LOTUS: DIFFUSION-BASED VISUAL FOUNDATION MODEL FOR HIGH-QUALITY DENSE PREDICTION

Anonymous authors
Paper under double-blind review

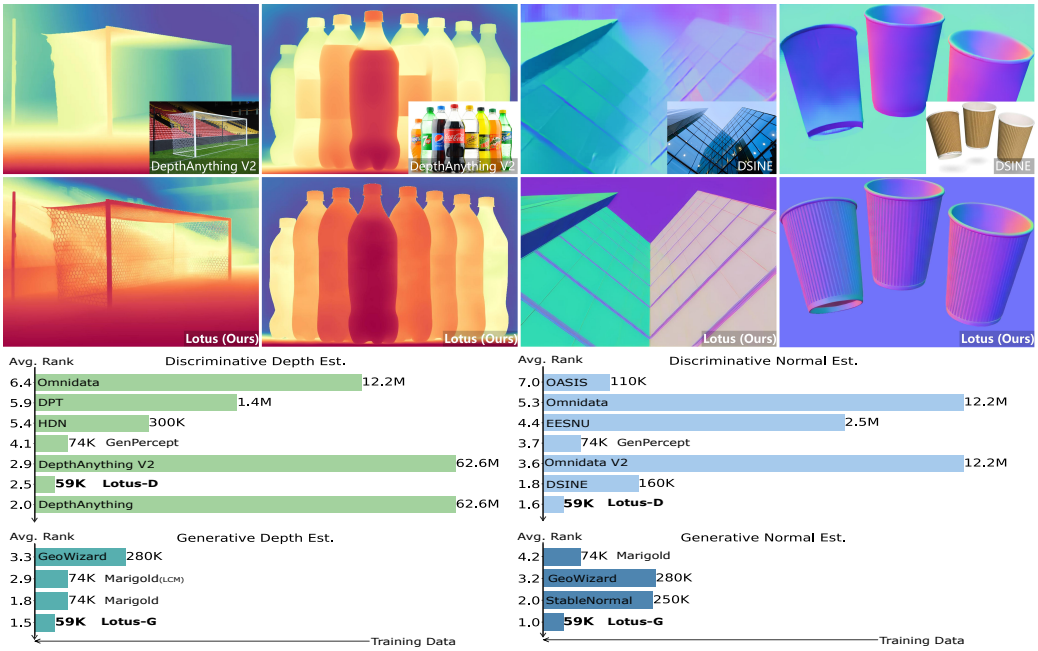

Figure 1: We present **Lotus**, a diffusion-based visual foundation model for dense geometry prediction. With minimal training data, Lotus achieves promising performance in zero-shot depth and normal estimation. "Avg. Rank" indicates the average ranking across all metrics, where lower values are better. Bar length represents the amount of training data used.

ABSTRACT

Leveraging the visual priors of pre-trained text-to-image diffusion models offers a promising solution to enhance zero-shot generalization in dense prediction tasks. However, existing methods often uncritically use the original diffusion formulation, which may not be optimal due to the fundamental differences between dense prediction and image generation. In this paper, we provide a systemic analysis of the diffusion formulation for the dense prediction, focusing on both quality and efficiency. And we find that the original parameterization type for image generation, which learns to predict noise, is harmful for dense prediction; the multi-step noising/denoising diffusion process is also unnecessary and challenging to optimize. Based on these insights, we introduce **Lotus**, a diffusion-based visual foundation model with a simple yet effective adaptation protocol for dense prediction. Specifically, Lotus is trained to directly predict annotations instead of noise, thereby avoiding harmful variance. We also reformulate the diffusion process into a single-step procedure, simplifying optimization and significantly boosting inference speed. Additionally, we introduce a novel tuning strategy called detail preserver, which achieves more accurate and fine-grained predictions. Without scaling up the training data or model capacity, Lotus achieves promising performance in zero-shot depth and normal estimation across various datasets. It also enhances efficiency, being significantly faster than most existing diffusion-based methods. Lotus' superior quality and efficiency enables a wide range of practical applications, such as joint estimation, single/multi-view 3D reconstruction, etc.

