International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01979-4

Towards Robust Monocular Depth Estimation: A New Baseline and Benchmark

Ke Xian[1,2] · **Zhiguo Cao**[3] · **Chunhua Shen**[4] · **Guosheng Lin**[2]

Received: 28 March 2023 / Accepted: 16 December 2023

Abstract
Before deploying a monocular depth estimation (MDE) model in real-world applications such as autonomous driving, it is critical to understand its generalization and robustness. Although the generalization of MDE models has been thoroughly studied, the robustness of the models has been overlooked in previous research. Existing state-of-the-art methods exhibit strong generalization to clean, unseen scenes. Such methods, however, appear to degrade when the test image is perturbed. This is likely because the prior arts typically use the primary 2D data augmentations (*e.g.*, random horizontal flipping, random cropping, and color jittering), ignoring other common image degradation or corruptions. To mitigate this issue, we delve deeper into data augmentation and propose utilizing strong data augmentation techniques for robust depth estimation. In particular, we introduce 3D-aware defocus blur in addition to seven 2D data augmentations. We evaluate the generalization of our model on six clean RGB-D datasets that were not seen during training. To evaluate the robustness of MDE models, we create a benchmark by applying 15 common corruptions to the clean images from IBIMS, NYUDv2, KITTI, ETH3D, DIODE, and TUM. On this benchmark, we systematically study the robustness of our method and 9 representative MDE models. The experimental results demonstrate that our model exhibits better generalization and robustness than the previous methods. Specifically, we provide valuable insights about the choices of data augmentation strategies and network architectures, which would be useful for future research in robust monocular depth estimation. Our code, model, and benchmark can be available at https://github.com/KexianHust/Robust-MonoDepth.

Keywords Monocular depth prediction · Generalization · Robustness · Strong data augmentation

1 Introduction

Monocular depth estimation is a fundamental yet challenging task in computer vision, which aims at inferring the depth value for each pixel in an image. It can be used in a broad range of applications, including 3D photo generation (Niklaus et al., 2019; Wang et al., 2022), shallow depth-of-field rendering (Wadhwa et al., 2018; Wang et al., 2018; Peng et al., 2022), and space-time view synthesis (Yoon et al., 2020; Li et al., 2021). Such applications typically deal with in-the-wild images and require depth estimation models with good generalization. To this end, previous methods collect large-scale in-the-wild data to train their models. For instance, (Li & Snavely, 2018) and (Xian et al., 2018) derive ground-truth depth maps from multi-view web photo collections and stereo web image pairs, respectively. Although these models can generalize to unseen scenes, they cannot generate accurate predictions due to a large amount of noise in the training labels and the lack of diversity of data modalities. To mitigate these issues, recent works improve the quality of data generation with stronger tools (Teed & Deng, 2020; Xian et al., 2020) and resort to mix-data training (Lasinger et al., 2020; Yin et al., 2021, 2022; Ranftl et al., 2021). By doing so, such methods exhibit strong generalization across various scenes. The success of these methods can be attributed to the training on large, heterogeneous datasets

Communicated by D. Scharstein.

✉ Guosheng Lin
gslin@ntu.edu.sg

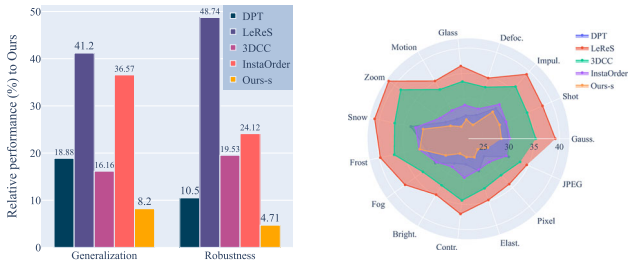
[1] EIC, Huazhong University of Science and Technology, Wuhan, China

[2] S-Lab, Nanyang Technological University, Nanyang, Singapore

[3] AIA, Huazhong University of Science and Technology, Wuhan, China

[4] Zhejiang University, Hangzhou, China

Published online: 20 January 2024



(a) Generalization & Robustness (b) Performance on DIODE-C

Fig. 1 **a** Illustrates the *generalization* and *robustness* of depth estimation models on clean and corrupted data. We report the average relative performance of the SOTA methods compared to Ours-ViT-hybrid across five RGB-D datasets (*i.e.*, NYUDv2, KITTI, ETH3D, DIODE, and TUM). **b** depicts the absrel performance of depth estimation models under 15 corruption types on DIODE-C. Our model (*i.e.*, Ours-s) outperforms models that use millions of data in terms of model generalization and robustness, with only 72K training images

for improving the generalization of the trained models. Thus, we also follow this promising direction to allow our model to generalize to in-the-wild images.

However, generalization is *not* equal to robustness. Even though the state-of-the-art (SOTA) depth estimator (Ranftl et al., 2021), which demonstrates strong generalization across multiple RGB-D benchmarks, would degrade significantly when the test image is corrupted by noise or blur. As a result, a true sense of robust depth model should take both generalization and robustness into account. Good generalization implies that the model can generalize to previously unseen scenes, whereas good robustness implies that the model can still work under various corruptions. While much progress has been made in investigating the robustness of models in image classification (Hendrycks & Dietterich, 2019) and semantic segmentation (Kamann & Rother, 2021), no comprehensive study of the robustness of depth estimation has been conducted.

In this work, we bridge this knowledge gap by (i) investigating data augmentation, which was overlooked in robust depth estimation, to improve the robustness of the trained model; (ii) creating a robustness benchmark to systematically study the robustness of depth estimation models. Data augmentation is a common technique to improve model robustness. Previous robust MDE methods (Xian et al., 2020; Lasinger et al., 2020; Ranftl et al., 2021; Yin et al., 2021), however, only use the primary data augmentation schemes such as random horizontal flipping, random cropping, and color jittering, ignoring more advanced data augmentation techniques. This limits their robustness and motivates us to investigate stronger data augmentation schemes. We consider 8 common image degradation in our daily photos, including scaling and cropping, horizontal flipping, rotation, color jittering, sharpness, gaussian blur, motion blur, and 3D-aware defocus blur. Unlike the layer-based defocus blur method (Kar et al., 2022), we implement the 3D-aware defocus blur in a continuous, scattering way. This helps us resolve artifacts at depth discontinuities caused by the layered method.

To validate the generalization of depth estimation models, we conduct zero-shot cross-dataset evaluation on six RGB-D datasets that were not seen during training, including NYUDv2 (Silberman et al., 2012), KITTI (Uhrig et al., 2017), ETH3D (Schöps et al., 2017), DIODE (Vasiljevic et al., 2019), TUM (Sturm et al., 2012), and OASIS (Chen et al., 2020). The experimental results show that our model outperforms the SOTA methods significantly (*cf.* Fig. 1), exhibiting better generalization. Even if we only use 72K images for training, which is almost one-twentieth of the training samples used in DPT, our model (Ours-s) still outperforms the previous methods.

To validate the robustness of depth estimation models, we create a robustness benchmark. This benchmark consists of six subsets (IBIMS-C, NYUDv2-C, KITTI-C, ETH3D-C, DIODE-C, and DIODE-C), where the IBIMS-C is used for validation, and the others are used for testing. In particular, we follow ImageNet-C (Hendrycks & Dietterich, 2019) to apply 15 image corruptions to each clean image of the raw RGB-D datasets. For each image corruption, we also define 5 different levels of severity. Therefore, we have a set of 75 corruptions for each image, resulting in 330, 300 images for benchmarking. We comprehensively investigate the robustness of depth estimation models by evaluating them on this benchmark. As shown in Fig. 1, our model still outperforms the prior arts under different corruptions, showing better robustness.

The contributions of this paper can be summarized as:

- We introduce strong data augmentation, consisting of commonly used 2D data augmentations and our 3D-aware defocus blur, to improve the generalization and robustness of the trained model. In particular, we implement the 3D-aware defocus blur in a scattering way to resolve artifacts at depth discontinuities caused by the layered method.
- We conduct zero-shot cross-dataset evaluation to evaluate the generalization of depth estimation models. Experimental results on six RGB-D datasets show that our method outperforms the prior arts by a large margin and sets new SOTA records. Notably, our model (Ours-ViT-hybrid) reduces the absrel by an average of 18.88% when compared to the prior art depth estimator DPT.
- We create a robustness benchmark to validate the robustness of depth estimation models. We comprehensively study the behaviors of robust depth models under different image corruptions. Despite using only one-eighth of our full training data, *i.e.*, 72K training images, our model augmented with strong data augmentations can still outperform the prior arts. This significantly reduces storage

space and training time compared to those methods that use millions of training data. In addition, instead of using unaccessible movie data, all of our training data comes from public resources. We will release our training data for reproducibility.

2 Related Work

2.1 Deep Learning-Based Depth Estimation

Deep learning-based depth estimation can be mainly categorized into supervised and self-supervised learning. Supervised learning methods (Eigen et al., 2014; Laina et al., 2016; Xu et al., 2017; Fu et al., 2018) use ground-truth depth maps as the supervisory signal to learn a mapping function from RGB images to depth maps. Although these methods can generate plausible predictions in specific environments (*e.g.*, indoor scenes), they cannot generalize well to in-the-wild scenes. Instead of directly using ground-truth depth maps for training, self-supervised learning methods (Godard et al., 2017; Zhou et al., 2017; Godard et al., 2019; Bian et al., 2021; Lee et al., 2022) use image reconstruction to supervise the model training. Typically, the self-supervisory signal comes from calibrated stereo pairs (Godard et al., 2017, 2019) or monocular videos (Zhou et al., 2017; Godard et al., 2019; Bian et al., 2021; Lee et al., 2022). However, such methods, always trained on a single dataset, fail to output reliable depth maps on a new dataset.

To enable the monocular depth estimator to work in unconstrained scenes, recent researches (Chen et al., 2016; Xian et al., 2018; Li & Snavely, 2018; Chen et al., 2019; Lasinger et al., 2020; Chen et al., 2020; Xian et al., 2020; Ranftl et al., 2021; Yin et al., 2021, 2022; Lee & Park, 2022) propose to train models on large-scale in-the-wild datasets. For example, (Chen et al., 2016) propose a dataset that consists of 495K internet images, where a point pair of ordinal relationship is manually annotated for each image. Even though the amount of training images is large, DIW (Chen et al., 2016) cannot generate structured depth maps due to the highly sparse supervision for each image. To alleviate this issue, OASIS (Chen et al., 2020) and InstaOrder (Lee & Park, 2022) manually annotate more pairs of ordinal relationships. Such sparse ordinal annotations, however, cannot well preserve the geometric properties of a scene. Therefore, the generation of denser relative depth annotations (Xian et al., 2018, 2020; Lasinger et al., 2020; Li & Snavely, 2018; Chen et al., 2019) has drawn more and more attention. The researchers propose to derive dense or semi-dense relative depth maps from internet stereo images (Xian et al., 2018, 2020), or 3D movies (Lasinger et al., 2020), or video sequences for multi-view reconstructions (Li & Snavely, 2018; Chen et al., 2019). To further improve the generalizability of depth estimation models, recent methods (Lasinger et al., 2020; Yin et al., 2022, 2021; Ranftl et al., 2021) propose to train models on multiple in-the-wild RGB-D datasets simultaneously. Despite the good generalization properties, these methods neglect the real robustness. They did not study the model robustness under different image corruptions.

