GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image

Xiao Fu[1*], Wei Yin[2*], Mu Hu[3*], Kaixuan Wang[3], Yuexin Ma[4],
Ping Tan[3,6], Shaojie Shen[3], Dahua Lin[1], and Xiaoxiao Long[5,6]

https://fuxiao0719.github.io/projects/geowizard/
[1]CUHK [2]The University of Adelaide [3]HKUST
[4]ShanghaiTech University [5]HKU [6]Light Illusions

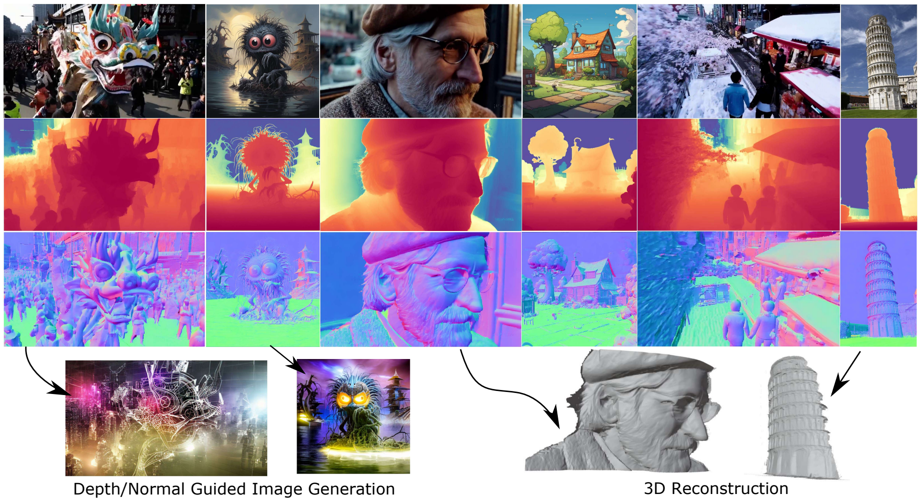

Fig. 1: We propose *GeoWizard*, an innovative foundation model for jointly estimating depth and surface normal from monocular images. Compared to prior discriminative counterparts, our work not only achieves surprisingly robust generalization on various types of real or unreal images but also faithfully captures intricate geometric details. The generated depth and normal could enhance many applications such as 2D content generation, 3D reconstruction and so on.

Abstract. We introduce *GeoWizard*, a new generative foundation model designed for estimating geometric attributes, e.g., depth and normals, from single images. While significant research has already been conducted in this area, the progress has been substantially limited by the low diversity and poor quality of publicly available datasets. As a result, the prior works either are constrained to limited scenarios or suffer from the inability to capture geometric details. In this paper, we demonstrate that generative models, as opposed to traditional discriminative models (e.g., CNNs and Transformers), can effectively address the inherently ill-posed problem. We further show that leveraging diffusion priors can markedly

[*] Equal contribution

improve generalization, detail preservation, and efficiency in resource usage. Specifically, we extend the original stable diffusion model to jointly predict depth and normal, allowing mutual information exchange and high consistency between the two representations. More importantly, we propose a simple yet effective strategy to segregate the complex data distribution of various scenes into distinct sub-distributions. This strategy enables our model to recognize different scene layouts, capturing 3D geometry with remarkable fidelity. *GeoWizard* sets new benchmarks for zero-shot depth and normal prediction, significantly enhancing many downstream applications such as 3D reconstruction, 2D content creation, and novel viewpoint synthesis.

Keywords: Monocular Images · Depth · Normal · Diffusion Models

1 Introduction

Estimating 3D geometry, e.g., depth and surface normal from monocular color images, is a fundamental but challenging problem in 3D computer vision, which plays essential roles in various downstream applications such as autonomous driving [12, 13], 3D surface reconstruction [31, 59, 69], novel view synthesis [29, 37], inverse rendering [51,68], and so on. Reverting the projection from a 3D environment to a 2D image presents a geometrically ambiguous challenge, necessitating the aid of prior knowledge. This may include understanding typical object dimensions and shapes, probable scene arrangements, as well as occlusion patterns.

The recent advancements in deep learning have significantly propelled the field of geometry estimation forward. Currently, this task is often approached as a neural image-to-image translation problem, where supervised learning techniques are employed. However, the progress in this area is constrained by two major shortcomings in the publicly available datasets: 1) **Low diversity.** Lacking efficient and reliable tools for data collection, most datasets are confined to specific scenarios, such as autonomous driving and indoor environments. Models trained on these datasets typically exhibit poor generalization capabilities when applied to out-of-domain images. 2) **Poor accuracy.** To enhance dataset diversity, some works generate pseudo labels for unlabeled data using methods like multi-view stereo (MVS) reconstruction or self-training techniques. Unfortunately, these pseudo-labels often suffer from being incomplete or of low quality. Consequently, while these approaches may improve model generalization, they still struggle in accurately capturing geometric details and require significantly more computational resources.

In this paper, our goal is to build a foundation model for monocular geometry estimation capable of producing high-quality depth and normal information for any images of any scenarios (even images generated by AIGC). Instead of employing straightforward data and computation scaling-up, our method proposes to unleash the diffusion priors for this ill-posed problem. The intuition is that stable diffusion models have been proven to inherently encode rich knowledge of the 3D world, and its strong diffusion priors pre-trained on billions of images could significantly facilitate potential 3D tasks.

Instead of tackling depth or normal estimation separately, *GeoWizard* jointly estimates depth and normal within a unified framework. Inspired by Wonder3D [31], we leverage ***geometry switcher*** to extend a single stable diffusion model to produce both depth and normal. The joint estimation allows mutual information exchange and high consistency between the two representations. However, direct training on mixed data encompassing various scenarios often leads to ambiguities in geometry estimation, potentially skewing the estimated depth/normal towards unintended layouts. To address this challenge, we propose a simple yet effective strategy, ***scene distribution decoupler***, to segregate the complex data distribution of different scenes into distinct sub-distributions (e.g., outdoor, indoor, and background-free objects). This strategic approach enables the diffusion model to discern different scene layouts, resulting in the capture of 3D geometry with remarkable fidelity. Consequently, *GeoWizard* achieves state-of-the-art performance in zero-shot depth and normal prediction, thereby significantly enhancing numerous downstream applications such as 3D reconstruction, 2D content creation, and novel viewpoint synthesis.

