Benchmarking Out-of-Distribution Depth Estimation under Corruptions

Lingdong Kong[1,2] Shaoyuan Xie[3] Hanjiang Hu[4] Lai Xing Ng[5] Benoit Cottereau[6] Wei Tsang Ooi[1,✉]
[1]National University of Singapore [2]CNRS@CREATE [3]Huazhong University of Science and Technology
[4]Carnegie Mellon University [5]Institute for Infocomm Research, A*STAR [6]CNRS
{lingdong,ooiwt}@comp.nus.edu.sg shaoyuanxie@hust.edu.cn hanjianghu@cmu.edu
ng_lai_xing@i2r.a-star.edu.sg benoit.cottereau@cnrs.fr

Abstract

Depth estimation from monocular images plays an important role in real-world visual perception systems. However, learning-based depth estimation models are trained and tested on clean data while ignoring out-of-distribution (OoD) situations. Common corruptions tend to happen in practical scenarios, especially for safety-critical applications like autonomous driving and robot navigation. To fill in this gap, we present a comprehensive robustness test suite dubbed ***RoboDepth***. *It consists of 18 corruptions from three categories: 1) weather and lighting conditions; 2) sensor failure and movement; and 3) data processing issues. Then, we conduct a comprehensive benchmark on 42 existing depth estimation models from indoor and outdoor scenes, to evaluate their robustness under corruptions. Our benchmark results indicate that, although promising results have been achieved, state-of-the-art depth estimation models are at risk of being vulnerable to corruptions. We further make in-depth discussions on the design considerations of building more robust depth estimation models, from aspects including pre-training, augmentation, modality, and learning paradigm. We hope our benchmark can lay a solid foundation for robust OoD depth estimation. The benchmark suite and toolkit are publicly available at* https://github.com/ldkong1205/RoboDepth.

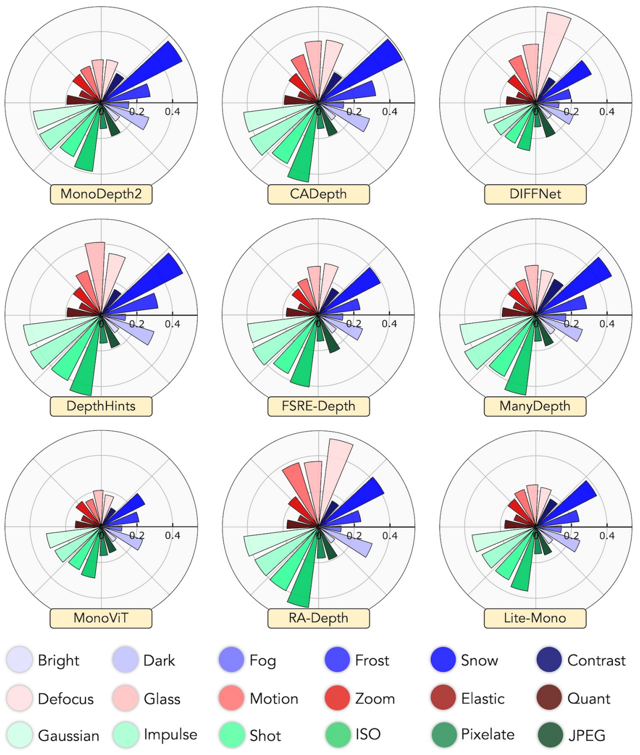


Figure 1. The depth estimation robustness (in terms of depth estimation error (DEE) defined in Sec. 3) under 18 corruptions in radar charts. Different MDE models exhibit diverse strengths and weaknesses against different corruptions that occur in the real world.

1. Introduction

Monocular depth estimation (MDE) is the task of predicting the depth of a scene using only a single image, typically captured using cameras equipped on drones, mobile robots, and vehicles [9, 21, 30, 49]. Accurate MDE is crucial for a wide range of applications, such as autonomous driving, augmented reality, and robot navigation. With the advent of recent learning-based paradigms, various MDE algorithms have been proposed and achieved promising performance on standard benchmark datasets [8, 12, 35, 37].

However, existing MDE models face significant challenges in terms of out-of-distribution (OoD) robustness under real-world corruptions, such as adverse weather [16] and sensor failure [20]. In particular, the existing learning-based visual perception models are highly sensitive to variations in lighting, noise, texture shift, and other factors that can distort the image and lead to inaccurate predictions [15, 19]. Furthermore, these models can struggle with the generalization of new scenes and objects that they have not encountered during training [29].

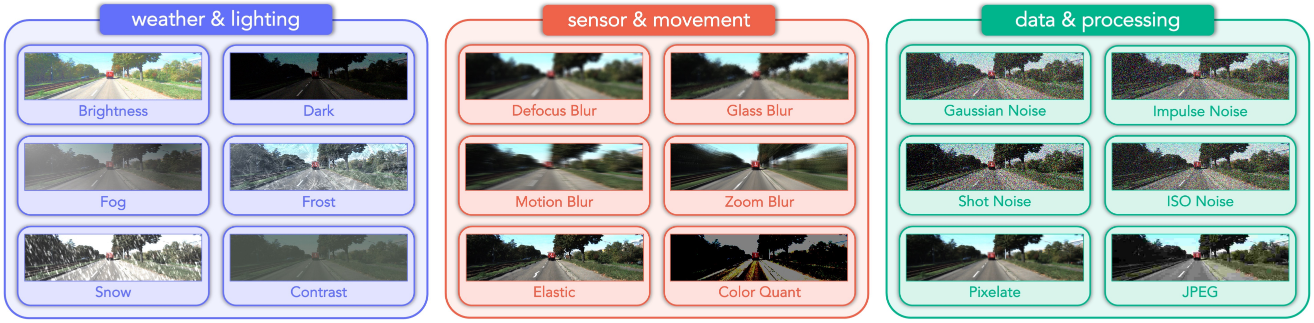


Figure 2. Corruption taxonomy. We break down common corruptions into three categories: 1) Weather and lighting conditions, such as sunny, low-light, fog, frost, snow, and contrast. 2) Sensor failure and movement, such as potential blurs (defocus, glass, motion, zoom) caused by motion. 3) Data processing issues, such as noises (Gaussian, impulse, ISO) happen due to hardware malfunctions.

As one of the basic visual perception tasks, models for MDE also likely learn the systematic errors in depth estimations; corruptions and perturbations such as lighting changes, motion blur, shadows, data compression, *etc*., are present in real images and rarely dealt adequately in current MDE systems [4, 24]. Despite the promising results constantly achieved on the relatively "clean" datasets [8,12,35], the absence of a suitable robustness benchmark hinders the further development of resilient and scalable MDE systems.

In this work, we make the first step towards robust MDE by establishing the *KITTI-C*, *NYUDepth2-C*, and *KITTI-S* benchmarks. Different from existing works on either merging multiple datasets for cross-domain MDE [26, 33, 46] or designing adversarial patches to attack MDE models [4, 7], our benchmarks aim at simulating common corruptions that have a high likelihood to occur in real-world environments. As shown in Fig. 2, we design 18 corruption types from three main categories: weather and lighting conditions, sensor failure and movement, and data processing issues. Each of these corruptions, which is further divided into five severity levels, mimics the wide range of scenarios that would cause distortions, texture shift, and/or degraded, lossy, and contaminated images [15, 29].