In this paper, we focus on the data itself, exploring the impact of data augmentation, diversity, and quantity on robust depth estimation. We systematically study the model robustness in addition to the model generalization.

Data augmentation. Data augmentation is a widely used technique in deep learning, which helps to avoid overfitting and improve robustness. Typically, substantial expert knowledge is required to design appropriate data augmentation operations for performance improvement. For instance, in object detection, one needs to rotate the ground-truth box if rotation is used in data augmentation. In image classification (Simonyan & Zisserman, 2014), the primary operations of data augmentation commonly involve flipping, scaling, cropping, rotation, color jittering, *etc*. In recent years, composition-based techniques, such as Mixup (Zhang et al., 2017), Cutout (DeVries & Taylor, 2017), CutMix (Yun et al., 2019), AutoAugment (Cubuk et al., 2019), and RandErasing Zhong et al. (2020), have been proposed to improve the model accuracy as well as the model robustness. Although these techniques have been successfully applied in image classification, object localization, and object detection, they are not appropriate for dense per-pixel prediction tasks like semantic segmentation and depth estimation. Such tasks require extracting features and outputting values for each pixel in an image. Therefore, recent dense prediction methods (Xian et al., 2020; Ranftl et al., 2021; Yin et al., 2021; Yuan et al., 2021) use per-pixel transformations as the primary operations of data augmentation, including random horizontal flipping, random scaling and cropping. However, all aforementioned augmentation methods are operated in 2D. In contrast, we explore 3D data augmentation in monocular depth estimation. We apply 2D and 3D data augmentations to improve the generalization and robustness of the trained model.

3 Method

3.1 Motivation

The prior arts (Ranftl et al., 2021; Yin et al., 2021) have demonstrated good generalization across various datasets. The test images from these datasets always appear clear object boundaries without obvious image degradation. However, our daily photos sometimes suffer from image degradation, *e.g.*, defocus blur and motion blur. The former degradation is due to the large aperture of camera sensors, while the latter is caused by the camera shake of hand-held

Fig. 2 Degradation example. When the test image is perturbed by defocus blur, one can find the SOTA model DPT (Ranftl et al., 2021) degrades significantly

cameras. Then, one may be curious about the performance of the SOTA depth estimator when the test image suffers from such image degradation. As shown in Fig 2, one can find that the SOTA depth estimator DPT (Ranftl et al., 2021) suffers from apparent degradation when the test image is perturbed by defocus blur. This is mainly because *the prior depth estimators neglect the role of data augmentation in robust depth estimation*. They only use the primary 2D data augmentations: random horizontal flipping, random scaling and cropping, and color jittering.

We consider common image degradation in our daily photos and propose a strong data augmentation method consisting of 2D and 3D data augmentations. Additionally, we create a robustness benchmark to study the robustness of depth estimation models under different image corruptions.

3.2 Pipeline of Data Fetching and Augmentation

Algorithm 1 summarizes the process of data fetching and augmentation. For each iteration, we evenly sample RGB images and the corresponding depth maps from different datasets. That is, each batch consists of the same number of RGB-D samples from each dataset. We apply primary data augmentation schemes to images and depth maps, *i.e.*, random scaling and cropping, random horizontal flipping, and random rotation. Notably, such schemes require the depth map to be transformed accordingly. Then, we apply strong data augmentations to images to improve the robustness of the trained model. We use N and P to control the number of RGB transformations and the possibility of applying data transformation, respectively. Next, we will describe our 3D-aware defocus blur in detail.

Algorithm 1: Data fetching and augmentation.

Inputs :

T_{rgbd}	rgb-d transformations applied to images and depths
T_{rgb}	rgb transformations applied to images
M,	multiple datasets for training
N	number of transformations
P	possibility of applying data transformation
I, G	image, ground-truth depth map
$\mathcal{F}, \mathcal{B}, \mathcal{S}$	functions of data fetching, rgb-d/rgb augmentations

// rgb-d transformations

T_{rgbd} = ['scaling/cropping', 'flipping', 'rotation']

// rgb transformations

T_{rgb} = ['color_jitter', 'sharpness', 'gaussian_blur', 'motion_blur', '3D_defocus_blur']

// multi-dataset

M = ['HRWSI', '3DKenBurns', 'DrivingStereo', 'MegaDepth', 'TartanAir', 'Taskonomy', 'Hypersim', 'IRS']

$I, G = \mathcal{F}(M)$ // evenly fetch data from M

$\hat{I}, \hat{G} = \mathcal{B}(I, G, T_{rgbd})$ // primary augmentations

$\hat{I} = \mathcal{S}(\hat{I}, \hat{G}, T_{rgb}, N, P)$ // strong augmentations

Output: Augmented image $\hat{I}$ and depth $\hat{G}$

3.3 3D-Aware Defocus Blur

Defocus blur is a common photographic effect that highlights the refocused object while blurring the rest areas. It is resulted from the large aperture of the thin lens. However, previous depth estimation methods (Yin et al., 2022, 2021) typically build their models on a pinhole imaging model, where the camera aperture is considered as a point, and no lenses are used to focus light. As shown in Fig. 3a, a clear inverted image is produced using the principle of pinhole imaging. Such a model does not include the blurring of unfocused objects caused by lenses. Modern imaging devices, however, are equipped with large-aperture lenses. Figure 3b and c illustrate the imaging process of a thin lens model.

In the thin lens model, light rays emitted from a point travel along paths through the lens, converging at a point behind the lens. When the point locates on the imaging plane, this can be called in-focus (*e.g.*, point a). In contrast, if the light rays converge in front of the imaging plane (*e.g.*, point b), the point B is projected to a circle of confusion (CoC) on the imaging plane with diameter D. Assuming the focal length F and aperture size L are known, the thin lens equation (Saleh & Teich, 2019) yields the following equation:

$$\frac{1}{d_A} + \frac{1}{z_A} = \frac{1}{F}, \tag{1}$$

where d_A and z_A denote the distances from the center of the lens to point A and converging point a, respectively. For the point B, d_B and z_B also satisfy the thin lens equation:

$$\frac{1}{d_B} + \frac{1}{z_B} = \frac{1}{F}. \tag{2}$$

In addition, we can derive the following equation according to the similar triangles in Fig. 1c:

$$\frac{z_A - z_B}{z_B} = \frac{D}{L}, \tag{3}$$

By eliminating z_A and z_B in Eq. 3 with Eq. 1 and Eq. 2, the diameter of CoC D can be obtained by:

$$D = \frac{LF(d_B - d_A)}{d_B(d_A - F)}. \tag{4}$$

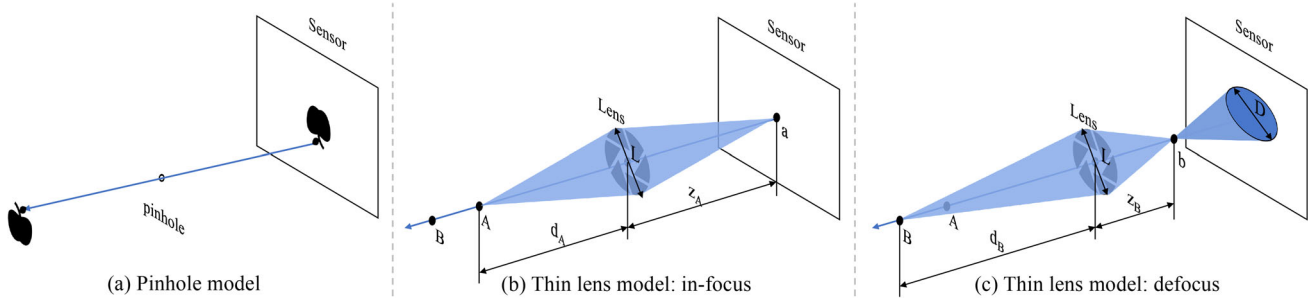


Fig. 3 Imaging principle. Previous monocular depth estimation methods, such as LeReS (Yin et al., 2021) and DiverseDepth (Yin et al., 2022), assume a pinhole camera model **a** However, modern cameras or mobile phones are equipped with large aperture lenses. One can adjust the focal point and the aperture size to refocus **b** at the target object and to defocus blur **c** the image

Considering that point A is the focal point, we use the focus distance d_f to replace d_A. Besides, the focus distance d_f is much larger than the focal length F in practice. For a point with a distance d from the optical center, the diameter of CoC D can be finally computed by:

$$D = \frac{LF|d - d_f|}{d(d_f - F)} \approx LF|\frac{1}{d_f} - \frac{1}{d}|. \tag{5}$$

One can find that the defocus blur can be controlled by the aperture size L, the focal length F, and the focus distance d_f. Therefore, given an all-in-focus image I, the corresponding ground-truth depth map G, and a focal distance d_f, we can obtain the signed defocus map $S = C|\frac{1}{d_f} - \frac{1}{G}|$, where $C = LF$ can be regarded as a factor that controls the maximum blur size. For each pixel i in the all-in-focus image, S_i denotes the blur size of pixel I_i.

To render realistic defocus blur, we implement it in a pixel-wise scattering way—each input pixel scatters the radiance to the neighborhood of it in the output. The formulation can be defined as:

$$\hat{I}(i) = \sum_{\triangle_i} I(i + \triangle_i) K_{i+\triangle_i}(-\triangle_i), \tag{6}$$

where K is the blur kernel which is a function of the pixel it is sampling. Here, we consider the simplest case, *i.e.*, the shape of the blur kernel is circular. The blur radius r_i is set to $\frac{CS_i}{\alpha}$, where α is a constant. In our experiment, α is set to 10, and C is randomly chosen from 3 to 20 for each image. Since we have to iterate over all pixels in the image, a large blur size would result in heavy computational costs. We thus implement the defocus blur with CUDA parallel programming to speed up the process.

3.4 Robustness Benchmark

Our robust depth estimation benchmark is derived from six RGB-D datasets (IBIMS (Koch et al., 2018), NYUDv2 (Silberman et al., 2012), KITTI (Uhrig et al., 2017), ETH3D (Schöps et al., 2017), DIODE (Vasiljevic et al., 2019), and TUM (Sturm et al., 2012)), where IBIMS is used for validation, and the others are used for testing. Following the ImageNet-C (Hendrycks & Dietterich, 2019), we apply 15 image corruptions for each clean image from these datasets, including gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, contrast, elastic transform, pixelate, jpeg compression, snow, fog, frost, and brightness. Each image corruption contains 5 levels of severity. We show some examples in Fig. 4.

