The Fourth Monocular Depth Estimation Challenge

Anton Obukhov[†1] Matteo Poggi[†2] Fabio Tosi[†2] Ripudaman Singh Arora[†3]
Jaime Spencer[†4] Chris Russell[†5] Simon Hadfield[†6] Richard Bowden[†6]
Shuaihang Wang[7] Zhenxin Ma[7] Weijie Chen[7] Baobei Xu[7] Fengyu Sun[7] Di Xie[7]
Jiang Zhu[7] Mykola Lavreniuk[8] Haining Guan[9] Qun Wu[9] Yupei Zeng[9] Chao Lu[9]
Huanran Wang[9] Guangyuan Zhou[4] Haotian Zhang[4] Jianxiong Wang[4] Qiang Rao[4]
Chunjie Wang[4] Xiao Liu[4] Zhiqiang Lou[10] Hualie Jiang[10] Yihao Chen[10] Rui Xu[10]
Minglang Tan[10] Zihan Qin[11] Yifan Mao[11] Jiayang Liu[11] Jialei Xu[11] Yifan Yang[11]
Wenbo Zhao[11] Junjun Jiang[11] Xianming Liu[11] Mingshuai Zhao[12] Anlong Ming[12]
Wu Chen[12] Feng Xue[13] Mengying Yu[12] Shida Gao[12] Xiangfeng Wang[12]
Gbenga Omotara[14] Ramy Farag[14] Jacket Demby's[14] Seyed Mohamad Ali Tousi[14]
Guilherme N. DeSouza[14] Tuan-Anh Yang[15,16] Minh-Quang Nguyen[15,16]
Thien-Phuc Tran[15,16] Albert Luginov[17] Muhammad Shahzad[17]

Abstract

This paper presents the results of the fourth edition of the Monocular Depth Estimation Challenge (MDEC), which focuses on zero-shot generalization to the SYNS-Patches benchmark, a dataset featuring challenging environments in both natural and indoor settings. In this edition, we revised the evaluation protocol to use least-squares alignment with two degrees of freedom to support disparity and affine-invariant predictions. We also revised the baselines and included popular off-the-shelf methods: Depth Anything v2 and Marigold. The challenge received a total of 24 submissions that outperformed the baselines on the test set; 10 of these included a report describing their approach, with most leading methods relying on affine-invariant predictions. The challenge winners improved the 3D F-Score over the previous edition's best result, raising it from 22.58% to 23.05%.

† Challenge Organizers

[1]Huawei Research Zürich [2]University of Bologna [3]Blue River Technology [4]Independent Researcher [5]Oxford Internet Institute [6]University of Surrey [7]Hikvision Research Institute [8]Space Research Institute NASU-SSAU, Kyiv, Ukraine [9]Megvii [10]Insta360 [11]Harbin Institute of Technology [12]Beijing University of Posts and Telecommunications [13]University of Trento [14]University of Missouri [15]University of Science, VNU-HCM, Vietnam [16]Vietnam National University, Ho Chi Minh, Vietnam [17]University of Reading

1. Introduction

Monocular depth estimation (MDE) is an important low-level vision task with applications in fields such as augmented reality, robotics, and autonomous vehicles. In 2024, the field was dominated by generative approaches, with DepthAnything [118, 119] representing the transformer-based solution and Marigold [49] being a denoising diffusion model based on the popular Text-to-Image LDM [85].

The MDE task aims to recover a notion of depth for each pixel of a single input image. This could take, for example, the form of a metric depth, where each value is a metric distance from the corresponding visible object in the scene to the camera capturing it. More often than that, disparity or inverse depth encodes reciprocal values of the metric depth up to some unknown scale, which is closely related to stereo vision systems. Another universal representation is affine-invariant depth, in which each value represents a linear interpolation between the scene's near and far planes.

The absence of multi-view geometric cues within a single image makes MDE highly ill-posed and, consequently, one of the longest-standing challenges in computer vision. Nevertheless, the emergence of deep learning has imbued this research with learned visual cues, enabling the development of increasingly accurate solutions year after year.

Early attempts to tackle this task were confined within single, limited domains, such as indoor environments [65] or

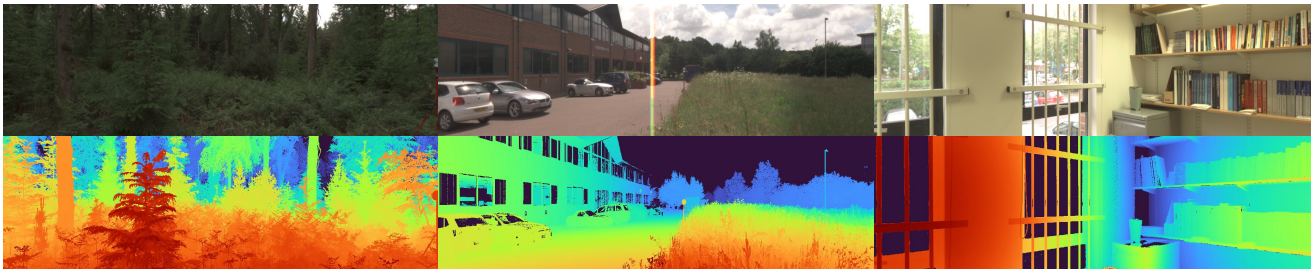

Figure 1. SYNS-Patches Dataset. The dataset contains samples from diverse scenes, including complex urban, natural, and indoor spaces. High-quality ground-truth depth measurements cover ∼78.20% of pixels; edges and object boundary annotations are also available.

urban driving scenarios [32], resulting in highly-specialized solutions that failed to generalize beyond their specific domain of expertise. To address this limitation, in the last three years, alongside to the development of more advanced frameworks and evaluation settings [81, 82], the Monocular Depth Estimation Challenge (MDEC) has pushed the research community toward more general models capable of cross-domain generalization, building upon a unique benchmark featuring a set of complex environments, comprising natural, agricultural, urban, and indoor settings. This benchmark includes a validation and a testing split, while not providing any training set – thus necessitating inherent generalization capabilities from all methods.

