The Second Monocular Depth Estimation Challenge

Jaime Spencer1 C. Stella Qian2 Michaela Tresckova3 Chris Russell4 Simon Hadfield1
Erich W. Graf3 Wendy J. Adams3 Andrew J. Schofield2 James Elder5 Richard Bowden1
Ali Anwar6 Hao Chen7 Xiaozhi Chen8 Kai Cheng9 Yuchao Dai10 Huynh Thai Hoa11
Sadat Hossain11 Jianmian Huang12 Mohan Jing9 Bo Li10 Chao Li13 Baojun Li12
Zhiwen Liu13 Stefano Mattoccia14 Siegfried Mercelis6 Myungwoo Nam11 Matteo Poggi14
Xiaohua Qi9 Jiahui Ren10 Yang Tang15 Fabio Tosi14 Linh Trinh6 S. M. Nadim Uddin11
Khan Muhammad Umair11 Kaixuan Wang8 Yafei Wang10 Yixing Wang13 Mochu Xiang10
Guangkai Xu9 Wei Yin8 Jun Yu9 Qi Zhang13 Chaoqiang Zhao15

Abstract

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

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

1. Introduction

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

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

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

The challenge was once again centered around SYNS-Patches [1,78]. This dataset was chosen due its diversity, which includes urban, residential, industrial, agricultural, natural and indoor scenes. Furthermore, SYNS-Patches contains dense high-quality LiDAR ground-
truth, which is exceedingly rare in outdoor environments. This ensures that the evaluations accurately reflect the capabilities of each model.

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

2. Related Work

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

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

**Self-Supervised.** Instead of relying on costly annotations, Garg et al. [25] proposed an algorithm based on view synthesis and the photometric consistency across stereo pairs. Monodepth [28] incorporated differentiable bilinear interpolation [42], virtual stereo prediction and a SSIM+L1 reconstruction loss. SfM-Learner [108] required only monocular video supervision by replacing the known stereo transform with a pose estimation network.

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

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

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

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

3. The Monocular Depth Estimation Challenge

The second edition of the Monocular Depth Estimation Challenge1 was organized on CodaLab [63] as part of a CVPR2023 workshop. The initial development phase lasted four weeks, using the SYNS-Patches validation split. The leaderboard for this phase was anonymous, where all method scores were publicly available, but usernames remained hidden. Each participant could see the metrics for their own submission.

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

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

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

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

3064
Table 1. SYNS-Patches. Distribution of images per category in the val/test splits.

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

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

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

Evaluation. Participants provided the unscaled disparity prediction for each dataset image. The evaluation server bilinearly upsampled the predictions to the target resolution and inverted them into depth maps. Self-supervised methods trained with stereo pairs and supervised methods using LiDAR or RGB-D data should be capable of predicting met-
ric depth. Despite this, in order to ensure comparisons are as fair as possible, the evaluation aligned predictions with the ground-truth using the median depth. We set a maximum depth threshold of 100 meters. 

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

4. Challenge Submissions

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

Baseline – S

| J. Spencer† | j.spencermartin@surrey.ac.uk |
| C. Russell‡ | cmrus@amazon.de |
| S. Hadfield§ | s.hadfield@surrey.ac.uk |
| R. Bowden∥ | rbowden@surrey.ac.uk |

Challenge organizers submission from the first edition.

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

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

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

**Team 1: DJI&ZJU – D**

| W. Yin§ | yvanwy@outlook.com |
| K. Cheng§ | chengkai21@mail.ustc.edu.cn |
| G. Xu§ | xugk@mail.ustc.edu.cn |
| H. Chen§ | haochen.cad@zju.edu.cn |
| B. Li‡ | libo@nwpu.edu.cn |
| K. Wang§ | wkx1993@gmail.com |
| X. Chen§ | xiaozhi.chen@dji.com |

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

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

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

**Team 2: Pokemon – D**

| M. Xiang§ | xiangmochu@mail.nwpu.edu.cn |
| J. Ren§ | renjiahui@mail.nwpu.edu.cn |
| Y. Wang§ | wangyufei777@mail.nwpu.edu.cn |
| Y. Dai§ | daiyuw@nwpu.edu.cn |