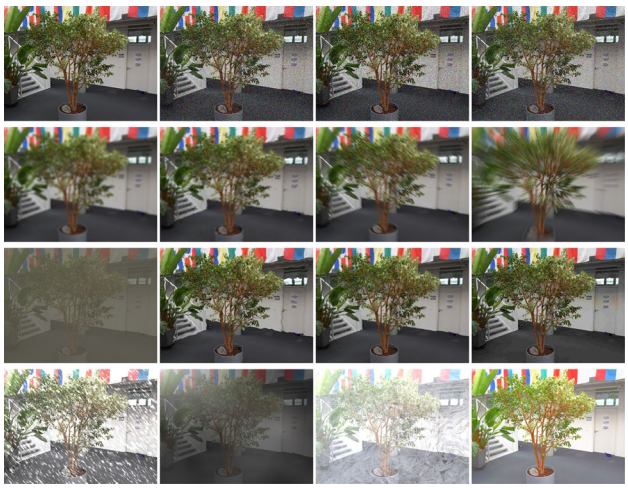


Fig. 4 Illustration of utilized image corruptions on IBIMS-C. We show the original clean image and its 15 corruption types (severity 3) from the IBIMS-C. From left-top to bottom-right: clean input, gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, contrast, elastic transform, pixelate, jpeg compression, snow, fog, frost, and brightness

4 Experiments

4.1 Experimental Setup

Datasets. We summarize the characteristics of our training data in Table 1. We collect 8 RGB-D datasets to train our model, including HRWSI (Xian et al., 2020), 3DKenBurns (Niklaus et al., 2019), DrivingStereo (Yang et al., 2019), MegaDepth (Li & Snavely, 2018), TartanAir (Wang et al., 2020), Taskonomy (Zamir et al., 2018), Hypersim (Roberts et al., 2021), and IRS (Wang et al., 2019). HRWSI (Xian et al., 2020) is a high-resolution relative depth dataset derived from uncalibrated web stereo images. We use the official training set for training. 3DKenBurns (Niklaus et al., 2019) is a synthetic dataset with accurate metric depth maps. We sampled around 76K RGB-D pairs for training. DrivingStereo (Yang et al., 2019) is a large-scale stereo dataset covering a diverse set of driving scenarios. We sampled around 87K left images and the corresponding ground truth labels for training. MegaDepth (Li & Snavely, 2018) is another relative depth dataset derived from web image sequences. Since these image sequences primarily feature outdoor static scenes, MegaDepth uses multi-view stereo reconstruction to obtain depth maps. We sampled around 102K RGB-D pairs for training. TartanAir (Wang et al., 2020) is collected in photo-realistic simulation environments featuring various weather, lighting conditions, and moving objects. We sampled around 92K left images and the corresponding depth maps for training. Taskonomy (Zamir et al., 2018) is an indoor dataset captured by a laser scanner. We use around 58K RGB-D pairs for training. Hypersim (Roberts et al., 2021) and IRS (Wang et al., 2019) are both synthetic indoor datasets, where the ground truth of the latter is inverse depth up to an unknown scale. In Summary, our full training data, consisting of 572K RGB-D pairs, covers a wide range of scenes.

To evaluate the generalization of depth estimation models, we conduct zero-shot cross-dataset evaluation on six RGB-D benchmarks, including NYUDv2 (Silberman et al., 2012), KITTI (Uhrig et al., 2017), ETH3D (Schöps et al., 2017), DIODE (Vasiljevic et al., 2019), TUM (Sturm et al., 2012), and OASIS (Chen et al., 2020). NYUDv2 is an indoor RGB-D dataset that consists of 654 test samples. Following common practice (Eigen & Fergus, 2015), we crop the predictions and the ground truth depth maps to size 427×561 before evaluation. KITTI is an outdoor RGB-D dataset, the ground truth depth maps of which are captured via a Velodyne laser scanner. Following (Godard et al., 2019), we evaluate depth estimation models using 652 test images. ETH3D, consisting of 412 test images, covers both indoor and outdoor scenes. We use the low-resolution version, *i.e.*, 504×756, for evaluation. DIODE also contains both indoor and outdoor scenes with high-quality ground-truth depth maps. We use the official test split (*i.e.*, 771 samples). TUM (Sturm et al., 2012) features people with complex movements. Following HRWSI (Xian et al., 2020), we use 1815 images for evaluation. OASIS is a large-scale dataset for monocular 3D in the wild that contains 10, 000 images with human annotations.

To evaluate the robustness of depth estimation models, we compare our models with 9 representative MDE methods on our robustness benchmark. Since each corruption has 5 different levels of severity, we thus have a set of 75 visual corruptions for each image. Considering that 1) OASIS would have 750, 000 samples if we apply these corruptions to the test images, which would be time-consuming to test on these images; 2) The metric depth datasets covering diverse scenes are enough to verify the model robustness, our robustness benchmark is thus composed of IBIMS-C, NYUDv2-C, KITTI-C, ETH3D-C, DIODE-C, and TUM-C, where we use IBIMS-C for validation and the others for testing.

Details of data augmentation. We introduce the hyperparameters used in our data augmentation schemes. Appropriate image perturbations can prevent model overfitting and improve the generalization and robustness of the model. We consider 8 data augmentation schemes, including scaling and cropping, horizontal flipping, rotation, color jittering, sharpness, gaussian blur, motion blur, and 3D-aware defocus blur. For the first three schemes, we need to transform RGB and depth simultaneously. For the rest five schemes, we only add appropriate defocus blur to RGB image, the depth map is untransformed. In particular, we random scale RGB and depth in the range of [0.6, 1.0] and crop them to the size of 384×384. We use horizontal flipping and rotation with a 50% chance, respectively. Notably, the maximum angle of rotation is set to 2.5.

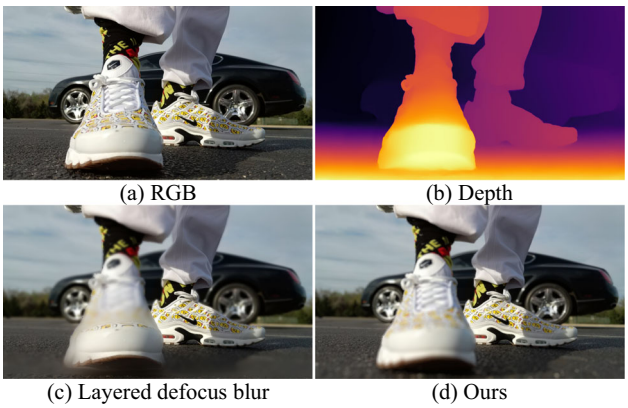
In addition to the rgb-d transformations, we also consider 5 rgb transformations. Considering that such transformations do not always exist in an image, we apply each of them with a 30% chance. We implement our scattering-based defocus blur with CUDA programming. For an image with resolution 360×640, our method only costs 0.77ms on an NVIDIA RTX A5000 GPU, while the layered-based one costs 3.21ms. Besides, we show a qualitative comparison in Fig. 5, one can find that our method can generate smoother and more realistic defocus blur. We implement the other four rgb transformations (color jittering, sharpness, gaussian blur, and motion blur) based on Kornia.[1] For color jittering, we use the default setting of Planckian jitter (Zini et al., 2022). For sharpness, we randomly select a sharpness strength from [0.1, 0.3]. For gaussian blur, we set the kernel size to 3×3 and and randomly choose the standard deviation of the kernel in the range of [0.1, 2.0]. For motion blur, we randomly choose the motion kernel size and the angle of the motion blur from {3, 5, 7} and [10.0, 30.0], respectively.

[1] https://kornia.readthedocs.io/en/latest/augmentation.module.html

Table 1 RGB-D datasets

Dataset	Indoor	Outdoor	Dynamic	Quality	Capture	Depth	# Images
HRWSI	√	√	√	Low	Stereo	No scale/shift	20,378
3DKenBurns	√	√		High	Synthetic	No scale	76,048
DrivingStereo		√	√	Medium	Laser	Metric	87,219
MegaDepth		√		Medium	SFM	No scale	102,339
TartanAir		√		High	Synthetic	Metric	91,648
Taskonomy	√			Medium	Laser	Metric	58,144
Hypersim	√			High	Synthetic	No scale	49,042
IRS	√			High	Synthetic	No scale	87,677

Our full training set, consisting of 572,495 RGBD images, covers a wide range of scenes, including indoor, outdoor, and dynamic scenes



(a) RGB (b) Depth (c) Layered defocus blur (d) Ours

Fig. 5 Defocus blur. Layered defocus blur is prone to produce artifacts around depth discontinuities, and the blur in the composite image looks distinctly layered. By contrast, our method synthesizes smoother and more realistic defocus blur

To study how data affect model accuracy and robustness, we investigate multi-dataset training and different combinations of data augmentation schemes. We show the multi-dataset training and data augmentation combinations in Table 2 and Table 3, respectively.

Details of model training. We implement our model based on DPT (Ranftl et al., 2021). Unless otherwise stated, we use ViT-hybrid (Dosovitskiy et al., 2020) as our encoder. We use the scale-shift invariant loss (Lasinger et al., 2020) combined with multi-scale gradient matching loss (Li & Snavely, 2018) to supervise the model training. The network is trained for 60 epochs using the AdamW optimizer with an initial learning rate of 5e-5, which is then multiplied by 0.1 after 45 epochs. We train our model on 8 V100 GPUs with a batch size of 16 per GPU. For each batch, we randomly sample the same number of images from different data sources. During training, the images fed to the network are at the size of 384×384. At the inference stage, we resize the network output to the original image size before evaluation.

Evaluation protocol. To evaluate the generalization of depth estimation models, we conduct zero-shot cross-dataset evaluation on six RGB-D benchmarks, including NYUDv2 (Silberman et al., 2012), KITTI (Uhrig et al., 2017), ETH3D (Schöps et al., 2017), DIODE (Vasiljevic et al., 2019), TUM (Sturm et al., 2012), and OASIS (Chen et al., 2020). Following DPT (Ranftl et al., 2021), we align the predictions and the ground truths before evaluation. All alignments are conducted in inverse depth space using the least-squares criterion and then converted back to depth space for evaluation. We use the absolute value of relative error absrel and the percentage of pixels $\delta_1 < 1.25$ for metric depth evaluation and the weighted human disagree rate WHDR (Chen et al., 2020) for ordinal depth evaluation.