1 INTRODUCTION

Dense prediction is a fundamental task in computer vision, benefiting a wide range of applications, such as 3D/4D reconstruction (Huang et al., 2024; Long et al., 2024; Wang et al., 2024; Lei et al., 2024), tracking (Xiao et al., 2024; Song et al., 2024), and autonomous driving (Yurtsever et al., 2020; Hu et al., 2023). Estimating pixel-level geometric attributes from a single image requires comprehensive scene understanding. Although deep learning has advanced dense prediction, progress is limited by the quality, diversity, and scale of training data, leading to poor zero-shot generalization. Instead of merely scaling data and model size, recent works (Lee et al., 2024; Ke et al., 2024; Fu et al., 2024; Xu et al., 2024) leverage diffusion priors for zero-shot dense prediction. These studies demonstrate that text-to-image diffusion models like Stable Diffusion (Rombach et al., 2022), pretrained on billions of images, possess powerful and comprehensive visual priors to elevate dense prediction performance. However, most of these methods directly inherit the pre-trained diffusion models for dense prediction tasks, without exploring more suitable diffusion formulations. This oversight often leads to challenging issues. For example, Marigold (Ke et al., 2024) directly fine-tunes Stable Diffusion for image-conditioned depth generation. While it significantly improves depth estimation, its performance is still constrained by overlooking the fundamental differences between dense prediction and image generation. Especially, its efficiency is also severely limited by standard iterative denoising processes and ensemble inferences.

Motivated by these concerns, we systematically analyze the diffusion formulation, trying to find a better formulation to fit the pre-trained diffusion model into dense prediction. Our analysis yields several important findings: ① The widely used parameterization, *i.e.*, noise prediction, for diffusion-based image generation is ill-suited for dense prediction. It results in large prediction errors due to harmful prediction variance at initial denoising steps, which are subsequently propagated and magnified throughout the entire denoising process (Sec. 4.1). ② Multi-step diffusion formulation is computation-intensive and is prone to sub-optimal with limited data and resources. These factors significantly hinder the adaptation of diffusion priors to dense prediction tasks, leading to decreased accuracy and efficiency (Sec. 4.2). ③ Though remarkable performance achieved, we observed that the model usually outputs vague predictions in highly-detailed areas (Fig. 8). This vagueness is attributed to catastrophic forgetting: the pre-trained diffusion models gradually lose their ability to generate detailed regions during fine-tuning (Sec. 4.3).

Following our analysis, we propose **Lotus**, a diffusion-based visual foundation model for dense prediction, featuring a simple yet effective fine-tuning protocol (see Fig. 2). First, Lotus is trained to directly predict annotations, thereby avoiding the harmful variance associated with standard noise prediction. Next, we introduce a one-step formulation, *i.e.*, one step between pure noise and clean output, to facilitate model convergence and achieve better optimization performance with limited high-quality data. It also considerably boosts both training and inference efficiency. Moreover, we implement a novel detail preserver through a task switcher, allowing the model either to generate annotations or reconstruct the input images. It can better preserve the fine-grained details in the input image during dense annotation generation, achieving