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

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

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

**Team 3: cv-challenge – D**

| C. Li‡ | lichao@vivo.com |
| Q. Zhang‡ | zhangqi.aiyj@vivo.com |
| Z. Liu‡ | zhiwen.liu@vivo.com |
| Y. Wang‡ | wangyixing@vivo.com |

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

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

**Training.** The original ZoeDepth [9] and DPT [66] were pretrained on a collection of 12 datasets. The models were then finetuned on Kitti (384 × 768) or NYUD-v2 (384 × 512).
for outdoor/indoor scenes, respectively. Different models were deployed on an automatic scene classifier. The fine-tuned models were combined with a content-adaptive multi-resolution merging method [60], where patches were combined based on the local depth cue density. Since the transformer-based backbone explicitly captured long-term structural information, the original double-estimation step was omitted.

**Team 4: DepthSquad – D**

M. Nam\(^{11}\) mwn0221@deltax.ai

H. T. Hoa\(^{11}\) hoahit@deltax.ai

K. M. Umair\(^{11}\) mumairkh@deltaix.ai

S. Hossain\(^{11}\) sadda@deltaix.ai

S. M. N. Uddin\(^{11}\) sayednadim@deltax.ai

**Network.** Based on the PixelFormer architecture [2] which used a Swin [55] encoder and self-attention decoder blocks with cross-attention skip connections. Disparity was predicted as a discrete volume [7], with the final depth map given as the weighted average using the bin probabilities.

**Supervision.** Supervised using the SILog loss w.r.t. the LiDAR ground-truth.

**Training.** The model was trained on the Kitti Eigen-Zhou (KEZ) split using images of size \(370 \times 1224\) for 20 epochs. Additional augmentation was incorporated in the form of random cropping and rotation, left-right flipping and Cut-Depth [41]. When predicting on SYNS-Patches, images were zero-padded to \(384 \times 1248\) to ensure the compatibility of the training resolution. These borders were removed prior to submission.

**Team 5: imec-IDLab-UAntwerp – MS**

L. Trinh\(^{6}\) khaclinh.trinh@student.uantwerpen.be

A. Anwar\(^{6}\) ali.anwar@uantwerpen.be

S. Mercelis\(^{6}\) siegfried.mercelis@uantwerpen.be

**Network.** Pretrained ConvNeXt-v2-Huge [87] encoder with an HR-Depth decoder [58], modified with deformable convolutions [19]. The pose network instead used ResNet-18 [35].

**Supervision.** Self-supervised using the photometric loss [29] and edge-aware smoothness.

**Training.** Trained on the Kitti Eigen-Benchmark (KEB) split with images of size \(192 \times 640\). The network was trained for a maximum of 30 epochs, with the encoder remaining frozen after 6 epochs.

**Team 6: GMD – MS**

B. Li\(^{12}\) 1966431208@qq.com

J. Huang\(^{12}\) huang176368745@gmail.com

**Network.** ConvNeXt-XLarge [56] backbone and an HR-Depth [58] decoder.

**Supervision.** Self-supervised based on the photometric loss [29].

**Training.** Trained on KEZ using a resolution of \(192 \times 640\).

**Team 7: MonoViTeam – MSD**

C. Zhao\(^{14}\) zhaoqc@mail.ecust.edu.cn

M. Poggi\(^{13}\) m.poggi@unibo.it

F. Tosi\(^{13}\) fabio.tosi@unibo.it

Y. Tang\(^{14}\) yangtang@ecust.edu.cn

S. Mattoccia\(^{13}\) stefano.mattoccia@unibo.it


**Supervision.** Self-supervised on Kitti Eigen (KE) using the photometric loss [29] (stereo and monocular support frames) and proxy depth regression. Regularized using edge-aware disparity smoothness [28] and depth gradient consistency w.r.t. the proxy labels.

