Zero-Shot Metric Depth with a Field-of-View Conditioned Diffusion Model

Saurabh Saxena[†], Junhwa Hur, Charles Herrmann, Deqing Sun, David J. Fleet

Google DeepMind

Abstract. Despite significant progress on domain and camera-specific models for monocular depth estimation, accurate *metric* depth estimation for images in the wild remains largely unsolved. Challenges include the joint modeling of indoor and outdoor scenes, which often exhibit significantly different distributions of RGB and depth, and the depth-scale ambiguity due to varying camera intrinsics. We propose a generic, task-agnostic diffusion model for monocular metric depth estimation, with several key advancements such as a log-scale depth parameterization to enable joint modeling of indoor and outdoor scenes and use of a diverse training data mixture with further synthetic augmentations to generalize beyond the limited camera intrinsics in training datasets. We show that conditioning on the field-of-view (FOV), instead of the much stronger entire intrinsics [23], is sufficient to handle scale ambiguity. Finally, we show that with an efficient parameterization of the reverse process, inference is remarkably fast, requiring just a few denoising iterations. Our method, dubbed DMD (Diffusion for Metric Depth), significantly outperforms recent methods on diverse indoor and outdoor zero-shot benchmarks.

1 Introduction

Monocular estimation of metric depth remains largely unsolved. Key challenges stem from (1) the large differences in RGB and depth distributions in indoor and outdoor data, and (2) depth scale ambiguity when one lacks knowledge of camera intrinsics. Not surprisingly, most recent works in monocular depth estimation train separate models for indoor and outdoor scenes, often on a single dataset with fixed camera intrinsics (e.g., an RGBD camera, or RGB+LIDAR for outdoor scenes). This avoids aforementioned challenges but at the cost of generality, i.e., overfitting to the camera intrinsics of the training data, and performing poorly on out-of-distribution data.

The predominant approach to jointly modeling indoor and outdoor data is to estimate scale- and shift-invariant depth, rather than metric depth (e.g., MiDaS [40]). Normalizing the depth distributions brings indoor and outdoor depth distributions closer and also avoids the problems of scale ambiguities in the presence of variable camera intrinsics. Recently there has been growing interest in bridging these different approaches, training joint indoor-outdoor models

[†] Correspondence to `srbs@google.com`

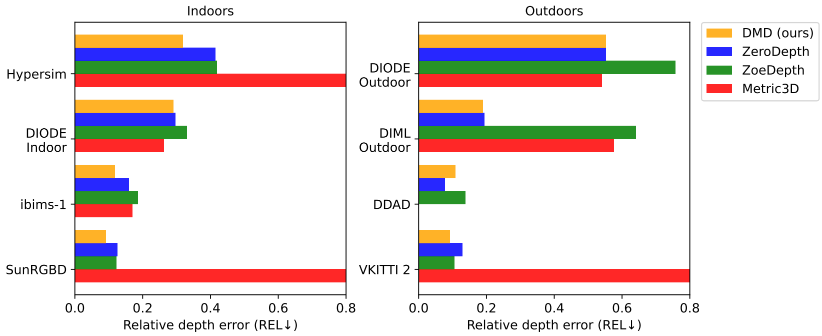


Fig. 1: Relative depth error for DMD (ours) compared to recent state-of-the-art joint indoor-outdoor models on zero-shot indoor and outdoor benchmarks. Overall DMD outperforms all baselines reducing relative error (REL) by 29% over ZoeDepth [5], by 9% over ZeroDepth [23] and by 49% over Metric3D [55]. Note that certain results for Metric3D are truncated for better visualization. Results for Metric3D on DDAD are omitted since DDAD is included in the training data for Metric3D.

that estimate metric depth. To cope with both indoor and outdoor domains, ZoeDepth [5] adds two heads to MiDaS [40], one for each domain, to convert from scale-invariant depth to metric depth. In order to handle diverse camera intrinsics, Metric3D [55] transforms images and depths to a canonical camera intrinsic. ZeroDepth [23] proposes conditioning on the entire intrinsics by building a dense ray map.

In this paper we introduce, for the first time, a diffusion model for zero-shot *metric* monocular depth estimation, outperforming previous approaches. To this end we leverage several key insights and innovations: *i)* representing depth in the log domain allocates model capacity in a more balanced way for indoor and outdoor scenes, greatly improving performance on regions with shallower depths; *ii)* Field-of-view (FOV) conditioning allows mixing datasets of varying intrinsics during training and helps resolve intrinsic scale ambiguities during inference; *iii)* FOV augmentation during training improves generalization to different camera intrinsics; *iv)* careful parameterization of the diffusion model dramatically reduces inference time; and *v)* using a diverse training mixture provides an additional boost in performance. The resulting model, dubbed DMD (Diffusion for Metric Depth), outperforms recently proposed indoor-outdoor monocular depth models [5], including those leveraging camera intrinsics [23, 55].

To summarize, we make the following contributions:

- We introduce DMD, a simple yet effective diffusion model for zero-shot metric depth estimation on general scenes, establishing a new SOTA.
- We identify key ingredients that greatly affect model performance, including log-scale depth, FOV conditioning, FOV augmentations, data mixtures, and denoising parameterization.
- For zero-shot metric depth estimation, DMD improves relative error (REL) by 29% over ZoeDepth [5], by 9% over ZeroDepth [23] and by 49% over Metric3D [55] using only a few denoising steps.

2 Related Work

Monocular depth (fine-tuned and evaluated in-domain). Given the challenges of learning a joint indoor-outdoor model, most approaches have restricted models to target a single dataset (either indoor or outdoor) with fixed intrinsics. In this setting, great progress has been made with advancements in specialized architectures [15, 16] such as the use of binning [1, 3, 8, 18, 34] or loss functions [15, 32] that are suited for this task. [2] proposed combining multiple training datasets with variable intrinsics by normalizing the images to the same camera intrinsic.

Joint indoor-outdoor models. To train joint indoor-outdoor models, one can mitigate the difficulty of learning diverse scene statistics by estimating scale- and shift-invariant depth instead. MiDaS [40] trains their model on diverse indoor-outdoor datasets and demonstrates good generalization to various unseen datasets. However, they do not provide metric depth. DPT [39] leverages this for pre-training and further fine-tunes separately for metric depth on NYU and KITTI. ZoeDepth [5] proposes adding a mixture-of-experts head, supervised by scene type, on top of a similarly pre-trained model, thereby handling indoor and outdoor scenes. In contrast, our model, DMD, uses a relatively generic framework, without domain-specific architectural components.