Table 2 Combinations of training sets

Dataset	M1	M2	M3	M4	M5	M6	M7	M8
HRWSI	✓	✓	✓	✓	✓	✓	✓	✓
3DKenBurns		✓	✓	✓	✓	✓	✓	✓
DrivingStereo			✓	✓	✓	✓	✓	✓
MegaDepth				✓	✓	✓	✓	✓
TartanAir					✓	✓	✓	✓
Taskonomy						✓	✓	✓
Hypersim							✓	✓
IRS								✓

Table 3 Combinations of data augmentation schemes

Scheme	DA1	DA2	DA3	DA4	DA5	DA6	DA7	DA8
Scaling/Cropping	✓	✓	✓	✓	✓	✓	✓	✓
Flipping		✓	✓	✓	✓	✓	✓	✓
Rotation			✓	✓	✓	✓	✓	✓
Color jittering				✓	✓	✓	✓	✓
Sharpness					✓	✓	✓	✓
Gaussian blur						✓	✓	✓
Motion blur							✓	✓
3D defocus blur								✓

Table 4 Zero-shot cross-dataset evaluation (%)

Method	#Training data	NYUDv2		KITTI		ETH3D		DIODE		TUM		OASIS	Ranking
		absrel↓	δ_1 ↑	absrel↓	δ_1↑	absrel↓	δ_1↑	absrel↓	δ_1↑	absrel↓	δ_1↑	WHDR↓	
MegaDepth (Li & Snavely, 2018) (CVPR'18)	130K	18.12	73.95	19.68	67.46	18.58	73.80	43.98	60.49	23.28	59.55	33.46	10.33
YouTube3D (Chen et al., 2019) (CVPR'19)	795K	16.36	77.65	30.04	49.21	21.30	70.07	43.56	61.75	24.15	57.46	33.54	10.50
HRWSI (Xian et al., 2020) (CVPR'20)	20K	15.82	76.30	19.52	67.78	18.30	78.43	33.95	65.58	15.45	78.77	32.52	8.67
MiDaSv2.1 (Lasinger et al., 2020) (TPAMI'20)	1.4M	10.16	89.96	12.35	84.86	11.98	87.54	26.60	71.36	12.14	87.41	30.63	5.42
DPT (Ranftl et al., 2021) (ICCV'21)	1.4M	9.88	90.34	12.00	86.47	10.77	90.92	27.72	72.20	10.79	88.84	30.72	4.08
LeReS (Yin et al., 2021) (CVPR'21)	360K	8.60	92.10	15.04	77.71	11.79	86.96	33.90	72.13	17.01	76.64	28.13	6.00
3DCC (Kar et al., 2022) (CVPR'22)	5M	**7.54**	**93.64**	13.72	83.13	**7.74**	**94.60**	31.72	**75.82**	12.78	86.77	28.31	3.50
DiverseDepth (Yin et al., 2022) (TPAMI'22)	300K	10.37	89.45	22.18	62.72	14.47	81.58	35.81	68.21	20.76	65.86	31.80	8.58
InstaOrder (Lee & Park, 2022) (CVPR'22)	101K	11.41	87.66	12.39	85.03	14.32	82.59	28.06	70.50	12.95	85.82	32.31	6.83
DPT* (Ranftl et al., 2021)(ICCV'21)	72K	9.03	91.61	11.97	85.88	9.62	91.20	25.40	72.15	11.05	88.38	28.42	3.67
Ours-ViT-hybrid†	72K	8.34	92.71	10.81	88.12	8.96	92.59	24.99	73.11	11.96	87.51	27.8635	2.33
Ours-ViT-hybrid	572K	7.81	93.58	**10.08**	**89.51**	8.30	94.12	**24.03**	74.26	**10.40**	**89.80**	**27.8632**	**1.33**

† indicates the model was trained on a small training subset. * indicates the model was trained with the same hardware setup, batch size, and training data as Ours-ViT-hybrid†. For absrel and WHDR, the lower the better. For δ_1, the higher the better. The best performance is **boldfaced**

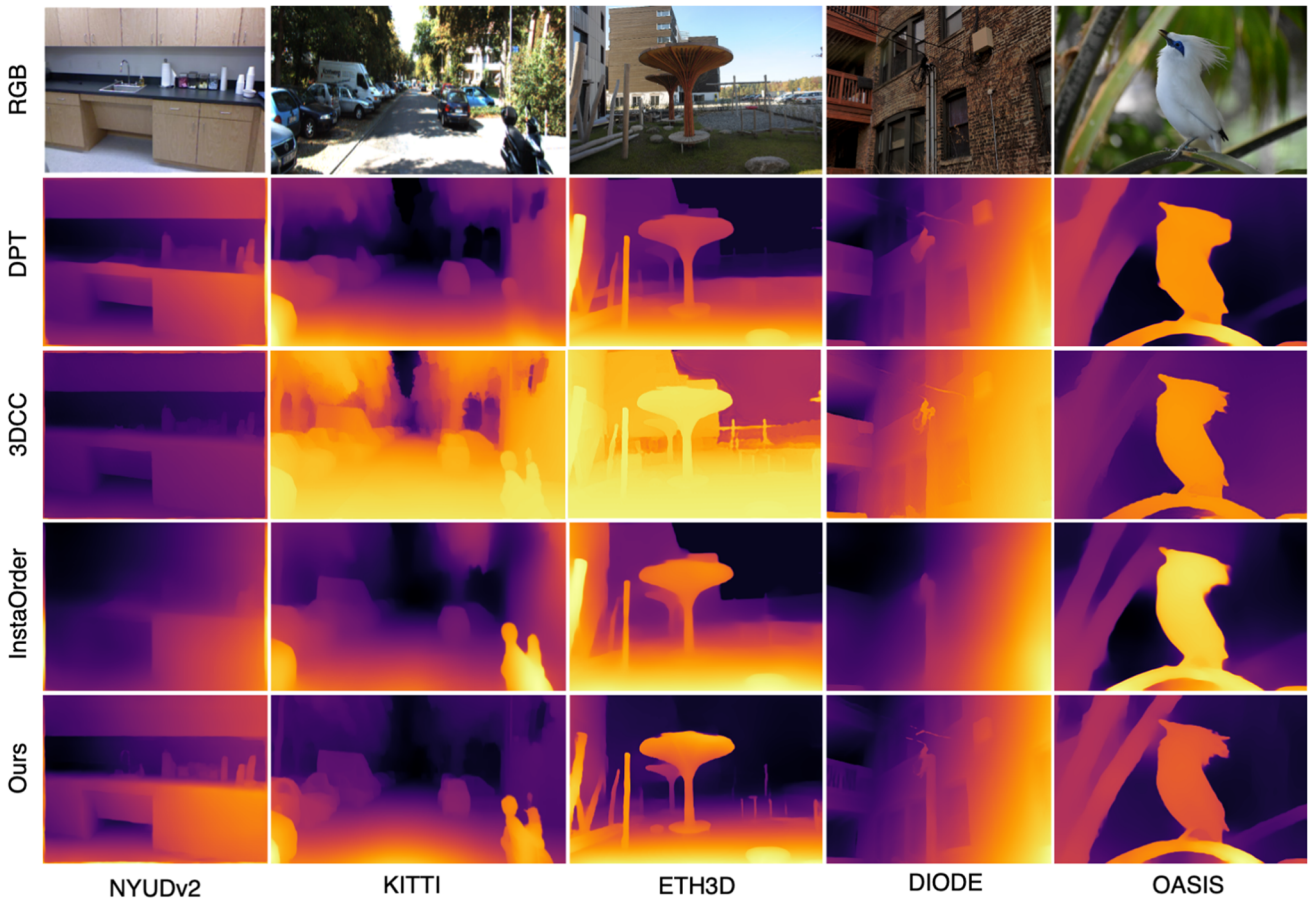


Fig. 6 Qualitative results on five RGB-D datasets. Compared to the prior arts, our predictions preserve finer-grained details

4.2 Zero-Shot Cross-Dataset Evaluation

We conduct zero-shot cross-dataset evaluation to evaluate the generalization of depth estimation models. Specifically, we test these models on six RGB-D datasets that were not seen during training. We compare our models with 9 robust monocular depth methods, including MegaDepth (Li & Snavely, 2018), YouTube3D (Chen et al., 2019), HRWSI (Xian et al., 2020), MiDaSv2.1 (Lasinger et al., 2020), DPT (Ranftl et al., 2021), LeReS (Yin et al., 2021), 3DCC (Kar et al., 2022), DiverseDepth (Yin et al., 2022), and InstaOrder (Lee & Park, 2022). To assess the impact of the proposed augmentation schemes, both the original DPT-Hybrid (Ranftl et al., 2021) and the augmented version are trained using the same hardware setup, including the same batch size (16 per GPU) and the same training data (our collected 72K samples). We use DPT* and Ours-ViT-hybrid† to represent these two models. We show the quantitative and qualitative results in Table 4 and Fig. 6, respectively. One can observe that: (i) Our models, *i.e.*, Ours-ViT-hybrid† and Ours-ViT-hybrid, demonstrate stronger generalization capability than the previous MDE methods. Specifically, our models outperform the other methods in terms of ordinal depth evaluation on OASIS, even if Ours-ViT-hybrid† is trained with only 72K images. (ii) DPT* outperforms the original DPT (average ranking: 3.67 vs. 4.08). The result indicates that our collected training data has higher quality, which helps reduce training time and improve model performance. By incorporating our proposed data augmentation strategies (Ours-ViT-hybrid†), one can find that the model generalization performance can be further improved (average ranking: 2.33 vs. 3.67). This verifies the effectiveness of our proposed data augmentation schemes. We analyze the reasons for the slight performance degradation on the TUM dataset. This is likely because the TUM dataset has low image resolution, a high level of noise, and contains a significant amount of motion blur. Introducing more and stronger perturbations will affect the training of the model. (iii) Other MDE methods may perform well in some specific scenarios but fail to generalize well to the other ones. For instance, DiverseDepth (Yin et al., 2022), LeReS (Yin et al., 2021), and 3DCC (Kar et al., 2022) work well in indoor scenes, but they can not obtain good performance in outdoor scenes like KITTI (Uhrig et al., 2017). This is likely due to the bias of training data since these methods

Fig. 7 More qualitative results on in-the-wild images. All images come from the OASIS (Chen et al., 2020) dataset

use many more indoor scenes for training. (iv) 3DCC (Kar et al., 2022) is prone to generate obvious artifacts around depth discontinuities (*e.g.*, the prediction on ETH3D). This may be caused by defocus blur based on layered blurring. By contrast, our models can preserve fine-grained details without obvious artifacts (see more examples in Fig. 7).

