

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

The Monocular Depth Estimation Challenge

This is the submitted version (pre peer-review, preprint) of the following publication:

*Published Version:*

Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, et al. (2023).  
The Monocular Depth Estimation Challenge. New York : IEEE Computer Society  
[10.1109/WACVW58289.2023.00069].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/906999> since: 2024-01-17

*Published:*

DOI: <http://doi.org/10.1109/WACVW58289.2023.00069>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

# The Monocular Depth Estimation Challenge

Jaime Spencer<sup>1</sup> C. Stella Qian<sup>2</sup> Chris Russell<sup>3</sup> Simon Hadfield<sup>1</sup> Erich Graf<sup>4</sup>  
Wendy Adams<sup>4</sup> Andrew J. Schofield<sup>2</sup> James Elder<sup>5</sup> Richard Bowden<sup>1</sup> Heng Cong<sup>6</sup>  
Stefano Mattoccia<sup>7</sup> Matteo Poggi<sup>7</sup> Zeeshan Khan Suri<sup>8</sup> Yang Tang<sup>9</sup> Fabio Tosi<sup>7</sup>  
Hao Wang<sup>6</sup> Youmin Zhang<sup>7</sup> Yusheng Zhang<sup>6</sup> Chaoqiang Zhao<sup>9</sup>

<sup>1</sup>University of Surrey <sup>2</sup>Aston University <sup>3</sup>Amazon <sup>4</sup>University of Southampton  
<sup>5</sup>York University <sup>6</sup>Independent <sup>7</sup>University of Bologna <sup>8</sup>DENSO ADAS Engineering  
Services GmbH <sup>9</sup>East China University of Science and Technology

## Abstract

*This paper summarizes the results of the first Monocular Depth Estimation Challenge (MDEC) organized at WACV2023. This challenge evaluated the progress of self-supervised monocular depth estimation on the challenging SYNS-Patches dataset. The challenge was organized on CodaLab and received submissions from 4 valid teams. Participants were provided a devkit containing updated reference implementations for 16 State-of-the-Art algorithms and 4 novel techniques. The threshold for acceptance for novel techniques was to outperform every one of the 16 SotA baselines. All participants outperformed the baseline in traditional metrics such as MAE or AbsRel. However, pointcloud reconstruction metrics were challenging to improve upon. We found predictions were characterized by interpolation artefacts at object boundaries and errors in relative object positioning. We hope this challenge is a valuable contribution to the community and encourage authors to participate in future editions.*

ture gradients, perspective distortions, stereo/motion parallax and more. Networks performing MDE must also learn these geometric cues, rather than just rely on correspondence matching.

The rise in popularity of this field has resulted in a plethora of contributions, including supervised [8, 46], self-supervised [9, 11, 12, 15] and weakly-supervised [47, 55, 59] approaches. Comparing these approaches in a fair and consistent manner is a highly challenging task, as it is the responsibility of each author to ensure they are following the same procedures as preceding methods. The need to provide this backward-compatibility can result in long-standing errors in the benchmarking procedure, ranging from incorrect metric computation and data preprocessing to incorrect ground-truths.

This paper covers the recent Monocular Depth Estimation Challenge (MDEC), organized as part of a workshop at WACV2023. The objective of this challenge was to provide an updated and centralized benchmark to evaluate contributions in a fair and consistent manner. This first edition focused on self-/weakly-supervised MDE, as they have the possibility to scale to larger amounts of data and do not require expensive LiDAR ground-truth. Despite this flexibility, the majority of published approaches train and evaluate only on automotive data. As part of this challenge, we tested the generalization of these approaches to a wider range of scenarios, including natural, urban and indoor scenes. This was made possible via the recently released SYNS-Patches dataset [1, 52]. In general, participants found it challenging to outperform the updated Garg baseline [9, 52] in pointcloud-based reconstruction metrics (F-Score), but generally improved upon traditional image-based metrics (MAE, RMSE, AbsRel).

## 1. Introduction

Depth estimation is a core computer vision task, allowing us to recover the 3-D geometry of the world. Whilst traditional approaches to depth estimation relied on stereo [17, 53, 5] or multi-view [2, 48, 31] matching, monocular approaches [8, 9, 12, 46] requiring only a single image have recently garnered much attention.

The task of Monocular Depth Estimation (MDE) is ill-posed, as an infinite number of scene arrangements with varying object sizes and depths could result in the same 2-D image projection. However, humans are capable of performing this task by relying on cues and priors such as absolute/relative object sizes, elevation within the scene, tex-

Table 1: **Dataset & Benchmark Comparison.** We summarize recent datasets commonly used in self-supervised monocular depth estimation. CityScapes represents a common pretraining dataset, while SYNS-Patches is testing-only. SYNS-Patches is the only dataset providing high-quality dense depth maps in a wide variety of environments.

