Video Generative Models as Geometry Learner

Haosen Yang[1], Jifei Song[1], Zhensong Zhang[2],
Xiatian Zhu[1⋆], and Jiankang Deng[3⋆]

[1] University of Surrey, Guildford, UK
[2] Independent Researcher
[3] Imperial College London, London, UK

Fig. 1: We present GeoNeXt, a unified geometry estimation model for both depth and surface normals, repurposed from a off-the-shelf generative video model. Our idea is to leverage the temporal coherence and rich priors of a pretrained video generative model, adapting them for image–geometry joint modeling. The predicted geometry provides strong structural cues that enable diverse applications, e.g., (1) controllable image generation, (2) image relighting, and (3) mesh reconstruction.

Abstract. Recent generative approaches to geometry estimation adapt pretrained image diffusion models and treat the task as image-conditioned generation. Leveraging off-the-shelf image diffusion models, they either (i) train task-specific geometry models (for depth and surface normal estimation) independently, losing the opportunity of exploring the intrinsic correlation of these geometric targets, or (ii) jointly fine-tune modified image diffusion backbones (e.g., altered self-attention), which typically demands substantial labeled data. To overcome these limitations in a principled fashion, we repurpose pretrained video generative models as a unified and data-efficient framework for geometry estimation, formulated innovatively as a *next-frames prediction* task. Our method, **GeoNeXt**, inherits naturally structured knowledge and richer priors from the video model, while further adapting them for joint modeling of images and geometry targets (image ↔ geometry) enabling more data efficient and effective learning of geometry. Extensive experiments validate our method

⋆ Corresponding authors.

for zero-shot monocular depth and surface normal estimation across diverse datasets, outperforming both previous task-specific and unified generative competitors while using substantially less training data. Notably, our method rivals discriminative state-of-the-art approaches trained on over 100× more data and even standouts on several benchmarks.

1 Introduction

Monocular 3D geometry estimation (e.g., depth and surface normal) is a fundamental problem in 3D vision, with applications spanning autonomous driving [18, 19], 3D surface reconstruction [63], and inverse rendering [30, 58]. However, recovering underlying 3D geometry from a single image remains challenging, as it requires comprehensive and precise geometric reasoning. Early discriminative approaches typically relied on supervised or semi-supervised learning with large collection of paired RGB images and depth maps. To this end, recent efforts [3, 13, 59, 60] focus on building data engines that leverage ever-growing amounts of unlabeled data to improve generalization. While straightforward, these approaches demand substantial compute and are sensitive to pseudo-label noise, leading to predictions lacking fine-grained, high-frequency detail.

More promisingly, several works [16, 29] have leveraged the priors of pretrained text-to-image (T2I) generative models (e.g., Stable Diffusion [43]) for zero-shot 3D geometry (e.g., depth and normal) estimation, attaining competitive performance with limited task-specific supervision. Within this paradigm, 3D geometry is typically treated as a "special image," and the problem is formulated as image-conditioned generation. Prior efforts [16, 17, 20, 22, 29, 32, 52] can be largely categorized into two tracks. *The first family* (e.g., [29]) trains task-specific models independently with minimal architectural modifications, thereby best leveraging pretrained capacity under limited supervision. However, this design precludes parameter sharing and fails to exploit inter-task dependencies, necessitating a distinct model for each geometric target.

The second family (e.g., [16]) avoids maintaining separate models, opting instead for a joint model that predicts both depth and normals by modifying the architecture of diffusion model (e.g., adding cross-attention or switcher modules). However, these architectural changes substantially increase data requirements, because they widen the gap between pretraining and finetuning, reducing transfer efficiency. Both families reuse *image diffusion models* for geometry estimation without explicitly modeling image–geometry interactions. We conjugate this strategy could lead to cross-modal inconsistencies and degraded geometric fidelity, with reduced clarity and loss of fine detail.

To address these limitations, we depart from image-generation models and explore pretrained video generative models to monocular geometry estimation. Our intuition is as follows: (**i**) Video diffusion models are pretrained on large-scale video datasets, in addition to images, yielding richer generative priors that enable high-quality synthesis across a wide range of domains. (**ii**) Their architectures and pretraining impart temporal-attention priors that capture cross-frame

dependencies, which we leverage for joint modeling between images and geometry targets (image $\leftrightarrow$ geometry).

(**iii**) Even without explicit geometric supervision during pretraining, such priors can be adapted to predict geometry [25, 27, 48, 61].

Building on this intuition, we propose **GeoNeXt**, a unified and data-efficient framework that reformulates geometry estimation as *next-frame prediction* in a video diffusion model, enabling coherent joint modeling of images and geometric modalities such as depth and surface normals. Given an RGB image, we cast geometry prediction as *next-frame* generation: the target geometry (depth and normals) is modeled as the subsequent frame conditioned on RGB image. Following Stable Video Diffusion [6], we replicate the input image into the geometry slots (depth and normals) to provide a strong, consistent conditioning signal for the geometry objective. Crucially, we preserve the pretrained model's native generation order and synthesize the image and its geometry in lockstep. Rather than training a geometry-only generator, GeoNeXt co-generates image and geometry along a shared denoising trajectory, which preserves fine-grained detail and enhances image–geometry consistency. Our adaptation is lightweight: we fine-tune only the denoising U-Net with minimal architectural changes (e.g., removing the text/CLIP branch), training on modest RGB–depth–normal triplets from datasets such as *Hypersim* and *Virtual KITTI* (e.g., Hypersim [42], Virtual KITTI [18]). Leveraging video-diffusion priors, GeoNeXt exhibits strong zero-shot generalization, effectively transferring image-to-video generative capability to image–geometry prediction without requiring large additional datasets.