Since MDE models rely upon sufficient and clear visual cues to infer accurate depth, the aforementioned corruptions tend to cause difficulties in depth predictions. A pilot study shown in Fig. 1 reveals that MDE models with unique architectures exhibit diverse behaviors under different corruptions. It is thus important to understand the root causes of performance degradation for each model so that we can design and build a robust and reliable MDE model. To achieve this goal, we benchmark intensively prior MDE methods based on our proposed datasets and conduct a comprehensive study on their robustness against corruptions. Based on our benchmarking results, we draw several interesting observations: **1)** We find that existing MDE models, either from indoor scenes or outdoor scenes, are at the risk of being vulnerable to corruptions. **2)** Models with monocular inputs are more stable than those trained with stereo pairs. **3)** Transferring knowledge from other related tasks helps MDE models preserve robustness. **4)** Training with high-resolution images yields more robust models against noise perturbations. **5)** The supervised and self-supervised MDE models have different sensitivities to corruptions. We will revisit these findings more formally in the following sections.

To sum up, this work has these key contributions:

- We introduce RoboDepth, the first systematically-designed robustness evaluation suite for monocular depth estimation under corruptions.
- We benchmark 42 models from indoor and outdoor scenes, on their robustness against corruptions.
- Based on our observations, we draw in-depth discussions on the design considerations of building more robust MDE models for practical applications.

2. Related Work

Monocular Depth Estimation (MDE). Since the pioneering works [10, 11, 13, 52] first adopts deep neural networks to perform monocular depth estimation, significant progress has been made in many aspects, as shown in Tab. 1. Notable innovations include network architectures [22, 32, 47, 50], optimization functions [6, 14, 48], internal constraints [44,51], multi-task learning [18], geometry constraint [40], and various sources of supervisions [26,33,36]. Based on the learning paradigm, most MDE methods can be split into supervised or self-supervised models. The former mainly focuses on indoor scenes and uses ground truth from RGB-D cameras or LiDAR sensors to train a regression model [1, 25]; while the latter formulates MDE as a novel view synthesis task to minimize the photometric loss between stereo pairs or from monocular video frames [52]. Although promising results have been achieved, the robustness of MDE models under adverse scenarios is still unknown. Due to the lack of relevant datasets, existing models are at risk of being vulnerable to corruptions. In this work, we fill in this gap by establishing comprehensive evaluation benchmarks and testing 42 MDE models from both indoor and outdoor environments to analyze their robustness.

Robust MDE. To the best of our knowledge, only a few works targeted robust learning of MDE and they focused on different aspects. Ranftl *et al.* [33] proposed a unified objective for merging multiple datasets with different depth scales and ranges for training robust models. Similar works [5, 26, 38, 42, 46] resort to web stereo data or 3D movies to train MDE models and adapt them to unseen datasets. Kopf *et al.* [20] estimate stable camera trajectories for hand-held cellphone videos. SC-DepthV3 [36] generates pseudo-depth to refine depth details for scenes with dynamic objects. Li *et al.* [23] proposed an attention module to choose scene-specific features for MDE on both indoor and outdoor scenes. SeasonDepth [16] contributed a dataset with depth maps under sunny, cloudy, and foliage weather. Most recently, there are works [4, 7] design adversarial patches to attack MDE models. Conversely, we aim to test the robustness of MDEs to corruptions that occur in the real-world environment. We establish the first benchmark of this kind and incorporate an ample number of MDE models for analysis and comparison.

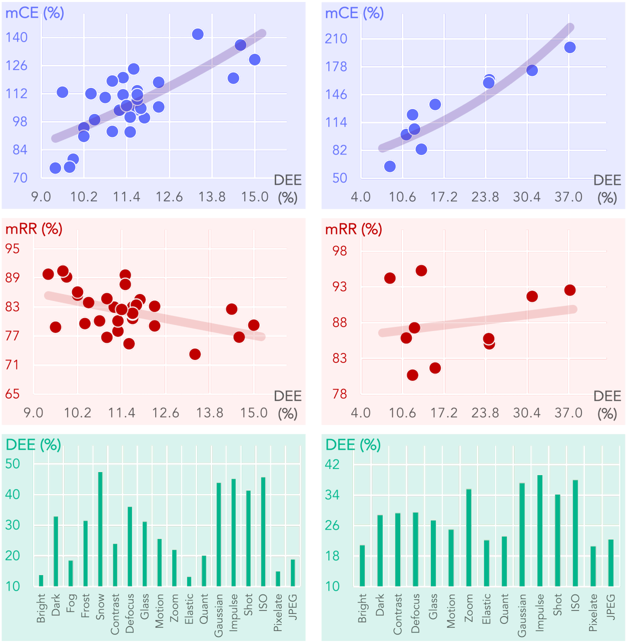
Corruption Robustness. ImageNet-C [15] is the pioneering work in this line of research which benchmarks classical image classification models to common corruptions and perturbations. Follow-up studies extend on the aspect to other visual perception tasks, *e.g.*, object detection [29], image segmentation [19], navigation [3], video classification [45], and pose estimation [39]. The essentiality of evaluating model robustness has been repeatedly validated. Since we are targeting a different task, *i.e.*, MDE, most of the well-studied corruption types become realistic or suitable for such a data format. This motivates us to explore new taxonomy for defining more proper corruption types for MDE.

3. RoboDepth Benchmark

In this section, we first introduce the taxonomy of corruptions included in our benchmarks (Sec. 3.1). We then elaborate on more details of the proposed datasets (Sec. 3.2) and corresponding robustness evaluation metrics (Sec. 3.3).

3.1. Corruption Type

Weather & Lighting Condition. The cameras on drones or vehicles operating under different weather and times of day capture distribution-shifted images which are rare or lacking in current MDE datasets. To probe the robustness of MDE models under these adverse weather and lighting conditions, we simulate six corruptions, *i.e.*, 'bright', 'dark', 'fog', 'frost', 'snow', and 'contrast', which commonly occur in the real-world environment. Compared to clean images, these corruptions tend to affect the intensity and color of the light source, leading to hazy, blurry, and noise-contaminated images, which increase the difficulties for MDE models to make accurate depth predictions.



(a) KITTI-C

(b) NYUDepth2-C

Figure 3. Benchmarking results of ***42*** monocular depth estimation models on *KITTI-C* and *NYUDepth2-C*. Figures from top to bottom: the depth estimation error (DEE) *vs.* **[first row]** mean corruption error (mCE), **[second row]** mean resilience rate (mRR), and **[third row]** sensitivity analysis among different corruption types.

Sensor Failure & Movement. An MDE system must behave robustly against motion perturbation and sensor failure to maintain safety requirements for practical applications. To achieve this goal, we mimic four motion-related corruptions, *i.e.*, 'defocus', 'glass', 'motion', and 'zoom' blurs; we also generate images under 'elastic transformation' and 'color quantization', which happen during sensor malfunction. These corruption types are often associated with issues including edge distortion, contrast loss, and pattern shift.

Data & Processing. Data collection and transmission are inevitably associated with various sources of noise and potential loss of information. We include four such random variations, *i.e.*, 'gaussian', 'impulse', 'shot', and 'ISO' noises. In addition, we investigate the degradation caused by 'pixelate' and 'JPEG compression' which are common corruptions in handling image data. Compared to clean images, the noise-contaminated data introduce errors in the intensity values of pixels, leading to a grainy or speckled appearance. The pixelation and lossy compression tend to lead to a loss of detail and clarity in the image and can result in visible artifacts, such as blockiness or blurring.

Table 1. Depth estimation model calibration from different aspects (modality, backbone, pertaining strategy, and loss function).