	Accuracy	Density (%)	Num Points	Urban	Natural	Indoor	Train	Val	Test
CityScapes [6]	✗	✗	✗	✓	✗	✗	88,250	✗	✗
Kitti Eigen-Zhou [10, 66]	✗	✗	✗	✓	✗	✗	39,810	4,424	✗
Kitti Eigen [10, 8]	Mid	4.10	19k	✓	✗	✗	45,200	1,776	697
Kitti Eigen-Benchmark [10, 56]	High	15.28	71k	✓	✗	✗	71,633	5,915	652
DDAD [15]	High	1.02	24k	✓	✗	✗	75,900	23,700	3,080
SYNS-Patches [1, 52]	High	78.30	365k	✓	✓	✓	✗	400	775

## 2. Related Work

To avoid using costly Light Detection and Ranging (LiDAR) annotations, self-supervised approaches to MDE instead rely on the proxy task of image reconstruction via view synthesis. The predicted depth is combined with a known (or estimated) camera transform to establish correspondences between adjacent images. This means that, whilst the network can predict depth from a single input image at test time, the training procedure requires multiple support frames to perform the view synthesis.

Methods can be categorized based on the source of these support frames. Stereo methods [9, 11, 47, 55] rely on stereo rectified images pairs with a known and fixed camera baseline. This allows the network to predict metric depth, but can result in occlusions artefacts if not trained carefully. On the other hand, monocular approaches [66, 23, 57] commonly use the previous and following frame from a monocular video. These approaches are more flexible, as no stereo data is required. However, they are sensitive to the presence of dynamic objects. Furthermore, depth is predicted only up to an unknown scale factor and requires median scaling during evaluation to align it with the ground-truth.

Garg [9] introduced the first approach to MDE via stereo view synthesis, using AlexNet [26] and an  $L_1$  reconstruction loss. Monodepth [11] drastically improved the performance through bilinear synthesis [18] and a weighted combination of SSIM [58] and  $L_1$ . It additionally incorporated virtual stereo supervision and a smoothness regularization weighted by the strength of the image edges. 3Net [45] extended this to a trinocular setting, while DVSO [47] and MonoResMatch [55] incorporated an additional residual refinement network.

SfM-Learner [66] introduced the first fully monocular framework, replacing the fixed stereo baseline with a Visual Odometry (VO) regression network. A predictive mask was introduced to downweigh the photometric loss at independently moving dynamic objects. Future methods refined this masking procedure via uncertainty estimation [21, 23], object motion prediction [28, 22, 30] and au-

tomasking [11, 4]. Monodepth2 [12] additionally proposed the minimum reprojection loss as a simple way of handling varying occlusions in a sequence of frames. Instead of averaging the reconstruction loss over the sequence, they proposed to take only the minimum loss across each image pixel, assuming this will select the frame with the non-occluded correspondence.

Subsequent approaches focused on improving the robustness of the photometric loss by incorporating feature descriptors [63, 51, 50], affine brightness changes [61], scale consistency [37, 4] or adversarial losses [3, 44, 33]. Meanwhile, the architecture of the depth prediction network was improved to target higher-resolution predictions by incorporating sub-pixel convolutions [49, 42], 3-D packing blocks [15], improved skip connections [60, 65, 36], transformers [64] and discrete disparity volumes [20, 13, 14].

Several methods incorporated additional supervision in the form of (proxy) depth regression from LiDAR [27, 16], synthetic [35], SLAM [23, 57, 47], hand-crafted stereo [55, 59, 33], the matted Laplacian [14] and self-distillation [43, 41]. One notable example is DepthHints [59], which combined hand-crafted disparity [17] with the min reprojection loss [12]. This provided a simple way of fusing multiple disparity maps into a single robust estimate.

### 2.1. Datasets & Benchmarks

This section reviews some of the most commonly used datasets and benchmarks used to evaluate MDE. Despite being a fundamental and popular computer vision task, there has not been a standard centralized challenge such as ImageNet [7], VOT [25] or IMC [19]. This makes it challenging to ensure that all methods use a consistent evaluation procedure. Furthermore, the lack of a withheld test set encourages overfitting due to repeated evaluation. Table 1 provides an overview of these datasets.

Kitti [10] is perhaps the most common training and testing dataset for MDE. It was popularized by the Kitti Eigen (KE) split [8], containing 45k images for training and 697 for testing. However, this benchmark contains some long-

Table 2: SYNS-Patches Category Distribution.