Our **contributions** are: (1) We introduce GeoNeXt, a novel generative approach for jointly estimating depth and surface normals, which repurposes a video generative model as a unified geometry learner, formulated as a next-frames prediction task. (2) We comprehensively study the optimal fine-tuning protocol for adapting the pretrained video generative model to geometry estimation. In particular, we analyze GeoNeXt's generative formulation, architectural design, and the influence of reconstruction order on model robustness and overall performance. (3) Extensive experiments validate the effectiveness of GeoNeXt for zero-shot monocular depth and surface normal estimation across diverse datasets, outperforming both task-specific and unified generative competitors even with substantially less training data, while competitively rivaling discriminative models trained with even larger data and compute.

2 Related Work

2.1 Monocular depth and normal estimation

Estimating depth and surface normals from a single image is an ill-posed but fundamental problem, as both capture complementary information about 3D scene geometry. Early approaches relied heavily on comprehensive scene understanding and were typically limited to specific domains [1, 2, 14, 15, 33, 36, 55, 64, 65]. Estimating depth and surface normals in unconstrained environments requires

models capable of generalizing beyond the training distribution, which remains a difficult challenge in practice.

To improve generalization, recent depth estimation works [35, 41, 59, 60] have focused on collecting large and diverse datasets and developing training strategies that leverage them effectively. Since metric depth depends on camera intrinsics, most approaches, including ours, predict affine-invariant depth, which is consistent across different imaging setups. In contrast, models that attempt to estimate absolute metric depth are often limited to fixed camera configurations [14] or must explicitly condition on known intrinsic parameters [5, 21]. A similar trend can be observed in monocular surface normal estimation, where large-scale pretrained models [2, 3, 13] have become the dominant paradigm, while others [3] further improve accuracy by incorporating task-specific inductive biases designed for normal estimation.

To avoid maintaining separate models, several studies have explored depth and surface normal estimation as a multi-task learning problem [24, 34, 56, 65]. These methods typically employ multiple task-specific decoder branches while sharing a common backbone to enable information exchange in-between. Although this design allows for joint estimation, the models remain discriminative. Moreover, existing approaches rely on limited training datasets and exhibit poor generalization, often missing fine geometric details in challenging cases. In contrast, our method tackles depth and surface normal joint estimation without large-scale annotated datasets. Instead, we repurpose the broad priors learned by pretrained video generative models and reformulate the task as next-frame prediction, enabling more efficient and effective geometry learning.

2.2 Diffusion Models for Geometry Estimation

Recently, diffusion models [23, 28, 51] have demonstrated remarkable capabilities in both image and video generation, producing high-quality results across a wide range of domains [4, 6, 7, 37, 38, 43, 57]. Several studies have explored extending diffusion models to geometry-related tasks, including depth estimation and surface normal prediction.

Recent attempts [26, 45, 46, 66] introduced diffusion-based frameworks for geometric prediction in collaboration with discriminative pipelines, achieving promising results but remaining limited to their specific training domains. More recent efforts [12, 17, 20, 22, 29, 62] have explored leveraging pretrained latent diffusion models (LDMs) for geometry estimation tasks, particularly for depth and surface normal prediction across diverse real-world settings, often requiring only modest amounts of fine-tuning data. Within this paradigm, 3D geometry (e.g., depth and surface normals) is typically represented as a "specialized image," framing the problem as image-conditioned generation. However, existing methods still train independent, task-specific models for each prediction objective, which limits scalability and cross-task consistency.

While GeoWizard [16] and Orchid [32] attempt joint depth–normal prediction by modifying the diffusion model architecture or retraining the entire model, including the VAE, these designs substantially increase data requirements

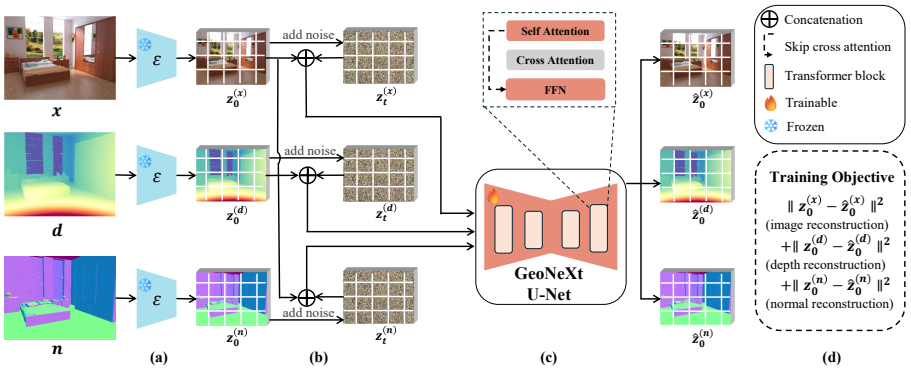

Fig. 2: Overview of the GeoNeXt *fine-tuning* protocol. (a) We start from the pretrained Stable Video Diffusion model and encode the RGB image $\mathbf{x}$, depth $\mathbf{d}$, and surface normal $\mathbf{n}$ into the latent space using the frozen Stable Diffusion VAE. (b) Noise is sampled and added to the latent $\mathbf{z}$. The image latent is replicated across the geometry slots (depth and normal), and further conditioned by concatenating the geometry latents. (c) We fine-tune only the GeoNeXt U-Net. (d) The model is optimized with the standard diffusion objective over image, depth, and normal latents to ensure fine-grained alignment and consistency between image and geometry.

and widen the gap between pretraining and fine-tuning. Inspired by monocular image-based geometry estimation, several works [25, 27, 48, 48] extend this paradigm to the video domain, leveraging pretrained video diffusion models to enhance temporal consistency and mitigate flickering artifacts across frame sequences. Our approach moves beyond these efforts by exploring the potential of pretrained video generative models for joint geometry estimation. **GeoNeXt** adapts the temporal priors learned by video generative models for unified modeling between images and geometry targets (image $\leftrightarrow$ geometry), enabling more efficient and effective geometry learning.