**Training.** Proxy depths were obtained by training a self-supervised RAFT-Stereo network [52] on the trinocular Multiscopic [101] dataset. The stereo network was trained for 1000 epochs using \(256 \times 480\) crops. The monocular network was trained on KE for 20 epochs using images of size \(320 \times 1024\).

**Team 8: USTC-IAT-United – MS**

J. Yu\(^{9}\) harryjun@ustc.edu.cn

M. Jing\(^{9}\) jing.mohan@mail.ustc.edu.cn

X. Qi\(^{9}\) xiaohua000109@163.com

**Network.** Predictions were obtained as a mixture of multiple networks: DiffNet [107],FeatDepth [74] and MonoDEVSNet [33]. DiffNet and FeatDepth used a ResNet backbone, while MonoDEVSNet used DenseNet [38].

**Supervision.** Self-supervised using the photometric loss [29].

**Training.** The three models were trained with different resolutions: \(320 \times 1024\), \(376 \times 1242\), \(384 \times 1248\), respectively. All predictions were interpolated to \(376 \times 1242\) prior to ensembling using a weighted average with coefficients \{0.35, 0.3, 0.35\}.

## 5. Results

Participant submissions were evaluated on SYNS-Patches [1, 78]. As previously mentioned, this paper only discusses submissions that outperformed the baseline in any pointcloud-image-based metric across the Overall dataset. Since both challenge phases ran independently and participants were responsible for generating the predictions, we cannot guarantee that the testing/validation metrics used the same model. We therefore report results only for the test split. All methods were median aligned w.r.t. the ground-truth, regardless of the supervision used. This ensures that the evaluations are identical and comparisons are fair.
Table 2. SYNS-Patches Results. We provide metrics across the whole dataset and per scene-category. As expected, supervised methods generally outperform self-supervised ones. The largest gap can be found in Indoor scenes, self-supervised methods were trained exclusively on automotive data. Teams DepthSquad & imec-IDLab-UAntwerp outperformed the challenge baseline [78] by incorporating more advanced network architectures.