Overall, our contributions are summarized as follows:

- We present *GeoWizard*, a new generative foundation model for joint depth and normal estimation that faithfully captures intricate geometric details.
- We propose a simple yet effective *scene distribution decoupler* strategy, aimed at guiding diffusion models to circumvent ambiguities that may otherwise lead to the conflation of distinct scene layouts.
- *GeoWizard* achieves SOTA performance in zero-shot estimation of both depth and normal, substantially enhancing a wide range of applications.

2 Related Work

Joint Depth and Normal Estimation. Estimating depth and normal from images is an ill-posed but important task, where depth and surface normal encode the 3D geometry in different aspects. Some existing approaches propose to explicitly acquire the surface normal from the depth map by using some geometric constraints, such as Sobel-like operator [16, 23], differentiable least square [32, 38], or randomly sampled point triplets [33, 63, 64]. IronDepth [2] propagates depth on pre-computed local surface. Zhao *et al.* [74] proposes to jointly refine depth and normal by a solver, but it conditions on multi-view prior and tedious post-optimization. On the other hand, several works [10, 24, 61, 73] create multiple branches for depth and normal, and enforce information exchange through propagating latent features. However, all the prior works tackle this problem using discriminative models and leverage limited scopes of training datasets, and therefore present poor generalization and fail to capture geometric details. In contrast, *GeoWizard* builds on generative models and fully leverage diffusion priors to tackle this problem, showing significantly improved generalization and ability to capture geometric details.

Diffusion Models for Geometry Estimation. Recently, diffusion models [15, 55] have shown supreme capabilities in 3D tasks, such as optical flow estimation [8, 48], view synthesis [27, 47, 52], depth estimation [18, 21, 75], and normal estimation [28, 31, 39] (in contrast to GAN [4]). For depth estimation, DDP [18] first introduces a unified diffusion architecture that blends the traditional perception pipeline to estimate the metric depth. DDVM [48] further boosts depth quality by training on synthetic data. Although they leverage improved diffusion process [56] or advanced perception backbone [26, 30] to speed up training, they still suffer from unaffordable low efficiency and slow convergence when scaled up to internet-scale data. A concurrent method Marigold [21] fine-tune the pretrained stable diffusion model for depth estimation and also try to leverage the diffusion priors. However, it suffers from ambiguities about mixed layouts of various scenarios and tends to produce depth maps with unintended layouts.

Diffusion-based methods are also applied to normal estimation. JointNet [71] attempts to connect multiple diffusion models to achieve multi-modality estimation (e.g., depth and normal), however their model size and resource costs will linearly increase depending on the number of modalities. Wonder3D [31] proposes to model joint color and normal distribution with a domain switcher to enhance geometric quality and consistency. Richdreamer [39] trains seperately depth and normal diffusion model on the LAION-2B [50] dataset with predictions from Midas [42]. However, these methods still struggle to capture geometric details. In contrast, to the best of our knowledge, *GeoWizard* reveals robust generalization and a significant ability to capture intricate geometric details.

3 Methodology

Given an input image $\mathbf{x}$, our goal is to generate its paired depth map $\hat{\mathbf{d}}$ and normal map $\hat{\mathbf{n}}$. Firstly, we delve into the problem with the diffusion paradigm (see Section 3.1). Secondly, we present our geometric diffusion model (see Section 3.2). The model uses a cross-domain geometry switcher to jointly generate the depth and normal using a single diffusion model. The mutual information exchange enhances geometric consistency. We further decouple the sophisticated scene distribution into several distinct sub-distributions (e.g., outdoor, indoor, and background-free objects) to avoid ambiguities of geometry estimation. The overview of *GeoWizard* is presented in Fig. 2.

3.1 Preliminaries on Geometric Distribution

Diffusion Probabilistic Models [15, 55] define a forward Markov chain that progressively transits the sample $\mathbf{x}$ drawn from data distribution $p(\mathbf{x})$ into noisy versions $\{\mathbf{x}_t, t \in (1, T) | \mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}\}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, T is the training step, α_t and σ_t are the noisy scheduler terms that control sample quality. In the reverse Markov chain, it learns a denoising network $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\cdot)$ parameterized by $\boldsymbol{\theta}$ usually structured as U-Net [45] to transform $\mathbf{x}_t$ into $\mathbf{x}_{t-1}$ from an initial Gaussian sample $\mathbf{x}_T$ through iterative denoising.

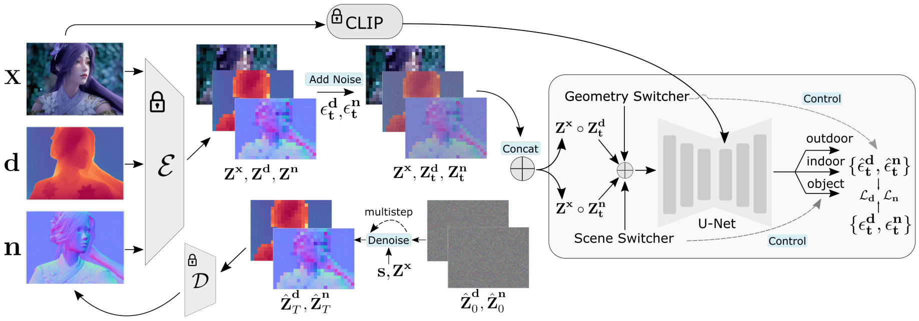
CLIP
x
d
n
$\mathcal{E}$
$\mathbf{Z^x}, \mathbf{Z^d}, \mathbf{Z^n}$
Add Noise
$\epsilon_\mathbf{t}^\mathbf{d}, \epsilon_\mathbf{t}^\mathbf{n}$
$\mathbf{Z^x}, \mathbf{Z_t^d}, \mathbf{Z_t^n}$
Concat
Geometry Switcher
Control
$\mathbf{Z^x} \circ \mathbf{Z_t^d}$
$\mathbf{Z^x} \circ \mathbf{Z_t^n}$
U-Net
outdoor
indoor
object
$\{\hat{\epsilon}_\mathbf{t}^\mathbf{d}, \hat{\epsilon}_\mathbf{t}^\mathbf{n}\}$
$\mathcal{L}_\mathbf{d}$, $\mathcal{L}_\mathbf{n}$
$\{\epsilon_\mathbf{t}^\mathbf{d}, \epsilon_\mathbf{t}^\mathbf{n}\}$
Scene Switcher
Control
multistep
Denoise
$\mathbf{s}, \mathbf{Z^x}$
$\mathcal{D}$
$\hat{\mathbf{Z}}_T^\mathbf{d}, \hat{\mathbf{Z}}_T^\mathbf{n}$
$\hat{\mathbf{Z}}_0^\mathbf{d}, \hat{\mathbf{Z}}_0^\mathbf{n}$