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

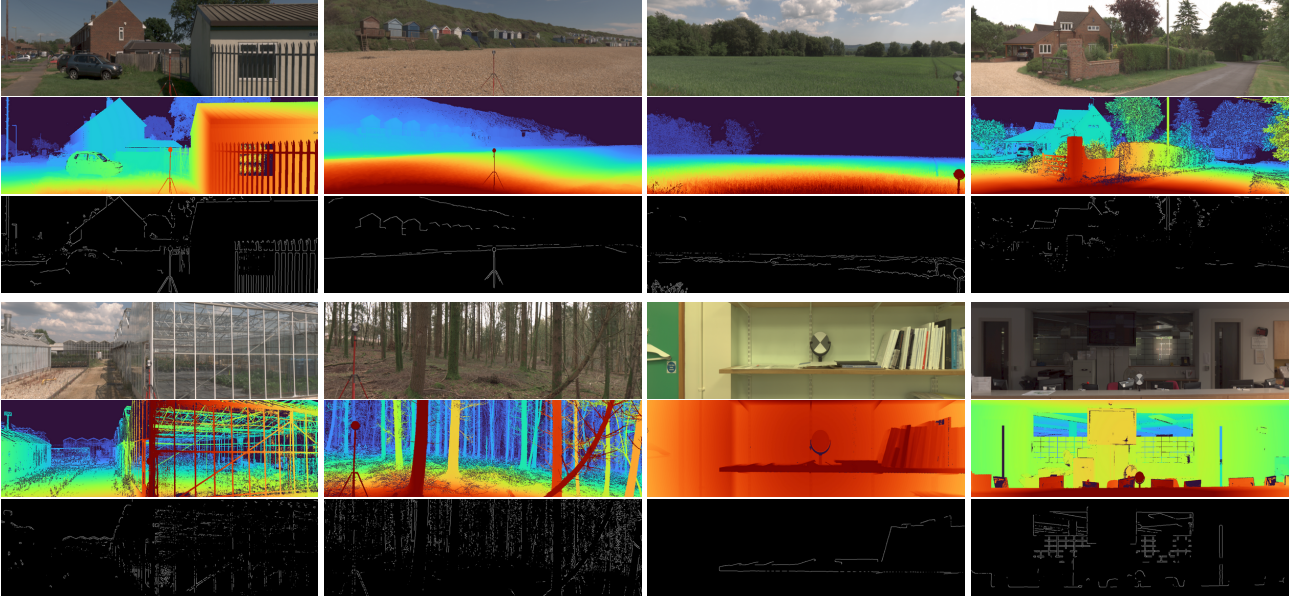


Figure 1: SYNS-Patches Challenge Dataset. We show some representative examples from the diverse set of testing categories. This includes complex urban, natural and indoor scenes with high-quality dense LiDAR. Depth boundaries were computed as Canny edges in the log-depth maps.

standing errors that heavily impact the accuracy of the results. The ground-truth depth suffers from background bleeding at object boundaries due to the different sensor viewpoints, coupled with the motion artefacts produced by the moving LiDAR. Furthermore, the data preprocessing omitted the transformation to the camera reference frame. These issues are further exacerbated by the sparsity of the ground-truth depth maps, which contain measurements for only 4.10% of the image pixels.

Uhrig *et al.* [56] aimed to correct these errors and provide a more reliable benchmark, dubbed the Kitti Eigen-Benchmark (KEB) split. The ground-truth density was improved to 15.28% by accumulating LiDAR data from  $\pm 5$  adjacent frames. This data was aggregated and refined by adding consistency checks using a hand-crafted stereo matching algorithm [17]. The main drawback is that this refinement procedure removes points at object boundaries, which are common sources of errors even in State-of-the-Art (SotA) approaches. However, despite providing a clear improvement over KE, adoption by the community has been slow. We believe this to be due to the need to provide consistent comparisons against previous methods that only

evaluate on KE, as this would require authors to re-run all preexisting approaches on this new baseline.

The DDAD dataset [15] contains data from multiple cities in USA and Japan and totalling to 76k training and 3k testing images. It provides an density of 1.03%, an average of 24k points per image and an increased depth range up to 250 meters. This dataset was the focus on the DDAD challenge organized at CVPR 2021, which featured additional fine-grained performance metrics on each semantic class. Similar to KEB, we believe that adoption of these improved datasets is hindered by the need to re-train and re-evaluate preexisting methods.

Spencer *et al.* [52] aimed to unify and update the training and benchmarking procedure for MDE. This was done by providing a public repository containing modernized SotA implementations of 16 recent approaches with common robust design decisions. The proposed models were evaluated on the improved KEB and SYNS-Patches, incorporating more informative pointcloud- [39] and edge-based [24] metrics. This modern benchmark procedure constitutes the basis of the Monocular Depth Estimation Challenge.



### 3. The Monocular Depth Estimation Challenge

The first edition of MDEC was organized as part of a WACV2023 workshop. The challenge was organized on CodaLab [40] due to its popularity and flexibility, allowing for custom evaluation scripts and metrics. We plan to arrange a permanent leaderboard on CodaLab that remains open to allow authors to continue evaluating on SYNS-Patches.

The first two weeks of the challenge constituted the development phase, where participants could submit predictions only on the validation split of SYNS-Patches. For the remainder of the challenge, participants were free to submit to either split. Participants only had access to the dataset images, while the ground-truth depth maps and depth boundaries were withheld to prevent overfitting.

#### 3.1. Dataset

The evaluation for the challenge was carried out on the recently introduced SYNS-Patches dataset [52], which is a subset of SYNS [1]. The original SYNS is composed of aligned image and LiDAR panoramas from 92 different scenes belonging to a wide variety of environments, such as Agriculture, Natural (*e.g.* forests and fields), Residential, Industrial and Indoor. This is a departure from the commonly used datasets in the field, such as Kitti [10], CityScapes [6] or DDAD [15], which focus purely on urban scenes collected by automotive vehicles. SYNS also provides dense LiDAR maps with 78.30% coverage and 365k points per image, which are exceptionally rare in outdoor environments. This allows us to compute metrics targeting complex image regions, such as thin structures and depth boundaries, which are common sources of error.