Model	Venue	Modality	Motivation	Backbone	Pretrain	Loss Function
MonoDepth2 [14]	ICCV'19	Mono & Stereo	Auto-masking & multi-scale cues	ResNet-18/50	ImageNet	photometric re-projection; per-pixel smoothness
DepthHints [41]	ICCV'19	Stereo	Complementary depth suggestions	ResNet-18	ImageNet	photometric re-projection; per-pixel smoothness
SC-Depth [2]	NeurIPS'19	Mono	Geometry consistency constraint	ResNet-50	CityScapes	consistency; photometric re-projection; smoothness
CADepth [44]	3DV'21	Mono & Stereo	Channel-wise structural attention	ResNet-50	ImageNet	photometric re-projection; per-pixel smoothness
HR-Depth [28]	AAAI'21	Mono	High-resolution features fusion	ResNet-18	Cityscapes	photometric re-projection; per-pixel smoothness
DIFFNet [51]	BMVC'21	Mono	Internel feature fusion mechanism	HRNet	ImageNet	photometric re-projection; per-pixel smoothness
ManyDepth [40]	CVPR'21	Multi-Mono	Sequential test-time information	ResNet-18	ImageNet	consistency; photometric re-projection; smoothness
FSRE-Depth [18]	ICCV'21	Mono	Semantics-guided triplet loss	ResNet-18	ImageNet	semantic triplet; photometric re-projection; smoothness
MonoViT [50]	3DV'21	Mono	Global reasoning via self-attention	MPViT	ImageNet	photometric re-projection; per-pixel smoothness
DynaDepth [48]	ECCV'22	Mono	Vision and IMU motion dynamics	ResNet-18/50	ImageNet	IMU photometric; cross-sensor photometric consistency
TriDepth [6]	WACV'23	Multi-Mono	Patch-based triplet optimizing strategy	ResNet-18	ImageNet	patch triplet; photometric re-projection; smoothness
Lite-Mono [47]	CVPR'23	Mono	Efficient mix of CNNs & attentions	ResNet-18	ImageNet	photometric re-projection; per-pixel smoothness

3.2. Robustness Benchmark

KITTI-C. Based on the KITTI Vision Suite [12], we establish a robustness benchmark for outdoor MDE. We simulate the defined 18 corruptions using data from the KITTI *val* set under Eigen's split. Similar to [15], we design five severity levels for each corruption to further consolidate the evaluation of robustness changes. As a result, this robustness probing dataset has a total number of $62,730$ RGB images with a size of 192×640. We also include the high-resolution version (320×1024) for evaluating the robustness of models with larger images as the input.

NYUDepth2-C. We construct a benchmark for robust indoor MDE based on NYU Depth V2 [35]. 15 of the defined corruptions are used, excluding 'fog', 'frost', and 'snow' which rarely occur in the indoor scenes. Since the indoor environments are less variant than outdoor ones, we only include four severity levels for each corruption. To sum up, this dataset contains $39,240$ images of size 480×640, which cover 23 different types of indoor scenes.

KITTI-S. To further investigate the root cause of robustness degradation, we form another robustness set based on KITTI [12], which is stylized via the style transfer model AdaIn [17]. This dataset has $8,364$ images from 12 styles, including 'cartoon', 'digital art', 'ink painting', 'kids' drawing', 'murals', 'oil painting', 'penciling', 'shadow play', 'sketch', 'stained glass', 'relief', and 'water color'.

3.3. Evaluation Metrics

Depth Estimation Error (DEE). We combine Abs Rel and δ_1, the two main measures defined in [10, 27], into a unified metric as DEE $=\frac{\text{Abs Rel}-\delta_1+1}{2}$, which is constantly used as the indicator of depth estimation error in our benchmark.

Corruption Error (CE). We follow [15] and use the mean CE (mCE) as the primary metric in comparing models' robustness. To normalize the severity effects, we choose MonoDepth2 [14] and AdaBins [1] as the baseline models for the *KITTI-C* and *NYUDepth2-C* benchmarks, respectively. The CE and mCE scores are calculated as follows:

$$\text{CE}_i=\frac{\sum_{l=1}^{5}(\text{DEE}_{i,l})}{\sum_{l=1}^{5}(\text{DEE}_{i,l}^{\text{baseline}})}~,\quad \text{mCE}=\frac{1}{N}\sum_{i=1}^{N}\text{CE}_i~. \quad (1)$$

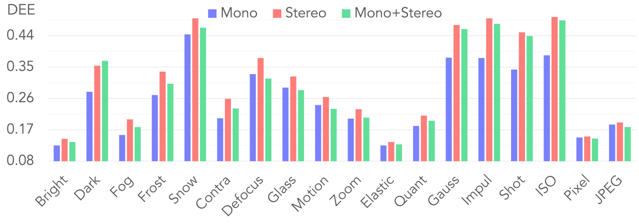


Figure 4. Depth estimation robustness comparisons among different modalities (Mono, Stereo, and Mono+Stereo).

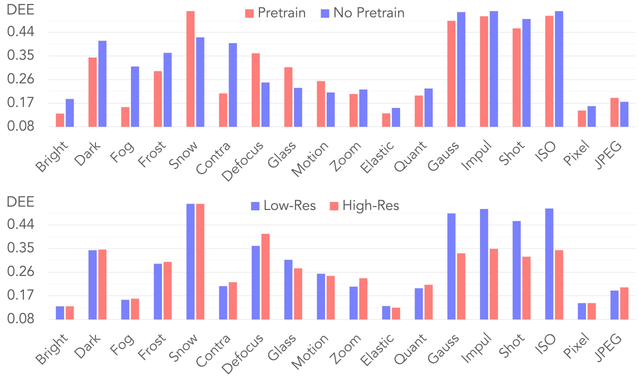


Figure 5. Depth estimation robustness of MonoDepth2 [14] under different training configurations. **[1st row]** Different pretrain techniques. **[2nd row]** Different input resolutions.

Resilience Rate (RR). We define mean RR (mRR) as the relative robustness indicator for measuring how much accuracy can a model retain when evaluated on the corruption sets. The RR and mRR scores are calculated as follows.

$$\text{RR}_i=\frac{\sum_{l=1}^{5}(1-\text{DEE}_{i,l})}{5\times(1-\text{DEE}_{\text{clean}})}~,\qquad \text{mRR}=\frac{1}{N}\sum_{i=1}^{N}\text{RR}_i~,\quad (2)$$

where $\text{DEE}_{\text{clean}}$ denotes the task-specific accuracy score on the "clean" evaluation set.

4. Experimental Analysis

4.1. Benchmark Configuration

Depth Estimation Model. We benchmark 42 depth estimation models and variants, which cover most of the open-

Table 2. The **Corruption Error (CE)** of 32 monocular depth estimation models on ***KITTI-C***. All scores are given in percentage (%). Blocks from top to bottom: **[1st]** the baseline MonoDepth2$_{\text{R18}}$ [14]; **[2nd]** methods *w/* monocular inputs; **[3rd]** methods *w/* stereo inputs; **[4th]** methods *w/* monocular + stereo inputs. **Bold**: Best in col. Underline: Second best in col. Blue : Best in row. Red : Worst in row.