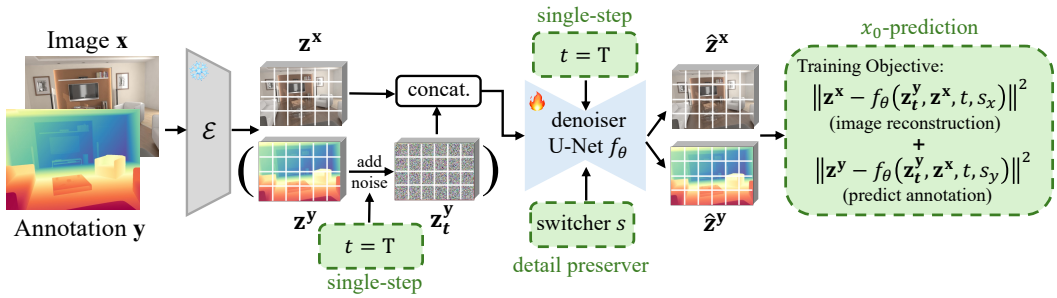

Figure 2: **Adaptation protocol of Lotus.** After the pre-trained VAE encoder $\mathcal{E}$ encodes the image $\mathbf{x}$ and annotation $\mathbf{y}$ to the latent space: ① the denoiser U-Net model f_θ is fine-tuned using x_0-prediction; ② we employ single-step diffusion formulation at time-step $t = T$ for better convergence; ③ we propose a novel detail preserver, to switch the model either to reconstruct the image or generate the dense prediction via a switcher s, ensuring a more fine-grained prediction. The noise $\mathbf{z}_{\mathbf{T}}^{\mathbf{y}}$ in bracket is used for our generative **Lotus-G** and is omitted for the discriminative **Lotus-D**.

higher performance without compromising efficiency, requiring additional parameters, or being affected by surface textures.

To validate Lotus, we conduct extensive experiments on two primary geometric dense prediction tasks: zero-shot monocular depth and normal estimation. The results demonstrate that Lotus achieves promising, and even superior, performance on these tasks across a wide range of evaluation datasets. Compared to traditional discriminative methods, Lotus delivers remarkable results with only 59K training samples. Among generative approaches, Lotus also outperforms previous methods in both accuracy and efficiency, being significantly faster than methods like Marigold (Ke et al., 2024) (Fig. 3). Beyond these improvements, Lotus seamlessly supports various applications, such as joint estimation, single/multi-view 3D reconstruction, an so on.

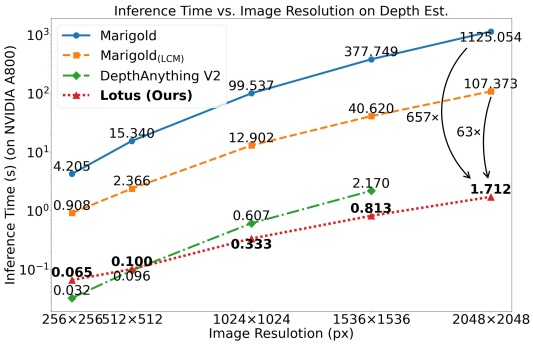

Figure 3: **Inference time comparison in depth estimation between Lotus and SoTA methods.** Lotus is hundreds of times faster than Marigold and slightly faster than DepthAnything V2 at high resolutions. DepthAnything V2's inference time at 2048×2048 is not plotted because it requires > 80GB graphic memory.

In conclusion, our key contributions are as follows:

- We systematically analyze the diffusion formulation and find their parameterization type, designed for image generation, is unsuitable for dense prediction and the computation-intensive multi-step diffusion process is also unnecessary and challenging to optimize.
- We propose a novel detail preserver that ensures more accurate dense predictions especially in detail-rich areas, without compromising efficiency, introducing additional network parameters, or being affected by surface textures.
- Based on our insights, we introduce **Lotus**, a diffusion-based visual foundation model for dense prediction with simple yet effective fine-tuning protocol. Lotus achieves promising performance on both zero-shot monocular depth and surface normal estimation. It also enables a wide range of applications.

2 Related Works

2.1 Text-to-Image Generative Models

In the field of text-to-image generation, the evolution of methodologies has transitioned from generative adversarial networks (GANs) (Goodfellow et al., 2014; Zhang et al., 2017; 2018; 2021; He et al., 2022; Karras et al., 2019; 2020; 2021; Zhang et al., 2017; 2018; Xu et al., 2018; Zhang et al., 2021) to advanced diffusion models (Ho et al., 2020; Ramesh et al., 2022; Saharia et al., 2022; Ramesh et al., 2021; Nichol et al., 2021; Chen et al., 2023; Rombach et al., 2022; Ramesh et al., 2021). A series of diffusion-based methods such as GLIDE (Nichol et al., 2021), DALL·E2 (Ramesh et al., 2022), and Imagen (Saharia et al., 2022) have been introduced, offering enhanced image quality and textual coherence. The Stable Diffusion (SD) (Rombach et al., 2022), trained on large-scale LAION-5B dataset (Schuhmann et al., 2022), further enhances the generative quality, becoming the community standard. In our paper, we aim to leverage the comprehensive and encyclopedic visual priors of SD to facilitate zero-shot generalization for dense prediction tasks.