Specifically, SYNS-Patches [1, 96] was selected for this purpose, offering a diverse collection of environments, including urban, residential, industrial, agricultural, natural, and indoor scenes, all annotated with dense high-quality LiDAR ground-truth, which is typically very challenging to acquire in outdoor settings. This key characteristic, showcased in Figure 1, allows for a fair evaluation of MDE models, potentially free from some of the biases affecting popular benchmarks [32].

The initial three editions of MDEC [99] benchmarked self-supervised approaches first [94], then expanded [95, 99] to include supervised methods. All approaches were evaluated on both image-based and pointcloud-based metrics after the predicted depth maps were aligned with ground truth through median scaling [140]. The fourth edition of MDEC, detailed in this paper, conducted in conjunction with CVPR'2025, refines the evaluation protocol used by the previous editions, primarily by introducing a robust least-squares alignment between predictions and ground-truth depth maps to adhere to the latest standards in the field [81, 82]. Additionally, given the recent advent of the first foundational models for depth estimation, we upgraded from the original baseline [29, 96] used in the first three editions with more contemporary models [49, 119], that better represent the latest advances in the field.

The availability of these models represents a strong starting point for any participant willing to compete in our challenge, which has already markedly increased overall participation in the third edition [99]. This trend continued this year, with 24 teams surpassing the State-of-the-Art (SotA) baseline in either pointcloud- or image-based metrics – compared to 19 teams outperforming the old baseline last year. Among these competitors, 10 submitted a report detailing their approach; however, only a single team managed to outperform the winning team of the third edition. In the subsequent sections, we provide concise overviews of each submission, examine their performance on SYNS-Patches, and explore promising future research directions.

2. Related Work

2.1. Single Domain MDE

Supervised Methods. The early advancements in monocular depth estimation primarily relied on single-domain supervised learning approaches with ground truth depth annotations. Eigen et al. [26] introduced the first convolutional neural network for depth estimation, featuring a coarse-to-fine architecture and scale-invariant loss function. This inspired subsequent developments incorporating structured prediction models, including Conditional Random Fields (CRFs) [56, 129] and regression forests [86]. The field witnessed significant performance improvements through architectural innovations such as deeper networks [82, 116], multi-scale fusion techniques [64], and transformer-based encoders [8, 16, 81]. Some works reframed depth estimation as a classification problem [6, 7, 27, 53], while others focused on developing novel loss functions, including gradient-based regression [55, 111], the berHu loss [52], ordinal relationship loss [14], and affine-invariant mechanisms [82].

Self-Supervised Methods. Advances in supervised methods are bound to extensive and costly data collection and annotation campaigns. Years ago, the limited availability of labeled datasets ignited the development of self-supervised approaches. These techniques replace the supervision from depth labels with image-reconstruction losses, given the availability of a second image [29] of the same scene. This paradigm consists of exploiting image warping as supervision to the model and can be implemented either by using paired images collected through a calibrated stereo

setup [29, 33, 77], or a single video collected by a moving camera [140]. Following these pioneering frameworks, researchers have pursued a number of directions to improve the accuracy of depth estimation. Some integrated feature-based reconstructions [93, 128, 134], semantic segmentation [131], adversarial losses [3], and proxy-depth representations [5, 18, 50, 71, 87, 102, 114]. Others explored trinocular supervision [78] and additional geometric constraints [9, 62, 110]. Significant effort was also directed toward improving depth estimates at object boundaries [101, 104]. The presence of moving objects in videos represents a serious concern for self-supervised methods. Some approaches dealt with it by using uncertainty [50, 76, 121], predicting motion masks [12, 22, 36, 103], incorporating optical flow estimation [60, 83, 127], or designing robust losses [34]. In a concurrent track, many works introduced improved architectural designs: on the one hand, to pursue higher accuracy [2, 35, 37, 47, 61, 74, 117, 138, 139], on the other hand, to deploy compact and efficient models [4, 19, 45, 69, 70, 75, 115] that could bring self-supervised depth estimation closer to the deployment in real-time applications.

2.2. Generalization and "In-the-Wild" MDE

Estimating depth in challenging, unconstrained environments demands methods that generalize to diverse, unknown settings [14, 15]. While the previously outlined methods trained on single datasets struggled with generalization, newer approaches addressed this limitation by incorporating diverse data sources including large-scale datasets [81, 82, 118], internet photo collections [55, 122], automotive LiDAR [32, 37, 44], RGB-D sensors [17, 65, 100], structure-from-motion [54, 55], optical flow [82, 116], and crowd-sourced annotations [14]. Among them, self-supervised approaches [125, 135] have also received increasing attention, with notable works like KBR(++) [97, 98] leveraging curated internet videos.

Furthermore, architectural improvements significantly advanced performance in this domain. The transition from CNNs to transformer-based models yielded substantial improvements, as demonstrated by DPT (MiDaS v3) [81] and Omnidata [25]. A significant turning point has been the development of foundational models such as Depth Anything [118] and its successor Depth Anything v2 [119], which use large unlabeled data ($\tilde{6}$2M images) to produce highly robust and generalizable depth estimators.