4.3 Robustness Under Image Corruptions

We create a robustness benchmark to systematically study the robustness of depth estimation models under different image corruptions. In particular, we use NYUDv2-C, KITTI-C, ETH3D-C, DIODE-C, and TUM-C for evaluation, which expands the clean data of the five metric depth datasets with 15 corruption types. We compare our models with 9 robust monocular depth estimation models. In Table 5, we report the absrel scores on the clean data, the average absrel scores of 15 corruption types, and the average absrel scores of each corruption type. One can conclude that the model generalization is `not` equal to the model robustness. For example, 3DCC (Kar et al., 2022) achieves the best performance on the clean data of NYUDv2, but it is less robust than our model trained with 72K images. Therefore, it is essential to consider both generalization and robustness to achieve the true sense of a robust depth model. This verifies the motivation of this paper. Besides, one can find that depth estimation models degrade when the test image is corrupted. Different image corruptions have different impacts on models:

- Performance w.r.t Noise. All kinds of noise degrade model performance. In general, impulse noise degrades each model the most among the three kinds of noise. 3DCC outperforms ours on NYUDv2-C and ETH3D-C. This is likely because 3DCC uses much more indoor data and augments with image noise for training. Therefore, 3DCC exhibits overfitting on indoor data, easily produces artifacts at depth discontinuities, and performs poorly in outdoor scenes (*e.g.*, KITTI-C and DIODE-C). DPT outperforms our model on the TUM-C dataset due to its training regimen leveraging a large amount of private movie data, predominantly characterized by dynamic human subjects. It is noteworthy that the TUM dataset itself features people with complex movements.
- Performance w.r.t Blur. Benefiting from our strong data augmentations (3D-aware defocus blur and motion blur), our models outperform the SOTA depth estimators in category blur. Specifically, our models are less affected by defocus blur. By contrast, DPT (Ranftl et al., 2021) degrades significantly when the test image is perturbed by defocus blur.
- Performance w.r.t Weather. Weather has a substantial impact on model robustness, especially in outdoor scenarios. Texture-corrupting distortions, such as snow and frost, degrade each model significantly. Thanks to the distribution of their respective training data, 3DCC and DPT outperform ours on ETH3D-C and TUM-C, respectively.
- Performance w.r.t Digital. Compared to the image distortions corrupting the texture of the test image, *e.g.*, snow and frost, MDE models are least affected by digital corruptions. This is because such corruptions do not change the texture of the test image too much. Similar to the image noise and weather conditions, 3DCC also outperforms ours on ETH3D-C because of the biased, large-scale training data (5M samples).

We also offer practical insights and implications based on our findings. This includes guidance on when and where our proposed model is most effective and under what circumstances other methods, such as 3DCC and DPT, might be preferable. In general, our models are the most robust, especially when the image contains defocus blur or motion blur. Note that with the popularity of mobile phones, defocus blur and motion blur are more common in our daily lives than other image corruptions. Other methods may perform well under certain circumstances but fail in more general or more complex conditions. For example, although 3DCC achieves good performance in indoor scenes like NYUDv2-C and ETH3D-C, it is prone to produce obvious artifacts around depth discontinuities. It cannot output satisfactory results in outdoor scenes or scenes with complex moving objects. If one only cares about depth accuracy in indoor and static scenes, 3DCC may be a good alternative.

We show qualitative results on some corrupted test images in Fig. 8. Overall, our model is more robust than the prior arts.

4.4 How Does Data Affect Accuracy and Robustness?

Here, we try to answer two questions: (1) How do different data augmentation strategies impact the accuracy and robustness of depth estimation models? (2) Does more training data necessarily mean better model performance?

Before answering the first question, we would like to explain why we do not train a model without any data preprocessing. Because different datasets contain images of different resolutions and mix-data training requires the resolution of images input to the network to be the *same* and *square*, resizing to a fixed square and scaling/cropping are two data pre-processing methods used before feeding images to the network. The reason we do not use "Resizing to a fixed square" is because the aspect ratio changes. This causes the shape distortion of objects in the image. So, we adopt the scaling/cropping strategy at training. The aspect ratio would be proportional scaling, thus the shape of objects in the image would be kept. To verify that scaling/cropping (DA1) is superior to "Resizing to a fixed square", we train a variant by

Table 5 Corruption robustness evaluation (%)

Dataset	Method	Clean	Average	Noise			Blur				Weather				Digital			
				Gauss	Shot	Impul	Defoc	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr	Elast	Pixel	JPEG
NYUDv2-C	MegaDepth (Li & Snavely, 2018)	18.12	21.58	22.82	21.72	22.40	23.08	21.97	21.25	22.52	25.12	24.17	19.34	20.00	20.90	19.63	18.87	19.88
	YouTube3D (Chen et al., 2019)	16.36	18.77	18.96	18.62	19.02	18.09	18.33	18.42	19.72	21.38	20.95	18.83	17.55	20.50	17.53	16.55	17.12
	HRWSI (Xian et al., 2020)	15.82	18.01	18.28	17.63	18.14	17.18	17.61	17.59	19.04	22.08	19.86	17.65	16.42	18.93	16.77	16.11	16.88
	MiDaSv2.1 (Lasinger et al., 2020)	10.16	13.93	14.88	14.49	15.51	13.75	15.80	13.83	15.64	18.69	15.95	11.95	11.15	12.49	12.36	10.66	11.74
	DiverseDepth (Yin et al., 2022)	10.37	16.50	20.51	20.82	21.51	14.60	17.41	14.12	18.45	19.58	18.25	13.99	12.46	16.68	13.63	11.34	14.20
	DPT (Ranftl et al., 2021)	9.88	12.63	12.52	12.17	12.70	13.62	12.46	12.11	14.58	17.98	**14.49**	11.73	10.70	11.53	10.86	10.44	11.52
	LeReS (Yin et al., 2021)	8.60	13.57	14.73	13.79	15.17	11.44	15.24	11.82	16.87	17.64	19.11	11.88	9.90	14.76	11.38	9.75	10.16
	3DCC (Kar et al., 2022)	**7.54**	11.28	**9.85**	**9.57**	**10.40**	9.96	17.80	10.46	14.33	15.85	14.60	**9.07**	8.83	10.94	10.32	**8.12**	**9.08**
	InstaOrder (Lee & Park, 2022)	11.41	14.47	14.74	14.35	15.36	14.12	14.97	14.25	15.76	18.30	17.12	13.48	12.51	14.10	13.15	11.88	12.92
	DPT* (Ranftl et al., 2021)	9.03	12.62	12.95	12.57	13.65	12.53	13.01	12.47	15.73	16.83	16.00	11.12	10.16	11.46	10.51	9.51	10.86
	Ours-ViT-hybrid†	8.34	11.24	11.43	10.97	11.97	9.66	10.88	9.98	14.67	**14.64**	15.79	10.54	9.23	**10.29**	9.81	8.67	10.12
	Ours-ViT-hybrid	7.81	**10.91**	11.00	10.62	11.50	**9.31**	**10.42**	**9.53**	**13.78**	15.03	15.40	10.38	**8.73**	10.63	**9.49**	8.18	9.65
KITTI-C	MegaDepth (Li & Snavely, 2018)	19.68	26.12	26.25	24.55	27.22	29.37	30.44	29.88	27.07	29.94	25.84	25.61	20.31	30.83	20.18	21.75	22.49
	YouTube3D (Chen et al., 2019)	30.04	47.20	46.10	46.99	46.29	45.63	47.50	46.68	45.17	46.33	47.28	47.25	49.16	45.45	49.53	49.02	49.61
	HRWSI (Xian et al., 2020)	19.52	23.99	25.51	24.48	25.14	22.37	23.39	24.23	26.52	27.99	24.26	25.50	18.96	26.14	19.38	21.47	24.58
	MiDaSv2.1 (Lasinger et al., 2020)	12.35	16.78	17.63	16.70	17.93	18.74	21.25	17.95	19.51	20.06	19.02	14.00	13.01	15.07	12.54	13.46	14.78
	DiverseDepth (Yin et al., 2022)	22.18	31.82	38.02	36.81	37.65	31.52	29.62	27.05	26.91	33.87	35.55	35.72	29.27	38.06	22.52	25.16	29.52
	DPT (Ranftl et al., 2021)	12.00	15.42	14.62	14.10	**15.36**	17.39	16.71	15.05	**16.21**	25.38	20.97	12.32	11.43	12.94	11.62	13.02	**14.21**
	LeReS (Yin et al., 2021)	15.04	25.79	32.24	28.69	32.12	29.54	27.86	19.45	29.54	37.35	31.64	24.70	19.89	23.07	15.32	18.20	17.20
	3DCC (Kar et al., 2022)	13.72	21.85	18.80	19.50	19.05	23.20	27.68	22.61	26.05	33.43	25.07	19.09	16.00	21.61	15.43	17.18	22.98
	InstaOrder (Lee & Park, 2022)	12.39	16.36	16.26	15.56	16.53	20.55	19.49	17.27	18.53	**18.78**	**17.38**	14.35	13.22	15.44	12.99	13.89	15.18
	DPT* (Ranftl et al., 2021)	11.97	16.18	17.44	16.60	18.32	15.90	16.38	17.96	19.00	23.94	18.71	13.42	11.98	13.77	12.01	12.73	14.62
	Ours-ViT-hybrid†	10.81	15.04	15.66	14.99	16.62	**12.89**	15.57	14.61	17.07	23.88	19.89	12.36	10.64	13.32	11.06	11.86	15.11
	Ours-ViT-hybrid	**10.08**	**14.47**	**15.05**	**13.78**	16.30	13.27	**14.77**	**13.78**	16.81	25.48	18.42	**11.67**	**9.80**	**12.23**	**10.06**	**11.13**	14.41
ETH3D-C	MegaDepth (Li & Snavely, 2018)	18.58	23.39	26.90	26.34	26.81	22.45	22.98	22.52	24.10	26.98	27.26	20.97	19.68	23.51	19.41	19.21	21.75
	YouTube3D (Chen et al., 2019)	21.30	24.31	25.14	25.08	25.08	23.21	23.30	23.91	24.88	26.52	26.58	25.24	21.89	27.14	22.30	21.56	22.78
	HRWSI (Xian et al., 2020)	18.30	21.09	22.01	21.09	21.76	20.17	20.47	20.21	22.85	25.27	23.50	20.72	19.41	20.93	18.98	18.65	20.36
	MiDaSv2.1 (Lasinger et al., 2020)	11.98	15.65	17.42	17.13	17.30	15.35	15.97	15.08	16.88	20.18	19.32	14.10	12.51	14.41	13.08	12.10	13.89
	DiverseDepth (Yin et al., 2022)	14.47	19.89	23.04	22.83	22.99	17.63	18.50	16.73	20.38	24.00	23.23	18.96	16.25	21.36	16.84	14.96	20.67
	DPT (Ranftl et al., 2021)	10.77	13.43	14.26	14.24	14.12	12.12	12.15	11.68	14.83	20.07	16.87	12.88	11.02	12.21	11.08	10.48	13.45
	LeReS (Yin et al., 2021)	11.79	15.22	17.36	16.95	17.17	12.83	15.14	12.59	18.00	21.37	20.27	13.77	12.16	14.92	12.18	11.34	12.24
	3DCC (Kar et al., 2022)	**7.74**	**10.76**	**11.15**	**11.13**	**11.00**	9.56	10.48	10.6	12.94	**16.17**	**14.59**	**9.03**	**8.21**	10.30	**8.47**	**8.17**	**10.20**
	InstaOrder (Lee & Park, 2022)	14.32	17.02	17.76	17.47	17.98	16.31	17.17	16.39	17.91	21.32	20.31	16.37	14.62	16.88	15.03	14.39	15.40