Fig. 2: The overall framework of GeoWizard. During fine-tuning, it first encodes the image $\mathbf{x}$, GT depth $\mathbf{d}$, and GT normal $\mathbf{n}$ through the original stable diffusion VAE $\boldsymbol{\varepsilon}$ into latent space, yielding latents $\mathbf{Z^x}$, $\mathbf{Z^d}$, and $\mathbf{Z^n}$ respectively. The two geometric latents are concatenated with $\mathbf{Z^x}$ to form two groups, $\mathbf{Z^x} \circ \mathbf{Z}_t^\mathbf{d}$ and $\mathbf{Z^x} \circ \mathbf{Z}_t^\mathbf{n}$. Each group is fed into the U-Net to generate the output in depth or normal domain in the guide of a geometry switcher. Additionally, the scene prompt $\mathbf{s}$ is introduced to produce results with one of three possible scene layouts (indoor/outdoor/object). During inference, given an image $\mathbf{x}$, a scene prompt $\mathbf{s}$, initial depth noise $\boldsymbol{\epsilon}_t^\mathbf{d}$ and normal noise $\boldsymbol{\epsilon}_t^\mathbf{n}$, GeoWizard can generate high-quality depth $\hat{\mathbf{d}}$ and normal $\hat{\mathbf{n}}$ jointly.

Unlike prior works that adopt CNN or transformer as architecture, we employ a diffusion-based scheme $f(\cdot)$ to model the joint depth and normal distribution $p(\mathbf{d}, \mathbf{n})$. A 3D asset $\mathbf{Z}$ possesses various attributes, such as albedo, roughness, and metalness, to describe its characteristics. We focus on depth and normal to represent the 3D spatial structure, approximating it to the distribution of a 3D asset $p_z \approx p(\mathbf{d}, \mathbf{n})$. Given a conditional input image $\mathbf{x}$, the depth map $\hat{\mathbf{d}}$ and the normal map $\hat{\mathbf{n}}$ can be obtained by the generative formulation $f(\cdot) : \mathbf{x} \in \mathbb{R}^3 \to (\hat{\mathbf{d}} \in \mathbb{R}^+, \hat{\mathbf{n}} \in \mathbb{R}^3)$, or in Markov probabilistic form:

$$f(\mathbf{x}) = p\left(\hat{\mathbf{d}}_T, \hat{\mathbf{n}}_T\right) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}\left(\hat{\mathbf{d}}_{t-1}, \hat{\mathbf{n}}_{t-1} \mid \hat{\mathbf{d}}_t, \hat{\mathbf{n}}_t, \mathbf{x}\right) \tag{1}$$

where $\hat{\mathbf{d}}_T, \hat{\mathbf{n}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

As shown in Fig. 2, the condition $\mathbf{x}$ is integrated into two ways: one is through the image embedding from CLIP [40] via cross-attention layers, and the other is by concatenating it in the latent space with geometric latents for more precise control. Our intuition is that the CLIP embeddings offer global-wise guidance, enhancing the model robustness and expressiveness under various Gaussian initialization, while the latent-wise concatenation further reduces randomness when generating $\hat{\boldsymbol{\epsilon}}_t^\mathbf{d}$ and $\hat{\boldsymbol{\epsilon}}_t^\mathbf{n}$. Our main challenge is to characterize the distribution $p_{\boldsymbol{\theta}}$ or specifically $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}$ to generate high-quality depth and normal maps.

3.2 Geometric Diffusion Model

We base our model on the pre-trained 2D latent diffusion model (Stable Diffusion [44]) so as to 1) utilize the strong, generalizable image priors learned from LAION-5B [50] 2) efficiently learn geometric priors in a low-dimensional latent space with minimum adjustments needed for U-Net architecture. However, this problem is non-trivial with two potential challenges: 1) the naive LDM is trained in the RGB domain, and thus may lack the capability to capture structural information and even impede it with reverse resistance. 2) The structure distributions are typically uniform, featuring similar values in localized areas, making them challenging for diffusion models to learn [25].

Joint Depth and Normal Estimation. To incorporate depth and normal for geometry estimation, one naive solution is to finetune two U-Nets ($f_{\mathbf{d}}$, $f_{\mathbf{n}}$) to model depth and normal distributions separately, i.e., $\hat{\mathbf{d}} = f_{\mathbf{d}}(\mathbf{x})$, $\hat{\mathbf{n}} = f_{\mathbf{n}}(\mathbf{x})$. However, this approach introduces extra parameters and overlooks the inherent connections between depth and normal, as both contribute to the unified geometric representation of a 3D shape. Normal describes surface variations and undulations , while depth outlines the spatial arrangement, guiding the orientation of normal. Our empirical experiment finds that this naive solution leads to geometric inconsistency in both depth and normal domain.

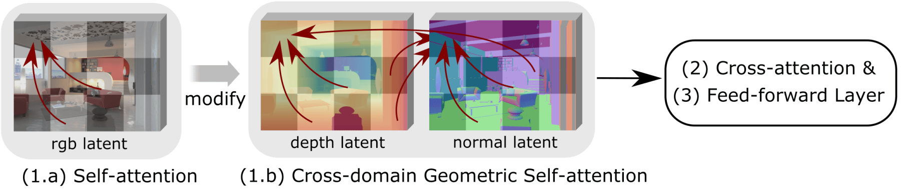

Fig. 3: The Structure of Geometric Transformer Block. Differing from the traditional self-attention layer (1.a) applied to RGB latent, we adapt it to a cross-domain geometric self-attention (1.b) that operates on depth latent and normal latent. This modification allows for mutual guidance and ensures geometric consistency.

Inspired by [31], we leverage a geometry switcher to enable a single stable diffusion model to generate depth or normal through indicators. Specifically, $\hat{\mathbf{d}} = f(\mathbf{x}, \mathbf{s_d})$, $\hat{\mathbf{n}} = f(\mathbf{x}, \mathbf{s_n})$, where $\mathbf{s_d}$ and $\mathbf{s_n}$ are one-dimensional vectors that control depth and normal domain, respectively. The switchers are encoded by the low-dimensional positional encoding and added with time embedding in the U-Net. We find that using switchers converges faster than shared modeling [28] or sequential modeling [31], and leads to more stable results.