<table>
<thead>
<tr>
<th>Scene-Category</th>
<th>Team</th>
<th>No.</th>
<th>F↑</th>
<th>F-Edges↑</th>
<th>MAE↓</th>
<th>RMSE↓</th>
<th>AbsRel↓</th>
<th>Acc-Edges↓</th>
<th>Comp-Edges↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>DJI&amp;ZJU</td>
<td>D 1</td>
<td>17.51</td>
<td>8.80</td>
<td>4.52</td>
<td>8.72</td>
<td>24.32</td>
<td>3.22</td>
<td>21.65</td>
</tr>
<tr>
<td></td>
<td>Pokemon</td>
<td>D 2</td>
<td>16.94</td>
<td>9.63</td>
<td>4.71</td>
<td>8.00</td>
<td>25.35</td>
<td>3.56</td>
<td>19.95</td>
</tr>
<tr>
<td></td>
<td>cv-challenge</td>
<td>D 3</td>
<td>16.70</td>
<td>9.36</td>
<td>4.91</td>
<td>8.63</td>
<td>24.33</td>
<td>3.02</td>
<td>18.07</td>
</tr>
<tr>
<td></td>
<td>imec-IDLab-UAntwerp</td>
<td>MS 4</td>
<td>16.00</td>
<td>8.49</td>
<td>5.08</td>
<td>8.96</td>
<td>28.46</td>
<td>3.74</td>
<td>11.32</td>
</tr>
<tr>
<td></td>
<td>GMD</td>
<td>MS 5</td>
<td>14.71</td>
<td>8.13</td>
<td>5.17</td>
<td>8.97</td>
<td>29.43</td>
<td>3.75</td>
<td>17.29</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>S 6</td>
<td>13.72</td>
<td>7.76</td>
<td>5.56</td>
<td>9.72</td>
<td>32.04</td>
<td>3.97</td>
<td>21.63</td>
</tr>
<tr>
<td></td>
<td>DepthSquad</td>
<td>D 7</td>
<td>12.77</td>
<td>7.68</td>
<td>5.17</td>
<td>8.83</td>
<td>29.92</td>
<td>3.56</td>
<td>35.26</td>
</tr>
<tr>
<td></td>
<td>MonoViTeam</td>
<td>MSD* 8</td>
<td>12.44</td>
<td>7.49</td>
<td>5.05</td>
<td>8.59</td>
<td>28.99</td>
<td>3.10</td>
<td>38.93</td>
</tr>
<tr>
<td></td>
<td>USTC-IAT-United</td>
<td>MS 9</td>
<td>11.29</td>
<td>7.18</td>
<td>5.81</td>
<td>9.58</td>
<td>32.82</td>
<td>3.47</td>
<td>43.38</td>
</tr>
<tr>
<td>Outdoor-Urban</td>
<td>Pokemon</td>
<td>D 2</td>
<td>16.94</td>
<td>8.80</td>
<td>4.52</td>
<td>8.72</td>
<td>24.32</td>
<td>3.22</td>
<td>21.65</td>
</tr>
<tr>
<td></td>
<td>cv-challenge</td>
<td>D 3</td>
<td>15.10</td>
<td>8.48</td>
<td>4.71</td>
<td>8.00</td>
<td>25.35</td>
<td>3.56</td>
<td>19.95</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>S 6</td>
<td>14.09</td>
<td>7.64</td>
<td>5.17</td>
<td>8.83</td>
<td>29.92</td>
<td>3.56</td>
<td>35.26</td>
</tr>
<tr>
<td></td>
<td>DepthSquad</td>
<td>D 7</td>
<td>12.90</td>
<td>5.92</td>
<td>4.49</td>
<td>7.80</td>
<td>27.44</td>
<td>3.26</td>
<td>40.33</td>
</tr>
<tr>
<td></td>
<td>MonoViTeam</td>
<td>MSD* 8</td>
<td>12.44</td>
<td>7.49</td>
<td>5.05</td>
<td>8.59</td>
<td>28.99</td>
<td>3.10</td>
<td>38.93</td>
</tr>
<tr>
<td></td>
<td>USTC-IAT-United</td>
<td>MS 9</td>
<td>11.31</td>
<td>7.37</td>
<td>5.14</td>
<td>8.69</td>
<td>30.64</td>
<td>3.13</td>
<td>40.15</td>
</tr>
<tr>
<td>Outdoor-Natural</td>
<td>Pokemon</td>
<td>D 2</td>
<td>14.90</td>
<td>6.75</td>
<td>6.26</td>
<td>10.47</td>
<td>28.40</td>
<td>3.54</td>
<td>14.44</td>
</tr>
<tr>
<td></td>
<td>cv-challenge</td>
<td>D 3</td>
<td>14.66</td>
<td>6.79</td>
<td>6.35</td>
<td>10.86</td>
<td>27.09</td>
<td>3.08</td>
<td>19.73</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>S 6</td>
<td>12.10</td>
<td>5.82</td>
<td>5.47</td>
<td>7.62</td>
<td>25.90</td>
<td>3.17</td>
<td>17.67</td>
</tr>
<tr>
<td></td>
<td>DepthSquad</td>
<td>D 7</td>
<td>11.54</td>
<td>5.02</td>
<td>4.61</td>
<td>7.02</td>
<td>23.77</td>
<td>2.83</td>
<td>35.26</td>
</tr>
<tr>
<td></td>
<td>MonoViTeam</td>
<td>MSD* 8</td>
<td>10.98</td>
<td>5.38</td>
<td>5.66</td>
<td>8.69</td>
<td>21.18</td>
<td>3.13</td>
<td>30.33</td>
</tr>
<tr>
<td></td>
<td>USTC-IAT-United</td>
<td>MS 9</td>
<td>9.26</td>
<td>4.92</td>
<td>6.11</td>
<td>8.69</td>
<td>30.64</td>
<td>3.13</td>
<td>40.15</td>
</tr>
<tr>
<td>Outdoor-Agriculture</td>
<td>Pokemon</td>
<td>D 2</td>
<td>15.58</td>
<td>6.40</td>
<td>6.26</td>
<td>10.47</td>
<td>28.40</td>
<td>3.54</td>