Method	mCE	Bright	Dark	Fog	Frost	Snow	Contr	Defoc	Glass	Motio	Zoom	Elast	Quant	Gaus	Impul	Shot	ISO	Pixel	JPEG
MonoDepth2$_{\text{R18}}$ [14]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MonoDepth2$_{\text{nopt}}$ [14]	119.8	140.8	122.5	200.7	112.6	78.1	222.5	104.1	95.9	92.1	103.0	114.7	109.8	114.8	116.2	118.2	116.8	105.5	87.2
MonoDepth2$_{\text{HR}}$ [14]	106.1	99.2	134.3	100.0	97.8	113.9	114.4	161.1	106.2	106.5	115.4	95.4	111.4	84.9	90.5	93.2	88.7	95.2	101.0
MonoDepth2$_{\text{R50}}$ [14]	113.4	97.7	105.0	100.0	103.6	96.3	124.6	175.0	162.0	128.2	103.5	100.8	102.6	106.5	103.6	108.2	109.5	106.9	107.7
MaskOcc$_{\text{R18}}$ [34]	104.1	100.0	101.8	98.7	102.2	96.3	107.0	130.3	121.9	105.6	100.0	100.0	95.3	105.0	105.4	107.1	107.5	98.6	90.3
DNet$_{\text{R18}}$ [43]	104.7	98.5	94.3	100.7	114.4	98.6	111.8	142.6	132.2	112.0	107.0	101.6	97.9	94.3	94.1	95.9	92.0	100.0	96.9
CADepth [44]	110.1	93.1	107.1	91.6	117.0	103.5	103.2	145.9	143.4	131.9	103.5	93.8	99.5	110.2	111.3	112.7	115.5	99.3	99.5
HR-Depth [28]	103.7	93.1	103.2	97.4	100.7	94.1	113.9	145.9	124.0	121.8	111.4	96.1	96.9	94.5	95.9	98.8	96.4	**93.1**	89.8
DIFFNet [51]	95.0	85.4	**79.3**	84.5	**71.8**	68.9	86.1	210.3	136.4	129.6	98.0	**88.4**	85.5	76.0	68.4	75.0	69.6	**93.1**	103.1
ManyDepth$_{\text{R18}}$ [40]	105.4	103.9	97.9	109.0	104.0	93.7	121.4	104.1	115.3	97.7	96.5	103.9	97.9	112.0	115.7	113.8	116.5	101.4	92.9
FSRE-Depth [18]	99.1	98.5	93.2	89.7	85.6	76.9	90.9	119.3	112.8	99.1	92.0	92.3	92.8	104.2	106.4	108.8	104.9	101.4	114.3
MonoViT [50]	79.3	81.5	86.8	74.8	76.9	53.8	63.6	73.8	84.3	75.5	89.1	91.5	75.7	80.7	75.3	79.7	74.7	111.7	78.6
MonoViT$_{\text{HR}}$ [50]	**75.0**	78.5	85.0	**73.6**	81.2	52.6	**62.6**	**59.4**	**70.7**	**67.1**	91.5	83.7	**75.1**	78.7	71.2	76.2	73.5	**93.1**	75.5
DynaDepth$_{\text{R18}}$ [48]	110.4	98.5	103.2	100.7	104.3	99.6	111.2	205.3	143.4	141.2	103.0	98.5	96.4	98.7	97.4	98.8	97.7	97.2	91.8
DynaDepth$_{\text{R50}}$ [48]	120.0	98.5	106.4	98.1	117.0	107.4	107.5	218.0	187.6	147.2	108.5	96.9	102.1	108.9	112.3	112.4	115.5	105.5	110.2
RA-Depth [31]	112.7	86.9	112.1	81.9	86.3	80.8	88.2	204.5	152.1	175.0	106.5	94.6	92.2	110.2	103.6	118.2	117.3	120.7	98.0
TriDepth$_{\text{R18}}$ [6]	109.3	100.8	107.1	121.3	122.0	97.5	141.7	109.8	124.4	98.2	94.5	97.7	103.1	108.9	112.6	111.8	112.9	97.9	104.6
Lite-Mono$_{\text{Tiny}}$ [47]	92.9	97.7	91.8	101.3	81.2	69.3	102.1	105.3	102.5	91.7	92.5	98.5	82.4	93.2	87.9	98.8	92.8	101.4	82.1
Lite-Mono$_{\text{Small}}$ [47]	100.3	97.7	89.6	104.5	90.6	84.2	127.3	144.7	116.5	113.9	101.5	99.2	83.4	91.2	86.4	93.8	91.8	106.2	83.7
Lite-Mono$_{\text{Base}}$ [47]	93.2	91.5	92.5	92.9	88.5	75.2	94.7	91.8	97.9	102.3	97.5	100.0	90.7	94.0	87.4	98.2	93.6	104.1	84.2
Lite-Mono$_{\text{Large}}$ [47]	90.8	84.6	81.1	81.3	92.1	84.7	79.7	91.0	93.0	101.9	95.5	93.8	76.7	94.5	89.5	96.8	93.3	110.3	93.9
MonoDepth2$_{\text{R18}}$ [14]	117.7	102.3	124.3	103.9	110.1	100.8	125.1	159.8	137.2	122.2	104.0	104.7	103.6	128.1	130.9	136.2	127.1	99.3	99.0
MonoDepth2$_{\text{nopt}}$ [14]	129.0	139.2	150.7	188.4	127.1	85.1	182.9	109.0	95.9	100.5	113.9	120.9	122.3	140.4	145.0	153.2	143.3	113.1	90.8
MonoDepth2$_{\text{HR}}$ [14]	111.5	101.5	101.8	107.7	128.5	103.5	127.3	177.1	128.9	129.2	122.4	100.8	106.7	89.3	88.2	94.7	88.7	103.5	106.6
DepthHints [41]	111.4	95.4	110.7	88.4	115.9	100.8	87.7	143.4	169.4	121.8	97.5	100.8	99.5	114.6	114.9	121.2	117.3	108.3	98.0
DepthHints$_{\text{nopt}}$ [41]	141.6	133.1	170.0	194.2	135.0	90.6	210.2	146.3	119.4	111.6	114.9	110.1	128.0	159.6	169.2	176.2	178.4	104.8	97.5
DepthHints$_{\text{HR}}$ [41]	112.0	93.9	100.7	91.0	114.4	93.9	96.3	188.1	150.0	148.2	130.4	91.5	94.8	103.4	108.2	111.8	109.3	97.2	93.4
MonoDepth2$_{\text{R18}}$ [14]	124.3	97.7	144.3	96.8	106.5	104.9	106.4	183.2	143.0	131.0	101.5	99.2	105.2	150.3	155.5	165.0	162.1	93.8	91.3
MonoDepth2$_{\text{nopt}}$ [14]	136.3	148.5	164.3	211.6	152.0	83.8	235.3	93.4	91.3	100.0	114.4	118.6	118.7	148.4	153.2	161.5	156.2	111.0	90.3
MonoDepth2$_{\text{HR}}$ [14]	106.1	99.2	134.3	100.0	97.8	113.9	114.4	161.1	106.2	106.5	115.4	95.4	111.4	84.9	90.5	93.2	88.7	95.2	101.0
CADepth [44]	118.3	94.6	127.5	88.4	112.3	108.8	90.4	138.5	170.3	120.4	96.0	97.7	96.4	142.2	143.7	154.1	150.0	100.0	98.0
MonoViT [50]	75.4	**80.0**	87.5	78.7	76.9	**42.1**	70.1	73.4	76.0	74.5	**83.6**	86.8	76.2	**72.1**	**66.1**	**71.2**	**67.0**	101.4	**73.5**

Table 3. The **Corruption Error (CE)** of 10 monocular depth estimation models on the ***NYUDepth2-C*** dataset. All scores are given in percentage (%). **Bold**: Best in col. Underline: Second best in col. Blue : Best in row. Red : Worst in row.