SYNS-Patches represents the subset of patches from each scene extracted at eye level at 20 degree intervals of a full horizontal rotation. This results in 18 images per scene and a total dataset size of 1656. Since the data collection procedure is highly sensitive to dynamic objects, additional manual verification is required. The final dataset consists of 1175 images, further separated into validation and testing splits of 400 and 775 images. We show some representative testing images in Figure 1 and the distribution of images categories per split in Table 2

#### 3.2. Training procedure

The first edition of MDEC focused on evaluating the State-of-the-Art in self-supervised monocular depth estimation. This included methods complemented by hand-crafted proxy depth maps or synthetic data. We expected most methods to be trained on Kitti [10] due to its widespread use. However, we placed no restrictions on the training dataset (excluding SYNS/SYNS-Patches) and encouraged participants to use additional training sources.

To aid participants and give a strong entry point, we provided a public starting kit on GitHub<sup>1</sup>. This repository contained the training and evaluating code for 16 recent SotA contributions to MDE. The baseline submission was the top F-Score performer out of all SotA approaches in this starting kit [9, 52]. This consisted of a ConvNeXt [34] backbone and DispNet [38] decoder. The model was trained on the Kitti Eigen-Zhou split with an image resolution of  $192 \times 640$  using only stereo view synthesis, the vanilla photometric loss and edge-aware smoothness regularization.

#### 3.3. Evaluation procedure

Participants provided their unscaled disparity predictions at the training image resolution. Our evaluation script bilinearly upsampled the predictions to the full image resolution and applied median scaling to align the predicted and ground-truth depths. Finally, the prediction and ground-truth were clamped to a maximum depth of 100m. We omit test-time stereo blending [11] and border cropping [8].

#### 3.4. Performance metrics

The predictions were evaluated using a wide variety of image/pointcloud/edge-based metrics. Submissions were ranked based on the F-Score performance [39], as this targets the structural quality of the reconstructed pointcloud. We provide the units of each metric, as well as an indication if lower ( $\downarrow$ ) or higher ( $\uparrow$ ) is better.

##### 3.4.1 Image-based

**MAE.** Absolute error ( $m\downarrow$ ) as

$$\sum |\hat{y} - y|, \quad (1)$$

where  $y$  represents the ground-truth depth at a single image pixel  $\mathbf{p}$  and  $\hat{y}$  is the predicted depth at that pixel.

**RMSE.** Absolute error ( $m\downarrow$ ) with higher outlier weight as

$$\sqrt{\sum (\hat{y} - y)^2}. \quad (2)$$

**AbsRel.** Range-invariant relative error ( $\%\downarrow$ ) as

$$\sum \frac{|\hat{y} - y|}{y}. \quad (3)$$

##### 3.4.2 Pointcloud-based

**F-Score.** Reconstruction accuracy ( $\%\uparrow$ ) given by the harmonic mean of **Precision** and **Recall** as

$$2 \cdot \frac{P \cdot R}{P + R}, \quad (4)$$

<sup>1</sup>[https://github.com/jspenmar/monodepth\\_benchmark](https://github.com/jspenmar/monodepth_benchmark)