To further enable mutual-guided geometric optimization, we modify the self-attention layer in U-Net to a cross-domain geometric self-attention layer to encourage spatial alignment, as shown in Fig. 3. This operator not only improves geometric consistency between depth and normal but also leads to faster convergence. We compute queries, keys, and values as follows:

$$\begin{aligned} \mathbf{q_d} = \mathbf{Q} \cdot \hat{\mathbf{z}}^{\mathbf{d}}, \mathbf{k_d} = \mathbf{K} \cdot (\hat{\mathbf{z}}^{\mathbf{d}} \oplus \hat{\mathbf{z}}^{\mathbf{n}}), \mathbf{v_d} = \mathbf{V} \cdot (\hat{\mathbf{z}}^{\mathbf{d}} \oplus \hat{\mathbf{z}}^{\mathbf{n}}) \\ \mathbf{q_n} = \mathbf{Q} \cdot \hat{\mathbf{z}}^{\mathbf{n}}, \mathbf{k_n} = \mathbf{K} \cdot (\hat{\mathbf{z}}^{\mathbf{n}} \oplus \hat{\mathbf{z}}^{\mathbf{d}}), \mathbf{v_n} = \mathbf{V} \cdot (\hat{\mathbf{z}}^{\mathbf{n}} \oplus \hat{\mathbf{z}}^{\mathbf{d}}) \end{aligned} \tag{2}$$

where $\hat{\mathbf{z}}^{\mathbf{d}}$ and $\hat{\mathbf{z}}^{\mathbf{n}}$ are latent depth and normal embeddings in transformer blocks, $\oplus$ denotes concatenation, and $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are query, key and value embeddings matrices. The cross-domain features are $\mathbf{Att}(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i), i = \{\mathbf{d}, \mathbf{n}\}$, where $\mathbf{Att}(\cdot)$ denotes softmax attention.

Scene Distribution Decoupler. As we explore diverse scenarios, we encounter situations where the estimated geometry shows a bias towards unintended layouts, leading to significant compression of foreground elements. This occurs because stable diffusion models may struggle with figuring out the correct spatial layouts of the captured scenes due to the varied spatial structures depicted in the training data. For example, outdoor scenes often feature an infinite depth range, indoor scenes have a constrained depth range and background-free objects exhibit even narrower depth ranges.

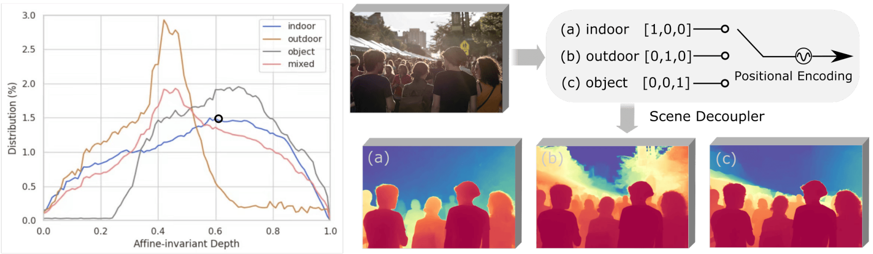

Fig. 4: Scene Distributions (left) and Decoupler Structure as Guider (right). We analyze the distributions of affine-invariant depth across three types of scenarios: indoor scenes, outdoor scenes, and background-free objects on our training dataset, where 'mixed' refers to the mixture of the three types. To clarify, the black circle dot indicates that the proportion of affine-invariant depth in [0.595, 0.605] is 1.5%. The Scene Decoupler encodes the one-hot domain vectors into positional embedding, which guides the stable diffusion to recognize the spatial layouts of different scene types.

A statistical analysis of scale-invariant depth distributions across different scene types is presented in Fig. 4, which shows that three types of scenes present different spatial structures. If we adopt Gaussian distribution to model the spatial layouts, the depth distributions of the outdoor, indoor and object scenarios have different means and variances (μ_1, σ_1^2), (μ_2, σ_2^2) and (μ_3, σ_3^2), respectively. The depth distribution of the mixed-up scenes tends to be a unified and neutralized distribution (red line) with $(\mu_1 + \mu_2 + \mu_3, \sigma_1^2 + \sigma_2^2 + \sigma_3^2)$. However, directly learning such a mixed distribution proves to be challenging.

To address the problem of layout ambiguity, we propose to learn the distinct three sub-distributions separately instead of directly learning the whole mixed distribution. To achieve this, we introduce a Scene Distribution Decoupler to guide the diffusion model toward learning different distributions. Specifically, $(\hat{\mathbf{d}}, \hat{\mathbf{n}}) = f(\mathbf{x}, \mathbf{s_i}), i = \{0, 1, 2\}$, where $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2$ denote the one-hot vectors of indoor, outdoor, object scene types, respectively. Resembling geometry switcher, these one-dimensional vectors are processed by positional encoding and are then element-wisely added to the time embedding.

Loss Function. We adopt multi-resolution noises [20,21] to preserve low-frequency details in the depth and normal maps, as similar values will frequently appear in local geometric regions. This deviation proves to be more efficient than a single-scale noise schedule. We perturb the two geometry branches with the same time-step scheduler to decrease the difficulty when learning more modalities. Finally, we utilize the v-prediction [46] as the learning objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\mathbf{d},\mathbf{n},\boldsymbol{\epsilon},t,s}[\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}\left(\mathbf{Z}_t^{\mathbf{d}}; \mathbf{x}, \mathbf{s_d}, \mathbf{s}_i\right) - \mathbf{v}_t^{\mathbf{d}}\right\|_2^2 + \|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}\left(\mathbf{Z}_t^{\mathbf{n}}; \mathbf{x}, \mathbf{s_n}, \mathbf{s}_i\right) - \mathbf{v}_t^{\mathbf{n}}\|_2^2] \tag{3}$$

where $\mathbf{v}_t^{\mathbf{d}} = \alpha_t \boldsymbol{\epsilon}_t^{\mathbf{d}} - \sigma_t \mathbf{Z}^{\mathbf{d}}$ and $\mathbf{v}_t^{\mathbf{n}} = \alpha_t \boldsymbol{\epsilon}_t^{\mathbf{n}} - \sigma_t \mathbf{Z}^{\mathbf{n}}$; $\boldsymbol{\epsilon}_t^{\mathbf{d}}$ and $\boldsymbol{\epsilon}_t^{\mathbf{n}}$ are two Gaussian noises independently sampled from multi-scale noise sets for depth and normal, respectively. The unified denoising network $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}$ with annealed noise scheduler generates the desired geometry noises conditioned by hierarchical switchers $(\mathbf{s_d}, \mathbf{s_n}, \mathbf{s}_i)$ and input image $\mathbf{x}$.