<td>14.44</td>
</tr>
<tr>
<td></td>
<td>cv-challenge</td>
<td>D 3</td>
<td>14.66</td>
<td>6.79</td>
<td>6.35</td>
<td>10.86</td>
<td>27.09</td>
<td>3.08</td>
<td>19.73</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>S 6</td>
<td>12.10</td>
<td>5.82</td>
<td>5.47</td>
<td>7.62</td>
<td>25.90</td>
<td>3.17</td>
<td>17.67</td>
</tr>
<tr>
<td></td>
<td>DepthSquad</td>
<td>D 7</td>
<td>11.54</td>
<td>5.02</td>
<td>4.61</td>
<td>7.02</td>
<td>23.77</td>
<td>2.83</td>
<td>35.26</td>
</tr>
<tr>
<td></td>
<td>MonoViTeam</td>
<td>MSD* 8</td>
<td>10.98</td>
<td>5.38</td>
<td>5.66</td>
<td>8.69</td>
<td>21.18</td>
<td>3.13</td>
<td>30.33</td>
</tr>
<tr>
<td></td>
<td>USTC-IAT-United</td>
<td>MS 9</td>
<td>9.26</td>
<td>4.92</td>
<td>6.11</td>
<td>8.69</td>
<td>30.64</td>
<td>3.13</td>
<td>40.15</td>
</tr>
<tr>
<td>Indoor</td>
<td>Pokemon</td>
<td>D 2</td>
<td>16.36</td>
<td>5.24</td>
<td>5.17</td>
<td>10.13</td>
<td>29.07</td>
<td>3.43</td>
<td>38.13</td>
</tr>
<tr>
<td></td>
<td>cv-challenge</td>
<td>D 3</td>
<td>14.68</td>
<td>5.82</td>
<td>5.61</td>
<td>10.02</td>
<td>25.90</td>
<td>3.17</td>
<td>17.67</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>S 6</td>
<td>12.10</td>
<td>5.82</td>
<td>5.17</td>
<td>7.62</td>
<td>25.90</td>
<td>3.17</td>
<td>17.67</td>
</tr>
<tr>
<td></td>
<td>DepthSquad</td>
<td>D 7</td>
<td>11.54</td>
<td>5.02</td>
<td>4.61</td>
<td>7.02</td>
<td>23.77</td>
<td>2.83</td>
<td>35.26</td>
</tr>
<tr>
<td></td>
<td>MonoViTeam</td>
<td>MSD* 8</td>
<td>10.98</td>
<td>5.38</td>
<td>5.66</td>
<td>8.69</td>
<td>21.18</td>
<td>3.13</td>
<td>30.33</td>
</tr>
<tr>
<td></td>
<td>USTC-IAT-United</td>
<td>MS 9</td>
<td>9.26</td>
<td>4.92</td>
<td>6.11</td>
<td>8.69</td>
<td>30.64</td>
<td>3.13</td>
<td>40.15</td>
</tr>
</tbody>
</table>

5.1. Quantitative Results

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

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

As expected, supervised approaches using ground-truth depth generally outperformed self-supervised approaches based on the photometric error. However, it is interesting to note that supervising a model with only automotive data (e.g. Team DepthSquad, trained on KEZ) was not sufficient to guarantee generalization to other scene
types. Meanwhile, as discussed in [78], improving the pre-trained backbone (Teams imec-IDLab-UAntwerp & GMD) is one of the most reliable ways of increasing performance. Alternative contributions, such as training with proxy depths (MonoViTeam) or ensembling different architectures (USTC-IAT-United), can improve traditional image-based results but typically result in slightly inferior reconstructions.

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

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

5.2. Qualitative Results

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

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

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

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

6. Conclusions & Future Work

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

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

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

Acknowledgements

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

References


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


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


 Maria Klodt and Andrea Vedaldi. Supervising the New with the Old: Learning SFM from SFM. In European Conference on Computer Vision, pages 713–728, 2018. 2
 Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single
image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015. 2