Table 5 continued

Dataset	Method	Clean	Average	Noise			Blur				Weather				Digital			
				Gauss	Shot	Impul	Defoc	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr	Elast	Pixel	JPEG
	DPT* (Ranftl et al., 2021)	9.62	13.00	15.03	14.89	14.93	10.81	11.69	11.24	15.34	18.47	17.35	12.00	10.15	11.03	10.59	9.71	11.81
	Ours-ViT-hybrid†	8.96	12.04	13.42	13.63	13.49	9.56	10.50	10.28	14.18	17.64	16.51	11.88	9.46	10.33	9.74	9.12	10.88
	Ours-ViT-hybrid	8.30	11.36	12.81	12.53	12.37	**8.80**	**9.75**	**9.75**	**14.03**	16.43	15.66	11.10	8.93	**10.02**	9.16	8.51	10.46
DIODE-C	MegaDepth (Li & Snavely, 2018)	43.98	46.17	46.42	45.65	46.49	46.21	46.15	46.34	49.72	47.65	47.25	46.52	44.00	47.14	44.67	44.02	44.26
	YouTube3D (Chen et al., 2019)	43.56	45.09	46.13	45.76	46.31	44.14	44.28	44.37	45.63	46.40	46.19	45.33	43.94	46.18	44.15	43.61	43.92
	HRWSI (Xian et al., 2020)	33.95	33.74	32.41	32.26	32.59	32.87	32.87	33.14	33.50	37.30	35.89	33.37	34.20	32.97	33.77	33.67	35.31
	MiDaSv2.1 (Lasinger et al., 2020)	26.60	28.21	29.16	29.11	30.38	27.32	27.86	27.09	27.96	30.93	30.78	27.74	26.49	27.92	26.66	26.49	27.34
	DiverseDepth (Yin et al., 2022)	35.81	40.02	42.96	42.09	43.54	37.01	37.51	36.06	42.05	42.68	43.44	41.95	37.92	41.71	37.05	35.62	38.64
	DPT (Ranftl et al., 2021)	27.72	28.85	29.89	30.06	30.11	26.92	26.62	26.49	27.57	33.65	**30.62**	30.12	27.60	27.05	28.69	26.63	30.71
	LeReS (Yin et al., 2021)	33.90	37.19	39.19	38.04	39.22	34.67	36.52	35.24	41.48	40.95	39.80	37.48	34.78	36.97	34.75	34.06	34.64
	3DCC (Kar et al., 2022)	31.72	34.34	35.32	34.72	35.93	32.77	33.42	33.30	38.54	36.90	36.99	33.12	32.63	34.27	32.31	31.67	33.19
	InstaOrder (Lee & Park, 2022)	28.06	29.69	30.24	30.18	31.16	28.24	28.72	28.49	28.96	32.69	31.94	29.82	28.39	29.72	28.43	28.16	30.24
	DPT* (Ranftl et al., 2021)	25.40	27.57	29.72	29.44	30.00	26.05	25.57	25.75	27.12	31.14	31.44	27.01	25.38	26.24	25.85	25.24	27.56
	Ours-ViT-hybrid†	24.99	27.10	28.52	28.75	29.18	24.88	25.81	24.78	**26.34**	31.16	31.85	27.53	25.34	25.62	25.81	24.77	26.21
	Ours-ViT-hybrid	**24.03**	**26.54**	**27.80**	**27.88**	**28.83**	**23.95**	**24.71**	**24.65**	27.31	**30.79**	31.19	**26.61**	**24.40**	**25.45**	**24.66**	**24.42**	**25.47**
TUM-C	MegaDepth (Li & Snavely, 2018)	23.28	26.51	29.12	29.00	28.93	27.32	24.79	26.58	28.03	27.06	25.73	26.14	24.65	26.97	23.76	24.88	24.76
	YouTube3D (Chen et al., 2019)	24.15	25.75	26.81	27.03	26.85	25.80	25.10	25.24	25.95	26.37	26.01	25.87	24.50	27.34	24.02	24.66	24.66
	HRWSI (Xian et al., 2020)	15.45	18.20	18.52	18.75	18.24	17.35	17.86	17.75	19.21	20.79	19.32	17.78	16.10	19.41	18.26	16.63	16.98
	MiDaSv2.1 (Lasinger et al., 2020)	12.14	16.47	16.22	16.71	16.92	17.58	19.26	16.85	18.88	18.93	17.78	15.03	13.31	15.69	16.68	13.14	14.07
	DiverseDepth (Yin et al., 2022)	20.76	24.68	25.97	26.50	25.73	24.02	23.49	24.48	26.97	25.38	25.97	24.89	22.21	25.67	22.88	22.50	23.57
	DPT (Ranftl et al., 2021)	10.79	14.56	**14.61**	**14.61**	**14.52**	15.77	15.43	14.46	**17.21**	18.86	**15.95**	**13.96**	11.60	12.25	14.05	12.11	12.98
	LeReS (Yin et al., 2021)	17.01	23.54	22.83	24.04	23.70	22.89	24.52	23.42	27.70	24.83	26.64	23.94	19.64	24.19	22.22	21.58	20.90
	3DCC (Kar et al., 2022)	12.78	16.80	16.57	16.64	16.50	14.41	17.66	17.21	21.30	22.04	20.74	17.38	14.67	15.72	13.43	12.65	15.10
	InstaOrder (Lee & Park, 2022)	12.95	15.96	15.27	15.50	15.66	16.57	17.48	16.35	17.98	18.02	16.98	15.46	13.75	16.25	16.19	13.68	14.23
	DPT* (Ranftl et al., 2021)	12.57	15.51	15.91	16.28	16.77	15.65	16.21	15.90	18.98	18.63	17.56	14.95	11.93	13.28	13.99	12.55	14.11
	Ours-ViT-hybrid†	11.96	15.30	16.08	15.94	16.01	14.08	15.86	14.91	18.60	**17.38**	16.97	15.88	12.32	13.54	14.78	13.08	14.00
	Ours-ViT-hybrid	**10.40**	**14.10**	15.09	15.05	15.29	**12.44**	**14.00**	**13.66**	17.97	17.68	16.01	14.38	**11.23**	**11.81**	**13.15**	**11.46**	**12.38**

We report the average absrel scores on different corrupted images, where each corruption type has five levels of severity. The best performance per corruption is **boldfaced**

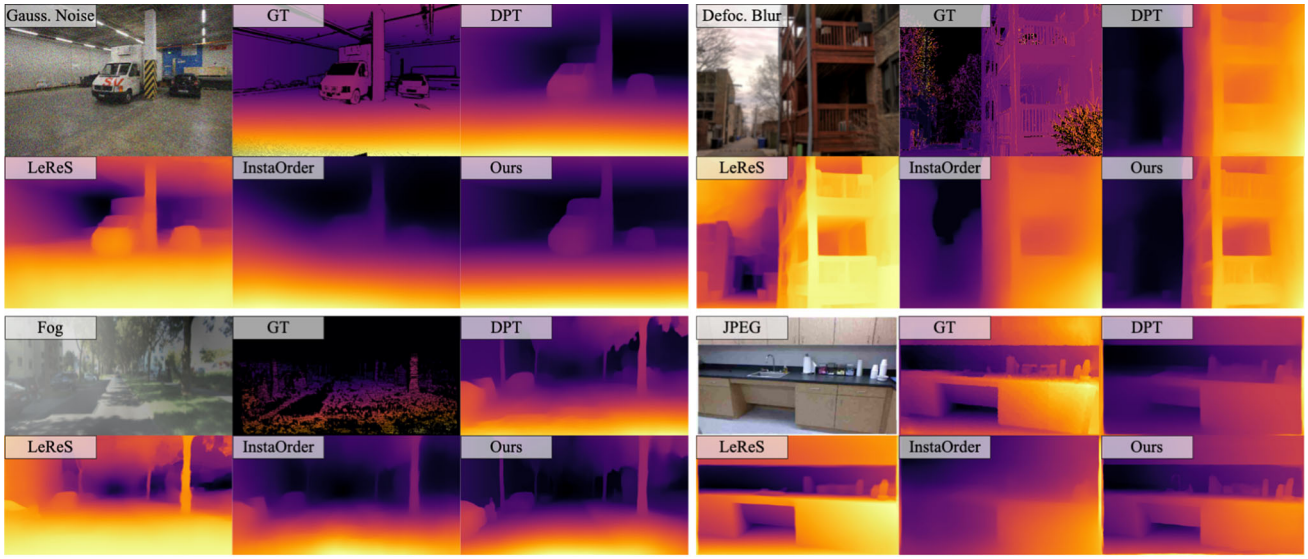


Fig. 8 Qualitative results on various corrupted images. The test images from the top left to bottom right are from ETH3D, DIODE, KITTI, and NYUDv2, respectively. We show four categories of image corruption (*i.e.*, noise, blur, weather, and digital). In particular, we visualize the results of images corrupted by gaussian noise, defocus blur, fog, and jpeg compression. The severities of corruption are level 3

resizing all images to a fixed square (384 × 384). For generalization, the absrel scores on the clean data of "Resizing to a fixed square" and scaling/cropping are 8.48 and 8.10, respectively. For robustness, the average absrel scores of 15 corruption types are 10.75 and 10.33, respectively.

