Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

Bingxin Ke Anton Obukhov Shengyu Huang
Nando Metzger Rodrigo Caye Daudt Konrad Schindler
Photogrammetry and Remote Sensing, ETH Zürich

Figure 1. **We present Marigold, a diffusion model and associated fine-tuning protocol for monocular depth estimation.** Its core principle is to leverage the rich visual knowledge stored in modern generative image models. Our model, derived from Stable Diffusion and fine-tuned with synthetic data, can zero-shot transfer to unseen datasets, offering state-of-the-art monocular depth estimation results.

Abstract

Monocular depth estimation is a fundamental computer vision task. Recovering 3D depth from a single image is geometrically ill-posed and requires scene understanding, so it is not surprising that the rise of deep learning has led to a breakthrough. The impressive progress of monocular depth estimators has mirrored the growth in model capacity, from relatively modest CNNs to large Transformer architectures. Still, monocular depth estimators tend to struggle when presented with images with unfamiliar content and layout, since their knowledge of the visual world is restricted by the data seen during training, and challenged by zero-shot generalization to new domains. This motivates us to explore whether the extensive priors captured in recent generative diffusion models can enable better, more generalizable depth estimation. We introduce Marigold*, a method for affine-invariant monocular depth estimation that is derived from Stable Diffusion and retains its rich prior knowledge. The estimator can be fine-tuned in a couple of days on a single GPU using only synthetic training data. It delivers state-of-the-art performance across a wide range of datasets, including over 20% performance gains in specific cases. Project page: https://marigoldmonodepth.github.io.*

1. Introduction

Monocular depth estimation aims to transform a photographic image into a depth map, *i.e.*, regress a range value for every pixel. The task arises whenever the 3D scene structure is needed, and no direct range or stereo measurements are available. Clearly, undoing the projection from the 3D world to a 2D image is a geometrically ill-posed problem and can

only be solved with the help of prior knowledge, such as typical object shapes and sizes, likely scene layouts, occlusion patterns, *etc*. In other words, monocular depth implicitly requires scene understanding, and it is no coincidence that the advent of deep learning brought about a leap in performance. Depth estimation is nowadays cast as neural image-to-image translation and learned in a supervised (or semi-supervised) fashion using collections of paired, aligned RGB images and depth maps. Early methods of this type were limited to a narrow domain defined by their training data, often indoor [47] or driving [18] scenes. More recently, there has been a quest to train generic depth estimators that can be either used off-the-shelf across a broad range of scenes or fine-tuned to a specific application scenario with a small amount of data. These models generally follow the strategy first employed by MiDAS [35] to achieve generality, namely to train a high-capacity model with data sampled from many different RGB-D datasets (respectively, domains). The latest developments include moving from convolutional encoder-decoder networks [35] to increasingly large and powerful vision transformers [36], and training on more and more data and with additional surrogate tasks [13] to amass even more knowledge about the visual world, and consequently to produce better depth maps. Importantly, visual cues for depth depend not only on the scene content but also on the (generally unknown) camera intrinsics [58]. For general in-the-wild depth estimation, it is often preferred to estimate affine-invariant depth (*i.e*., depth values up to a global offset and scale), which can also be determined without objects of known sizes that could serve as "scale bars".

The intuition behind our work is the following: Modern image diffusion models have been trained on internet-scale image collections specifically to generate high-quality images across a wide array of domains [3, 38, 41]. If the cornerstone of monocular depth estimation is indeed a comprehensive, encyclopedic representation of the visual world, then it should be possible to derive a broadly applicable depth estimator from a pretrained image diffusion model. In this paper, we set out to explore this option and develop **Marigold**, a latent diffusion model (LDM) based on Stable Diffusion [38], along with a fine-tuning protocol to adapt the model for depth estimation. The key to unlocking the potential of a pretrained diffusion model is to keep its latent space intact. We find this can be done efficiently by modifying and fine-tuning only the denoising U-Net. Turning Stable Diffusion into Marigold requires only synthetic RGB-D data (in our case, the Hypersim [37] and Virtual KITTI [7] datasets) and a few GPU days on a single consumer graphics card. Empowered by the underlying diffusion prior of natural images, Marigold exhibits excellent zero-shot generalization: Without ever having seen real depth maps, it attains state-of-the-art performance on several real datasets. To summarize, our contributions are:

1. A simple and resource-efficient fine-tuning protocol to convert a pretrained LDM image generator into an image-conditional depth estimator;
2. Marigold, a state-of-the-art, versatile monocular depth estimation module that offers excellent performance across a wide variety of natural images.

2. Related Work

2.1. Monocular Depth

At the technical level, monocular depth estimation is a dense, structured regression task. The pioneering work of Eigen et al. [14] introduced a multi-scale network and showed that the result can be converted to metric depth for a dataset recorded with a single sensor. Successive improvements have come from various fronts, including ordinal regression [15], planar guidance maps [24], neural conditional random fields [59], vision transformers [1, 27, 54], a piecewise planarity prior [34], first-order variational constraints [28] and variational autoencoders [32]. Some authors treat depth estimation as a combined regression-classification task, using various binning strategies like AdaBins [4] or BinsFormer [26] to discretize depth range. A notable recent trend involves training generative models, especially diffusion models [20, 49] for monocular depth estimation [12, 22, 43, 44]. Recently, a few works [19, 58] have revisited absolute depth estimation, by explicitly feeding camera intrinsics as additional input.