Intrinsics-conditioning for monocular depth. Incorporating camera intrinsics for depth estimation has been briefly explored in previous work [17, 24]. They argue that intrinsics-conditioning allows one to train on multiple datasets with varying intrinsics, but this is only demonstrated with small-scale experiments. Recently, ZeroDepth [23] introduces an intrinsic-conditioned metric-scale depth estimator via a variational latent representation and trains it on large-scale training datasets. However, using the full camera intrinsics limits the types of data augmentation. Metric3D [55] transforms images and depths to a canonical camera intrinsic, a hyperparameter that needs tuning.

In this paper, we condition on a weaker signal, *i.e.*, the input field of view (FOV), and introduce a novel FOV augmentation scheme that augments training data by cropping or uncropping to simulate diverse FOVs.

Diffusion for vision. Denoising diffusion models [25, 48] have recently emerged as a powerful class of generative models. Although initially proposed for natural image generation [13, 26, 37, 43], they have recently been shown to be effective for several computer vision tasks such as semantic segmentation [28], panoptic segmentation [9], optical flow [45] and monocular depth estimation [14, 28, 45]. Ours is the first demonstration that diffusion models can also support state-of-the-art zero-shot metric depth estimation for general indoor or outdoor scenes.

3 Diffusion for Metric Depth (DMD)

We cast monocular depth estimation as a generative RGB-to-depth translation task using probabilistic denoising diffusion models. To this end, we introduce several technical innovations to conventional diffusion models and training procedures to accommodate zero-shot, metric depth. In what follows we provide the

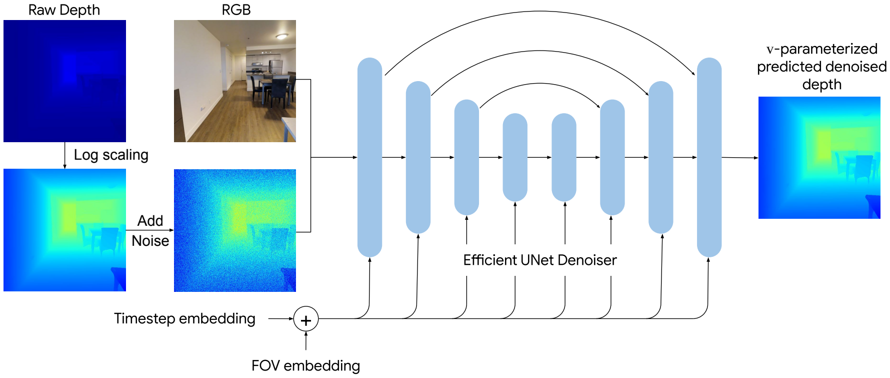


Fig. 2: The architecture of DMD consists of an Efficient UNet [43] backbone that denoises depth conditioned on an RGB image. Depth is parameterized in log-scale to equitably allocate representation capacity for shallow and deep depths. Each layer of the UNet is modulated by the timestep and FOV embeddings using FiLM [38] layers. FOV conditioning enables our models to predict the correct depth scales across diverse camera intrinsics. Furthermore, the denoising process is $\boldsymbol{v}$-parameterized which greatly reduces the number of sampling steps needed for inference.

necessary context and the design decisions that significantly impact the performance of DMD on zero-shot metric depth estimation for indoor and outdoor scenes.

3.1 Diffusion models

Diffusion models are probabilistic models that assume a forward process that gradually transforms a target distribution into a tractable noise distribution. A learned neural denoiser is trained to reverse this process, iteratively converting a noise sample to a sample from the target distribution. They have been shown to be remarkably effective with images and video, and they have recently begun to see use for dense vision tasks like segmentation, tracking, optical flow, and depth estimation. They are attractive as they exhibit strong performance on regression tasks, capturing posterior uncertainty, without task specific architectures, loss functions and training procedures.

For DMD we build on the task-agnostic Efficient U-Net architecture from DDVM [45]. While DDVM used the ϵ-parameterization for training the neural depth denoiser, here instead we use the $\boldsymbol{v}$-parameterization [44]. We find that the $\boldsymbol{v}$-parameterization yields remarkably efficient inference, using as few as one or two refinement steps, without the need for progressive distillation [44].

Under the $\boldsymbol{v}$-parameterization, the denoising network is given a noisy target image (in our case a depth map), $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, where $\mathbf{x}$ is the noiseless target input (depth map), $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, $t \sim \mathcal{U}(0,1)$, $\sigma_t^2 = 1 - \alpha_t^2$, and $\alpha_t > 0$ is computed with a pre-determined noise schedule, and the denoising network predicts $\boldsymbol{v} \equiv \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}$. From the output of the denoising network, i.e., $\boldsymbol{v}_\theta(\mathbf{z}_t, \boldsymbol{y}, t)$, where $\boldsymbol{y}$ is an optional conditioning signal (RGB image in this case), one obtains an estimate of $\mathbf{x}$ at step t, i.e., $\hat{\mathbf{x}}_t = \alpha_t \mathbf{z}_t - \sigma_t \boldsymbol{v}_\theta(\mathbf{z}_t, \boldsymbol{y}, t)$, and the corresponding estimate of the noise, denoted $\hat{\boldsymbol{\epsilon}}_t$. Under this parameterization, with a conventional L2 norm, the training objective is based on the expected 'truncated SNR weighting' loss, i.e., $\max(\|\mathbf{x} - \hat{\mathbf{x}}_t\|_2^2, \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_t\|_2^2)$ [44]. Motivated by the superior performance of the L1 loss in training DDVM [45] compared to the L2, we similarly employ a L1 loss for DMD as well, yielding the objective

$$\mathbb{E}_{\mathbf{x},\boldsymbol{y},t,\boldsymbol{\epsilon}} \left[\max(\|\mathbf{x} - \hat{\mathbf{x}}_t\|_1, \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_t\|_1) \right] \ . \tag{1}$$

3.2 Joint indoor-outdoor modeling

Training a joint indoor-outdoor model can be difficult because of the large differences in depth distributions one finds in indoor and outdoor scenes. Much of the available indoor training data have depths up to $10m$, while outdoor scenes include ground truth depths up to $80m$. Further, training data is often lacking the variation in camera intrinsics needed for robustness to images from different cameras. Rather, many datasets are captured with a fixed camera. To mitigate these issues we propose three innovations: the use of log depth; field of view augmentation; and field of view conditioning.