To investigate how different data augmentation schemes affect the accuracy and robustness of depth estimation models, we train the ViT-hybrid-based models with different data augmentation configurations on a small training subset that consists of 72K image and depth pairs. These models are trained for the same iterations and evaluated on IBIMS and IBIMS-C. As shown in Fig. 9a, one can observe that the accuracy and robustness of our model can be improved by using our strong data augmentation schemes. However, not all augmentation schemes play a positive role in model training. The accuracy and robustness of our model become worse if we add rotation (DA2 → DA3) and 2D gaussian blur (DA5 → DA6) to augment our data. In other words, rotation and 2D gaussian blur play a negative role in model training, whereas rotation has the most severe impact on model performance. This is because when we rotate the camera in-plane, the image content has changed but the corresponding ground-truth depth remains unchanged. Such a misalignment of affine transformation would confuse the model training, leading to worse performance. Besides, it is interesting to find that color jittering (DA3 → DA4) has the greatest impact on the generalization of the model, but does not bring significant improvement to the robustness. Motion blur (DA6 → DA7) and 3D defocus blur (DA7 → DA8) have the most significant improvement in model robustness because some of the test

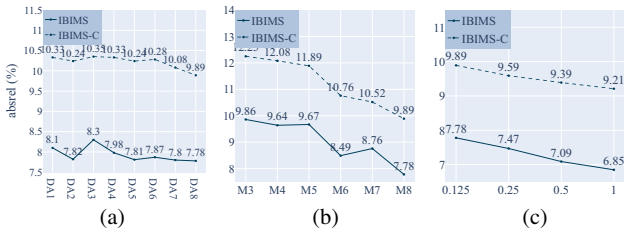


Fig. 9 Impact of data pre-processing on model accuracy and robustness. From left to right: **a** data augmentation schemes, **b** multi-dataset training, and **c** different percentages of each subset. We plot the absrel scores of the MDE models on IBIMS and IBIMS-C. The lower, the better

images are also corrupted by motion blur and 2D defocus blur. Since rotation and 2D gaussian blur play a negative part in robust depth estimation, we remove these two from our strong data augmentations and validate the model on IBIMS and IBIM-C. The absrel on IBIMS reduces from 7.78 to 6.90, and the absrel on IBIMS-C reduces from 9.89 to 9.24. Unless otherwise stated, our strong data augmentation method indicates the combinations of data augmentation schemes except for rotation and gaussian blur.

We show qualitative results under different image degradations in Fig. 10. One can find that the model would generate blurry results if the test image is corrupted by different kinds of noise, such as gaussian noise, shot noise, and impulse noise. Additionally, if the test image is corrupted by zoom blur, snow, and frost, the order of objects in the estimated depth map may be wrong. This is due to the geometric distortion of the corrupted images. To comprehensively understand

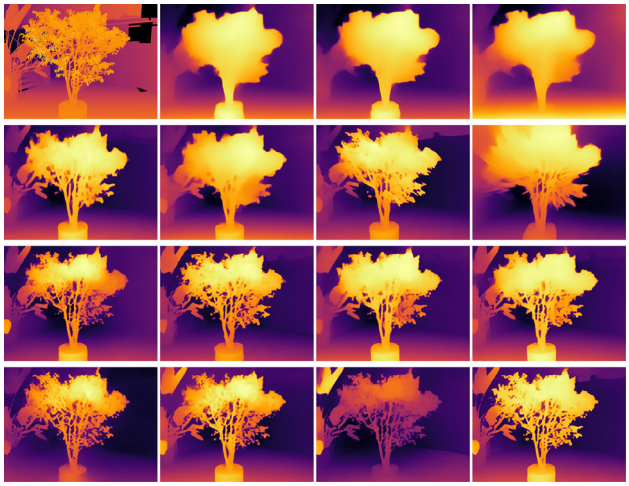


Fig. 10 The GT and our predictions under 15 corruption types (severity 3) from the IBIMS-C. From left-top to bottom-right: GT, gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, contrast, elastic transform, pixelate, jpeg compression, snow, fog, frost, and brightness

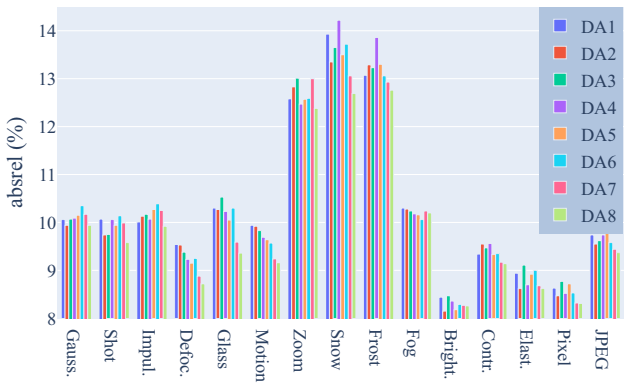


Fig. 11 Effect of different data augmentation combinations under various image corruptions on IBIMS-C. One can find that data augmentation indeed improves the robustness of MDE models, especially for category blur (defocus blur, frosted glass blur, and motion blur)

the importance of data augmentation in model robustness, we show the effect of different data augmentation combinations under various image corruptions on IBIMS-C in Fig. 11. Overall, our strong data augmentations can improve model robustness under various image corruptions. The improvements are significant, especially for category blur, such as defocus blur, frosted glass blur, and motion blur.

Aside from data augmentation, we investigate how data diversity and quantity affect the accuracy and robustness of depth estimation models. First, we train the ViT-hybrid-based models with our data augmentation schemes (DA8) on various dataset combinations. As shown in Table 2, M3 and M4 indicate that we use three and four datasets for training, respectively. It is worth noting that we sample every eight RGB-D pairs from each dataset in this experiment. As shown in Fig. 9b, the increase in data diversity can improve performance. If we train the model on the small training subset for a longer period of time (*i.e.*, 200 epochs), Ours-ViT-hybrid† can achieve competitive results with Ours-ViT-hybrid across five metric depth datasets (*cf.* Table 5) despite using only one-eighth of the full training data. *This suggests that our strong data augmentation schemes can compensate for the lack of data diversity and quantity to some extent.*

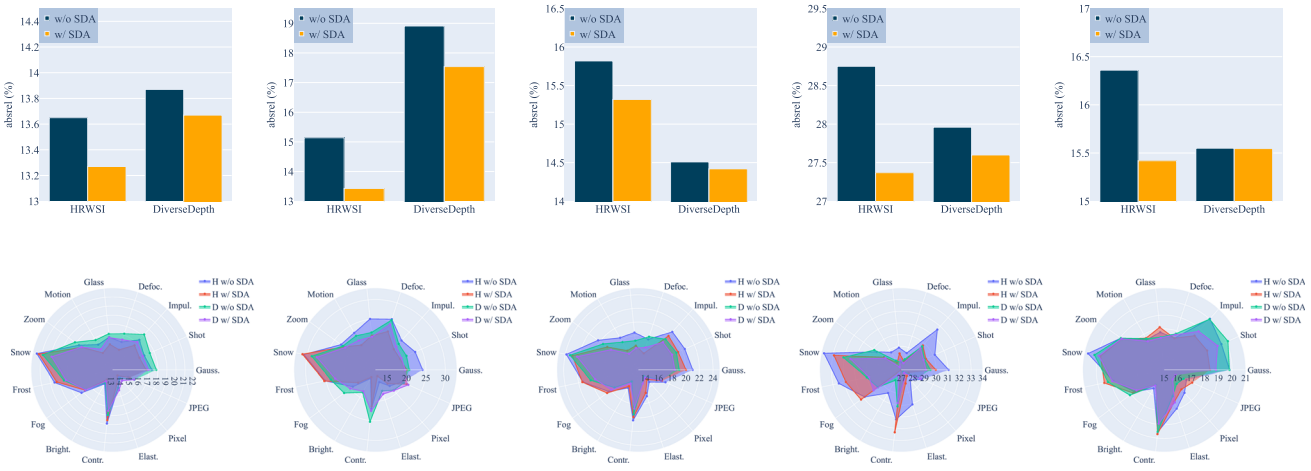
We further explore the question of whether and how data requirements can be reduced using augmentation. To answer this question, we train the ViT-hybrid-based models with our data augmentation schemes (DA8) on various percentage subsets of the full data, such as 12.5%, 25%, 50%, and 100%, to assess the model's performance. The results are shown in Fig. 9c. As the data continues to increase, the accuracy and robustness of the model improve. This implies that although data augmentation schemes can compensate for the lack of data diversity and quantity, the diversity and the amount of training data play significant roles in robust depth estimation. In this paper, we reveal that even if the amount of training data is not as large as other methods, on the basis of ensuring data quality and diversity, coupled with our proposed data augmentation schemes, the performance of the model is enough to match the state-of-the-art models. We believe that after our data and code are released, researchers can easily train state-of-the-art models under limited computing resources, which will promote the development of the community.

4.5 Does Strong data Augmentation also Benefit Other MDE Methods?

To verify the effectiveness of our strong data augmentation method, we apply it to two baseline methods (*i.e.*, HRWSI (Xian et al., 2020) and DiverseDepth (Yin et al., 2022)) with the same network architecture. We evaluate the models with and without strong data augmentation on five RGB-D datasets. We plot the results in Fig. 12. As one can see, adding strong data augmentation consistently improves the generalization and robustness of MDE models. It is worth noting that the fewer training data, the more significant the improvement of model performance by our strong data augmentation method. For example, performance improvements are more significant when adding strong data augmentation to HRWSI, which uses fewer training data than DiverseDepth. This justifies the effectiveness of our strong data augmentation method.

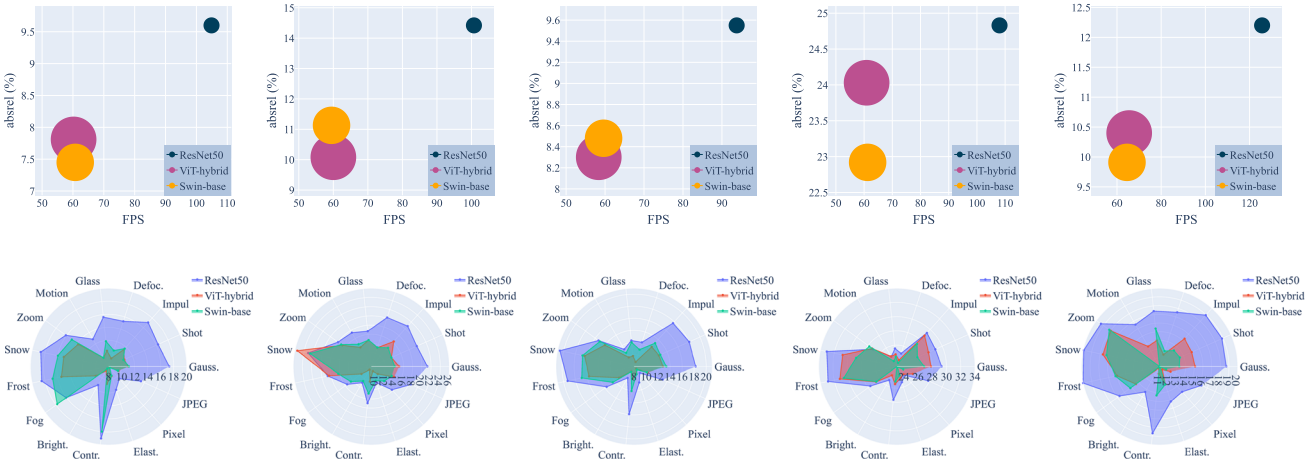
4.6 How Does Backbone Network Affect Generalization and Robustness?