Another breakthrough emerged with generative approaches, particularly diffusion models [40, 92]. Early applications include Ji et al. [46], Duan et al. [24], and Saxena et al. [88, 89]. Afterwards, Marigold [49] showed how diffusion-based generators can be adapted for MDE with rich prior knowledge. GeoWizard [28] extended an LDM [85] to jointly predict depth and normal maps, enabling information exchange between these representations. [136] introduced a diffusion-based depth refiner, while [109] explored guidance with sparse LiDAR inputs. [39] and [63] developed simplified protocols for dense prediction.

Other approaches have focused on recovering metric depth by incorporating camera information, including [10, 38, 41, 123]. Meanwhile, works on monocular point map estimation [72, 73, 112, 113, 124, 126] have explored direct 3D point recovery from single images.

Finally, video-based depth estimation has emerged as another direction, with approaches like [13, 42, 48, 91] addressing temporal consistency.

2.3. Overcoming Challenging Conditions

Adverse Weather. Monocular networks have historically struggled in adverse weather conditions. Various techniques have addressed these challenges, including methods for low visibility scenarios [93], day-night branches using GANs [106, 137], additional sensor integration [30], and performance trade-offs [107]. A notable advance came with [31], which enabled robustness across varying conditions without compromising performance in ideal settings.

Transparent and Specular Surfaces. A separate challenge involves estimating depth for transparent or mirror (ToM) surfaces [130, 132, 133]. Costanzino et al. [21] pioneered work dedicated to this problem, introducing specialized datasets [79, 80]. Their approach leveraged segmentation maps or pre-trained networks, generating pseudo-labels by inpainting ToM objects and processing them with pre-trained depth models [82], enabling existing networks to be fine-tuned for handling transparent and specular surfaces. Similarly, [105], instead, presented a novel approach using depth-aware diffusion models to systematically generate challenging scenes (both for adverse weather and ToM objects) with associated depth information and fine-tune networks through self-distillation.

3. Monocular Depth Estimation Challenge

The fourth edition of the Monocular Depth Estimation Challenge was hosted on CodaLab[1] [68] as part of the CVPR'2025 workshop. The participation instructions included a new starter pack[2] supporting train-free minimal effort inference with public foundation models.

The development phase lasted four weeks, using the SYNS-Patches validation split. During this phase, the leaderboard was public, but the participants' usernames were anonymized. Each participant could see the results achieved by their submissions. The final phase of the challenge was open for three weeks. At this stage, the leaderboard was completely private, disallowing participants to see their scores to avoid overfitting to the test split.

[1] https://codalab.lisn.upsaclay.fr/competitions/21305
[2] https://github.com/toshas/mdec_benchmark

Following the third edition [99], any form of supervision was allowed; as long as only one input image is used to obtain the prediction corresponding to the input image. This measure prevents mixing in methods with strong support of geometric priors. In what follows, we report results only for submissions that outperform the specified baselines in at least one point cloud or image-based metric in the testing phase.

Dataset. The challenge is based on the SYNS-Patches dataset [1, 96], chosen due to the diversity of scenes and environments. A few representative examples are shown in Figure 1. 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 through edge detection on the log-depth maps, allowing us to compute additional edge-based metrics. As outlined in [95, 96, 99], the images were manually checked to remove dynamic object artifacts.

Evaluation. Participants were asked to provide predictions for each dataset image. This year, we switched to LSE-based alignment between predictions and ground truth maps to accept various types of predictions. In addition to previously accepted `disparity` prediction methods, we welcomed `affine-invariant`, `scale-invariant`, and `metric` types. The evaluation server bilinearly upsampled the predictions to the target resolution. Submissions with disparity images (including previous years' submissions) were inverted into scale-invariant depth maps. Together with the other prediction types, we performed least squares alignment with two degrees of freedom, the affine-invariant scale and shift, and computed multiple pairwise metrics.

Metrics. Following the previous editions of the challenge [94, 95, 99], we use a combination of image-, pointcloud-, and edge-based metrics. Image-based metrics are the most common (MAE↓, RMSE↓, AbsRel↓, δ↑) and are computed using pixel-wise comparisons between the predicted and ground-truth depth map. Pointcloud-based metrics [67] (F-Score↑) instead bring the evaluation into the 3D domain. Among these, we select the reconstruction F-Score as the leaderboard ranking metric. Finally, edge-based metrics are computed only at depth boundary pixels. This includes F-Score at edges (F-Edges↑), as well as edge accuracy and completion metrics from IBims-1 [51].

4. Challenge Submissions

We now report the main technical details for each submission, as provided by the participants themselves.

Baseline #1 – Marigold [49]

Baseline model for this year's challenge.

Network. Based on the convolutional UNet LDM [85], repurposed for depth estimation. The model gradually denoises a random latent code conditioned on the input image to produce a depth map. It uses the frozen VAE component from the LDM for encoding and decoding, with the depth map replicated into three channels to match the RGB input format. The denoising UNet is modified to accept both image and depth latent codes concatenated along the feature dimension.

Supervision. The model is fine-tuned from the LDM [85] using only synthetic datasets (Hypersim [84], Virtual KITTI 2 [11]) to ensure dense, complete, and clean depth maps. It uses affine-invariant depth normalization through a linear transformation that maps depth values to the range $[-1, 1]$.

Training. Trained for 18K iterations with batch size 32 (accumulated over 16 steps to fit on a single GPU) using Adam optimizer with a learning rate of $3e^{-5}$ and random horizontal flipping augmentation. The training takes approximately 2.5 days on a single consumer GPU.