Log depth. Diffusion models assume data are normalized to $[-1, 1]$. While outdoor datasets comprise depths up to $80m$, depths in indoor scenes are usually less than $10m$. One might compress depth d linearly, with $d_{\text{lin}} = \text{normalize}(d/d_{\max})$, where $\text{normalize}(d) = \text{clip}(2 * d - 1, -1, 1)$. This, however, allocates little representation capacity to indoor scenes with a small depth range.

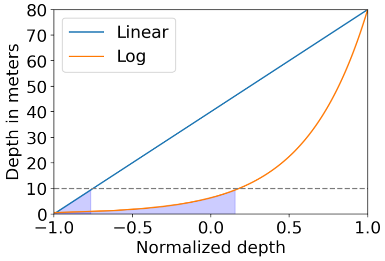
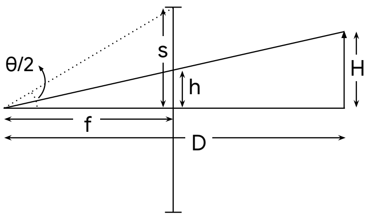


Fig. 3: Modelling depth in the log space allocates more representation capacity to regions with smaller depths, say <10 meters (shown by shaded region), which improves performance on such regions.

Instead, we compress depth on a log scale as the target for inference; i.e.,

$$d_{\log} = \text{normalize}\left(\frac{\log(d/d_{\min})}{\log(d_{\max}/d_{\min})}\right) \ , \tag{2}$$

where $d_{\min}$ and $d_{\max}$ denote the minimum and maximum supported depths (e.g., $0.5m$ and $80m$). This provides more representational capacity to small depths (Fig. 3). It also helps to account for non-stationary noise in depth data, where noise in depth maps increases with depth [30, 35]; this is at odds with the least-squares objective in diffusion model training that which assumes IID noise. But when the variance of depth errors increases with squared depth, log depth compression produces (approximately) additive IID noise. That is, if depth measurements have the form $d = \hat{d}\,(1 + \alpha\eta)$, where $\hat{d}$ is the true depth, $\eta \sim \mathcal{N}(0, 1)$ and α is a small positive constant, then the effective noise variance increases with $\hat{d}^2$. To first order, $\log d \approx \log \hat{d} + \alpha\eta$, in which case the noise variance becomes constant across the depth map, consistent with a simple squared loss. This yields a better balance between loss terms for pixels at different depths.



Field-of-view conditioning. Metric depth estimation from a single image is ill-posed when camera intrinsics are unknown. As depicted in the figure to the left, consider an object of height H at distance D, and a pinhole camera with focal length f, sensor half-height s and pixel size ρ. Then

$$D \;=\; H \, \frac{f}{h} \ . \tag{3}$$

where h/ρ is the object height in pixels. And if H was learned from training data (e.g., for common objects), and one were to condition on f/ρ (focal length in pixels), then this should be sufficient to infer D. However, while this conditioning signal depends on pixel size ρ, we would rather the conditioning signal be independent of the training image resolution (e.g., so that fine-tuning at larger resolutions does not require re-learning the intrinsic embeddings). Rewriting depth in quantities relative to the sensor half-height yields

$$D \;=\; H \, \frac{f/s}{h/s} \ , \tag{4}$$

where h/s is the size of the object relative to the image size, which can be inferred since the model has global context (via global self-attention layers). We thus use $s/f = \tan(\theta/2)$ as the conditioning signal where θ is the vertical angular FOV.

Compared with the full camera intrinsics, FOV is a weaker conditioning signal. This allows us to use simpler augmentations, as discussed below. Future work could investigate even weaker or ideally no conditional signals. We explored conditioning on the horizontal FOV as well, but that did not improve results substantially.

Field-of-view augmentation. Because datasets for depth estimation often have little or no variation in the field of view, it is easy for models to over-fit and thus generalize poorly to images with different camera intrinsics. To encourage models to generalize well to different fields of view, we propose to augment training data by cropping or uncropping to simulate diverse FOVs. Given a fixed focal length, this effectively changes the sensor size and thus the FOV. Given an image, one can straightforwardly simulate a smaller FOV via centered image crops. Simulating larger FOVs, e.g., via uncrop, is not as straightforward. Our preliminary experiments used generative uncropping for RGB images with Palette [42]; however, we found that padding the RGB image with Gaussian noise (mean-zero, variance 1) works as well, is simpler, and more efficient.

For handling missing ground truth depth with uncropping augmentation, we adopt the approach in [45], using a combination of near-neighbor in-filling and step-unrolled denoising during training. It is shown in [45] that this technique is effective in coping with the inherent distribution shift between training and testing when ground truth data are noisy or incomplete. This allows our models to support even larger FOV than the training datasets without requiring any sophisticated generative uncropping techniques. Note that while FOV augmentation does help simulate some variation in camera intrinsics, variations in other factors, like focal length, are much harder to simulate well.

4 Experiments

4.1 Training data

DDVM [45] shows that using large amounts of diverse training data is important for generic models. We follow the pre-training strategy proposed in [45]: initializing the model with unsupervised pre-training on ImageNet [12] and Places365 [58], with tasks proposed in [42], followed by supervised pre-training on ScanNet [11], SceneNet-RGBD [36], and Waymo [50]. Unlike [45], we also include the DIML Indoor [10] dataset for more diversity. No FOV augmentation or conditioning is used at this pre-training stage.

For the final training stage we train on a mixture of NYU [47], Taskonomy [57], KITTI [19] and nuScenes [7]. At this stage we apply FOV augmentation to NYU, KITTI and nuScenes, but not Taskonomy as it is large and has substantial FOV diversity.

4.2 Design choices

Denoiser Architecture. We adopt the modifications of the *Efficient U-Net* [43] proposed in DDVM [45], with one further modification to support FOV conditioning. The FOV embedding, like the timestep embedding, is constructed by first building a sin-cos positional embedding [52] followed by linear projection. The sum of these two embeddings is used to modulate different layers of the denoiser backbone using FiLM [38] layers. The predicted depth maps are resized to the ground-truth resolution for evaluation, following prior work [5]. Other training hyper-parameters such as the batch size and optimizer details are like those in [45].