2.2 Generative Models for Dense Perception

Currently, a notable trend involves adopting pre-trained generative models, particularly diffusion models, into dense prediction tasks. Marigold (Ke et al., 2024) and GeoWizard (Fu et al., 2024) directly apply the standard diffusion formulation and the pre-trained parameters, without addressing the inherent differences between image generation and dense prediction, leading to constrained performance. Their efficiency is also severely limited by standard iterative denoising processes and ensemble inferences. In this paper, we propose a novel diffusion formulation tailored to the dense prediction. Aiming to fully leveraging the pre-trained diffusion's powerful visual priors, Lotus enables more accurate and efficient predictions, finally achieving promising performance.

More recent works, GenPercept (Xu et al., 2024) and StableNormal (Ye et al., 2024), also adopted single-step diffusion. However, GenPercept (Xu et al., 2024) first removes noise input for deter-

ministic characteristic based on DMP (Lee et al., 2024), and then adopts one-step strategy to avoid surface texture interference. It lacks systematic analysis of the diffusion formulation, only treats the U-Net as a deterministic backbone and still falls short in performance. In contrast, Lotus systematically analyzes the standard stochastic diffusion formulation for dense prediction and proposes innovations such as the detail preserver to improve accuracy especially in detailed area, finally delivering much better results (Tab. 1). Additionally, Lotus is a stochastic model. In contrast to GenPercept's deterministic nature, Lotus enables uncertainty predictions. StableNormal (Ye et al., 2024) predicts normal maps through a two-stage process. While the first stage produces coarse normal maps with single-step diffusion, the second stage performs refinement still with iterative diffusion which is computation-intensive. In comparison, Lotus not only achieves fine-grained predictions thanks to our novel detail preserver without extra stages or parameters, but also delivers much superior results (Tab. 2) thanks to our designed diffusion formulation that better fits the pre-trained diffusion for dense prediction. Recently, a concurrent work, Diffusion-E2E-FT (Garcia et al., 2024), has also achieved promising results in a single step. Its main contribution lies in addressing the issue where Marigold (Ke et al., 2024) and similar models (Fu et al., 2024) use inconsistent pairings of time-step and noise, resulting in poor predictions. By setting the "time-step spacing" to "trailing" mode in schedulers, it prevents "GT" signal leakage during inference, improving accuracy. While the performance of Lotus-D and Diffusion-E2E-FT is similar, Lotus is based on a systematic analysis of stochastic diffusion for dense prediction, with innovations like the detail preserver to enhance accuracy, particularly in detailed areas. Additionally, unlike the deterministic Diffusion-E2E-FT, Lotus (Lotus-G) is a stochastic model that enables uncertainty predictions.