Inference. Off-the-shelf `prs-eth/marigold-depth-v1-0` checkpoint is applied at the native input resolution with 10 DDIM steps and ensemble size 1[3].

Baseline #2 – Depth Anything v2

Additional baseline model for this year's challenge.

Network. DPT decoder [81] built on DINOv2 encoders [66]. Multiple variants are available, from lightweight (ViT-S, 25M parameters) to high-capacity (ViT-G, 1.3B parameters).

Supervision. Three-stage training pipeline: 1) Train a teacher model (DINOv2-G) purely with high-quality synthetic images from five datasets (595K images); 2) Produce precise pseudo-labels on large-scale unlabeled real images dataset; 3) Train a final student model on 62M pseudo-labeled images from eight datasets. Training losses include a scale- and shift-invariant loss, gradient matching, and feature alignment to preserve semantics from pre-trained DINOv2 encoders.

Training. Trained at resolution 518×518 by resizing input to match the shorter side, followed by random cropping. The teacher model is trained with batch size 64 for 160K iterations. Student models are trained with batch size 192 for 480K iterations. Adam optimizer is used with learning rates $5e^{-6}$ (encoder) and $5e^{-5}$ (decoder).

Inference. Off-the-shelf `depth-anything/Depth-Anything-V2-Large-hf` checkpoint is applied at native resolution[3].

Team 1: HRI

S. Wang *wangshuaihang@hikvision.com*
Z. Ma *mazhenxin@hikvision.com*
W. Chen *chenweijie5@hikvision.com*
B. Xu *xubaobei@hikvision.com*
F. Sun *sunfengyu@hikvision.com*
D. Xie *xiedi@hikvision.com*
J. Zhu *zhujiang.hri@hikvision.com*

Network. ViT-G encoder with DPT decoder head. The encoder is initialized with pre-trained weights from Metric3D

[3] https://github.com/toshas/mdec_benchmark/blob/main/mdec_2025

v2 [41], while the decoder uses random initialization.
Supervision. Supervised learning with SILog, SSIL, and Feature Alignment losses. Feature Alignment uses KL divergence between the network encoder output and frozen pre-trained DINOv2 (ViT-G) encoder features for semantic knowledge distillation. The upper quarter and lateral sixteenths of label maps were excluded from loss computation due to low-quality annotations.
Training. Cityscapes dataset with pixel-wise depth ground truth. Learning rates $1e^{-5}$ for encoder and $1e^{-4}$ for decoder, batch size 32 across 8 GPUs, with random horizontal flipping. Early stopping at 3 epochs to mitigate overfitting. Training images were resized to 384×768, testing images to 768px on the short side maintaining aspect ratio.

Team 2: Lavreniuk

M. Lavreniuk *nick_93@ukr.net*

Network. Based on Depth Anything v2 [119].
Supervision. Supervised training with ground-truth depth using scale-invariant logarithmic loss, scale-and-shift invariant loss, and random proposal normalization loss.
Training. Trained on outdoor datasets (KITTI [32], Virtual KITTI 2 [11], DIODE outdoor [108], Cityscapes [20]) with random crop size 518×1078, batch size 1, learning rate $1e^{-6}$, maximum depth threshold 80 meters, for 5 epochs. Standard augmentations were applied during training and avoided during testing due to performance degradation.

Team 3: Mach-Calib

H. Guan *guanhaining@megvii.com*
Q. Wu *wuqun@megvii.com*
Y. Zeng *zengyupei@megvii.com*
C. Lu *luchao03@megvii.com*
H. Wang *wanghuanran@megvii.com*

Network. Depth-Anything-V2 (ViT-L) [119], using the pre-trained outdoor metric depth estimation model.
Supervision. Fine-tuned on ground truth metric depths using a composite loss comprising SiLog, SSIL, Virtual Normal, and Sky Regularization.
Training. Fine-tuned on Cityscapes [20], KITTI [32], and VKITTI2 [11] with 392×770 resolution, random horizontal flip, and canonical space camera transformation. Depth prediction up to 120m with dataset-specific truncation. Sky regions forced to maximum depth. Trained for 5 epochs with batch size 8, learning rates $5e^{-6}$ for encoder and $5e^{-5}$ for decoder. Inference uses test-time resolution scaling to 518px on the shorter side.

Team 4: EasyMono

G. Zhou *zhouguangyuan@buaa.edu.cn*
H. Zhang *zht199599@gmail.com*
J. Wang *doublewjx@gmail.com*
Q. Rao *erdosv001@gmail.com*
C. Wang *chunjiewang1993@gmail.com*
X. Liu *liuxiao@foxmail.com*

Network. Based on Depth-Anything [118] with a BEiT384-L backbone.
Supervision. Supervised using stereo disparities from Cityscapes [20]. The final loss is composed of the SSIL loss, Gradient Angle loss, and random proposal normalization loss (RPNL) loss.
Training. The network was fine-tuned at a resolution of 384×768 with the ground truth depth resolution of 1024×2048. Only random flipping was used as the data augmentation strategy. The training was restricted to 4 epochs, a strategic choice to prevent overfitting and ensure the model's robustness to new data.