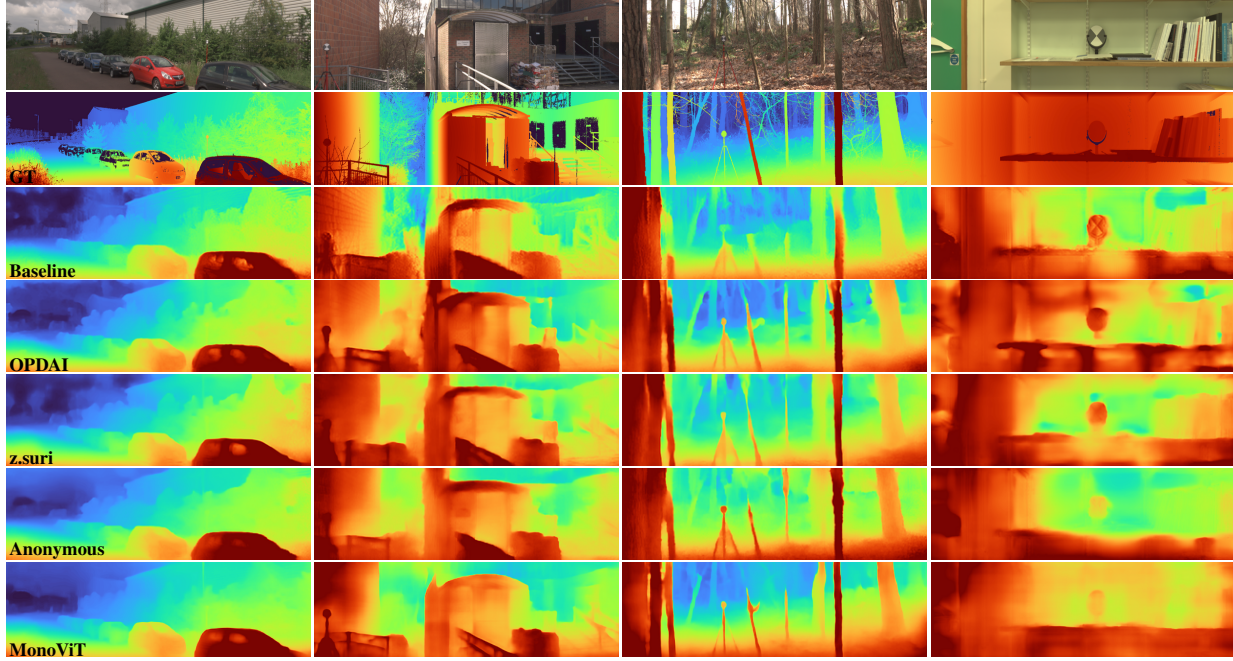


Figure 2: **SYNS-Patches Depth Visualization.** Models perform significantly better in urban environments that resemble the training automotive data. Thin structures, such as the railings and branches, are highly challenging to predict accurately and are commonly merged together.

**Precision.** Percentage ( $\% \uparrow$ ) of predicted 3-D points  $\hat{\mathbf{q}}$  within a threshold  $\delta$  of a ground-truth point  $\mathbf{q}$  as

$$\sum_{\hat{\mathbf{q}} \in \hat{Q}} \left[ \left[ \min_{\mathbf{q} \in Q} \|\mathbf{q} - \hat{\mathbf{q}}\| < \delta \right] \right], \quad (5)$$

where  $[\cdot]$  represents the Iverson brackets.

**Recall.** Percentage ( $\% \uparrow$ ) of ground-truth 3-D points within a threshold of a predicted point as

$$\sum_{\mathbf{q} \in Q} \left[ \left[ \min_{\hat{\mathbf{q}} \in \hat{Q}} \|\mathbf{q} - \hat{\mathbf{q}}\| < \delta \right] \right]. \quad (6)$$

Following Örnek *et al.* [39], the threshold for a correctly reconstructed point is set to 10 cm *i.e.*  $\delta = 0.1$ . Note that Precision and Recall are only used to compute the F-Score and are not reported in the challenge leaderboard.

### 3.4.3 Edge-based

**F-Score.**

Pointcloud reconstruction accuracy ( $\% \uparrow$ ) computed only at ground-truth  $\mathbf{M}$  and predicted  $\hat{\mathbf{M}}$  depth boundaries, represented by binary masks.

**Accuracy.** Distance ( $\text{px} \downarrow$ ) from each predicted depth boundary to the closest ground-truth boundary as

$$\sum EDT(\hat{\mathbf{M}}(\mathbf{p})) \mathbf{M}(\mathbf{p}), \quad (7)$$

where  $EDT$  represents the Euclidean Distance Transform.

**Completeness.** Distance ( $\text{px} \downarrow$ ) from each ground-truth depth boundary to the closest predicted boundary as

$$\sum EDT(\mathbf{M}(\mathbf{p})) \hat{\mathbf{M}}(\mathbf{p}). \quad (8)$$

These metrics were proposed as part of the IBims-1 [24] benchmark, which features dense indoor depth maps.

## 4. Challenge Submissions

### Baseline

<i>J. Spencer</i> <sup>1</sup>	<i>j.spencermartin@surrey.ac.uk</i>
<i>C. Russell</i> <sup>3</sup>	<i>cmruss@amazon.de</i>
<i>S. Hadfield</i> <sup>1</sup>	<i>s.hadfield@surrey.ac.uk</i>
<i>R. Bowden</i> <sup>1</sup>	<i>r.bowden@surrey.ac.uk</i>

Challenge organizers submission. Re-implementation of Garg [9] from the updated monocular depth benchmark[52]. Trained with stereo photometric supervision with edge-aware smoothness regularization. Network is composed of a ConvNeXt-B backbone [34] with a DispNet [38] decoder. Trained for 30 epochs on Kitti Eigen-Zhou with an image resolution of  $192 \times 640$ .

### 4.1. Team 1 - OPDAI

<i>H. Wang</i> <sup>6</sup>	<i>hwscut@126.com</i>
<i>Y. Zhang</i> <sup>6</sup>	<i>yusheng.z1995@gmail.com</i>
<i>H. Cong</i> <sup>6</sup>	<i>congheng@outlook.com</i>

Based on a ConvNext-B [34] with an HRDepth [36] decoder. Trained with monocular and stereo data, along with

Table 3: **SYNS-Patches Results.** Each sections reports results over a different set of scene categories. The updated SotA models from [52] provide a strong baseline for the challenge, outperforming all submissions in pointcloud reconstruction F-Score. However, all submissions significantly improved upon traditional image-based metrics by up to 6.47% in MAE and 7.42% in AbsRel. All approaches adapt better to unseen outdoor urban environments than the challenging natural and agricultural scenes.