3 Method

In this work, we adapt a pretrained video diffusion model for monocular geometry estimation jointly, leveraging its strong generative priors learned from large-scale, high-quality data. Given an input image $\mathbf{x}$, our objective is to predict its corresponding geometric information—namely, a depth map $\mathbf{d}$ and a surface-normal map $\mathbf{n}$. Section 3.1 provides a brief overview of the notation for image and video diffusion, followed by a detailed description of GeoNeXt in Section 3.2. Finally, we describe the inference strategy in Section 3.3. An overview of **GeoNeXt** training is presented in Fig. 2.

3.1 Preliminaries of Diffusion Models

Diffusion models [23,43,50,51] model a data distribution via (i) a forward noising process that gradually maps data to a simple reference distribution (e.g., Gaussian), and (ii) a learned reverse process that denoises samples back to the data manifold. In *latent* image diffusion, one first maps an image $\mathbf{x}_0$ to a compact latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ to reduce computation. The forward (Markov) process then perturbs $\mathbf{z}_0$ according to a noise schedule $\{\beta_t\}_{t=1}^{T}$:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) \;=\; \mathcal{N}\big(\mathbf{z}_t \,\big|\, \sqrt{1-\beta_t}\,\mathbf{z}_{t-1},\, \beta_t \mathbf{I}\big), \tag{1}$$

and the reverse process uses a learned denoiser $\mathcal{Z}_\theta = (\mu_\theta, \Sigma_\theta)$ to progressively reconstruct:

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) \;=\; \mathcal{N}\big(\mathbf{z}_{t-1} \,\big|\, \mu_\theta(\mathbf{z}_t, t),\, \Sigma_\theta(\mathbf{z}_t, t)\big). \tag{2}$$

Prior works [22,29] formulates monocular geometry estimation as *image–conditioned* diffusion: a geometry target: depth $\mathbf{d}$ or normals $\mathbf{n}$, is encoded to a latent, noise is added, and the RGB image provides conditioning during denoising (often by concatenation).

In contrast, we build on the video diffusion model Stable Video Diffusion (SVD) [6], which denoises a compact latent sequence and follows the EDM formulation [28]. In EDM, the forward process adds i.i.d. Gaussian noise with variance σ_t^2 to a clean sample $\mathbf{z}_0 \sim p(\mathbf{z})$:

$$\mathbf{z}_t \;=\; \mathbf{z}_0 + \sigma_t\, \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $\mathbf{z}_t \sim p(\mathbf{z}; \sigma_t)$ denotes the noised sample at noise level σ_t. As $\sigma_t \to \sigma_{\max}$, the sample distribution becomes effectively Gaussian; generation then proceeds by progressively denoising from $\sigma_{\max}$ to $\sigma_0 = 0$. EDM adopts a *preconditioned* denoiser to maintain stability across noise levels [28, 44]. Rather than predicting directly from $\mathbf{z}_t$, the model combines the noisy input with the network output using noise–dependent weights:

$$\begin{aligned} \mathcal{Z}_\theta(\mathbf{z}_t, \sigma_t, \mathbf{c}) = c_{\text{skip}}(\sigma_t)\, \mathbf{z}_t \\ + c_{\text{out}}(\sigma_t)\, F_\theta\left(c_{\text{in}}(\sigma_t)\, \mathbf{z}_t,\; c_{\text{noise}}(\sigma_t),\; \mathbf{c}\right) \end{aligned} \tag{4}$$

where F_θ is a learnable network (e.g., a UNet), and $c_{\text{in}}, c_{\text{out}}, c_{\text{skip}}, c_{\text{noise}}$ are scalar functions of σ_t that control input scaling, output scaling, the skip path, and the noise embedding.

3.2 Geometry as Next Frames

Generative Formulation To jointly estimate depth and surface normals, we formulate monocular geometry prediction as an image-conditioned, image-to-video diffusion problem, where geometry ($\mathbf{d}$ for depth and $\mathbf{n}$ for surface normals) is modeled as the subsequent frame(s) following the input image. The model learns the conditional distribution $p(\mathbf{g} \mid \mathbf{x})$, where $\mathbf{g} = \{\mathbf{d}, \mathbf{n}\}$ and $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$.

We define the clean *image–geometry triplet*

$$\mathcal{G} = (\mathbf{x},\, \mathbf{d},\, \mathbf{n}). \tag{5}$$

The triplet component is corrupted to $\mathcal{G}_t$ by adding Gaussian noise according to Eq. (3), while $\mathbf{x}$ is replicated and kept fixed as the conditioning signal. During the reverse process, the conditional denoiser $\mathcal{Z}_\theta$ iteratively removes noise, mapping $\mathcal{G}_t \mapsto \mathcal{G}_{t-1}$ conditioned on $\mathbf{x}$, while simultaneously denoising the image to maintain consistency between the image and geometry. Additional conditioning is introduced by concatenating the latent image features with the noisy geometry input before feeding them into the denoiser.

Following EDM [28], training minimizes the standard noise-prediction objective using the preconditioned parameterization in Eq. (4). At inference time, $(\mathbf{d}, \mathbf{n})$ are reconstructed by first initializing with Gaussian noise $\mathbf{g}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then progressively denoising it through iterative applications of $\mathcal{Z}_\theta(\mathbf{g}_t, \mathbf{x}, t)$ from timestep T down to 0.