4 Experiment

4.1 Implementation Details and Datasets

Implementation Details. We finetune the whole U-Net from the pre-trained Stable Diffusion V2 Model [44], which has been finetuned with image conditions. We use an image size of 576 × 768 and train the model for 20,000 steps with a total batch size of 256. This entire training procedure typically requires 2 days on a cluster of 8 Nvidia Tesla A100-40GB GPUs. We use the Adam optimizer with a learning rate of 1×10^{-5}. Additionally, to enhance dataset diversity, we apply random horizontal flipping, crop, and photometric distortion (contrast, brightness, saturation, and hue) to the 2D image collection during training.

Training Datasets. We train our model on three categories: 1)Indoor: *Hypersim* [43] is a photorealistic synthetic dataset with 461 indoor scenes. We filter out 191 scenes without tilt-shift photography. We further cull out incomplete images and finally obtain 25,463 samples. *Replica* [57] is a dataset of high-quality reconstructions of 18 indoor spaces. We filter out 50,884 samples with complete context. 2)Outdoor: *3D Ken Burns* [36] provides a large-scale synthetic dataset with 76,048 stereo pairs in 23 in-the-wild scenes. We further incorporate 39,630 synthetic city samples in 1440×3840 high resolutions from our own simulation platform. The normal GT is derived from the depth maps. (See Supp. for visualization) 3)Background-free Object: *Objaverse* [7,39] is a massive dataset of over 10 million 3D objects. We filter out 85,997 high-quality objects as training data.

4.2 Evaluation

Evaluation Datasets. We assess our model's efficacy across six zero-shot relative depth benchmarks, including NYUv2 [54], KITTI [11], ETH3D [49], ScanNet [6], DIODE [58], and OmniObject3D [60]. For surface normal estimation, we employ in-total five benchmarks on NYUv2 [38,54], ScanNet [6,17], iBim-1 [2,22], DIODE-outdoor [58], and OmniObject3D [60] for zero-shot evaluation.
Baselines. For affine-invariant depth estimation, we select baselines from state-of-the-art methods that demonstrate generalizability through training on diverse datasets. These methods are specialized in predicting either depth (DiverseDepth [65], LeReS [67], HDN [70], Marigold [21]) or disparity (MiDaS [42], DPT [41], Omnidata [9]). For surface normal estimation, the field has seen fewer works [9, 19, 71] addressing zero-shot estimation specifically. Hereby, We choose both SoTA in-domain (EENSU [1]) and zero-shot methods (Omnidata v1 [9], v2 [19], and the ultra-recent DSINE [3]) as the baselines.

Metrics. Building upon prior research [66], we assess the performance of depth estimation methods using the absolute relative error (AbsRel) and accuracy within a threshold $\delta^{\mathbf{1}} = 1.25$. For surface normal estimation, we evaluate using the Mean angular error and accuracy within $11.25°$, aligning with established methods [1]. We evaluate Geometric Consistency (GC) between depth and normal as follows: we first estimate the pseudo scale and shift of the estimated depth using GT depth, and then convert the estimated depth into metric depth. We calculate the Mean angular error of the normal difference between predicted normal and normal calculated from the metric depth to evaluate the consistency between estimated depth and normal.

4.3 Comparison

Depth Estimation. We present the quantitative evaluations of zero-shot affine-invariant depth in Table 1. DepthAnything [62] achieves the best quantitative numbers across three real datasets but presents a significant performance drop on unreal images (see Fig. 5 and Fig. 6). This may be because although DepthAnything is trained on 63.5M images, its discriminative nature limits its ability to generalize on images that significantly differ from training images. On the other hand, its results fail to capture rich geometric details. Compared to the robust depth estimator Marigold [21], GeoWizard shows more correct foreground-background relationships, especially in outdoor scenarios.

Normal Estimation. We present the quantitative evaluations of surface normal in Table 2, where our method achieves superior performance. When compared with the SoTA normal approach DSINE [3], our method recovers much finer-grained details and is more robust to unseen terrain in the Fig. 5. We further provide more out-of-domain comparisons in Fig. 6, where *GeoWizard* surprisingly generates astonishing details and correct spatial structures. DSINE [3]

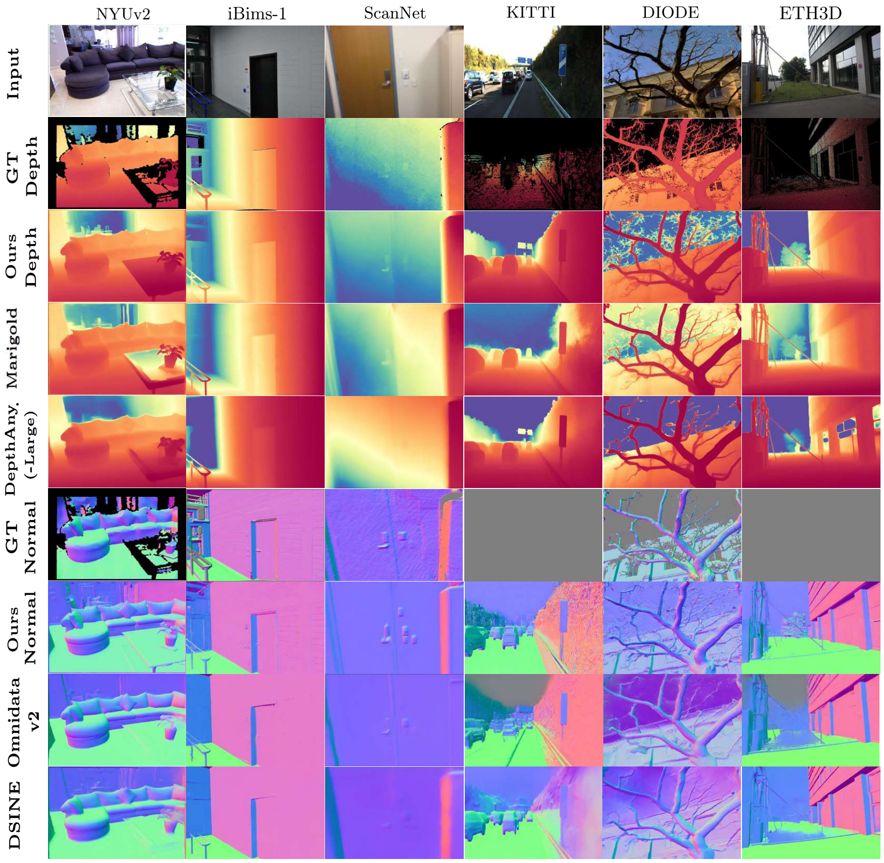

Fig. 5: Qualitative comparison on zero-shot depth and normal benchmarks.