Estimating depth "in the wild" refers to methods that are successful across a wide range of (possibly unfamiliar) settings, a particularly challenging task. It has been addressed by constructing large and diverse depth datasets and designing algorithms to handle that diversity. DIW [8] was perhaps the earliest work to introduce an uncontrolled dataset and to predict *relative (ordinal) depth* for it. OASIS [9] introduced relative depth and normals to better generalize across scenes. However, relative depth predictions (depth ordering) are of limited use for many downstream tasks, which has led several authors to consider *affine-invariant depth*. In that setting, depth is estimated up to an unknown (global) offset and scale. It offers a viable compromise between the ordinal and metric cases: on the one hand, it can handle general scenes consisting of unfamiliar objects; on the other hand, depth differences between different objects or scene parts are still geometrically meaningful relative to each other. MegaDepth [25] and DiverseDepth [56] utilize large internet photo collections to train models that can adapt to unseen data, while MiDaS [35] achieves generality by training on a mixture of multiple datasets. The step from CNNs to vision transformers has further boosted performance, as evidenced by DPT [36] and Omnidata [13]. LeReS [57] proposed a two-stage framework that first predicts affine-invariant depth, then upgrades it to metric depth by estimating the shift and

focal length. HDN [60] introduced multi-scale depth normalization to improve the prediction details and smoothness further. While this enables the depth estimator to handle images captured with different known cameras, it does not include the true in-the-wild setting, where the camera intrinsics of the test images are unknown. Our method addresses affine-invariant depth estimation but does not focus on compiling an extensive, annotated training dataset. Rather, we utilize the broader image priors in image LDMs and apply fine-tuning.

2.2. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [20] have emerged as a powerful class of generative models. They learn to reverse a diffusion process that progressively degrades images with Gaussian noise so that they can draw samples from the data distribution by applying the reverse process to random noise. This idea was extended to DDIMs [49], which provide a non-Markovian shortcut for the diffusion process. *Conditional diffusion models* are an extension of DDPMs [20, 49] that ingest additional information on which the output is then conditioned, similar to cGAN [29] and cVAE [48]. Conditioning can take various forms, including text [41], other images [40], or semantic maps [61].

In the realm of text-based image generation, Rombach et al. [38] have trained a diffusion model on the large-scale image and text dataset LAION-5B [46] and demonstrated image synthesis with previously unattainable quality. The cornerstone of their approach is a latent diffusion model (LDM), where the denoising process is run in an efficient latent space, drastically reducing the complexity of the learned mapping. Such models distill internet-scale image sets into model weights, thereby developing a rich scene understanding prior, which we harness for monocular depth estimation.

2.3. Diffusion for Monocular Depth Estimation

Several methods have tried to use DDPMs for metric depth estimation. The DDP approach [22] proposes an architecture to encode the image but decode a depth map and has obtained state-of-the-art results on the KITTI dataset. DiffusionDepth [12] performs diffusion in the latent space, conditioned on image features extracted with a SwinTransformer. DepthGen [44] extends a multi-task diffusion model to metric depth prediction, including handling noisy ground truth. Its successor DDVM [43] emphasizes pretraining on synthetic and real data for enhanced depth estimation. Finally, VPD [64] employs a pretrained Stable Diffusion model as an image feature extractor with additional text input.

Our approach advances beyond these methods, which perform well but only in their specific training domains. We explore the potential of pretrained LDMs for single-image depth estimation across diverse, real-world settings.

2.4. Foundation Models

Vision foundation models (VFMs) are large neural networks trained on internet-scale data. The extreme scaling leads to the emergence of high-level visual understanding, such that the model can then be used as is [52] or fine-tuned to a wide range of downstream tasks with minimal effort [6]. *Prompt tuning* methods [2, 55, 63] can efficiently adapt VFMs towards dedicated scenarios by designing suitable prompts. *Feature adaptation* methods [5, 16, 33, 62, 64, 65] can further pivot VFMs towards different tasks. *E.g.*, VPD [64] showed the potential to extract features from a pre-trained text-to-image model for (domain-specific) depth estimation. Concurrently, I-LoRA [11] demonstrated the multi-modal capabilities of pre-trained image generators. *Direct tuning* enables more flexible adaptation, not only for few-shot customization scenarios like DreamBooth [39] but also for object detection, as in 3DiffTection [53].

The Marigold depth estimator proposed here can be interpreted as an instance of such direct tuning, where StableDiffusion plays the role of the foundation model. With as few as 74k synthetic depth samples, we obtain state-of-the-art depth estimates on real image datasets, and convincing performance in the wild (*cf*. Fig. 1).

3. Method

3.1. Generative Formulation

We pose monocular depth estimation as a conditional denoising diffusion generation task and train Marigold to model the conditional distribution $D(\mathbf{d} \mid \mathbf{x})$ over depth $\mathbf{d} \in \mathbb{R}^{W \times H}$, where the condition $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$ is an RGB image.

In the *forward* process, which starts at $\mathbf{d}_0 := \mathbf{d}$ from the conditional distribution, Gaussian noise is gradually added at levels $t \in \{1, ..., T\}$ to obtain noisy samples $\mathbf{d}_t$ as

$$\mathbf{d}_t = \sqrt{\bar{\alpha}_t}\mathbf{d}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, $\bar{\alpha}_t := \prod_{s=1}^{t} 1 - \beta_s$, and $\{\beta_1, \ldots, \beta_T\}$ is the variance schedule of a process with T steps. In the *reverse* process, the conditional denoising model $\boldsymbol{\epsilon}_\theta(\cdot)$ parameterized with learned parameters θ gradually removes noise from $\mathbf{d}_t$ to obtain $\mathbf{d}_{t-1}$.