Augmentations. In addition to the FOV augmentation (Sec. 3.2) we use random horizontal flip augmentation, like many prior works.

Training details. We perform supervised depth pre-training for a total of 1.5M steps. We train at a low resolution of 240×320 for the first 1.4M steps and then finetune at 384×512 for 100k steps for compute efficiency. The models are then trained for 50k steps for the final supervised training stage.

Sampler. We use the DDPM [25] sampler with eight denoising steps for indoor datasets. For outdoor datasets we find that two denoising steps suffice. We report results using a mean of eight samples, following [45].

Evaluation. We adopt the evaluation protocol of ZoeDepth [5]. We report in-distribution performance on the NYU [47] and KITTI [19] datasets, and generalization performance on eight unseen datasets [5], namely, SunRGBD [49], iBims-1 [31], DIODE Indoor [51], Hypersim [41] for indoors, and Virtual KITTI

2 [6], DDAD [20], DIML Outdoor [10], DIODE Outdoor [51] for outdoors. We closely follow the evaluation protocol, including depth ranges and cropping, used in [5] and report results using the standard error and accuracy metrics that are used in literature.

Table 1: Zero-shot results on unseen indoor datasets. **Best** and second-best results amongst joint indoor-outdoor models are highlighted. Performance of domain-specific models trained on NYU are also provided for reference. DMD outperforms all baselines on all datasets except DIODE Indoor where it outperforms all baselines except Metric3D. Note, however, that overall DMD substantially outperforms Metric3D.

Method	**SUN RGB-D**		**iBims-1**		**DIODE Indoor**		**HyperSim**	
	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓
Domain-specific models:								
BTS [33]	0.172	0.515	0.231	0.919	0.418	1.905	0.476	6.404
AdaBins [3]	0.159	0.476	0.212	0.901	0.443	1.963	0.483	6.546
LocalBins [4]	0.156	0.470	0.211	0.880	0.412	1.853	0.468	6.362
NeWCRFs [56]	0.151	0.424	0.206	0.861	0.404	1.867	0.442	6.017
DDVM [45]	0.123	0.350	0.169	0.719	0.339	1.557	0.363	2.175
Joint indoor-outdoor models:								
ZoeDepth [5]	0.123	0.356	0.186	0.777	0.331	1.598	0.419	5.830
Metric3D [55]	1.457	3.043	0.169	0.535	**0.263**	**1.087**	1.082	8.199
ZeroDepth [23]	0.126	0.372	0.159	0.630	0.297	1.435	0.415	5.978
DMD	**0.091**	**0.275**	**0.118**	**0.447**	0.291	1.292	**0.318**	**4.394**

Table 2: Zero-shot results on four unseen outdoor datasets. **Best** and second-best results amongst joint indoor-outdoor models are highlighted. Performance of domain-specific models trained on KITTI are also provided for reference. Metric3D's DDAD result is omitted because their training datasets include DDAD. DMD outperforms baselines on the Virtual KITTI 2 and DIML Outdoor datasets. On DIODE Outdoor DMD outperforms ZoeDepth and performs competitively with ZeroDepth and Metric3D on relative error. On DDAD, ZeroDepth significantly outperforms which may be attributable to their use of a more diverse outdoor training mixture including Parallel Domain [21, 22], TartanAir [53], and proprietary Large-Scale Driving (LSD) data.

	VKITTI 2		**DDAD**		**DIML Outdoor**		**DIODE Outdoor**	
Method	REL↓	RMSE ↓	REL↓	RMSE ↓	REL↓	RMSE ↓	REL↓	RMSE ↓
Domain-specific models:								
BTS [33]	0.115	5.368	0.147	7.550	1.785	5.908	0.837	10.48
AdaBins [3]	0.122	5.420	0.154	8.560	1.941	6.272	0.863	10.35
LocalBins [4]	0.127	5.981	0.151	8.139	1.820	6.706	0.821	10.27
NeWCRFs [56]	0.117	5.691	0.119	6.183	1.918	6.283	0.854	9.228
DDVM [45]	0.098	4.963	0.126	7.083	2.000	7.302	0.660	8.766
Joint indoor-outdoor models:								
ZoeDepth [5]	0.105	5.095	0.138	7.225	0.641	3.610	0.757	7.569
Metric3D [55]	1.437	24.316	-	-	0.576	3.066	**0.540**	**7.300**
ZeroDepth [23]	0.129	5.668	**0.077**	**5.168**	0.194	2.117	0.552	8.943
DMD	**0.092**	**4.387**	0.108	5.365	**0.190**	**2.089**	0.553	8.943

4.3 Results

Zero-shot. Tables 1 and 2 report *zero-shot* performance on eight OOD datasets. In the indoors setting, DMD outperforms all baselines with the single exception of Metric3D on DIODE Indoors; however, note that Metric3D underperforms on

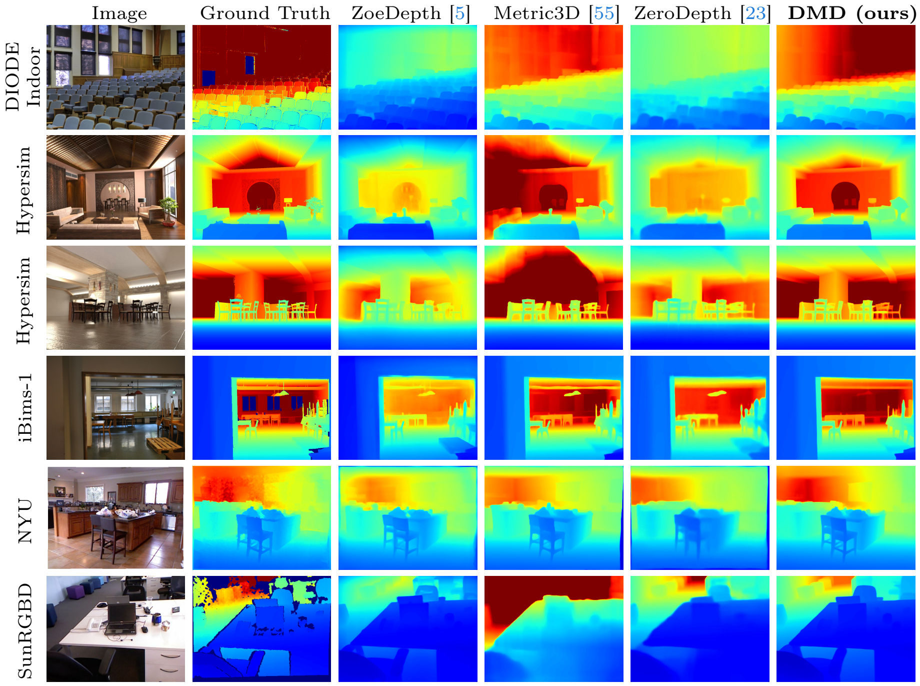


Fig. 4: Qualitative comparison between ZoeDepth [5], Metric3D [55], ZeroDepth [23], and ours on indoor scenes. Compared to other methods, our method estimates depths at more accurate scale over diverse datasets.