Method	mCE	Bright	Dark	Contr	Defoc	Glass	Motio	Zoom	Elast	Quant	Gaus	Impul	Shot	ISO	Pixel	JPEG
AdaBins$_{\text{EB5}}$ [1]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
BTS$_{\text{R50}}$ [22]	122.8	112.9	138.7	125.0	143.4	127.2	125.5	96.9	119.0	119.2	113.3	136.9	133.3	124.7	112.1	113.6
AdaBins$_{\text{R50}}$ [1]	134.7	135.6	151.0	136.3	144.3	135.9	133.2	101.6	133.3	143.1	117.4	138.8	136.4	126.6	150.0	137.0
DPT$_{\text{ViT-B}}$ [32]	83.2	102.3	93.8	84.9	65.5	**80.6**	84.2	**60.4**	90.9	109.3	51.3	57.0	65.0	52.4	137.9	113.0
SimIPU$_{\text{nopt}}$ [24]	200.2	293.9	220.1	211.3	177.0	194.7	217.4	112.8	249.0	258.9	119.2	125.9	153.1	121.3	302.4	245.5
SimIPU$_{\text{ImageNet}}$ [24]	163.1	203.8	190.7	177.4	160.4	163.6	176.1	109.9	200.0	191.4	114.1	123.8	140.8	118.2	199.2	176.6
SimIPU$_{\text{KITTI}}$ [24]	173.8	247.0	192.3	191.5	153.2	161.7	182.1	100.5	206.5	220.5	110.8	118.2	143.5	116.6	253.2	209.1
SimIPU$_{\text{WaymoOpen}}$ [24]	159.5	203.8	179.4	187.7	161.7	158.7	170.1	105.5	167.3	190.1	112.6	123.3	141.5	119.7	198.4	172.1
DepthFormer$_{\text{SwinT-1k}}$ [25]	106.3	111.4	143.8	110.9	93.6	126.2	103.8	78.1	114.4	127.2	75.4	85.8	98.3	80.3	129.8	116.2
DepthFormer$_{\text{SwinT-22k}}$ [25]	**63.5**	**75.0**	**77.3**	**58.0**	**54.0**	83.5	**64.7**	61.7	**73.2**	**78.8**	**40.8**	**41.7**	**50.3**	**41.3**	**81.5**	**70.1**

source models so far. 32 of them are for outdoor MDE and the remaining 10 are for indoor MDE.

Evaluation Protocol. To avoid any unfairness in robustness comparisons, we unify the common configurations among candidate models, such as backbones and data augmentations. We use public checkpoints whenever possible and reproduce the reported results based on official settings. More details on this aspect are included in the Appendix.

4.2. Benchmark Analysis

Observation 1: MDE Robustness - *existing outdoor and indoor MDE models are at the risk of being vulnerable to real-world corruptions*. We show the robustness of different models in terms of mCE, mRR, and corruption sensitivity in Fig. 3. As can be seen from the first two rows, existing MDE models, either from indoor or outdoor scenes, are showing clear relationships between the DEE scores and robustness metrics. Specifically, higher DEE scores correlate to higher mCE scores and are less robust compared to the baseline models. From Tab. 2 and Tab. 3, we observe that the Transformers-based MDE models are significantly more robust compared to conventional CNNs models. This is also observed for the mRR metrics, which are shown in Tab. 4 and Tab. 5. The qualitative results shown in Fig. 6 and Fig. 7 further validate that models with long-range receptive fields, such as MonoViT [50] and Lite-Mono [47], can better maintain accurate depth predictions under edge distortion, texture shift, and noise contamination.

Observation 2: Modality - *the learning paradigm plays a vital role in depth estimation robustness*. We analyze the ro-

Table 4. The **Resilience Rate (RR)** of 32 monocular depth estimation models on ***KITTI-C***. All scores are given in percentage (%). Blocks from top to bottom: **[1st]** the baseline MonoDepth2$_{\text{R18}}$ [14]; **[2nd]** methods *w/* monocular inputs; **[3rd]** methods *w/* stereo inputs; **[4th]** methods *w/* monocular + stereo inputs. **Bold**: Best in col. Underline: Second best in col. Blue : Best in row. Red : Worst in row.