At training time, parameters θ are updated by taking a data pair $(\mathbf{x}, \mathbf{d})$ from the training set, noising $\mathbf{d}$ with sampled noise $\boldsymbol{\epsilon}$ at a random timestep t, computing the noise estimate $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{d}_t, \mathbf{x}, t)$ and minimizing one of the denoising diffusion objective functions. The canonical standard noise objective $\mathcal{L}$ is given as follows [20]:

$$\mathcal{L} = \mathbb{E}_{\mathbf{d}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(0,I), t \sim \mathcal{U}(T)} \left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}} \right\|_2^2 . \tag{2}$$

At inference time, $\mathbf{d} := \mathbf{d}_0$ is reconstructed starting from a normally-distributed variable $\mathbf{d}_T$, by iteratively applying the learned denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{d}_t, \mathbf{x}, t)$.

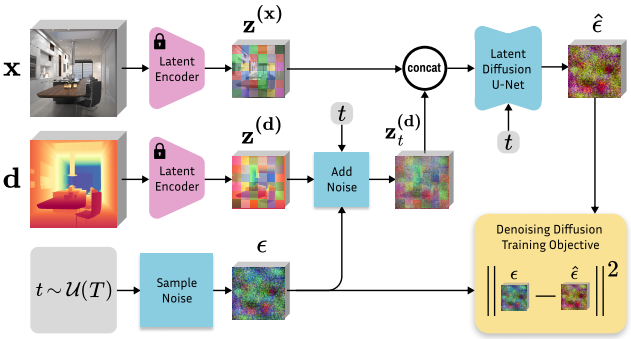

Figure 2. **Overview of the Marigold fine-tuning protocol.** Starting from pretrained Stable Diffusion, we encode the image $\mathbf{x}$ and depth $\mathbf{d}$ into the latent space using the original Stable Diffusion VAE. We fine-tune just the U-Net by optimizing the standard diffusion objective relative to the depth latent code. Image conditioning is achieved by concatenating the two latent codes before feeding them into the U-Net. The first layer of the U-Net is modified to accept concatenated latent codes. See details in Sec. 3.2 and Sec. 3.3.

Unlike diffusion models that work directly on the data, latent diffusion models perform diffusion steps in a low-dimensional latent space, offering computational efficiency and suitability for high-resolution image generation [38]. The latent space is constructed in the bottleneck of a variational autoencoder (VAE) trained independently of the denoiser to enable latent space compression and perceptual alignment with the data space. To translate our formulation into the latent space, for a given depth map $\mathbf{d}$, the corresponding latent code is given by the encoder $\mathcal{E}$: $\mathbf{z}^{(\mathbf{d})} = \mathcal{E}(\mathbf{d})$. Given a depth latent code, a depth map can be recovered with the decoder $\mathcal{D}$: $\hat{\mathbf{d}} = \mathcal{D}(\mathbf{z}^{(\mathbf{d})})$. The conditioning image $\mathbf{x}$ is also naturally translated into the latent space as $\mathbf{z}^{(\mathbf{x})} = \mathcal{E}(\mathbf{x})$. The denoiser is henceforth trained in the latent space: $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{(\mathbf{d})}, \mathbf{z}^{(\mathbf{x})}, t)$. The adapted inference procedure involves one extra step – the decoder $\mathcal{D}$ reconstructing the data $\hat{\mathbf{d}}$ from the estimated clean latent $\mathbf{z}_0^{(\mathbf{d})}$: $\hat{\mathbf{d}} = \mathcal{D}(\mathbf{z}_0^{(\mathbf{d})})$.

3.2. Network Architecture

One of our main objectives is training efficiency since diffusion models are often extremely resource-intensive to train. Therefore, we base our model on a pretrained text-to-image LDM (Stable Diffusion v2 [38]), which has learned very good image priors from LAION-5B [46]. With minimal changes to the model components, we turn it into an image-conditioned depth estimator. Fig. 2 contains an overview of the proposed fine-tuning procedure.

Depth encoder and decoder. We take the frozen VAE to encode both the image and its corresponding depth map into a latent space for training our conditional denoiser. Given that the encoder, which is designed for 3-channel (RGB) inputs, receives a single-channel depth map, we replicate the depth map into three channels to simulate an RGB image. At this point, the data range of the depth data plays a significant role in enabling affine-invariance. We discuss our normalization approach in Sec. 3.3. We verified that without any modification of the VAE or the latent space structure, the depth map can be reconstructed from the encoded latent code with a negligible error, *i.e.*, $\mathbf{d} \approx \mathcal{D}(\mathcal{E}(\mathbf{d}))$. At inference time, the depth latent code is decoded once at the end of diffusion, and the average of three channels is taken as the predicted depth map.

Adapted denoising U-Net. To implement the conditioning of the latent denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{(\mathbf{d})}, \mathbf{z}^{(\mathbf{x})}, t)$ on input image $\mathbf{x}$, we concatenate the image and depth latent codes into a single input $\mathbf{z}_t = \text{cat}(\mathbf{z}_t^{(\mathbf{d})}, \mathbf{z}^{(\mathbf{x})})$ along the feature dimension. The input channels of the latent denoiser are then doubled to accommodate the expanded input $\mathbf{z}_t$. To prevent inflation of activations magnitude of the first layer and keep the pretrained structure as faithfully as possible, we duplicate the weight tensor of the input layer and divide its values by two.