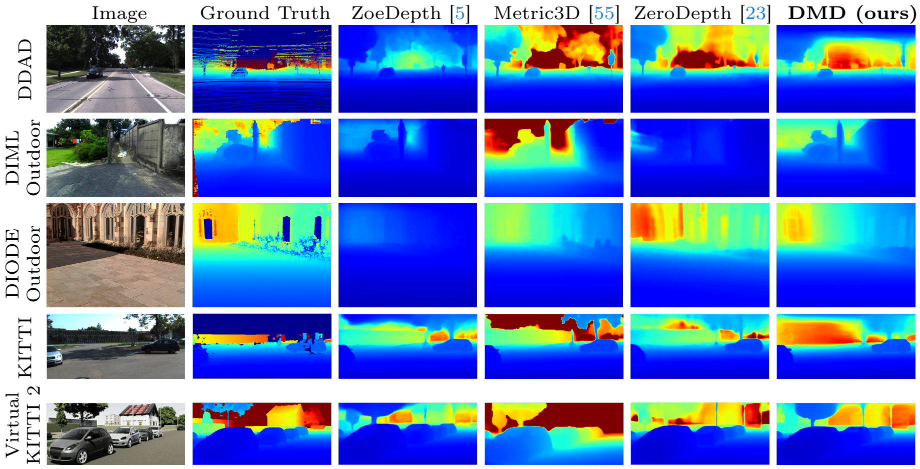
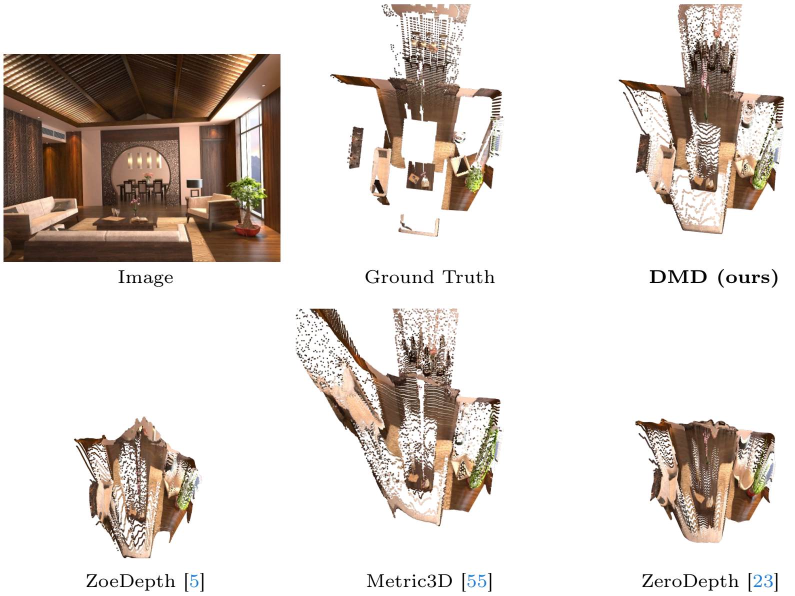


Fig. 5: Qualitative comparison between ZoeDepth [5], Metric3D [55], and ZeroDepth [23], and ours on outdoor scenes. Compared with other methods, our method is able to estimate a more accurate depth scale.



Image

Ground Truth

DMD (ours)

ZoeDepth [5]

Metric3D [55]

ZeroDepth [23]

Fig. 6: Qualitative comparison between ZoeDepth [5], Metric3D [55], and ZeroDepth [23], and ours by visualizing 3D point clouds obtained using the predicted depth map for a scene in the Hypersim dataset (the roof is omitted from the point clouds for better visualization). All renders use the same camera intrinsics and pose. ZoeDepth and ZeroDepth fail to recover depth for the dining area at the far end of the room and have distorted (wavy) walls. Metric3D fails to recover correct depth for the left half of the scene. Ours is the most faithful to the ground truth.

all other datasets. In the outdoors setting, DMD outperforms the baselines on the majority of datasets. Figures 4 and 5 compare depth maps from DMD against other baselines on indoor and outdoor datasets respectively. Fig. 6 further visualizes a point cloud for a scene in the Hypersim dataset illustrating the ability of DMD to recover better depth scale and overall detail compared to baselines.

In-distribution. On NYU, DMD outperforms both ZoeDepth and ZeroDepth on relative error. On KITTI, DMD outperforms ZoeDepth and is competitive with ZeroDepth on relative error. See Tab. 3 for detailed results.

4.4 Ablations

We next consider several ablations to test different components of the model. All models reported in the ablations were only fine-tuned on NYU and KITTI for expedience unless otherwise specified.

Log vs linearly scaled depth. Table 4 shows that parameterizing depth in log scale (Sec. 3.2) improves quantitative performance. As expected, this is beneficial for datasets of indoor scenes and also for datasets of outdoor scenes with shallower depths, like DIML Outdoor and DIODE Outdoor.

Field-of-view conditioning. Table 5 shows that FOV conditioning achieves the best performance. Fig. 7 perturbs the conditioning FOV signal during in-

Table 3: In-domain results showing that DMD outperforms both ZoeDepth and ZeroDepth on NYU for relative error (top) and is competitive with ZeroDepth on KITTI for relative error (bottom). **Best** results (amongst indoor-outdoor models only) are bolded. For reference, we include baselines trained separately for the indoor and outdoor domains showing that DMD is competitive despite being a more general model.