Team 5: Insta360-Percep

Z. Lou *johnlou@insta360.com*
H. Jiang *jianghualie@insta360.com*
Y. Chen *chenyihao@insta360.com*
R. Xu *jerry1@insta360.com*
M. Tan *tmltan@insta360.com*

Network. Pre-trained ViT-Large variant of Depth-Anything-V2 [119] with the final ReLU layer replaced by Softplus activation for more stable training. Final output converted to affine-invariant depth values using shifted inverse function: $\text{Depth} = \frac{200}{1+\text{DaV2(img)}}$.
Supervision. Uses scale-invariant log loss, scale-shift-invariant loss and multi-scale gradient matching loss. For synthetic datasets (Virtual KITTI 2 [11], Hypersim [84]), official ground truth depth is used. For real-world datasets (CityScapes, KITTI), dense and detailed depth maps are generated using FoundationStereo and DeformStereo.
Training. Trained on CityScapes, KITTI, VKITTI2, and Hypersim, preprocessed to match SYNS-Patches vertical-to-horizontal FOV ratio. Input size 392×784, batch size 8, with learning rates $1.5e^{-5}$ for encoder and $1e^{-4}$ for decoder, fine-tuned for 5 epochs.

Team 6: HIT-AIIA

Z. Qin *24s103315@stu.hit.edu.cn*
Y. Mao *24s103391@stu.hit.edu.cn*
J. Liu *1190200503@stu.hit.edu.cn*
J. Xu *xujialei@stu.hit.edu.cn*
Y. Yang *2022111217@stu.hit.edu.cn*
W. Zhao *wbzhao@hit.edu.cn*
J. Jiang *jiangjunjun@hit.edu.cn*
X. Liu *csxm@hit.edu.cn*

Network. Based on Depth Anything v2 [119] with a ViT-L

backbone, initialized with the authors' pre-trained weights for indoor metric depth. The entire pre-trained backbone is frozen to preserve its generalization capability while a parallel trainable branch is added. This trainable branch is a full copy of the original model, connected via zero-initialized 1×1 convolutions to the frozen backbone, outputting per-pixel scale (α) and shift (β) maps to refine the base prediction through $D_{\text{final}} = (1 + \alpha) \odot D_{\text{base}} + \beta$.

Supervision. Fine-tuned in a supervised manner using the SSIL loss as the loss function.

Training. Fine-tuned on CityScapes [20] and DIODE [108] datasets with inputs randomly cropped to 768×768 and augmented via random horizontal flipping. Trained with batch size 32 and AdamW optimizer with weight decay 0.01. Learning rates set to $5e^{-6}$ for pre-trained parameters and $5e^{-5}$ for newly introduced ones. OneCycleLR schedule adopted, starting at $1.25e^{-6}$ (one-fourth of base rate), increasing to $5e^{-6}$ after 20% warmup phase, and subsequently decaying to $2.5e^{-6}$.

Team 7: Robot02-vRobotit

M. Zhao *mingshuai_z@bupt.edu.cn*
A. Ming *mal@bupt.edu.cn*
W. Chen *chenw@bupt.edu.cn*
F. Xue *feng.xue@unitn.it*
M. Yu *yumengying@bupt.edu.cn*
S. Gao *gaostar2024@bupt.edu.cn*
X. Wang *xiangfeng_w@foxmail.com*

Network. Improved SM4Depth [57] based on Booster [80]. The architecture includes FOV alignment preprocessing, an encoder from Depth Anything [118], pyramid scene transformer (PST) with three parallel transformer encoders, online domain-aware bin mechanism, and a decoder with hierarchical scale constraints (HSC-Decoder). The encoder is frozen while only the PST and the Decoder are fine-tuned.

Supervision. Multiple loss functions: SILog loss, gradient matching loss, pixel-wise loss, virtual normal loss, bidirectional Chamfer loss, and cross entropy loss combined as $L = L_{pixel} + \sum_{s=2}^{4} L_{pixel_s} + \mu L_{vnl} + \gamma L_{bin} + L_{cel}$ where $\mu = 5$ and $\gamma = 0.1$.

Training. DinoV2 [66] Base backbone, trained for 20 epochs with batch size 10 on a single consumer GPU. Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$, initial learning rate $2e^{-5}$ gradually reduced to $2e^{-6}$. Fixed FOV set to $(\omega'_x, \omega'_y) = (58°, 45°)$ and fixed resolution to 564×424.

Team 8: ViGIR LAB

G. Omotara *goowfd@missouri.edu*
R. Farag *rmf3mc@missouri.edu*
J. Demby's *udembys@missouri.edu*
S. M. A. Tousi *stousi@missouri.edu*
G. N. DeSouza *desouzag@missouri.edu*

Network. Builds on Marigold [49] by integrating semantic information. While the Marigold network remains frozen, a semi-supervised model (UniMatch-V2 [120]) generates semantic segmentation masks that capture rich scene semantics. A new DenseNet-like [43] model is trained to incorporate the extracted semantics to refine and enhance the depth map.

Supervision. Fully supervised training using SILog loss to optimize the depth predictions.

Training. Trained on DIODE [108] dataset. Ground truth depth maps are first scaled, shifted, and normalized to $[0, 1]$ to match Marigold's affine-invariant representation. Training was conducted for 50 epochs with learning rate $3e^{-5}$ to ensure a smooth loss trajectory.

Team 9: HCMUS-DepthFusion

T. A. Yang *ytanh21@apcs.fitus.edu.vn*
M. Q. Nguyen *nmquang21@apcs.fitus.edu.vn*
T. P. Tran *ttphuc21@apcs.fitus.edu.vn*

Network. Hybrid approach combining Felzenszwalb segmentation and segment-wise confidence heuristic to improve depth accuracy. Segments images into consistent regions, facilitating depth estimation within structurally coherent areas. Combines outputs from diffusion-based (Marigold [49]) and transformer-based (Depth Anything v2 [119]) models.