3.3. Fine-Tuning Protocol

Affine-invariant depth normalization. For the ground truth depth maps $\mathbf{d}$, we implement a linear normalization such that the depth primarily falls in the value range $[-1, 1]$, to match the designed input value range of the VAE. Such normalization serves two purposes. First, it is the convention for working with the original Stable Diffusion VAE. Second, it enforces a canonical affine-invariant depth representation independent of the data statistics – any scene must be bounded by near and far planes with extreme depth values. The normalization is achieved through an affine transformation computed as

$$\tilde{\mathbf{d}} = \left(\frac{\mathbf{d} - \mathbf{d}_2}{\mathbf{d}_{98} - \mathbf{d}_2} - 0.5 \right) \times 2, \tag{3}$$

where $\mathbf{d}_2$ and $\mathbf{d}_{98}$ correspond to the 2% and 98% percentiles of individual depth maps. This normalization allows Marigold to focus on pure affine-invariant depth estimation.

Training on synthetic data. Real depth datasets suffer from missing depth values caused by the physical constraints of the capture rig and the physical properties of the sensors. Specifically, the disparity between cameras and reflective surfaces diverting LiDAR laser beams are inevitable sources of ground truth noise and missing pixels [21, 51]. In contrast to prior work that utilized diverse real datasets to achieve generalization [13, 35], we train exclusively with synthetic depth datasets. As with the depth normalization rationale, this decision has two objective reasons. First, synthetic depth is inherently dense and complete, meaning that every pixel has a valid ground truth depth value, allowing us to feed such data into the VAE, which can not handle data with invalid pixels. Second, synthetic depth is the cleanest

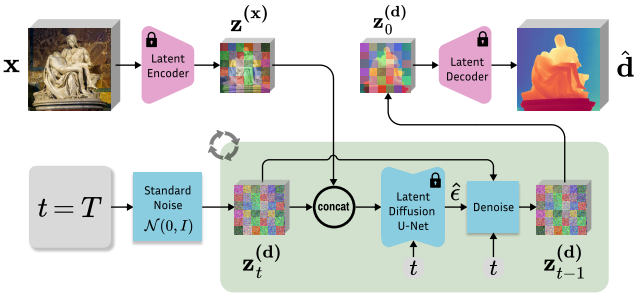

Figure 3. **Overview of the Marigold inference scheme.** Given an input image $\mathbf{x}$, we encode it with the original Stable Diffusion VAE into the latent code $\mathbf{z}^{(\mathbf{x})}$, and concatenate with the depth latent $\mathbf{z}_t^{(\mathbf{d})}$ before giving it to the modified fine-tuned U-Net on every denoising iteration. After executing the schedule of T steps, the resulting depth latent $\mathbf{z}_0^{(\mathbf{d})}$ is decoded into an image, whose 3 channels are averaged to get the final estimation $\hat{\mathbf{d}}$. See Sec. 3.4 for details.

possible form of depth, which is guaranteed by the rendering pipeline. If our assumption about the possibility of fine-tuning a generalizable depth estimation from a text-to-image LDM is correct, then synthetic depth gives the cleanest set of examples and reduces noise in gradient updates during the short fine-tuning protocol. Thus, the remaining concern is the sufficient diversity or domain gaps between synthetic and real data, which sometimes limits generalization ability. As demonstrated in our experiments, our choice of synthetic datasets leads to impressive zero-shot transfer.

Annealed multi-resolution noise. Previous works have explored deviations from the original DDPM formulations, such as non-Gaussian noise [30] or non-Markovian schedule shortcuts [49]. Our proposed setting and the fine-tuning protocol outlined above are permissive to changes to the noise schedule at the fine-tuning stage. We identified a combination of multi-resolution noise [23] and an annealed schedule to converge faster and substantially improve performance over the standard DDPM formulation. The multi-resolution noise is composed by superimposing several random Gaussian noise images of different scales, all upsampled to the U-Net input resolution. The proposed annealed schedule interpolates between the multi-resolution noise at $t = T$ and standard Gaussian noise at $t = 0$.

3.4. Inference

Latent diffusion denoising. The overall inference pipeline is presented in Fig. 3. We encode the input image into the latent space, initialize depth latent as standard Gaussian noise, and progressively denoise it with the same schedule as during fine-tuning. We empirically find that initializing with standard Gaussian noise gives better results than with multi-resolution noise, although the model is trained on the latter. We follow DDIM's [49] approach to perform non-Markovian sampling with re-spaced steps for accelerated inference. The final depth map is decoded from the latent code using the VAE decoder and postprocessed by averaging channels.

Test-time ensembling. The stochastic nature of the inference pipeline leads to varying predictions depending on the initialization noise in $\mathbf{z}_T^{(\mathbf{d})}$. Capitalizing on that, we propose the following test-time ensembling scheme, capable of combining multiple inference passes over the same input. For each input sample, we can run inference N times. To aggregate these affine-invariant depth predictions $\{\hat{\mathbf{d}}_1, \ldots, \hat{\mathbf{d}}_N\}$, we jointly estimate the corresponding scale $\hat{s}_i$ and shift $\hat{t}_i$, relative to some canonical scale and range, in an iterative manner. The proposed objective minimizes the distances between each pair of scaled and shifted predictions $(\hat{\mathbf{d}}'_i, \hat{\mathbf{d}}'_j)$, where $\hat{\mathbf{d}}' = \hat{\mathbf{d}} \times \hat{s} + \hat{t}$. In each optimization step, we calculate the merged depth map $\mathbf{m}$ by the taking pixel-wise median $\mathbf{m}(x, y) = \text{median}(\hat{\mathbf{d}}'_1(x, y), \ldots, \hat{\mathbf{d}}'_N(x, y))$. An extra regularization term $\mathcal{R} = |\min(\mathbf{m})| + |1 - \max(\mathbf{m})|$, is added to prevent collapse to the trivial solution and enforce the unit scale of $\mathbf{m}$. Thus, the objective function can be written as follows:

$$\min_{\substack{s_1,\ldots,s_N \\ t_1,\ldots,t_N}} \left(\sqrt{\frac{1}{b} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\hat{\mathbf{d}}'_i - \hat{\mathbf{d}}'_j\|_2^2} + \lambda \mathcal{R} \right) \quad (4)$$

where the binominal coefficient $b = \binom{N}{2}$ represents the number of possible combinations of image pairs from N images. After the iterative optimization for spatial alignment, the merged depth $\mathbf{m}$ is taken as our ensembled prediction. Note that this ensembling step requires no ground truth for aligning independent predictions. This scheme enables a flexible trade-off between computation efficiency and prediction quality by choosing N accordingly.