Latent space transformation Instead of operating in pixel space, we follow *latent diffusion models* (LDMs) [43], which perform the diffusion process in a compact, low-dimensional latent space. This offers substantial computational savings and scales to high-resolution synthesis. The latent space is provided by the bottleneck of a VAE trained independently of the denoiser, yielding representations that are compact yet perceptually aligned with image space.

To express our formulation in this latent space, we encode the image–geometry triplet with the VAE encoder $\mathcal{E}$ and decode with $\mathcal{D}$:

$$\mathbf{z}^{(\mathcal{G})} = \mathcal{E}(\mathcal{G}), \qquad \hat{\mathcal{G}} = \mathcal{D}\big(\mathbf{z}^{(\mathcal{G})}\big), \tag{6}$$

where $\mathcal{G} = (\mathbf{x},\, \mathbf{d},\, \mathbf{n})$ denotes the image–geometry triplet (depth $\mathbf{d}$ and surface normals $\mathbf{n}$). Here, $\mathbf{z}^{(\mathcal{G})}$ is the latent code of the triplet and $\hat{\mathcal{G}}$ is its reconstruction; $\mathcal{E}$ and $\mathcal{D}$ are the VAE encoder and decoder, respectively.

For depth, which is single-channel, we tile the map to three channels to match the VAE's RGB encoder; at decoding, we average the three output channels to recover a single-channel depth prediction. We adopt a *frozen image VAE* [43] to encode both the image and its corresponding geometry into the latent space used to train our conditional denoiser. Both depth and normal maps are preprocessed by normalizing the depth values to $[0, 1]$, followed by applying $(x - 0.5) \times 2$ to align them with the VAE input range of $[-1, 1]$.

Adapted denoising U-Net Stable Video Diffusion is an image-to-video diffusion model that synthesizes a video sequence conditioned on a single input image. The conditioning image can be injected to the UNet by (i) concatenating its latent with the input noise latent, and (ii) providing its CLIP [39] embedding to intermediate features via cross-attention. In our paradigm, geometry (e.g., depth and surface normal) is modeled as the next frames in a image to geomery, so we adapt the conditioning (i) to suit image-to-geomerty generation. Additionally, the temporal attention mechanism enables the image condition to be propagated

to the geometry branch, enhancing structural consistency. Furthermore, as illustrated in Fig. 2, given the encoded latents of the image, depth, and normal, we concatenate the geometry latents (depth and normal) with the image latent to form the final conditioning input. Instead of generating geometry alone, the image is also reconstructed to provide a consistent conditioning signal. We disable conditioning (ii) because it requires resizing the image, which leads to geometry distortions in the depth and normal maps.

3.3 Inference

At inference, we initialize the latent variables $\mathbf{z}_t^{(x)}$, $\mathbf{z}_t^{(d)}$, and $\mathbf{z}_t^{(n)}$ from standard Gaussian noise, and encode the input image into its latent $\mathbf{z}_0^{(x)}$. The initialized noises and the image latent $\mathbf{z}_0^{(x)}$ are concatenated and fed into the denoiser U-Net, which iteratively predicts the latents of the image, depth, and surface normal across diffusion steps. After denoising, the image latent is discarded, and the final depth map and surface normal are decoded from their corresponding latents via the VAE decoder. By jointly generating the image, depth, and surface normal during inference, GeoNeXt enhances consistency between image appearance and geometric structure.

4 Experiments

Implementation We implement **GeoNeXt** in PyTorch, building upon the image-to-video variant of Stable Video Diffusion [6] as a representative choice. As our method is architecture-agnostic and can be seamlessly integrated with more advanced DiT-based models (e.g., WAN [54]), we report the corresponding results and implementation details in the supplementary material. The original cross-attention conditioning in SVD is disabled. During training, we adopt the standard EDM noise schedule [13], where the noise level σ_t is sampled from a distribution $p(\sigma) = \mathcal{N}(0.7, 1.6)$, following SVD. At inference, we use the EDM sampler with only 5 denoising steps. Following [29], we ensemble the outputs from 5 inference runs initialized with different noise seeds. We train using the Adam optimizer with a learning rate of 5×10^{-6} and apply random horizontal flipping for data augmentation. For depth estimation, we operate in disparity space, *i.e.*, $d = 1/d'$, where d denotes the predicted disparity and d' represents the corresponding true depth.

Training datasets We employ two synthetic datasets that together cover both indoor and outdoor environments. **(1)** ***Hypersim*** [42] contains 461 photorealistic indoor scenes. We use the official training split with approximately 54K samples and remove incomplete ones, resulting in about 39K valid samples. All RGB images, depth maps, and normal maps are resized to 576×768. **(2)** ***Virtual KITTI 2*** [9] is a synthetic driving dataset comprising five urban scenes captured under diverse illumination and weather conditions. We utilize four scenes for training, providing roughly 20K samples, which are cropped to 352×1216

Table 1: Quantitative comparison on zero-shot affine-invariant depth estimation. We compare GeoNeXt with representative discriminative and generative methods across five datasets. The upper block lists large-scale discriminative approaches, the middle block reports generative methods trained for depth estimation only, and the lower block presents unified geometry estimation methods, which serve as our main comparison focus. * indicates results reproduced by us. **Bold** marks the best performance.