Method	**NYU**					
	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	REL $\downarrow$	RMS $\downarrow$	$\log_{10}\downarrow$
Domain-specific models:						
BTS [33]	0.885	0.978	0.994	0.110	0.392	0.047
DPT [39]	0.904	0.988	0.998	0.110	0.357	0.045
AdaBins [3]	0.903	0.984	0.997	0.103	0.364	0.044
NeWCRFs [56]	0.922	0.992	0.998	0.095	0.334	0.041
BinsFormer [34]	0.925	0.989	0.997	0.094	0.330	0.040
PixelFormer [1]	0.929	0.991	0.998	0.090	0.322	0.039
IEBins [46]	0.936	0.992	0.998	0.087	0.314	0.038
MIM [54]	0.949	0.994	0.999	0.083	0.287	0.035
DDVM [45]	0.946	0.987	0.996	0.074	0.315	0.032
Joint indoor-outdoor models:						
ZoeDepth [5]	0.953	**0.995**	0.999	0.077	0.277	0.033
ZeroDepth [23]	**0.954**	**0.995**	**1.000**	0.074	**0.269**	-
DMD (ours)	0.953	0.989	0.996	**0.072**	0.296	**0.031**

Method	**KITTI**						
	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	REL $\downarrow$	Sq-rel $\downarrow$	RMS $\downarrow$	RMS log $\downarrow$
Domain-specific models:							
BTS [33]	0.956	0.993	0.998	0.059	0.245	2.756	0.096
DPT [39]	0.959	0.995	0.999	0.062	–	2.573	0.092
AdaBins [3]	0.964	0.995	0.999	0.058	0.190	2.360	0.088
NeWCRFs [56]	0.974	0.997	0.999	0.052	0.155	2.129	0.079
BinsFormer [34]	0.974	0.997	0.999	0.052	0.151	2.098	0.079
PixelFormer [1]	0.976	0.997	0.999	0.051	0.149	2.081	0.077
IEBins [46]	0.978	0.998	0.999	0.050	0.142	2.011	0.075
MIM [54]	0.977	0.998	1.000	0.050	0.139	1.966	0.075
DDVM [45]	0.965	0.994	0.998	0.055	0.292	2.613	0.089
Joint indoor-outdoor models:							
ZoeDepth [5]	0.966	0.993	0.996	0.057	0.204	2.362	0.087
ZeroDepth [23]	**0.968**	**0.995**	**0.999**	**0.053**	**0.164**	**2.087**	**0.083**
DMD (ours)	0.967	**0.995**	**0.999**	**0.053**	0.203	2.411	0.084

ference, showing that optimal performance occurs at or close to the true FOV.

No FOV augmentation or conditioning. ZoeDepth found that without scene-type supervision for the experts (i.e., *Auto Router*), ZoeDepth's performance degrades, even for in-domain data. To compare against ZoeDepth in this setting, we fine-tune a model on NYU and KITTI without FOV augmentations or conditioning. Interestingly, DMD performs relatively well in this setting for in-domain data (Table 7). Nevertheless, as shown in Table 6, OOD performance is better with FOV augmentation and conditioning.

$\boldsymbol{\epsilon}$ **vs** $\boldsymbol{v}$ **diffusion parameterization.** Inference latency is a concern with diffusion models for vision. DDVM [45], for example, uses 128 denoising steps for depth estimation which can be prohibitive. We find that using the $\boldsymbol{v}$ parameterization dramatically reduces the number of denoising steps required for good performance. As shown in Table 8, $\boldsymbol{\epsilon}$-parameterization requires 64 denoising steps to match the performance of a model with $\boldsymbol{v}$-parameterization using only 1 de-

Table 4: Ablation showing that log depth improves quantitative performance on indoor datasets (top) since log-scaling increases the share of representation capacity allocated to shallow depths. On outdoor datasets (bottom) the performance remains mostly unchanged except for DIML Outdoor which has outdoor scenes with shallow depths and benefits from log-parameterized depth.

	Experiment	NYU	SunRGBD	DIODE Indoor	iBims-1	Hypersim
REL	Linear scaling	0.082	**0.108**	0.324	0.146	0.398
	Log scaling	**0.076**	0.109	**0.298**	**0.130**	**0.382**
RMS	Linear scaling	0.340	0.314	1.526	0.612	5.693
	Log scaling	**0.313**	**0.306**	**1.407**	**0.563**	**5.527**
	Experiment	KITTI	DIML Outdoor	DIODE Outdoor	Virtual KITTI 2	DDAD
REL	Linear scaling	0.056	0.467	0.630	**0.092**	**0.122**
	Log scaling	**0.055**	**0.300**	**0.628**	0.093	**0.122**
RMS	Linear scaling	**2.516**	3.126	10.129	**4.788**	**6.288**
	Log scaling	2.527	**2.522**	**9.577**	4.828	6.740

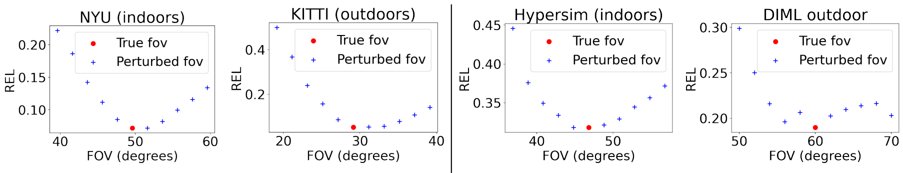
Table 5: Depth errors for models trained with and without field-of-view conditioning on zero-shot indoors (top) and outdoors (bottom) datasets. Results shows that FOV conditioning provides a substantial boost in performance. DIML Outdoor benefits the most, which is understandable given its large FOV, for which generalization is a major challenge for simple FOV augmentation.

Metric	Experiment	NYU	SunRGBD	DIODE Indoor	iBims-1	Hypersim
REL	No FOV cond	0.081	0.116	0.316	0.18	0.400
	With FOV cond	**0.076**	**0.109**	**0.298**	**0.130**	**0.382**
RMS	No FOV cond	0.319	0.325	1.474	0.712	**5.196**
	With FOV cond	**0.313**	**0.306**	**1.407**	**0.563**	5.527
Metric	Experiment	KITTI	DIML Outdoor	DIODE Outdoor	VKITTI 2	DDAD
REL	No FOV cond	0.057	1.257	**0.613**	0.100	**0.121**
	With FOV cond	**0.055**	**0.300**	0.628	**0.093**	0.122
RMS	No FOV cond	2.574	5.382	**8.582**	5.021	6.826
	With FOV cond	**2.527**	**2.522**	9.577	**4.828**	**6.740**

noising step. Intuitively, $\boldsymbol{v}$-parameterization ensures that the model accurately recovers the signal at both ends of the noise schedule, unlike $\boldsymbol{\epsilon}$-parameterization where estimating the noise is easy for low SNR inputs.