Method	mRR	Bright	Dark	Fog	Frost	Snow	Contr	Defoc	Glass	Motio	Zoom	Elast	Quant	Gaus	Impul	Shot	ISO	Pixel	JPEG
MonoDepth$_2$ $_{\text{R18}}$ [14]	84.5	98.8	81.7	95.9	82.1	55.5	92.3	85.8	86.0	89.0	90.7	98.9	91.6	69.9	69.4	74.9	69.5	97.1	91.3
MonoDepth$_2$ $_{\text{nopt}}$ [14]	82.5	95.4	76.8	80.5	80.4	70.2	68.2	87.2	89.7	93.6	**92.6**	**99.5**	92.1	65.3	64.0	69.9	63.9	**99.0**	**96.9**
MonoDepth$_2$ $_{\text{HR}}$ [14]	82.4	98.3	70.4	95.4	82.3	47.2	88.7	68.5	83.9	86.9	86.7	99.0	88.6	76.1	73.1	77.1	74.0	97.3	90.5
MonoDepth$_2$ $_{\text{R50}}$ [14]	80.6	98.9	80.0	95.7	80.8	57.5	86.9	64.9	68.9	81.9	89.7	98.5	90.8	66.9	67.6	71.6	65.1	95.7	89.4
MaskOcc $_{\text{R18}}$ [34]	83.0	98.5	81.0	95.9	81.2	57.5	90.6	77.2	79.8	87.4	90.5	98.6	92.4	67.6	66.8	72.0	66.0	97.1	93.2
DNet $_{\text{R18}}$ [43]	83.3	98.9	83.5	95.7	77.4	56.2	89.7	73.9	77.1	85.9	89.0	98.5	92.0	72.3	71.9	76.4	72.9	96.9	91.8
CADepth [44]	80.1	98.5	78.5	96.2	75.8	52.8	90.5	72.2	73.2	80.2	88.8	98.5	90.6	64.7	63.6	69.2	61.9	96.0	90.3
HR-Depth [28]	82.9	99.0	80.1	95.6	81.2	58.5	88.6	72.5	78.8	83.0	87.4	98.7	91.6	71.7	70.6	74.8	70.5	97.4	92.8
DIFFNet [51]	85.4	99.0	**86.6**	96.8	**89.2**	72.2	93.4	54.2	74.6	80.2	89.4	98.7	93.0	78.8	81.7	83.0	81.3	96.3	88.9
ManyDepth [40]	83.1	98.6	82.8	94.8	81.2	59.4	88.1	85.1	82.2	90.0	91.9	98.8	92.5	65.0	62.7	69.9	62.5	97.3	93.3
FSREDepth [18]	83.9	97.9	82.9	96.6	85.6	68.1	93.2	79.6	81.6	88.2	91.5	98.9	92.1	67.3	65.8	70.7	66.6	95.7	87.1
MonoViT [50]	89.2	**99.2**	84.0	**98.1**	87.4	80.5	**97.8**	91.0	88.4	92.9	91.1	97.9	94.8	76.6	78.5	80.9	78.8	93.0	93.9
MonoViT $_{\text{HR}}$ [50]	89.7	99.1	84.1	97.8	85.5	80.7	97.5	**94.4**	**91.5**	**94.4**	90.1	98.5	94.4	77.0	79.8	81.8	78.9	95.5	94.0
DynaDepth $_{\text{R18}}$ [48]	81.5	98.8	80.5	95.6	80.5	55.6	89.7	56.5	74.0	78.7	89.8	98.9	92.2	70.3	70.3	75.2	70.3	97.3	92.9
DynaDepth $_{\text{R50}}$ [48]	78.0	98.3	79.1	95.6	76.2	50.9	90.1	52.8	61.6	76.9	88.2	98.7	90.5	65.6	63.5	69.7	62.2	95.5	88.4
RA-Depth [31]	78.8	98.1	75.9	96.6	84.2	64.9	92.4	55.4	69.9	68.8	87.0	97.1	90.9	63.8	66.0	66.2	60.3	91.3	89.4
TriDepth $_{\text{R18}}$ [6]	81.6	98.4	79.3	92.0	75.0	56.9	83.2	82.9	79.2	89.2	91.7	99.0	90.7	65.9	63.7	70.2	63.7	97.2	90.0
Lite-Mono $_{\text{Tiny}}$ [47]	86.7	98.6	84.0	95.3	87.6	73.0	91.4	84.0	85.0	90.6	92.0	98.6	95.0	72.5	74.4	75.0	72.3	96.4	94.8
Lite-Mono $_{\text{Small}}$ [47]	84.7	98.6	84.6	94.7	84.6	64.4	86.1	73.1	81.1	85.2	89.9	98.5	94.8	73.5	75.0	77.0	72.8	95.6	94.5
Lite-Mono $_{\text{Base}}$ [47]	86.0	99.0	83.3	96.2	84.8	69.2	92.5	87.2	85.7	87.5	90.3	97.9	92.7	71.8	74.2	74.8	71.6	95.4	93.8
Lite-Mono $_{\text{Large}}$ [47]	85.5	99.1	86.1	97.3	83.0	63.1	94.8	86.6	86.3	86.9	90.0	97.9	**94.9**	70.9	72.6	74.7	71.1	93.5	90.9
MonoDepth$_2$ $_{\text{R18}}$ [14]	79.1	98.9	74.3	95.7	79.3	55.3	87.3	69.6	76.2	83.9	90.2	98.6	91.2	57.9	56.0	61.2	57.8	97.6	91.9
MonoDepth$_2$ $_{\text{nopt}}$ [14]	79.2	96.4	68.0	83.3	76.2	66.5	77.4	86.4	90.4	92.1	90.7	99.3	89.9	54.2	51.3	56.4	52.2	98.4	96.7
MonoDepth$_2$ $_{\text{HR}}$ [14]	81.6	98.3	81.0	94.3	72.9	53.3	86.3	64.3	77.9	81.7	85.4	98.5	89.9	74.4	74.4	76.8	74.3	96.3	89.6
DepthHints [41]	80.1	98.8	77.8	97.3	76.6	54.7	94.3	73.3	66.5	83.1	90.6	98.1	91.1	63.1	62.3	66.3	61.4	95.0	91.1
DepthHints $_{\text{nopt}}$ [41]	79.5	98.0	80.1	95.9	76.2	58.0	91.5	60.4	71.1	75.9	82.4	98.4	91.2	67.3	64.6	69.2	64.3	95.9	91.2
DepthHints $_{\text{HR}}$ [41]	73.2	95.5	60.5	80.7	72.3	62.0	70.1	74.3	82.1	87.6	88.8	99.1	87.0	44.7	39.5	46.3	35.6	97.9	93.4
MonoDepth$_2$ $_{\text{R18}}$ [14]	75.4	98.8	67.4	96.2	79.8	52.5	90.6	62.6	74.0	81.1	90.1	98.6	90.2	47.9	44.7	49.7	42.0	97.7	92.9
MonoDepth$_2$ $_{\text{nopt}}$ [14]	76.7	94.5	63.2	78.7	67.8	67.0	65.6	90.4	91.2	91.8	90.2	99.2	90.3	50.4	47.3	52.8	46.1	98.2	96.4
MonoDepth$_2$ $_{\text{HR}}$ [14]	82.4	98.3	70.4	95.4	82.3	47.2	88.7	68.5	83.9	86.9	86.7	99.0	88.6	76.1	73.1	77.1	74.0	97.3	90.5
CADepth [44]	76.7	98.5	72.3	97.0	77.4	49.9	93.4	74.4	66.1	83.2	90.7	98.2	91.5	51.0	49.6	53.5	47.0	96.1	90.8
MonoViT [50]	**90.4**	**99.2**	83.6	97.2	87.2	**86.9**	96.2	90.9	90.4	92.9	92.1	98.3	94.5	**80.1**	**82.3**	**83.9**	**82.0**	94.5	94.8

Table 5. The **Resilience Rate (RR)** of 10 monocular depth estimation models on the ***NYUDepth2-C*** dataset. All scores are given in percentage (%). **Bold**: Best in col. Underline: Second best in col. Blue : Best in row. Red : Worst in row.

Method	mRR	Bright	Dark	Contr	Defoc	Glass	Motio	Zoom	Elast	Quant	Gaus	Impul	Shot	ISO	Pixel	JPEG
AdaBins $_{\text{EB5}}$ [1]	85.8	97.8	90.8	88.7	86.2	89.4	91.9	69.4	95.4	95.6	68.7	70.5	79.5	69.8	98.7	95.3
BTS $_{\text{R50}}$ [22]	80.6	96.9	83.3	83.7	75.5	84.1	87.6	71.5	93.2	93.4	63.6	55.6	69.3	59.9	98.1	94.0
AdaBins $_{\text{R50}}$ [1]	81.6	97.5	84.0	84.4	78.5	85.5	89.7	72.5	94.5	93.1	64.4	57.1	71.1	61.6	96.7	93.7
DPT $_{\text{ViT-B}}$ [32]	**95.3**	**100.0**	**94.7**	94.9	**97.9**	96.5	**97.8**	88.9	**99.7**	96.6	**92.6**	91.1	**93.6**	**92.7**	96.0	95.6
SimIPU $_{\text{nopt}}$ [24]	92.5	97.5	91.2	87.9	93.0	95.4	95.5	**90.3**	98.6	**97.0**	85.2	84.2	87.6	85.8	99.5	**99.0**
SimIPU $_{\text{ImageNet}}$ [24]	85.0	96.7	83.3	82.5	82.4	87.7	89.4	76.5	91.8	94.1	73.4	71.0	77.5	72.9	99.6	96.3
SimIPU $_{\text{KITTI}}$ [24]	91.6	98.0	91.1	86.3	93.0	**97.0**	96.7	89.2	99.4	**97.0**	82.6	81.1	84.0	81.0	**99.7**	98.6
SimIPU $_{\text{WaymoOpen}}$ [24]	85.7	96.6	86.1	79.5	81.9	88.9	90.8	78.6	98.3	94.2	74.1	71.2	77.2	72.0	99.6	97.1
DepthFormer $_{\text{SwinT-1k}}$ [25]	87.3	97.5	82.4	87.4	89.1	84.6	92.5	80.0	94.3	92.3	80.7	77.6	81.3	79.4	95.9	93.8
DepthFormer $_{\text{SwinT-22k}}$ [25]	94.2	98.6	93.0	**96.0**	95.5	90.6	96.4	83.5	97.2	96.4	92.0	**92.3**	93.2	92.2	98.4	97.6

bustness of different variants of MonoDepth2 [14] and show their DEE scores in Fig. 4. We observe that the pure monocular input is helping the model retain resilience again most of the corruptions. Using stereo inputs, however, lead to robustness degradation compared to the monocular and mix versions. We conjecture that this is mainly because MDE models trained with stereo pairs rely more on the scene structural consistency between left and right images, where such requirements could be destroyed by feeding in corrupted data. Such constraints could be relaxed if the model is trained on monocular sequences.