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE-Full		OmniObject3D	
	AbsRel ↓	δ1 ↑	AbsRel ↓	δ1 ↑	AbsRel ↓	δ1 ↑	AbsRel ↓	δ1 ↑	AbsRel ↓	δ1 ↑	AbsRel ↓	δ1 ↑
DiverseDepth [65]	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	-	-
MiDaS [42]	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	-	-
LeReS [67]	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	-	-
Omnidata v2 [19]	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	3.0	**99.9**
HDN [70]	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	78.0	-	-
DPT [41]	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	**18.2**	75.8	-	-
Metric3D [66]	5.8	96.3	**5.8**	**97.0**	6.6	96.0	7.4	94.1	22.4	78.5	-	-
DepthAnything [62]	**4.3**	**98.1**	7.6	94.7	12.7	88.2	**4.2**	**98.0**	27.7	75.9	1.8	**99.9**
Marigold [21]	5.5	96.4	9.9	91.6	6.5	96.0	6.4	95.1	30.8	77.3	3.0	99.8
GeoWizard (Ours)	5.2	96.6	9.7	92.1	**6.4**	**96.1**	6.1	95.3	29.7	**79.2**	**1.7**	**99.9**

Table 1: Quantitative comparison on 6 zero-shot affine-invariant depth benchmarks. We mark the best results in bold and the second best underlined. Discriminative methods are colored in blue while generative ones in green. Please note that DepthAnything is trained on 63.5M images while ours is only trained on 0.28M images.

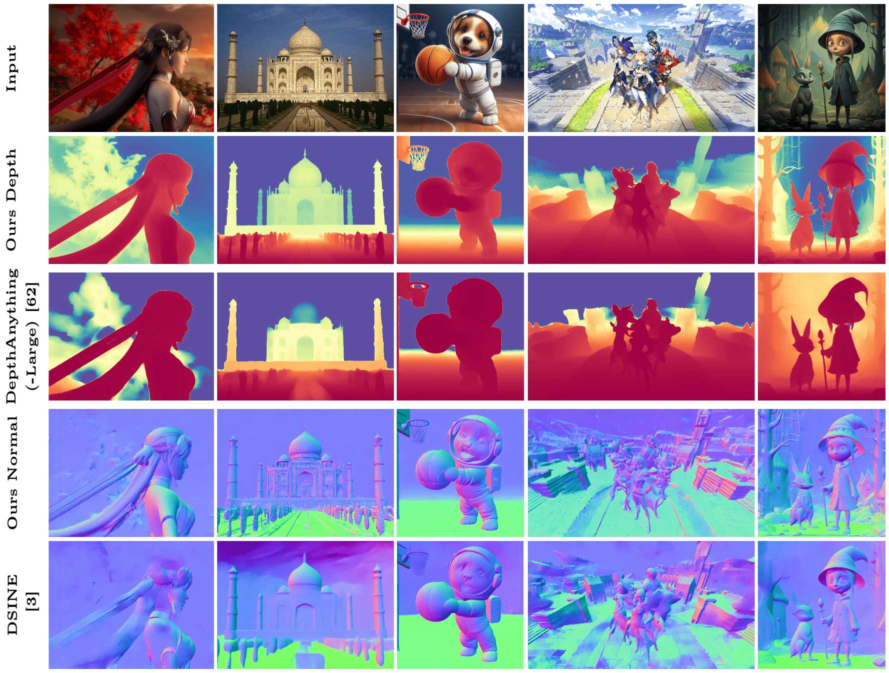

Fig. 6: Geometry comparison on in-the-wild collections. As discriminative models, DepthAnything and DSINE show significant performance drop on in-the-wild images, especially for the unreal images that are barely included in the collected training datasets. Please check more examples in the supplementary materials.

can recover rough shape, but it struggles to produce high-frequency details, such as hairline, architectural texture, and limbs.

Method	NYUv2		ScanNet		iBims-1		DIODE-outdoor		OmniObject3D	
	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑
EESNU [1]	**16.2**	58.6	-	-	20.0	58.5	29.5	26.8	31.9	18.8
Omnidata v1 [9]	23.1	45.8	22.9	47.4	19.0	62.1	22.4	38.4	23.1	42.6
Omnidata v2 [19]	17.2	55.5	16.2	60.2	18.2	63.9	20.6	40.6	21.4	46.1
DSINE [3]	16.4	**59.6**	16.2	61.0	17.1	**67.4**	**19.3**	**44.1**	21.7	45.1
GeoWizard (Ours)	17.0	56.5	**15.4**	**61.6**	**13.0**	65.3	20.6	38.9	**20.8**	**47.8**

- : EENSU [1] is trained on ScanNet, thus the in-domain performance is omitted.

Table 2: Quantitative comparison across 5 zero-shot surface normal benchmarks.

4.4 Ablation Study

We collect zero-shot validation sets that incorporate depth and normal from three scene distributions - the official test split of NYUv2 [54], consisting of

654 images, and 138 high-quality samples from ScanNet [6] for indoor domain; the 432 in-the-wild samples from our simulation platform and filtered 86 images from DIODE [58] for outdoor domain; 300 randomly selected real-world samples (over 40 categories) of OmniObject3D [60] for object domain.

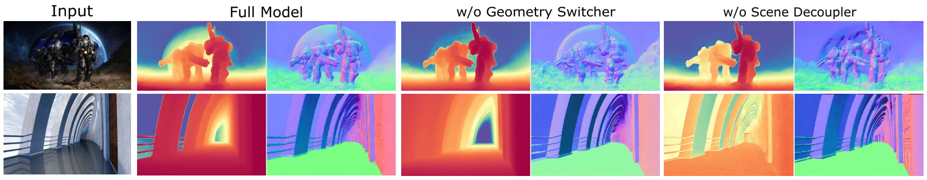

Fig. 7: Qualitative ablation. The geometric consistency decreases a lot, especially in far regions, when removing the cross-domain geometry switcher. Without the distribution decoupler, the estimated depth and normal mistakenly perceive the spatial layouts of the input images, like the Earth in the first row and the Sky in the second row.