2.3 Monocular Depth and Normal Prediction

Monocular depth and normal prediction are two crucial dense prediction tasks. Solving them typically demands comprehensive scene understanding capability. Starting from Eigen et al. (2014), early CNN-based methods for depth prediction, such as Fu et al. (2018), Lee et al. (2019), Yuan et al. (2022), focus only on specific domains. Subsequently, in pursuit of a generalizable depth estimator, many methods expand model capacity and train on larger and more diverse datasets, such as DiverseDepth (Yin et al., 2021a) and MiDaS (Ranftl et al., 2020). DPT (Ranftl et al., 2021) and Omnidata (Eftekhar et al., 2021) are further proposed based on vision transformer (Ranftl et al., 2021), significantly enhancing performance. LeRes (Yin et al., 2021b) and HDN (Zhang et al., 2022) further introduce novel training strategies and multi-scale depth normalization to improve predictions in detailed areas. More recently, the DepthAnything series (Yang et al., 2024a;b) and Metric3D series (Yin et al., 2023; Hu et al., 2024) collect and leverage millions of training data to develop more powerful estimators. Normal prediction follows the same trend. Starting with the early CNN-based methods like OASIS (Chen et al., 2020), EESNU (Bae & Davison, 2021) and Omnidata series (Eftekhar et al., 2021; Kar et al., 2022) expand the model capacity and scale up the training data. Recently, DSINE (Bae & Davison, 2024) achieves SoTA performance by rethinking inductive biases for surface normal estimation. In our paper, we focus on leveraging pre-trained diffusion priors to enhance zero-shot dense predictions, rather than expanding model capacity or relying on large training data, which avoids the need for intensive resources and computation.

3 Preliminaries

Diffusion Formulation for Dense Prediction. Following Ke et al. (2024) and Fu et al. (2024), we also formulate dense prediction as an image-conditioned annotation generation task based on Stable Diffusion (Rombach et al., 2022), which performs the diffusion process in low-dimensional latent space for computational efficiency. First, the auto-encoder, which consists an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$, is trained to map between RGB space and latent space, *i.e.*, $\mathcal{E}(\mathbf{x}) = \mathbf{z}^{\mathbf{x}}$, $\mathcal{D}(\mathbf{z}^{\mathbf{x}}) \approx \mathbf{x}$. The auto-encoder also maps between dense annotations and latent space effectively, *i.e.*, $\mathcal{E}(\mathbf{y}) = \mathbf{z}^{\mathbf{y}}$, $\mathcal{D}(\mathbf{z}^{\mathbf{y}}) \approx \mathbf{y}$ (Ke et al., 2024; Fu et al., 2024; Xu et al., 2024; Ye et al., 2024). Following Ho et al. (2020), Stable Diffusion establishes a pair of *forward* nosing and *reversal* denoising processes in latent space. In *forward* process, Gaussian noise is gradually added at levels $t \in [1, T]$ into sample $\mathbf{z}^{\mathbf{y}}$ to obtain the noisy sample $\mathbf{z}^{\mathbf{y}}_{\mathbf{t}}$:

$$\mathbf{z}^{\mathbf{y}}_{\mathbf{t}} = \sqrt{\overline{\alpha}_t}\mathbf{z}^{\mathbf{y}} + \sqrt{1 - \overline{\alpha}_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\overline{\alpha}_t := \prod_{s=1}^{t}(1 - \beta_s)$, and $\{\beta_1, \beta_2, \ldots, \beta_T\}$ is the noise schedule with T steps. At time-step T, the sample $\mathbf{z}^{\mathbf{y}}$ is degraded to pure Gaussian noise. In the *reversal* process, a neural network f_θ, usually a U-Net model (Ronneberger et al., 2015), is trained to iteratively remove noise from $\mathbf{z}^{\mathbf{y}}_{\mathbf{t}}$ to predict the clean sample $\mathbf{z}^{\mathbf{y}}$. The network is trained by sampling a random $t \in [1, T]$ and minimizing the loss function L_t.