Method	Training Data	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
		AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Large-scale discriminative depth models											
MiDaS [41]	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5
Omnidata [13]	12.2M	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2
DPT [40]	1.4M	9.8	90.3	10.0	90.1	**7.8**	**94.6**	8.2	93.4	**18.2**	75.8
DA [59]	62.6M	**4.3**	**98.1**	7.6	**94.7**	12.7	88.2	4.3	**98.1**	26.0	**75.9**
DA-V2 [60]	62.6M	4.5	97.9	**7.4**	94.6	13.1	86.5	**4.2**	97.8	26.5	73.4
Generative methods for depth estimation only											
Marigold [29]	74K	5.5	96.4	9.9	91.6	6.5	95.9	6.4	95.2	30.8	77.3
E2E-FT [17]	74K	**5.4**	96.5	9.6	92.1	6.4	95.9	**5.8**	**96.5**	30.3	77.6
DepthFM [20]*	63K	6.9	95.5	9.8	91.5	6.6	95.8	7.8	93.1	25.2	**77.9**
Lotus-G [22]*	59K	**5.4**	**96.6**	**8.5**	**92.2**	**5.9**	**97.0**	5.9	95.6	**23.0**	73.0
Generative methods for unified geometry estimation											
GeoWizard [16]*	208K	5.6	96.3	14.4	82.0	6.8	95.8	6.4	95.2	33.0	73.5
GeoNeXt	**59K**	**5.3**	**96.8**	**8.2**	**92.6**	**5.6**	**97.2**	**5.9**	**95.8**	**22.6**	**74.3**

with a far-plane depth limit of 80 m. Following Marigold [29], we sample each training batch from the two datasets with a 9:1 ratio, selecting from *Hypersim* with 90% probability and from *Virtual KITTI* with 10%.

Evaluation datasets We evaluate **GeoNeXt** on depth and surface normal estimation using real-world datasets that are not seen during training. For depth estimation, we test on five datasets: NYUv2 [49] and ScanNet [11], which include indoor scenes; KITTI [18], which captures diverse outdoor driving scenes; ETH3D [47], a high-resolution dataset covering both indoor and outdoor environments; and DIODE [53], another high-resolution dataset providing dense and accurate depth maps across varied conditions. For surface normal estimation, we evaluate on five datasets as well: NYUv2 [49], ScanNet [11], and iBims-1 [31], which contain real indoor scenes; Sintel [8], which features highly dynamic outdoor scenes; and OASIS [10], which consists of in-the-wild Internet images.

4.1 Results Analysis

Zero-shot Depth Estimation Comparison We present the quantitative results of zero-shot affine-invariant depth estimation in Table 1. The proposed GeoNeXt is compared with three categories of state-of-the-art methods: (1) large-scale discriminative depth estimation models; (2) generative methods for depth estimation only; and (3) generative methods for unified geometry estimation. Compared

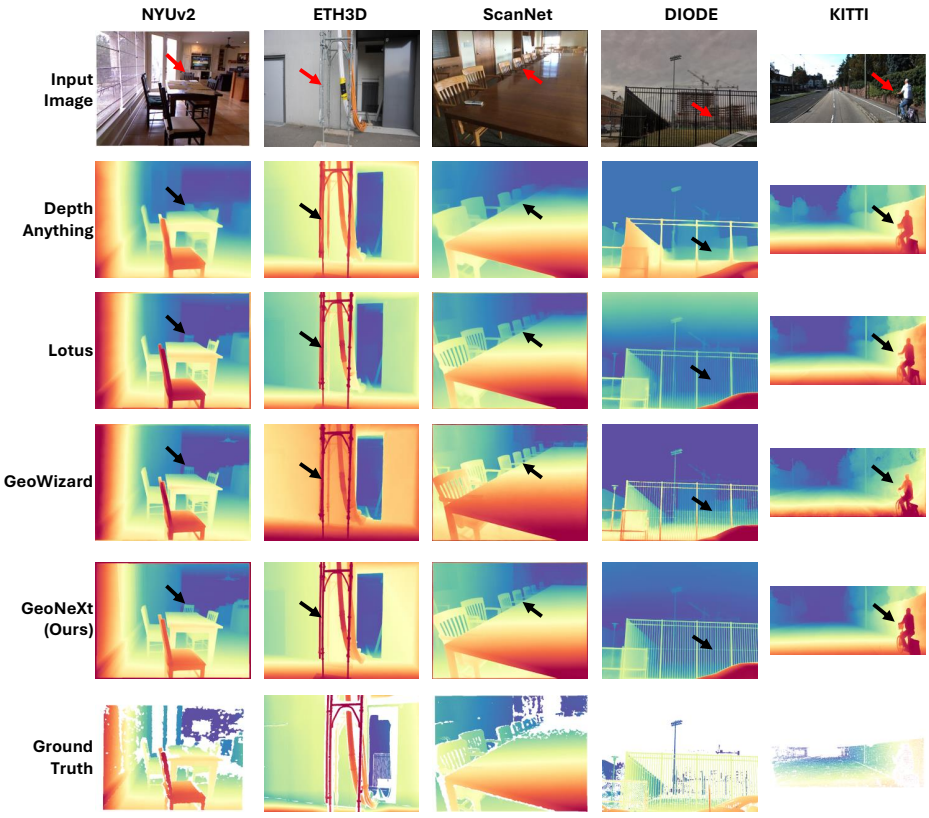

Fig. 3: Qualitative comparison of monocular depth estimation methods across diverse datasets. GeoNeXt demonstrates superior reconstruction of fine structures.

with discriminative models such as DepthAnything, which rely on 63.5M training images, our model uses nearly **100**× less data yet achieves highly competitive results. On several benchmarks, GeoNeXt even surpasses them, achieving, for example, an AbsRel of 13.1 vs. 5.6 on ETH3D. Next, we compare GeoNeXt with recent generative methods designed solely for depth estimation.