Joint Depth and Normal Estimation. We first investigate the effect of the geometry switcher. When removing the cross-domain geometry switcher (w/o Geometry Switcher), the overall geometric consistency drops significantly (16.2→18.1, also as illustrated in Fig. 7), verifying that cross-domain self-attention effectively correlates the two representations. We also train two diffusion models to separately learn depth and normal (Separate models), but this significantly reduces the performance across all evaluated metrics.

Method	Indoor			Outdoor			Object			Overall		
	AbsRel ↓	Mean ↓	GC ↓	AbsRel ↓	Mean ↓	GC ↓	AbsRel ↓	Mean ↓	GC ↓	AbsRel ↓	Mean ↓	GC ↓
Separate models	7.4	15.1	18.2	12.5	26.2	27.9	5.2	18.2	20.1	8.5	16.9	19.1
w/o Geometry Switcher	5.7	13.1	17.3	9.8	22.3	27.1	**3.3**	15.8	18.5	6.9	15.0	18.1
w/o Scene Decoupler	5.8	13.8	15.4	10.5	24.7	24.5	3.7	15.5	17.9	7.5	16.1	16.5
Full Model	**5.5**	**12.6**	14.7	**9.6**	**22.1**	**23.5**	3.5	**15.4**	**17.6**	**6.7**	**14.8**	**16.2**

Table 3: Quantitative ablation on geometry estimation across three types of scenarios.

Decoupling Scene Distributions. As we decouple the complex scene distribution into several sub-domains, GeoWizard can concentrate on a specific domain during in-the-wild inference. Therefore, it is not surprising that removing the decouple (w/o Scene Decoupler) leads to a performance drop across all domains (visually shown in Fig. 7). Interestingly, the impact on the object domain is minimal, suggesting that object-level distribution is simpler to learn.

4.5 Application

GeoWizard enables a wide range of downstream applications, including 3D reconstruction, novel view synthesis, and 2D content creation.

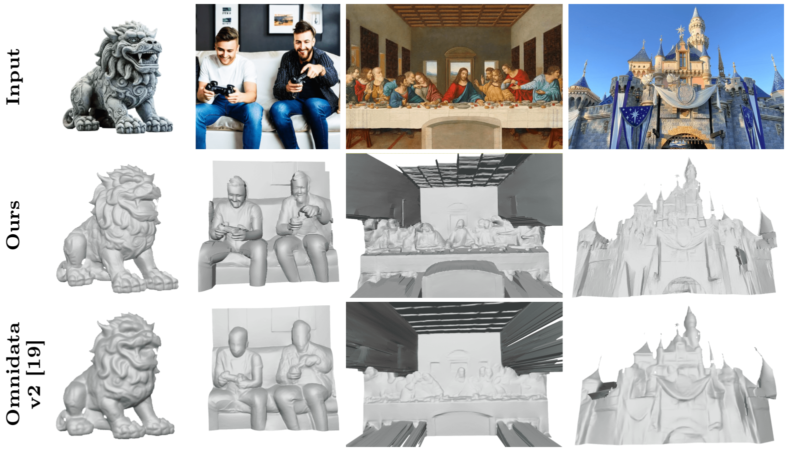

Fig. 8: Geometry comparison on different scene domains. Ours consistently achieves more fine-grained details and spatial structure over Omnidata v2.

3D Reconstruction with Geometric Cues. We can leverage the monocular geometric cues for surface reconstruction. Using BiNI algorithm [5], we can extract the 3D mesh directly. In Fig. 8, compared to Omnidata v2 [19], GeoWizard consistently generates finer details with higher fidelity and frequency detail (See the beard of the stone lion, and the two men's faces) and more accurate 3D spatial structure (see the Last Supper). Additionally, we can condition these geometric cues to help surface reconstruction method [69] to generate high-quality geometry. We conduct experiments on 4 scenes from ScanNet and employ evaluation metrics following [14,35,69]. Table 4 illustrates that our geometric guidance surpasses previous methods, particularly in terms of "Recall" and "F-score".

Geometric Cues	Acc↓	Comp↓	C-$\mathcal{L}_1$ ↓	Prec↑	Recall ↑	F-score↑
Omnidata v2 [19]	0.035	0.048	0.042	79.9	68.1	73.3
DSINE [3]	0.036	0.045	0.040	**80.1**	70.2	74.7
GeoWizard (Ours)	**0.033**	**0.042**	**0.038**	80.0	**70.7**	**75.1**

Table 4: Geometric guidance used for MonoSDF [69] on the ScanNet [6] dataset.

Depth-aware Novel View Synthesis. We can utilize the depth cue generated by our model to enhance depth-based inpainting methods [53]. As shown in Fig. 9, compared to Midas V3.1 [42], GeoWizard achieves better novel view synthesis results and enables more realistic 3D photography effect.

2D Content Generation. We adopt depth/normal conditioned ControlNet [72] (SD 1.5) that takes spatial structure as input to evaluate the geometry indirectly. As depicted in Fig. 10, the generated color images conditioned by our depth and

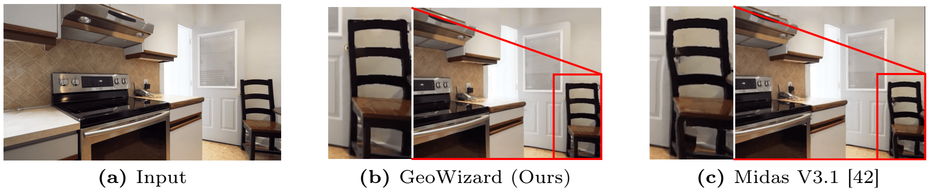

Fig. 9: Novel view synthesis comparison. GeoWizard guides the [53] to generate more coherent and plausible structures like the thin chair legs and doorways.

normal are more semantically coherent to the original input image. However, the generated images conditioned on depth map of DepthAnything [62] and normal map of DSINE [3] fail to keep similar 3D structures with the input image.