4. Experiments

4.1. Implementation

We implement Marigold using PyTorch and utilize Stable Diffusion v2 [38] as our backbone, following the original pre-training setup with a v-objective [42]. We disable text conditioning and perform the steps outlined in Sec. 3.2. During training, we apply the DDPM noise scheduler [20] with 1000 diffusion steps. At inference time, we apply the DDIM scheduler [49] and only sample 50 steps. For the final prediction, we aggregate results from 10 inference runs with varying starting noise. Training our method takes 18K iterations using a batch size of 32. To fit one GPU, we accumulate gradients for 16 steps. We use the Adam optimizer with a learning rate of $3 \cdot 10^{-5}$. Additionally, we apply random horizontal flipping augmentation to the training data. Training our method to convergence takes approximately 2.5 days on a single Nvidia RTX 4090 GPU card.

Table 1. **Quantitative comparison** of Marigold with SOTA affine-invariant depth estimators on several zero-shot benchmarks. All metrics[†] are presented in percentage terms; **bold** numbers are the best, underscored second best. Our method outperforms other methods on both indoor and outdoor scenes in most cases, without having seen a real depth sample.

Method	# Training samples		NYUv2		KITTI		ETH3D		ScanNet		DIODE		Avg. Rank
	Real	Synthetic	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	
DiverseDepth [56]	320K	—	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	7.6
MiDaS [35]	2M	—	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	7.3
LeReS [57]	300K	54K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	5.2
Omnidata [13]	11.9M	310K	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	4.8
HDN [60]	300K	—	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	**78.0**	3.2
DPT [36]	1.2M	188K	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	**18.2**	75.8	3.9
Ours (w/o ensemble)	—*	74K	6.0	95.9	10.5	90.4	7.1	95.1	6.9	94.5	31.0	77.2	2.5
Ours (w/ ensemble)			**5.5**	**96.4**	**9.9**	**91.6**	**6.5**	**96.0**	**6.4**	**95.1**	30.8	77.3	**1.4**

[†] Most baselines are sourced from Metric3D [58], except for the ScanNet benchmark. For ScanNet, Metric3D used a different random split that is not publicly accessible, therefore we re-ran all baselines on our split. For HDN [60] we show the ScanNet results from Metric3D, as no source code is available.
* Image-text data is used in the pretrained model.

4.2. Evaluation

Training datasets. We train Marigold on two synthetic datasets covering both indoor and outdoor scenes. **Hypersim** [37] is a photorealistic dataset with 461 indoor scenes. We use the official split with around 54K samples from 365 scenes for training. Incomplete samples are filtered out. RGB images and depth maps are resized to 480×640 size. Additionally, we transform the original distances relative to the focal point into conventional depth values relative to the focal plane. The second dataset, **Virtual KITTI** [7] is a synthetic street-scene dataset featuring 5 scenes under varying conditions like weather or camera perspectives. Four scenes containing a total of around 20K samples are used for training. We crop the images to the KITTI benchmark resolution [17] and set the far plane to 80 meters.

Evaluation datasets. We evaluate Marigold on 5 real datasets that are not seen during training. **NYUv2** [31] and **ScanNet** [10] are both indoor scene datasets captured with an RGB-D Kinect sensor. For NYUv2, we utilize the designated test split, comprising a total of 654 images. In the case of the ScanNet dataset, we randomly sampled 800 images from the 312 official validation scenes for testing. **KITTI** [17] is a street-scene dataset with sparse metric depth captured by a LiDAR sensor. We employ the Eigen test split [14] made of 652 images. **ETH3D** [45] and **DIODE** [50] are two high-resolution datasets, both featuring depth maps derived from LiDAR sensor measurements. For ETH3D, we incorporate all 454 samples with available ground truth depth maps. For DIODE, we use the entire validation split, which encompasses 325 indoor samples and 446 outdoor samples.

Evaluation protocol. Following the protocol of affine-invariant depth evaluation [35], we first align the estimated merged prediction $\mathbf{m}$ to the ground truth $\mathbf{d}$ with the least squares fitting. This step gives us the absolute aligned depth map $\mathbf{a} = \mathbf{m} \times s + t$ in the same units as the ground truth. Next, we apply two widely recognized metrics [35, 36, 57, 58] for assessing quality of depth estimation. The first is Absolute Mean Relative Error (*AbsRel*), calculated as: $\frac{1}{M}\sum_{i=1}^{M} |\mathbf{a}_i - \mathbf{d}_i|/\mathbf{d}_i$, where M is the total number of pixels. The second metric, $\delta1$ accuracy, measures the proportion of pixels satisfying $\max(\mathbf{a}_i/\mathbf{d}_i, \mathbf{d}_i/\mathbf{a}_i) < 1.25$.