Training data mixture. We compare between DMD-NK which is fine-tuned on NYU and KITTI, and DMD which adds Taskonomy and NuScenes to the fine-tuning mix. As shown in Table 9 and Fig. 8, these additional datasets significantly improve performance; DMD significantly improves the depth scale and fine-grained depth details near object boundaries.

Limitation. While RGB camera intrinsics are available for most practical uses of monocular depth estimators (e.g. cell phones, robot platforms or self-driving cars), they may sometimes be unknown (e.g. internet images or generative imagery). One solution to handle the unknown FOV would be to estimate the camera intrinsics from the RGB image. To test this, we train a simple neural



In-distribution

Zero-shot

Fig. 7: Plots showing the effect of perturbing the FOV during inference. Optimal performance is at or near the true FOV. Performance degrades with larger perturbation.

Table 6: Ablation showing that training without FOV augmentations and conditioning hurts generalization to out-of-domain indoor (top) and outdoor (bottom) datasets due to overfitting on the training data intrinsics.

Metric	Experiment	NYU	SunRGBD	DIODE Indoor	iBims-1	Hypersim
REL	No FOV aug or cond	**0.074**	0.124	0.337	0.180	0.479
	W/ FOV aug and cond	0.076	**0.109**	**0.298**	**0.130**	**0.382**
RMS	No FOV aug or cond	**0.310**	0.348	1.535	0.722	**5.247**
	W/ FOV aug and cond	0.313	**0.306**	**1.407**	**0.563**	5.527

Metric	Experiment	KITTI	DIML Outdoor	DIODE Outdoor	VKITTI 2	DDAD
REL	No FOV aug or cond	**0.055**	1.399	**0.615**	0.095	**0.116**
	W/ FOV aug and cond	**0.055**	**0.300**	0.628	**0.093**	0.122
RMS	No FOV aug or cond	2.597	5.919	**8.529**	4.874	**6.476**
	W/ FOV aug and cond	**2.527**	**2.522**	9.577	**4.828**	6.740

Table 7: Comparison against ZoeDepth without scene type supervision. ZoeDepth performance degrades significantly when the scene type (indoor or outdoor) is not provided. DMD learns well without such supervision.

	REL ↓		**RMSE ↓**	
	NYU	KITTI	NYU	KITTI
ZoeDepth w/o Auto Router	0.102	0.075	0.377	**2.584**
Ours w/o fov aug / cond	**0.074**	**0.055**	**0.310**	2.597

Table 8: Reduced inference latency: On the left, we report relative error on NYU and KITTI for DMD-NK (without FOV augmentation or conditioning). $\boldsymbol{v}$-parameterization achieves optimal performance with as few as 4 denoising steps whereas $\boldsymbol{\epsilon}$-parameterization requires 64 steps to reach the same performance. This dramatically reduces inference latency (right) with DMD being 6× faster than DDVM [45], despite the higher resolution, and 14× faster than the $\boldsymbol{\epsilon}$-parameterized version.

	NYU		KITTI	
Steps	$\boldsymbol{\epsilon}$	$\boldsymbol{v}$	$\boldsymbol{\epsilon}$	$\boldsymbol{v}$
1	2.374	0.077	0.596	0.056
4	1.484	0.075	0.406	0.055
16	0.409	0.074	0.141	0.055
64	0.077	0.074	0.056	0.055

Method	Steps	Resolution	Time [ms]
DDVM	64	240 × 320	248
DMD ($\boldsymbol{\epsilon}$)	64	384 × 512	543
DMD ($\boldsymbol{v}$)	4	384 × 512	38

network to regress to the FOV on a mixture of NYU and KITTI using the same FOV augmentations as DMD-NK. The results are promising, performing competitively with those using the true FOV (see appendix for details). We hy-

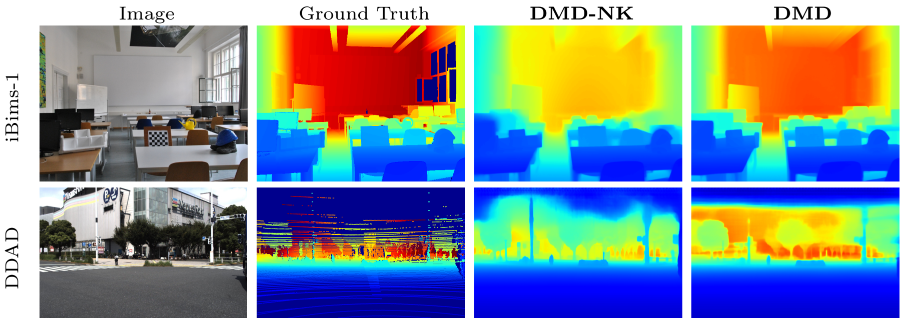


Fig. 8: Qualitative comparison between DMD-NK (fine-tuned on NYU and KITTI) and DMD (fine-tuned on KITTI, NYU, nuScenes, and Taskonomy). DMD further improves depth scale estimation as well as fine details on depth boundaries.

Table 9: Data mixture ablation: We compare DMD (trained on NYU, KITTI, Taskonomy and NuScenes) against DMD-NK (trained on NYU and KITTI alone). Increasing the diversity of training mixture significantly improves performance.

Metric	Experiment	NYU	SunRGBD	DIODE Indoor	ibims-1	Hypersim
REL	DMD-NK	0.076	0.109	0.298	0.130	0.382
	DMD	**0.072**	**0.091**	**0.291**	**0.118**	**0.318**
RMS	DMD-NK	0.313	0.306	1.407	0.563	5.527
	DMD	**0.296**	**0.275**	**1.292**	**0.447**	**4.394**

Metric	Experiment	KITTI	DIML Outdoor	DIODE Outdoor	VKITTI 2	DDAD
REL	DMD-NK	0.055	0.300	0.628	0.093	0.122
	DMD	**0.053**	**0.190**	**0.553**	**0.092**	**0.108**
RMS	DMD-NK	2.527	2.522	9.577	4.828	6.740
	DMD	**2.411**	**2.089**	**8.943**	**4.387**	**5.365**

pothesize that field-of-view estimates can be further improved using a better camera intrinsics estimators [27,29], thereby further improving depth estimates.

