2021.

[32] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust monocular depth estimation under challenging conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[40] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. 2023.

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[42] 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.

[43] Lam Huynh, Phong Nguyen, Jiří Matas, Esa Rahtu, and Janne Heikkilä. Lightweight monocular depth with a novel neural architecture search method. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3643–3653, 2022.

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

[45] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. 2023.

[46] 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.

[47] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition*, 2024. CVPR.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep con-

volutional 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.

[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.

[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.

[57] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. 2022.

[58] Albert Luginov and Ilya Makarov. Swiftdepth: An efficient hybrid cnn-transformer model for self-supervised monocular depth estimation on mobile devices. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 642–647, 2023.

[59] 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.

[60] 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.

[61] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. *Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[62] 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.

[63] 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.

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

[65] 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.

[66] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2024.

[67] 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.

[68] Valentino Peluso, Antonio Cipolletta, Andrea Calimera, Matteo Poggi, Fabio Tosi, Filippo Aleotti, and Stefano Mattoccia. Monocular depth perception on microcontrollers for edge applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1524–1536, 2021.

[69] Valentino Peluso, Antonio Cipolletta, Andrea Calimera, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Enabling energy-efficient unsupervised monocular depth estimation on armv7-based platforms. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1703–1708. IEEE, 2019.

[70] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15560–15569, 2021.

[71] 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.

[72] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5848–5854. IEEE, 2018.

[73] 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.

[74] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5314–5334, 2021.

[75] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning Monocular Depth Estimation with Unsupervised Trinocular Assumptions. In *International Conference on 3D Vision*, pages 324–333, 2018.

[76] Yiming Qian, Srikumar Ramalingam, and James H Elder. Ls3d: Single-view gestalt 3d surface reconstruction from manhattan line segments. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 399–416. Springer, 2019.

[77] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: a benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[78] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 21168–21178, 2022.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *arXiv preprint arXiv:2306.01923*, 2023.

[85] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023.

[86] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013.

[87] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[88] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[89] 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.

[90] 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.

[91] Jaime Spencer, C Stella Qian, Michaela Trescakova, Chris Russell, Simon Hadfield, Erich W Graf, Wendy J Adams, Andrew J Schofield, James Elder, Richard Bowden, et al. The second monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3063–3075, 2023.

[92] 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.

[93] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15768–15779, October 2023.

[94] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. *arXiv preprint arXiv:2403.01569*, 2024.

[95] 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.

[96] Lior Talker, Aviad Cohen, Erez Yosef, Alexandra Dana, and Michael Dinerstein. Mind the edge: Refining depth edges in sparsely-supervised monocular depth estimation. 2024. CVPR.

[97] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. *Conference on Computer Vision and Pattern Recognition*, 2019-June:9791–9801, 2019.

[98] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4654–4665, 2020.

[99] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8942–8952, 2021.

[100] Madhu Vankadari, Sourav Garg, Anima Majumder, Swagat Kumar, and Ardhendu Behera. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 443–459. Springer, 2020.

[101] Madhu Vankadari, Stuart Golodetz, Sourav Garg, Sangyun Shin, Andrew Markham, and Niki Trigoni. When the sun goes down: Repairing photometric losses for all-day depth estimation. In *Conference on Robot Learning*, pages 1992–2003. PMLR, 2023.

[102] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019.

[103] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning Depth from Monocular Videos Using Direct Methods. *Conference on Computer Vision and*

Pattern Recognition, pages 2022–2030, 2018.

[104] 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.

[105] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023.

[106] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

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

[108] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019.

[109] 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.

[110] 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.

[111] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[112] 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.

[113] 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.

[114] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. 2023.

[115] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023.

[116] 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.

[117] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2021.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. 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*, pages 298–313. Springer, 2019.

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

[124] 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.

[125] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. 35, 2022.

[126] Chaoqiang Zhao, Yang Tang, and Qiyu Sun. Unsupervised monocular depth estimation in highly complex environments. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1237–1246, 2022.

[127] 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 Confer-*

ence on 3D Vision, 2022.

[128] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

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

[130] 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.