To demonstrate the advantages of the next-frame prediction formulation within video diffusion models, we compare our approach with image diffusion based generative methods, including Marigold, E2E-FT, DepthFM, Lotus-G, and GeoWizard. In particular, Lotus-G-depth is trained on a similar scale (59K samples) but optimized only for depth prediction, requiring a separate model. In contrast, our method formulates geometry estimation as a next-frame prediction task and models geometric consistency within a unified framework. As shown in Table 1, our model achieves superior or comparable performance across datasets.

We further compare with unified geometry estimation methods such as GeoWizard, which also rely on image diffusion models but introduce additional

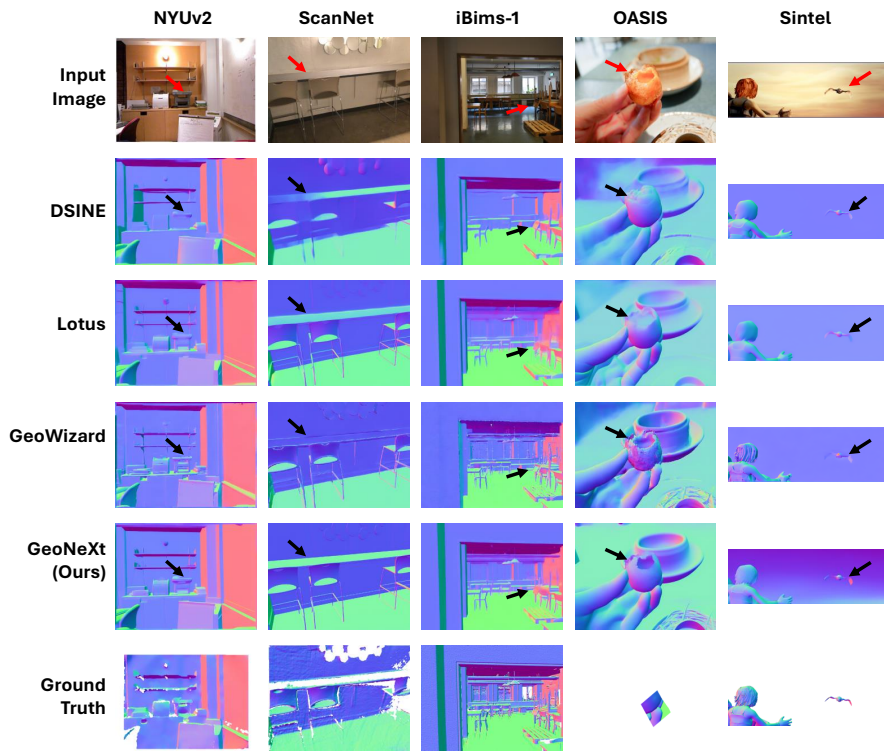

Fig. 4: Qualitative comparison of surface normal estimation methods across diverse datasets. GeoNeXt produces more accurate and detailed normal reconstructions, effectively capturing fine geometric structures and surface orientations.

architectural modifications such as cross-attention and switching modules. Although GeoWizard is trained on a substantially larger dataset (208K samples), GeoNeXt still maintains superior performance. For example, GeoNeXt surpasses GeoWizard by 6.2 in AbsRel and 9.4 in δ_1 on KITTI, and by 10.4 in AbsRel and 1.8 in δ_1 on DIODE, despite using significantly less training data.In addition, Figure 3 presents qualitative comparisons across multiple datasets. Our method demonstrates clearly superior visual results, particularly in reconstructing fine-grained structures (e.g., chair, poles, and fences), preserving object boundaries.

Zero-shot Normal Estimation Comparison To thoroughly evaluate our method, we conduct zero-shot surface normal estimation and compare its performance with three categories of state-of-the-art approaches, following the same experimental setup as used for depth estimation, as shown in Table 2. Despite being trained on only 59K samples, our model delivers competitive or superior accuracy relative to discriminative baselines such as Omnidata and EESNU.

Compared with normal-specific generative methods based on image diffusion models and trained on a similar data scale, such as Lotus-G-normal, GeoNeXt

Table 2: Quantitative results on zero-shot surface normal estimation. We evaluate **GeoNeXt** against representative discriminative and generative approaches across five benchmark datasets. The top section summarizes large-scale discriminative baselines, the middle section reports generative methods trained exclusively for surface normal prediction, and the bottom section highlights unified geometry estimation frameworks for direct comparison. * indicates results reproduced by us. **Bold** marks the best performance.

Method	Training Data	NYUv2		ScanNet		iBims-1		Sintel		OASIS	
		mean↓	11.25°↑	mean↓	11.25°↑	mean↓	11.25°↑	mean↓	11.25°↑	mean↓	11.25°↑
Large-scale discriminative normal models											
Omnidata [13]	12.2M	23.1	45.8	22.9	47.4	19.0	62.1	41.5	11.4	24.9	31.0
EESNU [2]	2.5M	**16.2**	58.6	-	-	20.0	58.5	42.1	11.5	27.7	24.0
DSINE [3]	160K	16.4	**59.6**	**16.2**	**61.0**	**17.1**	**67.4**	**34.9**	**21.5**	**24.4**	**28.8**
Generative methods for surface normal estimation only.											
Marigold [29]	74K	20.9	50.5	21.3	45.6	18.5	64.7	-	-	-	-
StableNormal [62]	250K	18.6	53.5	17.1	57.4	18.2	65.0	36.7	14.1	26.5	23.5
E2E-FT [17]	74K	**16.5**	**60.4**	**14.7**	**66.1**	**16.1**	**69.7**	**33.5**	**22.3**	23.2	**29.4**
Lotus-G [22]*	59K	16.6	59.4	15.1	63.8	17.2	66.3	33.6	21.0	**22.9**	29.3
Generative methods for unified geometry estimation.											
GeoWizard [16]*	208K	18.9	50.1	17.2	53.8	19.4	62.8	40.3	12.8	25.0	23.7
GeoNeXt	**59K**	**16.7**	**60.0**	**16.0**	**62.8**	**16.4**	**69.2**	**33.0**	**21.5**	**22.8**	**30.8**

achieves the best performance across most datasets. Moreover, compared to unified geometry estimation models like GeoWizard, GeoNeXt attains the overall strongest results, for example, mean angular errors of 16.4 on iBims-1 and 33.0 on Sintel, while jointly estimating both depth and normals using significantly less training data. These results quantitatively demonstrate the effectiveness and data efficiency of our approach.