Supervision. Segment-wise confidence computation using inverse depth variance to determine the reliability of depth estimates. Confidence metric defined as $c_i = \frac{1}{\sigma_i + \epsilon}$ where σ_i is the depth variance within segment i. Adaptive depth selection enables the fusion of different model outputs at the segment level based on the confidence threshold.

Training. Felzenszwalb segmentation parameters: scale=200, σ=0.8, min_size=50. Fusion strategy uses confidence threshold T=1.2, preferring Marigold predictions unless Depth Anything v2's confidence is significantly higher.

Team 10: ReadingLS

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

Network. Custom hybrid CNN-Transformer architecture with SwiftFormer-L1 [90] encoder and a two-level decoder similar to SwiftDepth [58]. A separate ResNet-18 camera network predicts pose and intrinsics during training only. Both encoders are pre-trained on ImageNet [23]. The depth network has 30M parameters, while the camera network has 15M parameters.

Supervision. Combined self-supervised learning (SSL) and pseudo-supervised learning (PSL) [59]. SSL uses minimum reprojection loss with auto-masking, PSL uses scale-and-shift invariant loss, with Depth-Anything-V2-Large generating pseudo-labels.

Training. Trained simultaneously on YouTube [59] (1.1M triplets) and KITTI [32] (40K triplets). Batch size 16, learning rate $1e^{-5}$, weight decay $1e^{-3}$, for 110 epochs on KITTI and 4 epochs on YouTube. The loss weighting parameter was adjusted from 0.9 to 0.95 during training.

Table 1. SYNS-Patches Results. We provide metrics across the whole test split of the dataset.

	Team	Rank	F↑	F-Edges↑	MAE↓	RMSE↓	AbsRel↓	Acc-Edges↓	Comp-Edges↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
$\mathcal{A}$	HRI	1	**23.05**	9.37	**3.19**	**5.64**	21.17	3.30	6.77	77.54	91.79	95.78
	Anonymous	2	22.95	9.21	3.44	6.48	24.65	3.27	**6.59**	76.17	90.33	94.73
	PICO-MR [99]	-	22.58	8.26	3.41	5.89	22.02	4.02	11.28	76.18	90.95	95.44
$\mathcal{A}$	Lavreniuk	3	20.81	9.12	3.40	5.77	**20.48**	3.20	7.27	75.33	91.33	**95.97**
$\mathcal{A}$	Mach-Calib	4	20.69	8.75	3.63	6.78	23.36	3.22	13.96	74.31	90.44	95.12
	Anonymous	5	20.50	8.63	3.68	7.07	23.96	3.17	13.66	74.00	90.32	95.01
$\mathcal{A}$	EasyMono	6	20.24	7.30	3.43	5.91	22.47	3.91	20.98	74.73	90.55	95.24
$\mathcal{A}$	Insta360-Percep	7	19.41	8.98	3.25	5.73	21.29	**2.75**	15.54	**77.91**	**92.03**	95.91
	Anonymous	8	18.85	8.27	4.06	7.06	26.35	3.11	13.75	67.57	87.98	94.18
$\mathcal{A}$	HIT-AIIA	9	18.53	**9.74**	3.83	6.29	23.96	3.06	9.90	69.12	88.49	94.31
	Anonymous	10	18.16	6.32	3.57	6.13	24.07	4.51	15.75	71.62	89.12	94.74
†	*Anonymous*	11	17.46	9.35	4.43	7.30	29.19	3.27	17.69	65.64	86.45	93.09
†	*Anonymous*	12	17.46	9.35	4.43	7.30	29.19	3.27	17.69	65.64	86.45	93.09
†	*Anonymous*	13	17.45	9.35	4.43	7.30	29.20	3.27	17.69	65.64	86.46	93.10
$\mathcal{A}$	Robot02-vRobotit	14	17.25	8.27	3.90	6.54	23.64	3.07	22.85	71.07	89.38	94.80
*	*Anonymous*	15	17.01	9.20	4.44	7.34	29.42	3.25	18.55	65.72	86.67	93.35
*	*Anonymous*	16	17.01	9.20	4.44	7.34	29.42	3.25	18.55	65.72	86.67	93.35
$\mathcal{A}$	**Marigold [49]**	-	17.01	9.19	4.44	7.34	29.42	3.25	18.56	65.72	86.67	93.35
	Anonymous	17	16.79	9.10	4.51	7.44	30.10	3.32	17.65	65.24	86.22	93.05
	Anonymous	18	16.67	7.49	4.01	6.92	26.34	3.07	26.32	68.17	88.47	94.42
	Anonymous	19	16.47	8.99	4.74	7.71	30.93	3.26	18.75	62.26	84.27	91.90
	ViGIR LAB	20	16.25	8.54	5.37	9.00	43.72	3.29	14.15	61.99	82.89	90.46
$\mathcal{A}$	*Anonymous*	21	15.31	8.46	4.62	7.46	30.63	3.42	15.59	62.41	84.89	92.43
	Anonymous	22	15.03	7.64	5.48	9.29	38.59	3.70	20.33	58.82	81.79	90.85
	Depth Anything v2 [119]	-	14.34	6.72	4.84	9.13	33.57	2.99	35.54	66.34	86.02	92.44
$\mathcal{A}$	HCMUS-DepthFusion	23	14.20	7.81	4.90	7.96	33.55	3.32	17.66	61.09	83.90	92.21
	ReadingLS	24	13.52	6.60	4.56	7.90	30.15	3.14	35.94	65.34	86.55	93.46

Legend: ***Baselines*** ; ***3rd MDEC winning team*** ; † – *likely the same method;* * – *likely minor variations of Marigold;* $\mathcal{A}$ – *affine-invariant predictions; for each metric we highlight* ***absolute best*** , *second best* , *third best*

5. Results

Following the protocol from previous editions, submitted predictions were evaluated on the testing split of SYNS-Patches [1, 96], after being aligned to ground-truth depths according to median depth scaling or least-squares alignment, as requested by the participants.