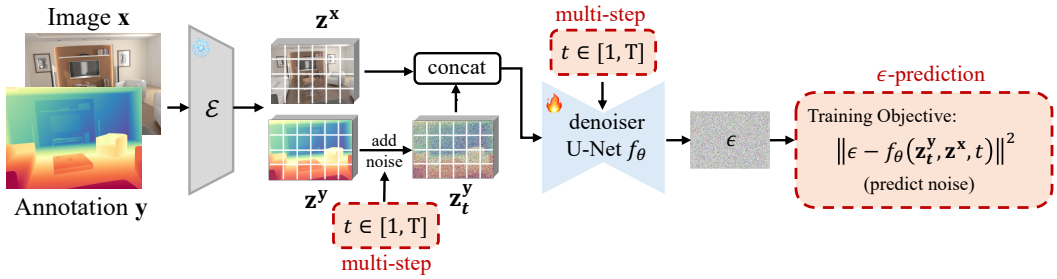

Figure 4: **Adaptation protocol of Direct Adaptation.** Starting with a pre-trained Stable Diffusion model, image $\mathbf{x}$ and annotation $\mathbf{y}$ are encoded using the pre-trained VAE. The noisy annotation $\mathbf{z}_\mathbf{t}^\mathbf{y}$ is obtained by adding noise at level $t \in [1, T]$. The U-Net input layer is coupled to accommodate the concatenated inputs and then fine-tuned using the standard diffusion objective, ϵ-prediction, under the original multi-step formulation.

Parameterization Types. To enable gradient computation for network training, there are two basic parameterizations of the loss function L_t. ① ϵ-prediction (Ho et al., 2020): the model f_θ learns to predict the added noise ϵ; ② x_0-prediction (Ho et al., 2020): the model f_θ learns to directly predict the clean sample $\mathbf{z}^\mathbf{y}$. The loss functions for these parameterizations are formulated as:

$$\begin{aligned} \epsilon\text{-prediction: } L_t^\epsilon &= ||\epsilon - f_\theta^\epsilon(\mathbf{z}_\mathbf{t}^\mathbf{y}, \mathbf{z}^\mathbf{x}, t)||^2, \\ x_0\text{-prediction: } L_t^\mathbf{z} &= ||\mathbf{z}^\mathbf{y} - f_\theta^\mathbf{z}(\mathbf{z}_\mathbf{t}^\mathbf{y}, \mathbf{z}^\mathbf{x}, t)||^2. \end{aligned} \tag{2}$$

where f_θ^* is the denoiser model to be learnt, $* \in \{\epsilon, \mathbf{z}\}$. ϵ-prediction is commonly chosen as the standard for parameterizing the denoising model, as it empirically achieves high-quality image generation with fine details and realism.

Denoising Process. DDIM (Song et al., 2020) is a key technique for multi-step diffusion models to achieve fast sampling, which implements an implicit probabilistic model that can significantly reduce the number of denoising steps while maintaining output quality. Formally, the denoising process from $\mathbf{z}_\tau^\mathbf{y}$ to $\mathbf{z}_{\tau-1}^\mathbf{y}$ is:

$$\mathbf{z}_{\tau-1}^\mathbf{y} = \sqrt{\overline{\alpha}_{\tau-1}}\hat{\mathbf{z}}_\tau^\mathbf{y} + \text{direction}(\mathbf{z}_\tau^\mathbf{y}) + \sigma_\tau \epsilon_\tau, \tag{3}$$

where $\hat{\mathbf{z}}_\tau^\mathbf{y}$ is the predicted clean sample at the denoising step τ, direction$(\mathbf{z}_\tau^\mathbf{y})$ represents the direction pointing to $\mathbf{z}_\tau^\mathbf{y}$ and σ_τ can be set to 0 if deterministic denoising is needed. And $\tau \in \{\tau_1, \tau_2, \ldots, \tau_S\}$, an increasing sub-sequence of the time-step set $[1, T]$, is used for fast sampling. During inference, DDIM iteratively denoises the sample from τ_S to τ_1 to obtain the clean one.

4 Methodology

We start our analysis by directly adapting the original diffusion formulation with minimal modifications as illustrated in Fig. 4. We call this starting point as "*Direct Adaptation*"[1]. Direct Adaptation is optimized using the standard diffusion objective as formulated in Eq. 2 (first row) and inferred by standard multi-step DDIM sampler. As shown in Tab. 3, Direct Adaptation fails to achieve satisfactory performance. In following sections, we will systematically analyze the key factors that affect adaptation performance step by step: parameterization types (Sec. 4.1); number of time-steps (Sec. 4.2); and the novel detail preserver (Sec. 4.3).