		F-Score $\uparrow$	F-Score (Edges) $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	AbsRel $\downarrow$	EdgeAcc $\downarrow$	EdgeComp $\downarrow$
<i>Overall</i>	Baseline	<b>13.72</b>	<b>7.76</b>	5.56	9.72	32.04	3.97	<b>21.63</b>
	OPDAI	<a href="#">13.53</a>	7.41	<b>5.20</b>	<a href="#">8.98</a>	<b>29.66</b>	<a href="#">3.67</a>	<a href="#">27.31</a>
	z.suri	13.08	7.46	5.39	9.27	29.96	3.81	32.70
	Anonymous	12.85	7.30	5.32	9.04	30.22	3.83	43.77
	MonoViT	12.66	<a href="#">7.51</a>	<a href="#">5.22</a>	<b>8.96</b>	<a href="#">29.70</a>	<b>3.36</b>	35.47
<i>Outdoor-Urban</i>	Baseline	<b>14.09</b>	<b>6.48</b>	4.77	8.43	29.10	3.89	<b>22.75</b>
	OPDAI	<a href="#">13.17</a>	<a href="#">5.99</a>	<b>4.53</b>	<a href="#">7.93</a>	<b>27.12</b>	<a href="#">3.47</a>	<a href="#">27.71</a>
	z.suri	12.72	5.97	4.77	8.25	27.99	3.64	34.31
	Anonymous	12.83	5.56	4.60	7.95	28.04	3.66	41.04
	MonoViT	12.00	5.87	<a href="#">4.54</a>	<b>7.85</b>	<a href="#">27.91</a>	<b>3.12</b>	35.24
<i>Outdoor-Natural</i>	Baseline	<b>12.11</b>	<a href="#">5.32</a>	7.46	12.86	36.89	3.84	<b>18.35</b>
	OPDAI	11.61	5.26	<b>6.82</b>	<b>11.52</b>	<a href="#">33.53</a>	<a href="#">3.52</a>	<a href="#">24.11</a>
	z.suri	11.40	5.25	7.14	12.07	34.43	3.61	30.96
	Anonymous	11.83	5.31	7.11	11.76	34.16	3.59	38.96
	MonoViT	<a href="#">11.84</a>	<b>5.72</b>	<a href="#">6.92</a>	<a href="#">11.72</a>	<b>33.33</b>	<b>3.30</b>	31.33
<i>Outdoor-Agriculture</i>	Baseline	<a href="#">12.26</a>	<b>4.77</b>	6.10	10.84	33.58	4.00	<b>18.73</b>
	OPDAI	12.26	4.47	<a href="#">5.78</a>	10.20	31.53	<a href="#">3.69</a>	<a href="#">27.38</a>
	z.suri	<b>12.75</b>	4.40	5.85	10.33	<b>30.52</b>	3.77	29.03
	Anonymous	11.53	4.20	5.78	<a href="#">10.12</a>	<a href="#">30.76</a>	3.87	41.89
	MonoViT	11.34	<a href="#">4.57</a>	<b>5.72</b>	<b>10.03</b>	30.99	<b>3.40</b>	33.53
<i>Indoor</i>	Baseline	<a href="#">21.11</a>	<b>28.96</b>	1.04	1.51	22.77	<a href="#">4.60</a>	<a href="#">37.09</a>
	OPDAI	<b>23.56</b>	27.95	<a href="#">1.00</a>	1.54	<b>21.12</b>	4.82	<b>36.28</b>
	z.suri	19.95	<a href="#">28.84</a>	<b>0.98</b>	<b>1.42</b>	21.44	5.20	43.93
	Anonymous	19.32	28.82	1.07	1.55	23.94	5.10	74.43
	MonoViT	20.45	27.65	1.01	<a href="#">1.49</a>	<a href="#">21.16</a>	<b>4.28</b>	55.52

proxy depth hints [59]. This submission uses a large combination of losses, including the photometric loss with an explainability mask [66], autoencoder feature-based reconstruction [50], virtual stereo [11], proxy depth regression, edge-aware disparity smoothness [11], feature smoothness [50], occlusion regularization [47] and explainability mask regularization [66]. The models were trained on Kitti Eigen-Zhou (KEZ) without depth hints and KEB with depth hints for 5 epochs and an image resolution of  $192 \times 640$ .

#### 4.2. Team 2 - z.suri

Z. K. Suri<sup>8</sup>      z.suri@eu.denso.com

The depth and pose estimation networks used ConvNeXt-B [34] as the encoder, with the depth network complemented by a DiffNet [65] decoder. Trained with both stereo and monocular inputs, using edge-aware regularization [11] and the min reconstruction photometric loss with automasking [12]. A strong pose network is essential for accurate monocular depth estimation. This submission introduced a

stereo pose regression loss. The pose estimation network was additionally given a stereo image pair and supervised w.r.t. the know ground-truth camera baseline between them. The networks were trained on Kitti Eigen-Zhou with an image resolution of  $192 \times 640$ .