Observation 3: Pretraining Strategy - *transferring knowledge from other tasks, such as ImageNet classification, brings strengths and weaknesses in the model's robustness*. The first row in Fig. 5 highlights that MDE models pretrained on object-centric datasets, *e.g.*, ImageNet, are more robust against corruptions by weather/lighting conditions (except for 'snow') and data processing noises, which are mostly texture-shifted corruptions. Motion and sensor corruptions, however, contain more edge and object distortions and could be eased by models without ImageNet pretraining. This implies that the CNN-based MDE models like MonoDepth2 [14] could become more shape-biased when pretrained on object-centric datasets. More evidence from the results on the *KITTI-S* dataset in the Appendix further verifies this finding, where the stylized data are causing less degradation for these MDE models.

Observation 4: Input Resolution - *training and testing*

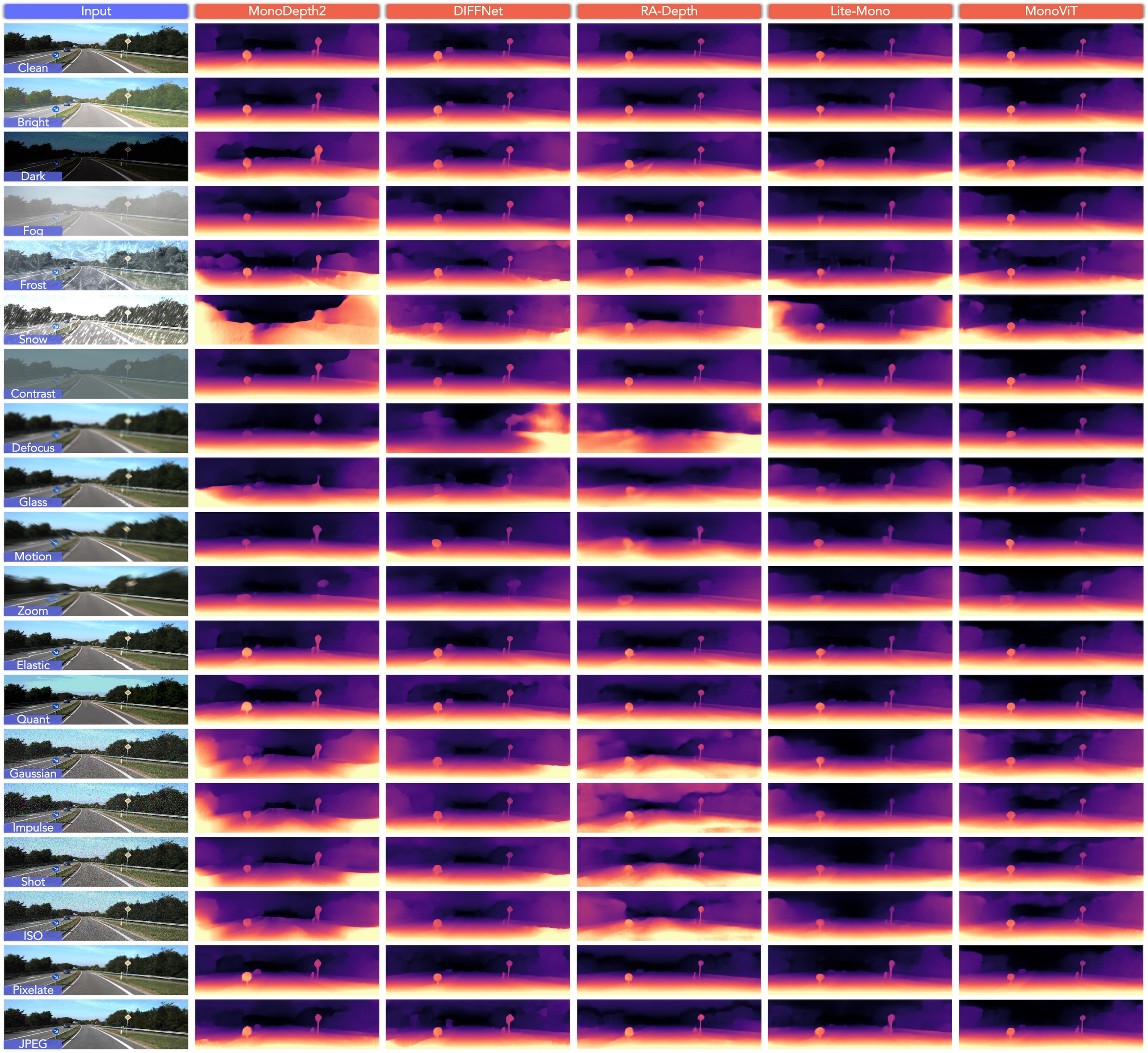


Figure 6. Qualitative results of different MDE models under defined corruption types (from severity level 3) in the *KITTI-C* dataset.

on images with higher resolutions tend to yield more robust MDE models. From the results shown in the second row of Fig. 5, we can observe that MDE models trained with high resolutions will likely yield more robust feature learning on noise-contaminated corruptions, including Gaussian, impulse, shot, and ISO noises. Since these noises affect the global pixel distribution instead of the local one, the CNN-based MDE models trained with high-resolution images will be able to capture more fine-grained information to suppress the degradation caused by noises.

Observation 5: Learning Paradigm - *the supervised and self-supervised learning result in different sensitivities in model's robustness*. From the third row of Fig. 3 we can observe that the self-supervised MDE models are less sensitive to lighting changes and motion blurs, compared to the supervised models. Both models suffer from the noises and behave robustly against lossy image compressions.

5. Conclusion

In this work, we establish the RoboDepth benchmark for probing the out-of-distribution robustness of monocular depth estimation models under corruptions. We introduce three new datasets and two metrics for evaluating the robustness of both indoor and outdoor MDE models. Our