5.1. Quantitative Results

Table 1 collects the results of this fourth edition of the challenge, ranking the submitted methods according to their F-Score performance. Within the table, we highlighted the new baselines introduced for this edition, as well as the results achieved by the winning team of the third edition according to the new aligning protocol – both excluded from the official ranking. We also grouped entries with close-to-identical results within brackets, likely representing duplicate team submissions.

Similarly to the previous edition [99], where the Depth Anything model [118] was widely adopted by participants, this year we observe a widespread use of its successor, Depth Anything v2 [119], among the leading teams – including Lavreniuk, Mach-Calib, Insta360-Percep, HIT-AIIA.

Team HRI secured the top position on the leaderboard with an F-score of 23.05, surpassing the best baseline method (Marigold) by about 35%. Among the teams submitting the report, HRI is the only one outperforming PICO-MR, the winning team from "The Third Monocular Depth Estimation Challenge" [99], in terms of F-score.

Team Lavreniuk, in the third position, achieved an F-score lower than what was achieved by PICO-MR (20.81) while outperforming it on several image-based metrics. Peculiarly, Lavreniuk achieves the absolute best results on AbsRel and $\delta < 1.25^3$ metrics.

Teams Mach-Calib and EasyMono, ranking fourth and sixth, respectively, obtained results very close to those of Lavreniuk, although not excelling in any particular metric. Nonetheless, they could still outperform Marigold by approximately 20% on the F-score. In contrast, team Insta360-Percep performed slightly worse in terms of F-score while securing the absolute best results on metrics such as Acc-Edges, $\delta < 1.25$ and $\delta < 1.25 < 2$. Similarly, team HIT-AIIA achieved the absolute best result for F-Edges. In contrast, team Robot02-vRobotit did not excel in most metrics while consistently outperforming the baselines in most metrics, such as F-score, MAE, RMSE, and AbsRel.

Finally, teams ViGIR LAB, HCMUS-DepthFusion, and ReadingLS managed to outperform the baselines on only one metric among the ten evaluated. We highlight the efforts by these teams to develop custom solutions not limited to fine-tuning existing foundational models. Their performance further underscores the importance of the web-scale data used to train such foundational models, giving them a clear advantage over any possible alternative that does not leverage similar quantities of training data.

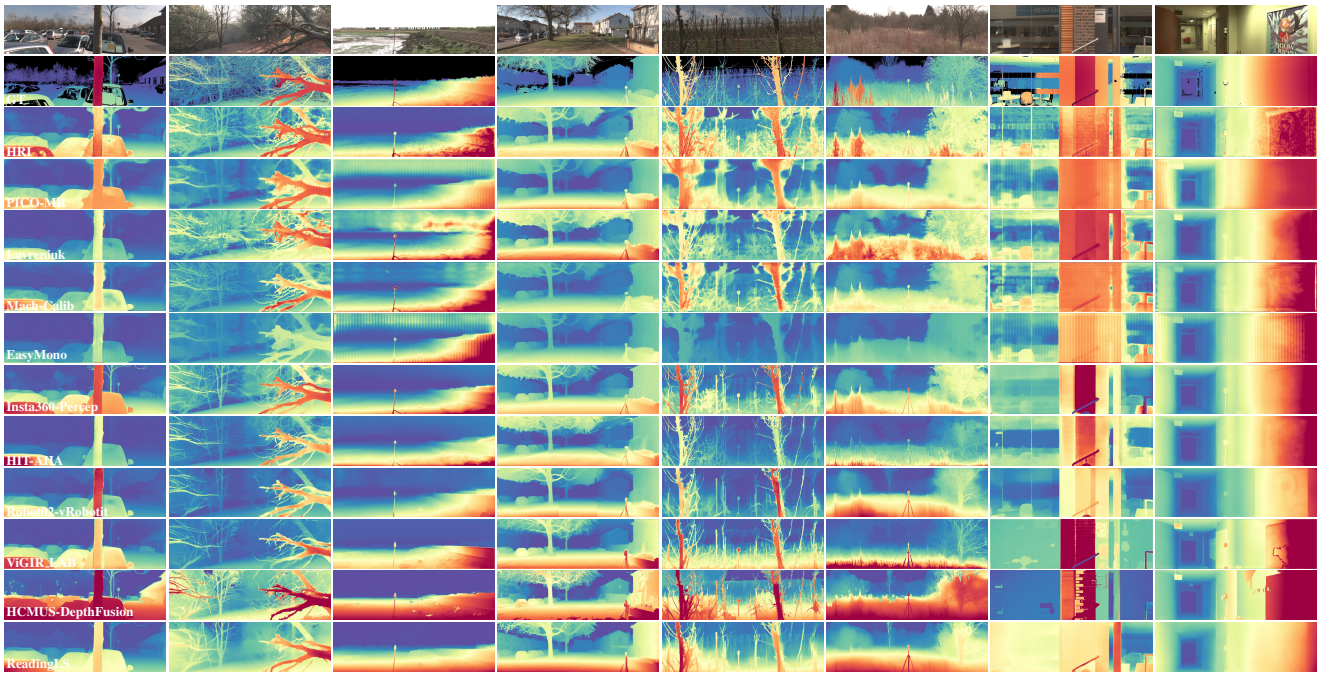

Figure 2. Qualitatives on SYNS-Patches. Best viewed in color and zoomed in. Methods are ranked based on their F-Score in Table 1.