#### 4.3. Team 3 - Anonymous

The author of this submission did not provide any details.

#### 4.4. Team 4 - MonoViT

C. Zhao<sup>9</sup>      y20180082@mail.ecust.edu.cn  
M. Poggi<sup>7</sup>      m.poggi@unibo.it  
F. Tosi<sup>7</sup>      fabio.tosi5@unibo.it  
Y. Zhang<sup>7</sup>      youmin.zhang2@unibo.it  
Y. Tang<sup>9</sup>      yangtang@ecust.edu.cn  
S. Mattoccia<sup>7</sup>      stefano.mattoccia@unibo.it

Trained on KE with an image resolution of  $320 \times 1024$ . The depth network used the MonoViT [64] architecture, combining convolutional and MPViT [29] encoder blocks.

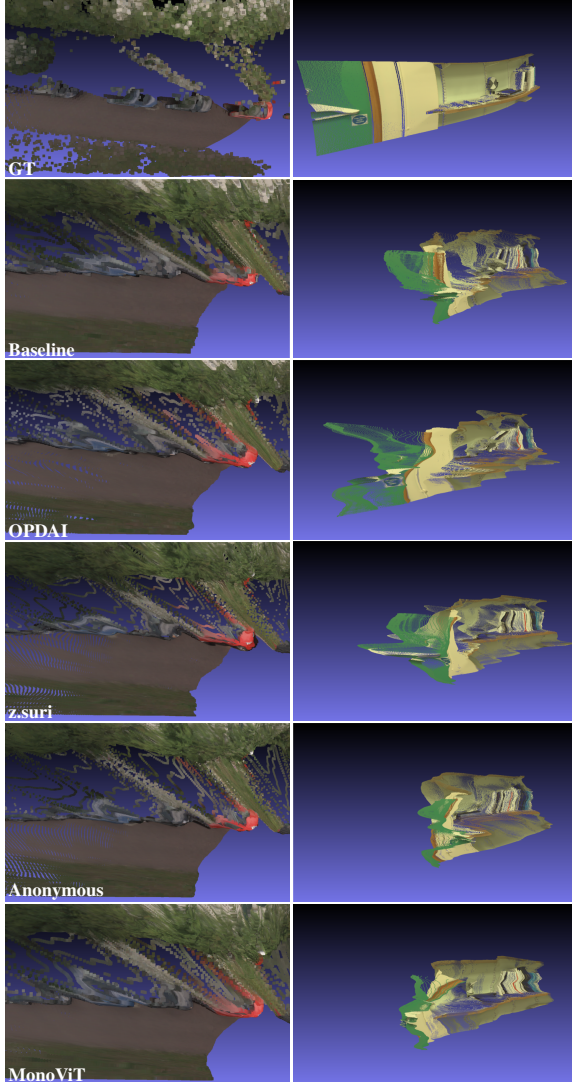


Figure 3: **SYNS-Patches Pointcloud Visualization.** Converting the depth maps into pointclouds allows us to evaluate the quality of the reconstructed scene. All approaches can reliably estimate the road surface ground plane. However, object boundaries exhibit smooth interpolation artefacts connecting them to background structures. Adapting to previously unseen indoor environments is still highly challenging.

The network was trained using stereo and monocular support frames, based on the minimum photometric loss [12], edge-aware smoothness [11] and  $L_1$  proxy depth regression. Proxy depth labels were obtained by training a self-supervised stereo network [32, 54] on the Multiscopic dataset [62]. This dataset provides three horizontally aligned images, allowing the network to compensate for occlusions. The pretrained stereo network was trained using Center and Right pairs, but used the full triplet when computing the per-pixel minimum photometric loss. It was trained for 1000 epochs using  $256 \times 480$  crops.

## 5. Results

Table 3 show the performance of the participants' submissions on the SYNS-Patches test set. As seen, most submissions outperformed the baseline in traditional image-based metrics (MAE, RMSE, AbsRel) across all scene types. However, the baseline still achieved the best performance in both pointcloud reconstruction metrics (F-Score (Edges)). We believe this is due to the fact that most existing benchmarks report only image-based metrics. As such, novel contributions typically focus on improving performance on only these metrics. However, we believe pointcloud-based reconstruction metrics [39] are crucial to report, as they reflect the true objective of monocular depth estimation.

As expected, all approaches transfer best to other Outdoor Urban environments, while the previously unseen Natural and Agriculture category provided a more difficult challenge. In most outdoor environments the baseline provides the best F-Score performance, while OPDAI & MonoViT improve on image-based metrics ( $> 0.5$  meter improvement in Outdoor Natural MAE). It is also interesting to note that all approaches improve the accuracy of the detected edges by roughly 15%. Meanwhile, edge completeness is drastically reduced, implying that participant submissions are more accurate at extracting strong edges, but oversmooth predictions in highly textured regions. Finally, it is worth noting that the increased metric performance in indoor environments is likely due to the significantly decreased depth range.