Comparison with other methods. We compare Marigold to six baselines, each claiming zero-shot generalization. DiverseDepth [56], LeReS [57] and HDN [60] estimate affine-invariant depth maps, while MiDaS [35], DPT [36], and Omnidata [13] produce affine-invariant disparities. As shown in Tab. 1, Marigold outperforms prior art in most cases and secures the highest overall ranking. Despite being trained solely on synthetic depth datasets, the model can well generalize to a wide range of real scenes. This successful adaptation of diffusion-based image generation models toward depth estimation confirms our initial hypothesis that a comprehensive representation of the visual world is the cornerstone of monocular depth estimation. It also shows that our fine-tuning protocol was successful in adapting Stable Diffusion for this task without unlearning such visual priors.

For a visual assessment, we present qualitative comparison in Fig. 4. Additionally, in Fig. 5, we provide 3D visualizations of reconstructed surface normals. Marigold not only correctly captures the scene layout, such as the spatial relationships between walls and furniture in the first example in Fig. 5, but also captures fine-grained details, as indicated by the arrows in Fig. 4. Moreover, the reconstruction of flat surfaces, especially walls, is significantly better (see Fig. 4). Furthermore, our method effectively models common shapes and their layouts, once again aligning with our expectations regarding the generative prior.

4.3. Ablation Studies

Two zero-shot validation sets are selected for the ablation studies – the official training split of NYUv2 [31], consisting of 785 samples, and a randomly selected subset of 800

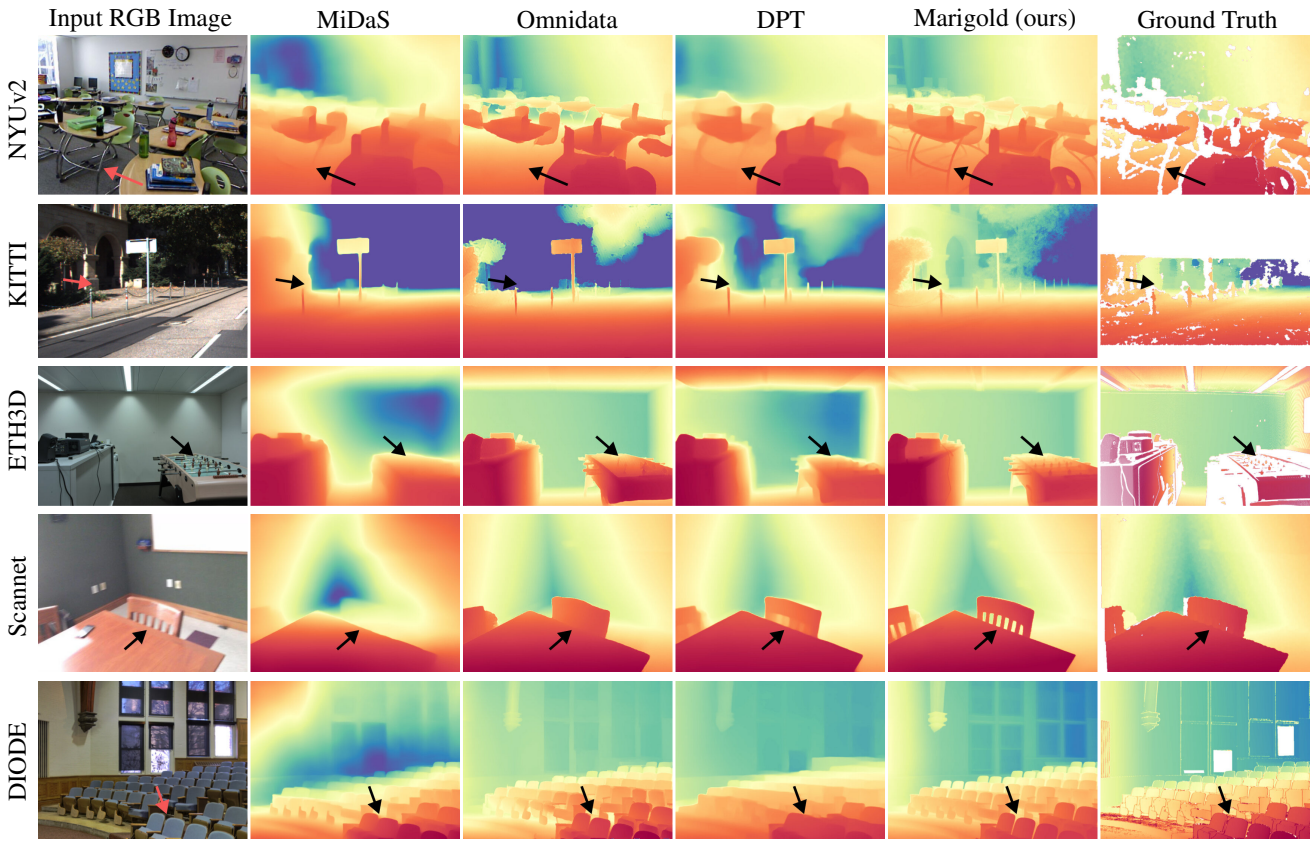

Figure 4. **Qualitative comparison (depth)** of monocular depth estimation methods across different datasets. Marigold excels at capturing thin structures (*e.g.*, chair legs) and preserving overall layout of the scene (*e.g.*, walls in ETH3D example and chairs in DIODE example).