5 Conclusion

We propose a generic diffusion-based monocular metric depth generator with no specialized architectures and minimal task-specific inductive biases for handling diverse indoor and outdoor scenes. Our log-scale depth parameterization adequately allocates representation capacity to different depth ranges. We advocate augmenting the FOV of training data through simple cropping/uncropping to enable generalization to fields-of-view beyond those in the training datasets and show that simply uncropping with noise padding is effective for simulating a larger FOV. We find that conditioning on the FOV is essential for disambiguating depth-scale. We further propose a new fine-tuning dataset mixture that dramatically improves performance. With these innovations combined, we establish a new state of the art outperforming the existing methods across diverse zero-shot and in-domain datasets by a substantial margin.

References

1. Agarwal, A., Arora, C.: Attention Attention Everywhere: Monocular depth prediction with skip attention. In: WACV (2023) 3, 11
2. Antequera, M.L., Gargallo, P., Hofinger, M., Bulò, S.R., Kuang, Y., Kontschieder, P.: Mapillary planet-scale depth dataset. In: ECCV. pp. 589–604 (2020) 3
3. Bhat, S.F., Alhashim, I., Wonka, P.: AdaBins: Depth estimation using adaptive bins. In: CVPR. pp. 4009–4018 (2021) 3, 8, 11
4. Bhat, S.F., Alhashim, I., Wonka, P.: LocalBins: Improving depth estimation by learning local distributions. In: ECCV. pp. 480–496 (2022) 8
5. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023) 2, 3, 7, 8, 9, 10, 11
6. Cabon, Y., Murray, N., Humenberger, M.: Virtual KITTI 2 (2020) 8
7. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 7
8. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE T-CSVT **28**(11), 3174–3182 (2017) 3
9. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. In: ICCV (2023) 3
10. Cho, J., Min, D., Kim, Y., Sohn, K.: DIML/CVL RGB-D dataset: 2M RGB-D images of natural indoor and outdoor scenes. arXiv preprint arXiv:2110.11590 (2021) 7, 8
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017) 7
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 7
13. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS (2022) 3
14. Duan, Y., Guo, X., Zhu, Z.: DiffusionDepth: Diffusion denoising approach for monocular depth estimation. arXiv preprint arXiv:2303.05021 (2023) 3
15. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV. pp. 2650–2658 (2015) 3
16. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS. vol. 27 (2014) 3
17. Facil, J.M., Ummenhofer, B., Zhou, H., Montesano, L., Brox, T., Civera, J.: CAM-Convs: Camera-Aware Multi-Scale Convolutions for Single-View Depth. In: CVPR (2019) 3
18. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR. pp. 2002–2011 (2018) 3
19. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets Robotics: The KITTI dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 7
20. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3D packing for self-supervised monocular depth estimation. In: CVPR (2020) 8
21. Guizilini, V., Lee, K.H., Ambrus, R., Gaidon, A.: Learning optical flow, depth, and scene flow without real-world labels. IEEE Robotics and Automation Letters (2022) 8

22. Guizilini, V., Li, J., Ambrus, R., Gaidon, A.: Geometric unsupervised domain adaptation for semantic segmentation. In: ICCV (2021) 8
23. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: ICCV. pp. 9233–9243 (2023) 1, 2, 3, 8, 9, 10, 11
24. He, L., Wang, G., Hu, Z.: Learning depth from single images with deep neural network embedding focal length. IEEE Transactions on Image Processing **27**(9), 4676–4689 (2018) 3
25. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. NeurIPS (2020) 3, 7
26. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022) 3
27. Hold-Geoffroy, Y., Sunkavalli, K., Eisenmann, J., Fisher, M., Gambaretto, E., Hadap, S., Lalonde, J.F.: A perceptual measure for deep single image camera calibration. In: CVPR (2018) 14
28. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: DDP: Diffusion model for dense visual prediction. In: ICCV (2023) 3
29. Jin, L., Zhang, J., Hold-Geoffroy, Y., Wang, O., Blackburn-Matzen, K., Sticha, M., Fouhey, D.F.: Perspective fields for single image camera calibration. In: CVPR (2023) 14
30. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors **12**(2), 1437–1454 (2012) 5
31. Koch, T., Liebel, L., Körner, M., Fraundorfer, F.: Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. Computer Vision and Image Understanding **191**, 102877 (2020) 7
32. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3DV. pp. 239–248 (2016) 3
33. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv:1907.10326 (2019) 8, 11
34. Li, Z., Wang, X., Liu, X., Jiang, J.: BinsFormer: Revisiting adaptive bins for monocular depth estimation. arxiv.2204.00987 (2022) 3, 11
35. Mallick, T., Das, P.P., Majumdar, A.K.: Characterizations of noise in kinect depth images: A review. IEEE Sensors journal **14**(6), 1731–1740 (2014) 5
36. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: SceneNet RGB-D: Can 5M synthetic images beat generic imagenet pre-training on indoor segmentation? In: ICCV (2017) 7
37. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021) 3
38. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer . In: AAAI (2018) 4, 7
39. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021) 3, 11
40. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI **44**(3), 1623–1637 (2020) 1, 2, 3
41. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021) 7

42. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-Image Diffusion Models. In: SIGGRAPH (2022) 6, 7
43. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: NeurIPS (2022) 3, 4, 7
44. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR (2022) 4, 5
45. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In: NeurIPS (2023) 3, 4, 5, 6, 7, 8, 11, 13
46. Shao, S., Pei, Z., Wu, X., Liu, Z., Chen, W., Li, Z.: IEBins: Iterative elastic bins for monocular depth estimation. In: NeurIPS (2023) 11
47. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV. pp. 746–760 (2012) 7
48. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265 (2015) 3
49. Song, S., Lichtenberg, S.P., Xiao, J.: Sun RGB-D: A RGB-D scene understanding benchmark suite. In: CVPR. pp. 567–576 (2015) 7
50. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020) 7
51. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. arXiv preprint arXiv:1908.00463 (2019) 7, 8
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS (2017) 7
53. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: TartanAir: A dataset to push the limits of visual SLAM. In: IROS. pp. 4909–4916. IEEE (2020) 8
54. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. In: CVPR (2023) 11
55. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3D: Towards zero-shot metric 3D prediction from a single image. In: ICCV. pp. 9043–9053 (2023) 2, 3, 8, 9, 10
56. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected CRFs for monocular depth estimation. In: CVPR. pp. 3916–3925 (2022) 8, 11
57. Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018) 7
58. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 7