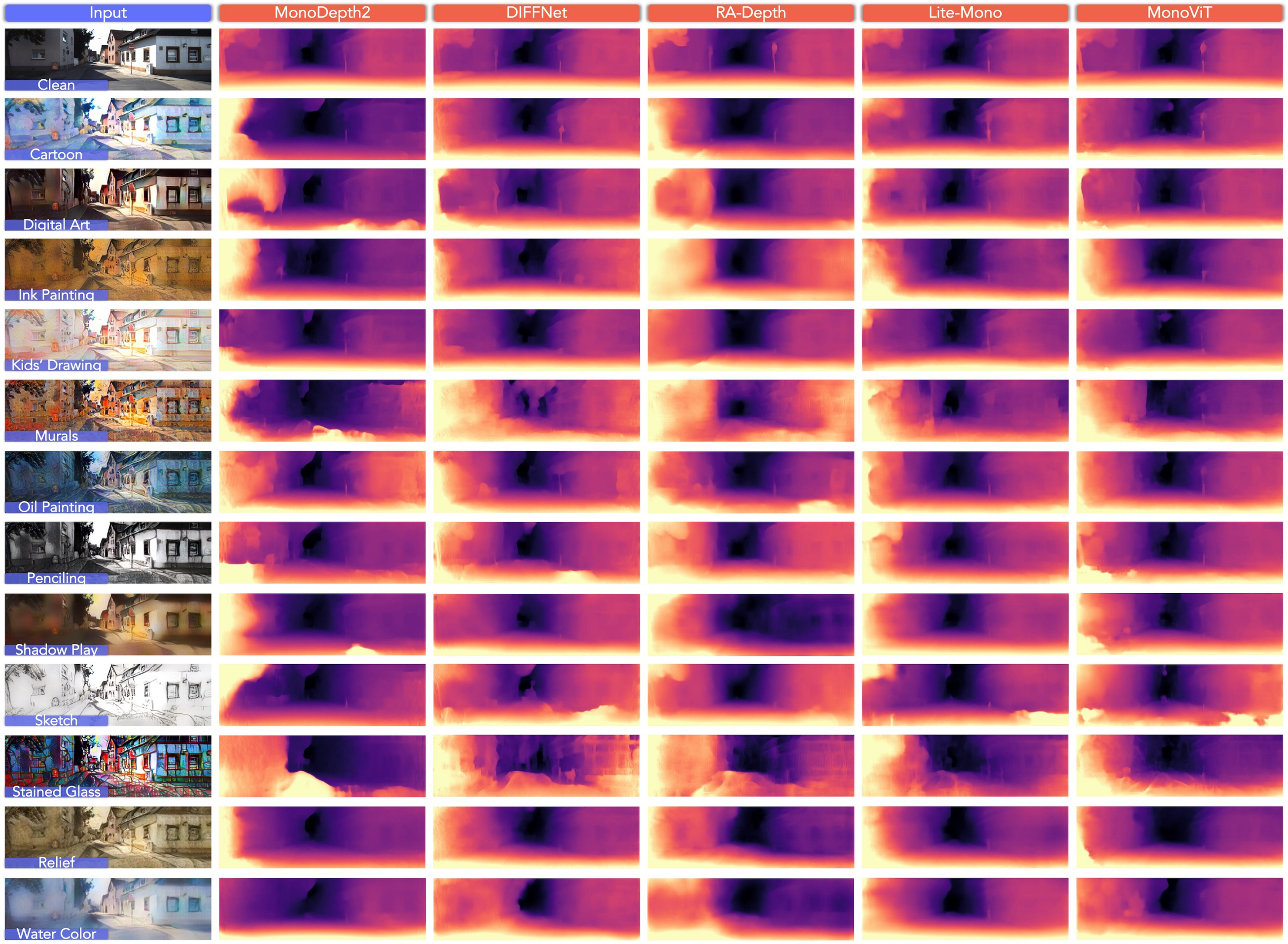


Figure 7. Qualitative results of five monocular depth estimation models on the ***KITTI-S*** dataset, including MonoDepth2 [14], DIFFNet [51], RA-Depth [31], Lite-Mono [47], and MonoViT [50]. The lighter regions correspond to near distances and vice versa. Best viewed in color. Zoomed-in for more details.

results reveal the importance of robustness probing among modern MDE algorithms and summarize the design considerations in terms of architecture, modality, pertaining, resolution, *etc*. We hope this work could lay a solid foundation for the development of robust MDE techniques.

Acknowledgements

This research is part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. This work is affiliated with the WP4 of the DesCartes programme, with an identity number: A-8000237-00-00.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4009–4018, 2021. 2, 4, 5, 6

[2] 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 *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2019. 4

[3] Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, and Aniruddha Kembhavi. Robustnav: Towards benchmarking robustness in embodied navigation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 15691–15700, 2021. 3

[4] Hemang Chawla, Kishaan Jeeveswaran, Elahe Arani, and Bahram Zonooz. Image masking for robust self-supervised monocular depth estimation. *arXiv preprint arXiv:2210.02357*, 2022. 2, 3

[5] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5604–5613, 2019. 3

[6] Xingyu Chen, Ruonan Zhang, Ji Jiang, Yan Wang, Ge Li, and Thomas H Li. Self-supervised monocular depth estimation: Solving the edge-fattening problem. In *IEEE/CVF Eur. Conf. Comput. Vis. (ECCV) Conf. Appli. Comput. Vis. (WACV)*, pages 5776–5786, 2023. 2, 4, 5, 6

[7] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 514–532, 2022. 2, 3

[8] 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3213–3223, 2016. 1, 2

[9] Xingshuai Dong, Matthew A. Garratt, Sreenatha G. Anavatti, and Hussein A. Abbass. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Trans. Intell. Transport. Syst. (TITS)*, 23(10):16940–16961, 2022. 1

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2014. 2, 4

[11] Ravi Garg, BG Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 740–756, 2016. 2

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3354–3361, 2012. 1, 2, 4

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 270–279, 2017. 2

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 3828–3838, 2019. 2, 4, 5, 6, 8

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Int. Conf. Learn. Represent. (ICLR)*, 2019. 1, 2, 3, 4

[16] Hanjiang Hu, Baoquan Yang, Zhijian Qiao, Shiqi Liu, Ding Zhao, and Hesheng Wang. Seasondepth: Cross-season monocular depth prediction dataset and benchmark under multiple environments. In *Int. Conf. Mach. Learn. Worksh. (ICMLW)*, 2022. 1, 3

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 1501–1510, 2017. 4

[18] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 12642–12652, 2021. 2, 4, 5, 6

[19] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8828–8838, 2020. 1, 3

[20] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1611–1621, 2021. 1, 3

[21] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 44(4):1738–1764, 2020. 1

[22] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2, 5, 6

[23] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In *Asian Conf. Comput. Vis. (ACCV)*, pages 663–678, 2019. 3

[24] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *AAAI Conf. Artifi. Intell. (AAAI)*, 2022. 2, 5, 6

[25] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2, 5, 6

[26] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2041–2050, 2018. 2, 3

[27] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 38(10):2024–2039, 2015. 4

[28] 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. In *AAAI Conf. Artifi. Intell. (AAAI)*, pages 2294–2301, 2021. 4, 5, 6

[29] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 1, 2, 3

[30] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A reviews. *Neurocomputing*, 438:14–33, 2021. 1

[31] He Mu, Hui Le, Bian Yikai, Ren Jian, Xie Jin, and Yang Jian. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 565–581, 2022. 5, 6, 8

[32] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 12179–12188, 2021. 2, 5, 6

[33] 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 Trans. Pattern Anal. Mach. Intell. (PAMI)*, 44(3):1623–1637, 2022. 2, 3

[34] Maarten Schellevis. Improving self-supervised single view depth estimation by masking occlusion. *arXiv preprint arXiv:1908.11112*, 2019. 5, 6

[35] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 746–760, 2012. 1, 2, 4

[36] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *arXiv preprint arXiv:2211.03660*, 2022. 2, 3

[37] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *Int. Conf. 3D Vis. (3DV)*, pages 11–20, 2017. 1

[38] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *Int. Conf. 3D Vis. (3DV)*, pages 348–357, 2019. 3

[39] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 11855–11864, 2021. 3

[40] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1164–1174, 2021. 2, 4, 5, 6

[41] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 2162–2171, 2019. 4, 5, 6

[42] 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 311–320, 2018. 3

[43] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H. Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 2330–2337, 2020. 5, 6

[44] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *Int. Conf. 3D Vis. (3DV)*, pages 464–473, 2021. 2, 4, 5, 6

[45] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2021. 3

[46] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 14128–14139, 2022. 2, 3

[47] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 2, 4, 5, 6, 8

[48] Sen Zhang, Jing Zhang, and Dacheng Tao. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 143–160, 2022. 2, 4, 5, 6

[49] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020. 1

[50] 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. In *Int. Conf. 3D Vis. (3DV)*, 2022. 2, 4, 5, 6, 8

[51] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *Brit. Mach. Vis. Conf. (BMVC)*, 2021. 2, 4, 5, 6, 8

[52] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1851–1858, 2017. 2