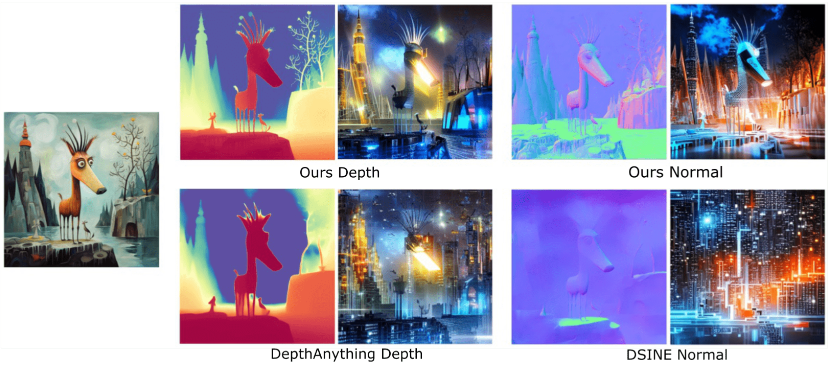

Fig. 10: Images generated by ControlNet conditioned on estimated depth maps and normal maps using text prompt *"futuristic technology"*.

5 Conclusion

In this work, we present *GeoWizard*, a holistic diffusion model for geometry estimation. We distill the rich knowledge in the pre-trained stable diffusion to boost the task of high-fidelity depth and normal estimation. Using the proposed geometry switcher, *GeoWizard* jointly produces depth and normal using a single model. By decoupling the mixed and sophisticated distribution of all scenes into several distinct sub-distributions, our model could produce 3D geometry with correct spatial layouts for various scene types. In the future, we plan to decrease the number of denoising steps to speed up the inference of our method. The latent consistency models [34] may be leveraged to train a few-step diffusion model so that the inference time may be decreased to less than 1 second.

Acknowledgments

We thank the experimental help from Yichong Lu and valuable suggestions from Yuwei Guo. Xiao Fu is supported by the Hong Kong PhD Fellowship Scheme (HKPFS) 2023-2024.

References

1. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: ICCV (2021)
2. Bae, G., Budvytis, I., Cipolla, R.: Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. In: BMVC (2022)
3. Bae, G., Davison, A.J.: Rethinking inductive biases for surface normal estimation. In: CVPR (2024)
4. Bhattad, A., McKee, D., Hoiem, D., Forsyth, D.: Stylegan knows normal, depth, albedo, and more. In: NeurIPS. vol. 36 (2023)
5. Cao, X., Santo, H., Shi, B., Okura, F., Matsushita, Y.: Bilateral normal integration. In: ECCV (2022)
6. 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. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR (2023)
8. Dong, Q., Zhao, B., Fu, Y.: Open-ddvm: A reproduction and extension of diffusion model for optical flow estimation. arXiv.org (2023)
9. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: ICCV (2021)
10. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013)
12. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
13. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
14. Guo, H., Peng, S., Lin, H., Wang, Q., Zhang, G., Bao, H., Zhou, X.: Neural 3d scene reconstruction with the manhattan-world assumption. In: CVPR (2022)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
16. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: WACV (2019)
17. Huang, J., Zhou, Y., Funkhouser, T., Guibas, L.J.: Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In: ICCV (2019)
18. 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. arXiv.org (2023)
19. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3d common corruptions and data augmentation. In: CVPR (2022)
20. Kasiopy: https://wandb.ai/johnowhitaker/multires_noise/reports/multi-resolution-noise-for-diffusion-model-training–vmlldzoznjyyotu2?s=31. arXiv.org (2023)

21. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024)
22. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: ECCVW (2018)
23. Kusupati, U., Cheng, S., Chen, R., Su, H.: Normal assisted stereo depth estimation. In: CVPR (2020)
24. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: CVPR (2015)
25. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: WACV (2024)
26. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
27. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023)
28. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. In: ICLR (2024)
29. Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., Wang, W.: Neural rays for occlusion-aware image-based rendering. In: CVPR (2022)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
31. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv.org (2023)
32. Long, X., Liu, L., Theobalt, C., Wang, W.: Occlusion-aware depth estimation with adaptive normal constraints. In: ECCV (2020)
33. Long, X., Zheng, Y., Zheng, Y., Tian, B., Lin, C., Liu, L., Zhao, H., Zhou, G., Wang, W.: Adaptive surface normal constraint for geometric estimation from monocular images. arXiv.org (2024)
34. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv.org (2023)
35. Lyu, X., Dai, P., Li, Z., Yan, D., Lin, Y., Peng, Y., Qi, X.: Learning a room with the occ-sdf hybrid: Signed distance function mingled with occupancy aids scene representation. In: ICCV (2023)
36. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. In: ACM TOG (2019)
37. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
38. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: CVPR (2018)
39. Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L., Han, X.: Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. arXiv.org (2023)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
41. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)

42. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In: IEEE TPAMI (2022)
43. 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)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
46. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv.org (2022)
47. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single real image. arXiv.org (2023)
48. 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)
49. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
50. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
51. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. In: ICCV (2019)
52. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv.org (2023)
53. Shih, M.L., Su, S.Y., Kopf, J., Huang, J.B.: 3d photography using context-aware layered depth inpainting. In: CVPR (2020)
54. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
55. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
57. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv.org (2019)
58. 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. CoRR (2019)
59. Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., Wang, W.: Neuris: Neural reconstruction of indoor scenes using normal priors. In: ECCV (2022)
60. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: CVPR (2023)
61. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: CVPR (2018)

62. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
63. Yin, W., Liu, Y., Shen, C.: Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction (2021)
64. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: ICCV (2019)
65. Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., Renyin, D.: Diversedepth: Affine-invariant depth prediction using diverse data. arXiv.org (2020)
66. 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 (2023)
67. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: CVPR (2021)
68. Yu, Y., Smith, W.A.: Inverserendernet: Learning single image inverse rendering. In: CVPR (2019)
69. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In: NeurIPS (2022)
70. Zhang, C., Yin, W., Wang, B., Yu, G., Fu, B., Shen, C.: Hierarchical normalization for robust monocular depth estimation. In: NeurIPS (2022)
71. Zhang, J., Li, S., Lu, Y., Fang, T., McKinnon, D., Tsin, Y., Quan, L., Yao, Y.: Jointnet: Extending text-to-image diffusion for dense distribution modeling. In: ICLR (2024)
72. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
73. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: CVPR (2019)
74. Zhao, W., Liu, S., Wei, Y., Guo, H., Liu, Y.J.: A confidence-based iterative solver of depths and surface normals for deep multi-view stereo. In: ICCV (2021)
75. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: ICCV (2023)