5.2. Qualitative Results

Figure 2 presents selected qualitative results from the SYNS-patches test set. From top to bottom, we report input images and ground truth depth maps, followed by the predictions from teams that submitted the report, ordered according to their achieved F-score. For reference, we also show qualitative results from PICO-MR [99] – characterized by noticeable grid artifacts.

We appreciate how most top-performing methods retain significantly more detailed predictions than PICO-MR. This enhanced level of detail is caused by the use of more modern foundational models, such as Metric3D v2 [41] and Depth Anything v2 [119], which inherently generate more refined depth maps than the original Depth Anything. Furthermore, we appreciate how the predictions by HRI and particularly Lavreniuk appear cleaner and exhibit fewer artifacts, pointing at a general qualitative improvement beyond the quantitative ones observed in the tables.

This widespread adoption of depth foundation models led to substantial improvements in both indoor and outdoor scenarios, proving once again the crucial role of web-scale data availability, further pushing the boundaries of MDE. However, similarly to previous editions, all methods still exhibit over-smoothing issues at depth discontinuities, which can be seen as content bleeding artefacts when depth maps are projected into point clouds.

To summarize, the qualitative results provide additional insights beyond those that emerged from the quantitative analysis, hinting at further improvements over the previous editions. Nonetheless, evident challenges persist when aiming at generalization, such as dealing with very thin structures or non-Lambertian surfaces, leaving room for further developments in the field.

6. Conclusions & Future Directions

This paper summarized the results of the fourth edition of MDEC. Compared to previous editions, we have seen a further increase in participation in our challenge, confirming that MDE is a vibrant research trend. We witnessed a drastic increase in reliance on the pre-trained foundational models and various prediction types.

Despite yielding marginal improvements over the previous years in terms of the F-score, this year's methods generally deliver sharper predictions and appear useful in downstream tasks. This suggests that, despite the significant improvements achieved by foundational models over previous approaches, the problem remains far from being solved. As the benchmark approaches saturation, further progress will likely require more than simply increasing the volume of training data or usage of foundation model priors. Next-generation advances in depth estimation may come from tackling more challenging settings, such as the perception of non-Lambertian surfaces [80], accurate metric depth estimation [41], or alternative scene representations, such as point maps [113].

We hope future editions of MDEC will attract even more participants to push the boundaries of monocular depth estimation (MDE) forward.

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] Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

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

[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.

[13] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025.

[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] 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. Ieee, 2009.

[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. In *ICCV*, 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] 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 recog-*

nition, pages 2002–2011, 2018.

[28] 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. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.

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

[41] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[42] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.

[43] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[44] 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.

[45] 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.

[46] 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. In *ICCV*, 2023.

[47] 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.

[48] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[49] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Mannequinchallenge: Learning the depths of moving people by watch-

ing frozen people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4229–4241, 2020.

[55] 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.

[56] 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.

[57] Yihao Liu, Feng Xue, Anlong Ming, Mingshuai Zhao, Huadong Ma, and Nicu Sebe. Sm4depth: Seamless monocular metric depth estimation across multiple cameras and scenes by one model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3469–3478, 2024.

[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. IEEE, 2023.

[59] Albert Luginov and Muhammad Shahzad. Nimbled: Enhancing self-supervised monocular depth estimation with pseudo-labels and large-scale video pre-training. *arXiv preprint arXiv:2408.14177*, 2024.

[60] 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.

[61] 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.

[62] 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.

[63] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[64] 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.

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

[66] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.

[73] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[86] 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.

[87] 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.

[88] 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.

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

[90] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17425–17436, 2023.

[91] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] Jaime Spencer, Fabio Tosi, Matteo Poggi, Ripudaman Singh Arora, Chris Russell, Simon Hadfield, Richard Bowden, Guangyuan Zhou, Zhengxin Li, Qiang Rao, Yiping Bao, Xiao Liu, Dohyeong Kim, Jinseong Kim, Myunghyun Kim, Mykola Lavreniuk, Rui Li, Qing Mao, Jiang Wu, Yu Zhu, Jinqiu Sun, Yanning Zhang, Suraj Patni, Aradhye Agarwal, Chetan Arora, Pihai Sun, Kui Jiang, Gang Wu, Jian Liu, Xianming Liu, Junjun Jiang, Xidan Zhang, Jianing Wei, Fangjun Wang, Zhiming Tan, Jiabao Wang, Albert Luginov, Muhammad Shahzad, Seyed Hosseini, Aleksander Trajcevski, and James H. Elder. The third monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–14, June 2024.

[100] 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.

[101] Lior Talker, Aviad Cohen, Erez Yosef, Alexandra Dana, and Michael Dinerstein. Mind the edge: Refining depth edges in sparsely-supervised monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition*,

2024. CVPR.

[102] 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.

[103] 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.

[104] 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.

[105] Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *European Conference on Computer Vision (ECCV)*, 2024.

[106] 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.

[107] 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.

[108] 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.

[109] Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion, 2024.

[110] 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.

[111] 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.

[112] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024.

[113] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

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

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[120] Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[121] 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.

[122] 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.

[123] 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 *ICCV*, 2023.

[124] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6480–6494, 2022.

[125] 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.

[126] 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 (CVPR)*, pages 204–213, June 2021.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2025 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*, 2025.

[134] 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.

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

[136] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. *arXiv preprint arXiv:2407.17952*, 2024.

[137] 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.

[138] 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.

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

[140] 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.