We further present qualitative comparisons in Figure 4, where our method produces more accurate surface normal predictions, particularly on curved and deformable surfaces. The results demonstrate strong cross-domain generalization, with our model achieving consistent and detailed normal reconstructions that faithfully capture fine geometric structures and surface orientations across both synthetic and real-world scenes.

4.2 Ablation Study

We conduct ablation studies for GeoNeXt on depth and surface normal estimation. For evaluation, we select two zero-shot sets for each task: NYU and ETH3D for depth estimation, and NYU and iBims-N for surface normal estimation.

Generative formulation We hypothesize that a key factor behind the effectiveness of GeoNeXt lies in its formulation and architecture, which jointly synthesize the image and its geometry in lockstep, rather than employing a geometry-only

Table 3: Ablation on the generative formulation, model architecture, and joint estimation. Upper block: Models trained for unified geometry estimation; **Lower** block: Mfiodels trained for depth or surface normal estimation individually.

Method	NYUv2		ETH3D		NYUv2		iBims-N	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Mean↓	11.25° ↑	Mean↓	11.25° ↑
Unified geometry estimation								
(a) w/o Image Recon.	6.5	95.2	6.7	96.0	17.9	55.8	18.5	65.2
(b) w/ CLIP Embed.	5.6	96.5	5.8	97.0	16.9	59.6	16.5	69.0
(c) Full model	**5.3**	**96.8**	**5.6**	**97.2**	**16.7**	**60.2**	**16.4**	**69.2**
Separated geometry estimation								
(d) Depth only	5.9	96.5	6.4	96.3	–	–	–	–
(e) Surface Normal only	–	–	–	–	17.2	58.8	17.3	67.1

generator. To validate this hypothesis, we first evaluate a variant without the image reconstruction branch. Following the setup of Marigold, which formulates the task as image-conditioned generation, we treat the first and second frames as depth and normal targets, respectively. As shown in Table 3(a), removing image reconstruction leads to a noticeable drop in performance, confirming that reconstructing the input image enhances image–geometry consistency and contributes significantly to the overall effectiveness of the generative formulation.

Model architecture Another important component of GeoNeXt is the CLIP embedding module in the U-Net, which resizes the input image to extract semantic features for conditioning the generation process. To evaluate its effect, we remove this CLIP-based conditioning. As shown in Table 3(b), excluding the CLIP embedding slightly improves performance, suggesting that resizing the image for CLIP feature extraction may distort intrinsic structural information, thereby hindering stable and consistent geometry generation.

Joint Depth and Normal Estimation We further investigate the effectiveness of joint geometry estimation. To this end, we train two separate models to estimate depth and surface normals independently. As shown in Table 3(d) and (e), this setup results in a clear performance drop across all evaluation metrics, indicating that joint training more effectively captures the correlation between the two geometric representations.

Depth and Normal Reconstruction Order We also analyze the impact of reconstruction order between depth and surface normals. As shown in Table 4, both reconstruction orders, where depth and normal frames are swapped, achieve nearly identical performance across all metrics. These marginal variations suggest that GeoNeXt successfully models the intrinsic relationships among image, depth, and normal representations, making the reconstruction order largely flexible without significantly affecting performance.

Table 4: Ablation on reconstruction order. We compare two configurations that assign depth and normal to different frames. The results are nearly identical, demonstrating the robustness of the model to reconstruction order.

Method	NYUv2		ETH3D		NYUv2		iBims-N	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Mean↓	11.25° ↑	Mean↓	11.25° ↑
normal-depth	5.7	96.6	**5.5**	**97.2**	**16.6**	60.0	16.5	**69.4**
depth-normal	**5.3**	**96.8**	5.6	**97.2**	16.7	**60.2**	**16.4**	69.2

4.3 Cost-effectiveness analysis

Table 5: Cost-effectiveness analysis. **Upper**: models trained on geometric targets; we report combined depth and normal inference time and parameter count. **Lower**: unified geometry estimation variants. **NFEs**: number of function evaluations required to obtain a prediction, defined as *diffusion steps* × *ensemble size* for diffusion models.

Method	ETH3D		iBims-N		NFEs	Inference time (s)	Param (B)
	AbsRel↓	δ_1 ↑	Mean↓	11.25° ↑			
Separated geometry estimation							
Marigold [29]	7.0	95.5	18.5	65.6	1×1	0.8	1.9
Marigold [29]	6.5	95.9	18.5	64.7	50×10	180.5	1.9
Lotus-G [22]	**5.9**	97.0	**17.2**	66.3	1×1	**0.8**	1.9
Unified geometry estimation							
GeoWizard [16]	7.3	96.3	19.9	82.0	1×1	1.2	**0.9**
GeoWizard [16]	6.8	96.8	19.4	82.5	50×10	272.1	**0.9**
GeoNeXt	5.8	97.0	16.8	68.9	1×1	**1.0**	1.5
GeoNeXt	**5.6**	**97.2**	**16.4**	**69.2**	5×5	10.0	1.5

We present a cost-effectiveness analysis of different approaches in Table 5. All inference times are measured on a single NVIDIA A5000 using an input resolution of 768×768. Compared with models trained separately for depth and surface normals, our unified approach achieves comparable inference time while delivering the best overall performance, even though Lotus-G leverages more advanced training strategies. Moreover, separated models require storing and maintaining two independent checkpoints, which increases the overall storage footprint. In real-world deployment, these approaches also introduce considerable I/O overhead, as separate models (e.g., for depth and normals) need to be loaded onto and off the GPU individually. The resulting data transfer latency can substantially exceed the actual computation time required for inference. We note that this I/O overhead is not included in the reported runtime measurements.