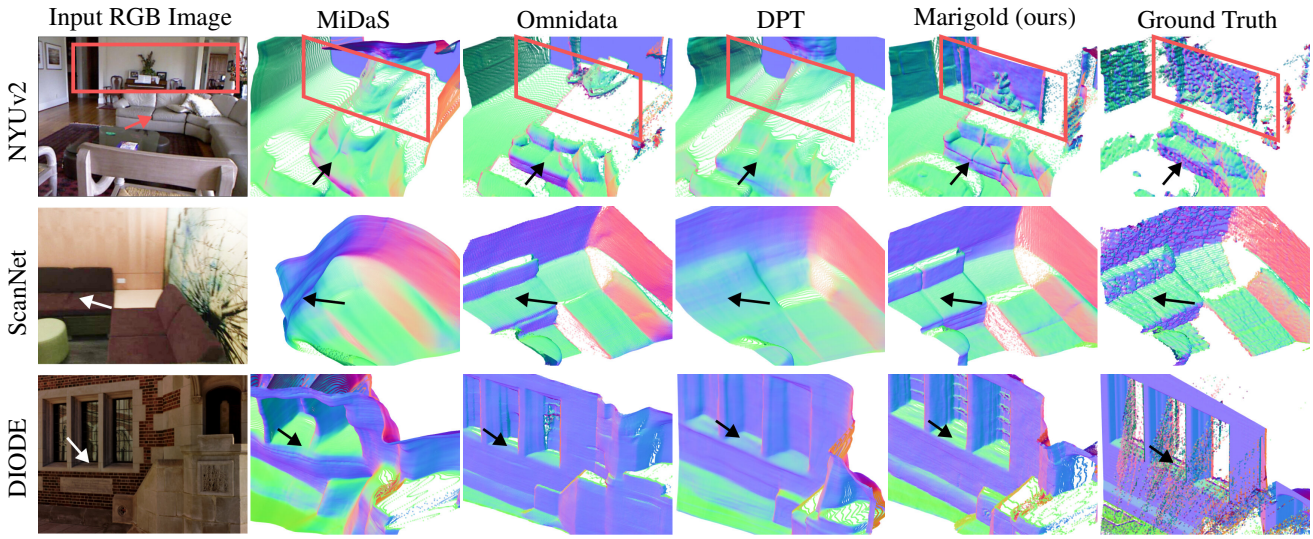

Figure 5. **Qualitative comparison (unprojected, colored as normals)** of monocular depth estimation methods across different datasets. Marigold stands out for its superior reconstruction of flat surfaces and detailed structures.

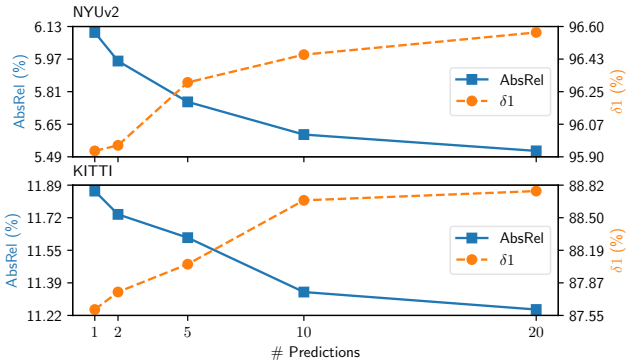

Figure 6. **Ablation of ensemble size.** We observe a monotonic improvement with the growth of ensemble size. This improvement starts to diminish after 10 predictions per sample.

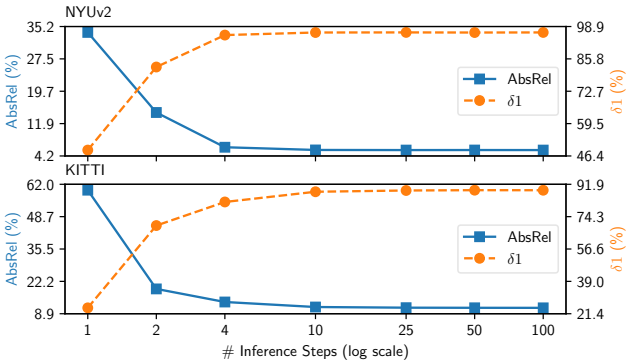

Figure 7. **Ablation of denoising steps.** The performance improves as the number of denoising steps increases, while we observe saturation after 10 steps.

Table 2. **Ablation of training noise.** Multi-resolution noise improves over Gaussian noise; annealing yields further improvement.

Multi-res. noise	Annealed	NYUv2 AbsRel↓	NYUv2 δ1↑	KITTI AbsRel↓	KITTI δ1↑
✗	-	7.7	93.4	14.2	82.1
✓	✗	5.8	96.1	12.1	87.1
✓	✓	**5.6**	**96.5**	**11.3**	**88.7**

Table 3. **Ablation of training datasets.** Hypersim [37] delivers strong results; Virtual KITTI [7] improves outdoor performance.

Hypersim	Virtual KITTI	NYUv2 AbsRel↓	NYUv2 δ1↑	KITTI AbsRel↓	KITTI δ1↑
✗	✓	13.9	83.4	15.4	79.3
✓	✗	5.7	96.3	13.7	82.5
✓	✓	**5.6**	**96.5**	**11.3**	**88.7**

images from the KITTI Eigen [14] training split. Refer to supplementary sections for extra ablations and discussion.

Training noise. We investigate the impact of three types of noise during the training phase. As shown in Tab. 2, training with multi-resolution noise significantly improves the depth prediction accuracy over using standard Gaussian noise. Furthermore, the gradual annealing of multi-resolution noise yields an additional improvement. We also noticed that training with multi-resolution noise leads to more consistent predictions given different initial noise at inference time and annealing further enhances this consistency.

Training data domain. To better understand the impact of the synthetic datasets used for our fine-tuning protocol, we ablate on a photorealistic street-scene Virtual KITTI [7], and a more diverse and higher-quality indoor dataset Hypersim [37]. The results are shown in Tab. 3. When fine-tuned on a single synthetic dataset, the pretrained LDM can already be adapted for monocular depth estimation to a certain degree, while the more diverse and photorealistic data leads to better performance on both indoor and outdoor scenes. Interestingly, adding additional training data from a different domain not only improves the performance on the new domain but also brings improvements in the original domain.