We show qualitative visualizations for the predicted depth maps and pointclouds in Figures 2 & 3, respectively. The target images were selected prior to evaluation to reflect the wide variety of available environments. Generally, we find that most predictions are oversmoothed and lack high-frequency detail. For instance, many models fill in gaps between thin objects, such as railings (second image) or branches (third image). As is expected, all submissions tend to perform better in urban settings, as they are more similar to the training distribution. The submission by MonoViT generally produces the highest-quality visualizations, with more detailed thin structures and sharper boundaries. This is reflected by the improved image-based metrics. However, as seen in the pointcloud visualizations in Figure 3, these predictions still suffer from boundary interpolation artefacts that are not obvious in the depth map visualizations. This reinforces the need for more detailed metrics in these complex image regions.

## 6. Conclusions & Future Work

This paper has presented the results for the first edition of Monocular Depth Estimation Challenge. It was interesting to note that, while most submissions outperformed the



baseline in traditional image-based metrics (MAE, RMSE, AbsRel), they did not improved pointcloud F-Score reconstruction. As expected, SYNS-Patches represents a challenging dataset for current monocular depth estimation systems. We believe this to be due to the over-reliance on automotive training data. Despite its availability and ease of collection, it does not contain varied enough scenarios to generalize to more complex natural scenes. As such, it is likely that additional sources of training data are required to develop truly generic perception systems.

Future editions of MDEC may expand to additionally evaluate supervised MDE approaches. This would help compare the SotA in both branches of research and help to determine the reliability of supervised networks. We hope this provides a valuable contribution to the community and strongly encourage authors in this field to participate in future editions of the challenge.

## Acknowledgements

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

## References

- [1] Wendy J Adams, James H Elder, Erich W Graf, Julian Leyland, Arthur J Lugtigheid, and Alexander Murry. The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 6(1):35805, 2016.
- [2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.
- [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] 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.
- [5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018.
- [6] 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.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Juan Luis Gonzalez Bello and Munchurl Kim. Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes. In *Advances in Neural Information Processing Systems*, volume 33, pages 12626–12637, 2020.
- [14] 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.
- [15] 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.
- [16] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11078–11088, June 2021.
- [17] Heiko Hirschmüller. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. *Conference on Computer Vision and Pattern Recognition*, II:807–814, jun 2005.
- [18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.
- [19] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021.
- [20] 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.



- [21] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [22] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020.
- [23] 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.
- [24] 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.
- [25] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Johanna Bjorklund, Yushan Zhang, Zhongqun Zhang, Song Yan, Wenyan Yang, Dingding Cai, Christoph Mayer, and Gustavo Fernandez. The tenth visual object tracking vot2022 challenge results, 2022.
- [26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [27] Yevhen Kuznetsov, Jörg Stückler, and Bastian Leibe. Semi-Supervised Deep Learning for Monocular Depth Map Prediction. In *Conference on Computer Vision and Pattern Recognition*, pages 2215–2223, 2017.
- [28] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):1863–1872, May 2021.
- [29] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7286, 2022.
- [30] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. In *Conference on Robot Learning*, pages 1908–1917. PMLR, 2021.
- [31] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5987–5997, October 2021.
- [32] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.
- [33] Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Self-supervised monocular depth estimation for all day images using domain separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12737–12746, October 2021.
- [34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.
- [35] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single View Stereo Matching. *Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.
- [36] 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.
- [37] 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.
- [38] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. *Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.
- [39] Evin Pinar Örneke, 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.
- [40] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letourne, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. CodaLab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022.
- [41] Andra Petrovai and Sergiu Nedevschi. Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1588, June 2022.
- [42] 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.
- [43] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [44] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *2018 international conference on 3D vision (3DV)*, pages 587–595. IEEE, 2018.
- [45] 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.

- [46] 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.
- [47] 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.
- [48] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [49] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *Conference on Computer Vision and Pattern Recognition*, 2016-Decem:1874–1883, 2016.
- [50] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-Metric Loss for Self-supervised Learning of Depth and Egomotion. In *European Conference on Computer Vision*, pages 572–588, 2020.
- [51] 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.
- [52] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *arXiv preprint arXiv:2208.01489*, 2022.
- [53] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, 2003.
- [54] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.
- [55] 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.
- [56] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity Invariant CNNs. *International Conference on 3D Vision*, pages 11–20, 2018.
- [57] 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.
- [58] Zhou Wang, A C Bovik, H R Sheikh, and E P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [59] Jamie Watson, Michael Firman, Gabriel Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. *International Conference on Computer Vision*, 2019-Octob:2162–2171, 2019.
- [60] 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.
- [61] 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.
- [62] Weihao Yuan, Yazhan Zhang, Bingkun Wu, Siyu Zhu, Ping Tan, Michael Yu Wang, and Qifeng Chen. Stereo matching by self-supervision of multiscopic vision. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5702–5709. IEEE, 2021.
- [63] 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.
- [64] 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.
- [65] Hang Zhou, David Greenwood, and Sarah Taylor. Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. In *British Machine Vision Conference*, 2021.
- [66] 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.