GeoWizard achieves unified geometry estimation by modifying the attention forward pass and introducing a geometry-switch mechanism. However, it requires

substantially more training data to converge. When applied with a single prediction, its performance is limited. Although increasing the number of denoising steps or applying large-scale ensembling can improve its accuracy, this comes at the cost of significantly higher inference time, making it less practical compared with our method. In contrast, our single-step prediction already surpasses all baseline methods. By incorporating only a few additional denoising iterations along with a lightweight ensemble strategy, we can further enhance performance while preserving fast and computationally efficient inference.

5 Conclusion

We propose GeoNeXt, a novel approach for jointly estimating geometry from a single image. The core idea is to repurpose pretrained video generative models as a unified and data-efficient framework for geometry estimation, formulated as a next-frame prediction task. GeoNeXt naturally inherits temporal coherence and rich generative priors from video models, while further adapting them for joint modeling between images and geometry targets (image $\leftrightarrow$ geometry). We systematically explore three key aspects: a joint generative formulation within the video generative framework, architectural adaptations for geometry estimation, and robustness to reconstruction order. Our results highlight the importance of preserving fine-grained details in both depth and surface normal estimation, supported by an effective training protocol. Extensive quantitative and qualitative evaluations demonstrate the effectiveness of our approach, with GeoNeXt achieving strong performance even with limited training data.

References

1. Aich, S., Vianney, J.M.U., Islam, M.A., Liu, M.K.B.: Bidirectional attention network for monocular depth estimation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 11746–11752. IEEE (2021)
2. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13137–13146 (2021)
3. Bae, G., Davison, A.J.: Rethinking inductive biases for surface normal estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9535–9545 (2024)
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
5. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
8. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. pp. 611–625. Springer (2012)
9. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
10. Chen, W., Qian, S., Fan, D., Kojima, N., Hamilton, M., Deng, J.: Oasis: A large-scale dataset for single image 3d in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 679–688 (2020)
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
12. Duan, Y., Guo, X., Zhu, Z.: Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In: European Conference on Computer Vision. pp. 432–449. Springer (2024)
13. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2021)
14. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
15. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
16. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: European Conference on Computer Vision. pp. 241–258. Springer (2024)
17. Garcia, G.M., Abou Zeid, K., Schmidt, C., De Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 753–762. IEEE (2025)
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
19. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
20. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast generative monocular depth estimation with flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3203–3211 (2025)
21. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023)
22. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124 (2024)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

24. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
25. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: Depthcrafter: Generating consistent long depth sequences for open-world videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2005–2015 (2025)
26. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21741–21752 (2023)
27. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4d: Leveraging video generators for geometric 4d scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20658–20671 (2025)
28. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)
29. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
30. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
31. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
32. Krishnan, A., Yan, X., Casser, V., Kundu, A.: Orchid: Image latent diffusion for joint appearance and geometry generation. arXiv preprint arXiv:2501.13087 (2025)
33. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
34. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1119–1127 (2015)
35. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
36. Li, Z., Chen, Z., Liu, X., Jiang, J.: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. Machine Intelligence Research **20**(6), 837–854 (2023)
37. Liu, H., Yang, H., Huijben, E.M., Schuiveling, M., Su, R., Pluim, J.P., Veta, M.: Pathopainter: Augmenting histopathology segmentation via tumor-aware inpainting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 408–417. Springer (2025)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
39. 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: International conference on machine learning. pp. 8748–8763. PmLR (2021)

40. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
41. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
42. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
45. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. Advances in Neural Information Processing Systems **36**, 39443–39469 (2023)
46. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
47. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3260–3269 (2017)
48. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22841–22852 (2025)
49. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
50. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
51. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
52. Sun, Y.T., Yu, X., Huang, Z., Huang, Y.H., Guo, Y.C., Yang, Z., Cao, Y.P., Qi, X.: Unigeo: Taming video diffusion for unified consistent geometry estimation. arXiv preprint arXiv:2505.24521 (2025)
53. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
54. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
55. Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., Yuille, A.L.: Surge: Surface regularized geometry estimation from a single image. Advances in Neural Information Processing Systems **29** (2016)

56. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 675–684 (2018)
57. Yang, H., Bulat, A., Hadji, I., Pham, H.X., Zhu, X., Tzimiropoulos, G., Martinez, B.: Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2459–2468 (2025)
58. Yang, H., Zhang, C., Wang, W., Volino, M., Hilton, A., Zhang, L., Zhu, X.: Improving gaussian splatting with localized points management. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21696–21705 (2025)
59. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
60. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
61. Yang, Z., Pan, Z., Yang, Y., Zhu, X., Zhang, L.: Driving scene synthesis on free-form trajectories with generative prior. arXiv preprint arXiv:2412.01717 (2024)
62. Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., Han, X.: Stablenormal: Reducing diffusion variance for stable and sharp normal. ACM Transactions on Graphics (TOG) **43**(6), 1–18 (2024)
63. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems **35**, 25018–25032 (2022)
64. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: New crfs: Neural window fully-connected crfs for monocular depth estimation. arXiv preprint arXiv:2203.01502 (2022)
65. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4106–4115 (2019)
66. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5729–5739 (2023)