Test-time ensembling. We test the effectiveness of the proposed test-time ensembling scheme by aggregating various numbers of predictions. As shown in Fig. 6, a single prediction of Marigold already yields reasonably good results. Ensembling 10 predictions reduces the absolute relative error on NYUv2 by $\approx 8\%$ and ensembling 20 predictions brings an improvement of $\approx 9.5\%$. It has been observed as a systematic effect that the performance is constantly improved as the number of predictions increases, while the marginal improvement diminishes with more than 10 predictions.

Number of denoising steps. We evaluate the effect of the re-spaced inference denoising steps driven by the DDIM scheduler [49]. The results are shown in Fig. 7. Although trained with 1000 DDPM steps, the choice of 50 steps is sufficient to produce accurate results during inference. As expected, we obtain better results when using more denoising steps. We observe that the elbow point of marginal returns given more denoising steps depends on the dataset but is always under 10 steps. This implies that one can further reduce the denoising steps to 10 or even less to gain efficiency while keeping comparable performance. Interestingly, this threshold is smaller than what is usually required for diffusion-based image generators [38, 49], *i.e.*, 50 steps.

5. Conclusion

We have presented Marigold, a fine-tuning protocol for Stable Diffusion and a model for state-of-the-art affine-invariant depth estimation. Our results confirm the importance of a detailed visual scene understanding prior for depth estimation, which we source from the pretrained text-to-image diffusion model. Future research directions to overcome current limitations include improving inference efficiency, ensuring that similar inputs yield consistent outputs despite the model's generative nature, and better handling of distant scene parts.

References

[1] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, Mannat Kaur, and Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *ICRA*, 2021. 2

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. 2

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2

[5] Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. Stylegan knows normal, depth, albedo, and more. In *Advances in Neural Information Processing Systems*, pages 73082–73103. Curran Associates, Inc., 2023. 3

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022. 3

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020. 2, 6, 8

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *NeurIPS*, 29, 2016. 2

[9] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. OASIS: A large-scale dataset for single image 3d in the wild. In *CVPR*, 2020. 2

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 6

[11] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv*, 2023. 3

[12] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 2, 3

[13] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 2, 4, 6

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 2, 6, 8

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 2

[16] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 2023. 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 6

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 2013. 2

[19] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2, 3, 5

[21] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. In *ICCV*, 2023. 4

[22] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. In *ICCV*, 2023. 2, 3

[23] Kasiopy. Multi-resolution noise for diffusion model training. https://wandb.ai/johnowhitaker/multires_noise/reports/Multi-Resolution-Noise-for-Diffusion-Model-Training--VmlldzozNjYyOTU2?s=31, 2023. last accessed 17.11.2023. 5

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

[25] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 2

[26] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 2

[27] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, pages 1–18, 2023. 2

[28] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. VA-DepthNet: A variational approach to single image depth prediction. In *ICLR*, 2023. 2

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[30] Eliya Nachmani, Robin San Roman, and Lior Wolf. Non Gaussian denoising diffusion models. *arXiv preprint arXiv:2106.07582*, 2021. 5

[31] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 6

[32] Jia Ning, Chen Li, Zheng Zhang, Chunyu Wang, Zigang Geng, Qi Dai, Kun He, and Han Hu. All in tokens: Unifying output space of visual tasks via soft token. In *ICCV*, pages 19900–19910, 2023. 2

[33] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. SVL-adapter: Self-supervised adapter for vision-language pretrained models. In *BMVC*, 2022. 3

[34] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *CVPR*, 2022. 2

[35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 2, 4, 6

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2, 6

[37] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 2, 6, 8

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 4, 5, 8

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 3

[40] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–10, 2022. 3

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35, 2022. 2, 3

[42] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 5

[43] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *arXiv preprint arXiv:2306.01923*, 2023. 2, 3

[44] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 2, 3

[45] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 6

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 3, 4

[47] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760, 2012. 2

[48] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 3

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 5, 8

[50] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arXiv preprint arXiv:1908.00463*, 2019. 6

[51] Wolfgang Wagner, Andreas Ullrich, Vesna Ducic, Thomas Melzer, and Nick Studnicka. Gaussian decomposition and calibration of a novel small-footprint full-waveform digitising airborne laser scanner. *ISPRS journal of Photogrammetry and Remote Sensing*, 60(2):100–112, 2006. 4

[52] Tianfu Wang, Menelaos Kanakis, Konrad Schindler, Luc Van Gool, and Anton Obukhov. Breathing new life into 3d assets with generative repainting. In *BMVC*. BMVA Press, 2023. 3

[53] Chenfeng Xu, Huan Ling, Sanja Fidler, and Or Litany. 3DiffTection: 3d object detection with geometry-aware diffusion features. *arXiv*, 2023. 3

[54] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 2021. 2

[55] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 3

[56] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 2, 6

[57] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. 2, 6

[58] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. 2, 6

[59] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. NeWCRFs: Neural window fully-connected CRFs for monocular depth estimation. In *CVPR*, 2022. 2

[60] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *NeurIPS*, 35, 2022. 3, 6

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3

[62] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free CLIP-adapter for better vision-language modeling. *arXiv:2111.03930*, 2021. 3

[63] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, pages 15211–15222, 2023. 3

[64] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv:2303.02153*, 2023. 3

[65] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *ECCV*, pages 696–712, 2022. 3