We train various backbones, *e.g.*, ResNet50, ViT-hybrid, and Swin-base, on our full training set. Figure 13 shows the impact of the backbone network on model generalization and robustness. One can observe that the ResNet50-based model



(a) NYUDv2 (b) KITTI (c) ETH3D (d) DIODE (e) TUM

Fig. 12 Impact of strong data augmentation on other MDE methods. The top row shows the generalization of MDE models with and without strong data augmentation, while the bottom row depicts the robustness of MDE models with and without strong data augmentation



(a) NYUDv2 (b) KITTI (c) ETH3D (d) DIODE (e) TUM

Fig. 13 Impact of the backbone network on model generalization and robustness. The parameters of ResNet50-, ViT-hybrid-, and Swin-base-based models are 43.28M, 123.15M, and 101.08M. The top row illustrates the generalization capability of MDE models employing various backbones, while the bottom row demonstrates the robustness of MDE models using different backbones

runs the fastest but demonstrates the worst generalization and robustness. ViT-hybrid- and Swin-base-based models have similar inference speeds (around 60 FPS). The ViT-hybrid-based model shows better performance on the clean data of outdoor scenes, *e.g.*, KITTI and ETH3D. However, it performs slightly worse than the Swin-base-based one in indoor scenes. Regarding the robustness, it is interesting to find that (i) even though the ViT-hybrid-based model performs worse than the Swin-base-based one on the clean data of NYUDv2, the former performs better under various image corruptions except for the category shot noise and brightness. (ii) the Swin-base-based model is more robust under the texture-corrupting distortions, *e.g.*, snow, frost, gaussian noise, shot noise, and impulse noise.

5 Conclusion

We investigate data augmentation in robust monocular depth estimation in this paper. We have proposed a strong data augmentation method consisting of 2D data augmentations and our 3D-aware defocus blur. We thoroughly investigated the generalization and robustness of our models and 9 SOTA depth estimation methods. Experiment results show that the proposed model outperforms the prior arts regarding both

(a) All-in-focus RGB (b) Estimated depth (c) Shallow depth-of-field

Fig. 14 Limitations. When dealing with high-resolution images (*e.g.*, 1000×1600), our model sometimes generates inconsistent depth maps, such as missing small parts of the object. This would result in artifacts in downstream applications

generalization and robustness. We hope that our robust depth model will be useful to the vision community and that our robustness benchmark will encourage future research into the robustness of depth estimation in order to benefit real-world applications.

5.1 Limitations and Future Work

As shown in Fig. 14, one can find that our model sometimes generates inconsistent depth maps (*e.g.*, missing small parts of the object) when dealing with high-resolution images. This is because some fine-grained details would be lost if we rescale the images to a small size before feeding them to our model. Such a limitation would lead to artifacts when applied to downstream tasks like 3D photo generation and shallow depth-of-field rendering. We plan to construct a refinement module augmented by extra knowledge (*e.g.*, instance segmentation) to alleviate this issue.

Acknowledgements This work was in part supported by the National Key R&D Program of China (No. 2022ZD0118700), and partly supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20220-0007). This work was also supported under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). Z. Cao was supported by the National Natural Science Foundation of China (No. U1913602).

Data Availability The data that supports our findings are all publicly available online:1. HRWSI (Xian et al., 2020): https://kexianhust.github.io/Structure-Guided-Ranking-Loss/. 2. 3DKenBurns (Niklaus et al., 2019): https://github.com/sniklaus/3d-ken-burns. 3. DrivingStereo (Yang et al., 2019): https://drivingstereo-dataset.github.io/. 4. MegaDepth (Li & Snavely, 2018): https://www.cs.cornell.edu/projects/megadepth/. 5. TartanAir (Wang et al., 2020): https://theairlab.org/tartanair-dataset/. 6. Taskonomy (Zamir et al., 2018): http://taskonomy.stanford.edu/. 7. Hypersim (Roberts et al., 2021): https://github.com/apple/ml-hypersim. 8. IRS (Wang et al., 2019): https://github.com/HKBU-HPML/IRS. 9. IBIMS (Koch et al., 2018): https://www.asg.ed.tum.de/lmf/ibims1/. 10. NYUDv2 (Silberman et al., 2012): https://cs.nyu.edu/\protect\unhbox\voidb@x\penalty\@M\{}silberman/datasets/nyu_depth_v2.html. 11. KITTI (Uhrig et al., 2017): https://www.cvlibs.net/datasets/kitti/index.php. 12. ETH3D (Schöps et al., 2017): https://www.eth3d.net/datasets. 13. DIODE (Vasiljevic et al., 2019): https://diode-dataset.org/. 14. TUM (Sturm et al., 2012): https://cvg.cit.tum.de/data/datasets/rgbd-dataset. 15. OASIS (Chen et al., 2020): https://oasis.cs.princeton.edu/download.

References

Bian, J.-W., Zhan, H., Wang, N., Li, Z., Zhang, L., Shen, C., et al. (2021). Unsupervised scale-consistent depth learning from video. *IJCV, 129*(9), 2548–2564.

Chen, W., Fu, Z., Yang, D., & Deng, J. (2016). Single-image depth perception in the wild. In *NeurIPS*. (pp. 730–738).

Chen, W., Qian, S., & Deng, J. (2019). Learning single-image depth from videos using quality assessment networks. In *CVPR*. (pp. 5604–5613).

Chen, W., Qian, S., Fan, D., Kojima, N., Hamilton, M., & Deng, J. (2020). Oasis: A large-scale dataset for single image 3d in the wild. In *CVPR*. (pp. 679–688).

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *CVPR*. (pp. 113–123).

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint* arXiv:1708.04552.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16

words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929.
Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*. (pp. 2650–2658).
Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, volume 27. (pp. 1–9).
Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *CVPR*. (pp. 2002–2011).
Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *CVPR*. (pp. 270–279).
Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *CVPR*. (pp. 3828–3838).
Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint* arXiv:1903.12261.
Kamann, C., & Rother, C. (2021). Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *IJCV, 129*, 462–483.
Kar, O. F., Yeo, T., Atanov, A., & Zamir, A. (2022). 3d common corruptions and data augmentation. In *CVPR*. (pp. 18963–18974).
Koch, T., Liebel, L., Fraundorfer, F., & Körner, M. (2018). Evaluation of CNN-based single-image depth estimation methods. In *ECCVW*. (pp. 331–348).
Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *Proceeding of IEEE International Conference 3D Vision*. (pp. 239–248).
Lasinger, K., Ranftl, R., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI, 44*(3), 1623–1637.
Lee, H., & Park, J. (2022). Instance-wise occlusion and depth orders in natural scenes. In *CVPR*. (pp. 21210–21221).
Lee, S., Rameau, F., Im, S., & Kweon, I. S. (2022). Self-supervised monocular depth and motion learning in dynamic scenes: Semantic prior to rescue. *IJCV, 130*(9), 2265–2285.
Li, Z., Niklaus, S., Snavely, N., & Wang, O. (2021). Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*. (pp. 6498–6508).
Li, Z., & Snavely, N. (2018). Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*. (pp. 2041–2050).
Niklaus, S., Mai, L., Yang, J., & Liu, F. (2019). 3D Ken burns effect from a single image. *ACM TOG, 38*(6), 1841–18415.
Peng, J., Cao, Z., Luo, X., Lu, H., Xian, K., & Zhang, J. (2022). Bokehme: When neural rendering meets classical rendering. In *CVPR*. (pp. 16283–16292).
Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *ICCV*. (pp. 12179–12188).
Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., et al. (2021). Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*. (pp. 10912–10922).
Saleh, B. E., & Teich, M. C. (2019). *Fundamentals of photonics*. London: Wiley.
Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., et al. (2017). A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*. (pp. 3260–3269).
Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *ECCV*. (pp. 746–760).
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556.
Sturm, J., Engelhard, N., Endres, F., Burgard, W., & Cremers, D. (2012). A benchmark for the evaluation of rgb-d slam systems. In *IROS*. (pp. 573–580).
Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*. (pp. 402–419).
Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., & Geiger, A. (2017). Sparsity invariant CNNS. In *Proceeding of IEEE International Conference of 3D Vision*. (pp. 11–20).
Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F. Z., et al. (1908). *2019* (p. 00463). DIODE: A dense indoor and outdoor depth dataset. arxiv.
Wadhwa, N., Garg, R., Jacobs, D. E., Feldman, B. E., Kanazawa, N., Carroll, R., et al. (2018). Synthetic depth-of-field with a single-camera mobile phone. *ACM TOG, 37*(4), 1–13.
Wang, L., Shen, X., Zhang, J., Wang, O., Lin, Z., Hsieh, C.-Y., et al. (2018). Deeplens: Shallow depth of field from a single image. *ACM TOG, 37*(6), 1–11.
Wang, Q., Li, Z., Salesin, D., Snavely, N., Curless, B., & Kontkanen, J. (2022). 3d moments from near-duplicate photos. In *CVPR*. (pp. 3906–3915).
Wang, Q., Zheng, S., Yan, Q., Deng, F., Zhao, K., & Chu, X. (2019). Irs: A large synthetic indoor robotics stereo dataset for disparity and surface normal estimation. *arXiv preprint* arXiv:1912.09678, *6*.
Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., & Wang, C., et al. (2020). Tartanair: A dataset to push the limits of visual slam. In *IROS*. (pp. 4909–4916).
Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., & Li, R., et al. (2018). Monocular relative depth perception with web stereo data supervision. In *CVPR*. (pp. 311–320).
Xian, K., Zhang, J., Wang, O., Mai, L., Lin, Z., & Cao, Z. (2020). Structure-guided ranking loss for single image depth prediction. In *CVPR*. (pp. 611–620).
Xu, D., Ricci, E., Ouyang, W., Wang, X., & Sebe, N. (2017). Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*. (pp. 5354–5362).
Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., & Zhou, B. (2019). Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*. (pp. 899–908).
Yin, W., Liu, Y., & Shen, C. (2022). Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE TPAMI, 44*(10), 7282–7295.
Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., & Chen, S., et al. (2021). Learning to recover 3d scene shape from a single image. In *CVPR*. (pp. 204–213).
Yoon, J. S., Kim, K., Gallo, O., Park, H. S., & Kautz, J. (2020). Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*. (pp. 5336–5345).
Yuan, J., Liu, Y., Shen, C., Wang, Z., & Li, H. (2021). A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *ICCV*. (pp. 8229–8238).
Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*. (pp. 6023–6032).
Zamir, A. R., Sax, A., Shen, W. B., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *CVPR*. (pp. 3712–3722).
Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint* arXiv:1710.09412.
Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. *AAAI, 34*(07), 13001–13008.

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *CVPR*. (pp. 1851–1858).

Zini, S., Buzzelli, M., Twardowski, B., & van de Weijer, J. (2022). Planckian jitter: enhancing the color quality of self-supervised visual representations. *arXiv preprint* arXiv:2202.07993.