4.1 Parameterization Types

The type of parameterization is crucial, it not only determines the loss function discussed in Sec. 3, but also influences the inference process (Eq. 3). During inference, the predicted clean sample $\hat{\mathbf{z}}_\tau^\mathbf{y}$, a key component in Eq. 3, is calculated according to different parameterizations [2].

$$\begin{aligned} \epsilon\text{-prediction: } \hat{\mathbf{z}}_\tau^\mathbf{y} &= \frac{1}{\sqrt{\overline{\alpha}_\tau}}(\mathbf{z}_\tau^\mathbf{y} - \sqrt{1-\overline{\alpha}_\tau} f_\theta^\epsilon(\mathbf{z}_\tau^\mathbf{y}, \mathbf{z}^\mathbf{x}, \tau)), \\ x_0\text{-prediction: } \hat{\mathbf{z}}_\tau^\mathbf{y} &= f_\theta^\mathbf{z}(\mathbf{z}_\tau^\mathbf{y}, \mathbf{z}^\mathbf{x}, \tau). \end{aligned} \tag{4}$$

In the community, ϵ-prediction is chosen as the standard for image generation. However, it is not effective for dense prediction task. In the following, we will discuss the impact of different parameterization types in denoising inference process for dense prediction task.

[1] Details of Direct Adaptation will be provided in the supplementary materials.

[2] The latest parameterization, v-prediction, combines ϵ-prediction and x_0-prediction, producing results that are intermediate between the two. Please see the supplementary materials for more details.

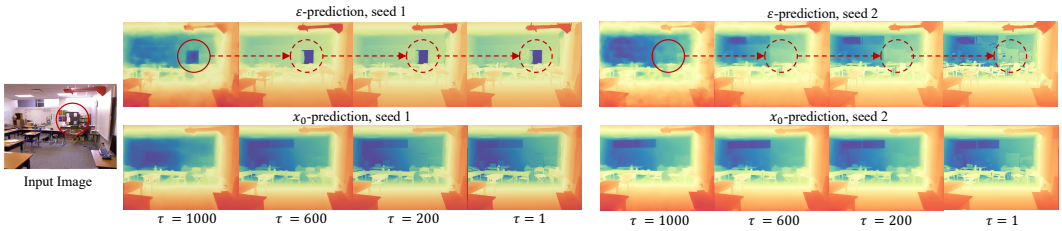

Figure 5: **Comparisons among different parameterizations using various seeds.** All models are trained on *Hypersim* (Roberts et al., 2021) and tested on the input image for depth estimation. The standard DDIM sampler is used with 50 denoising steps. Four steps are selected for clear illustration. From left (larger τ) to right (smaller τ) is the iterative denoising process.

Insights from the literature (Benny & Wolf, 2022; Salimans & Ho, 2022) reveal that ϵ-prediction introduces larger pixel variance compared to x_0-prediction, especially at the initial denoising steps (large τ). This variance mainly originates from the noise input. Specifically, for ϵ-prediction in Eq. 4, at initial denoising step, $\tau \to T$, the value $\frac{1}{\sqrt{\bar{\alpha}_\tau}} \to +\infty$. Thus, the prediction variance from $f_\theta^\epsilon(\mathbf{z}_\tau^\mathbf{y}, \mathbf{z}^\mathbf{x}, \tau)$ will be amplified significantly, resulting in large variance of predicted $\hat{\mathbf{z}}_\tau^\mathbf{y}$. In contrast, there is no coefficient for x_0-prediction to re-scale the model output, achieving more stable predictions of $\hat{\mathbf{z}}_\tau^\mathbf{y}$ at initial denoising steps. Subsequently, the predicted $\hat{\mathbf{z}}_\tau^\mathbf{y}$ is used in Eq. 3, where its coefficient $\sqrt{\bar{\alpha}_{\tau-1}}$ are same across the two parameterizations, and other terms are of the same order of magnitude. Therefore, the $\hat{\mathbf{z}}_\tau^\mathbf{y}$ predicted by ϵ-prediction, which has larger variance, exerts a more significant influence on denoising process. Since the process is iterative, this influence is continually preserved and maybe amplified.

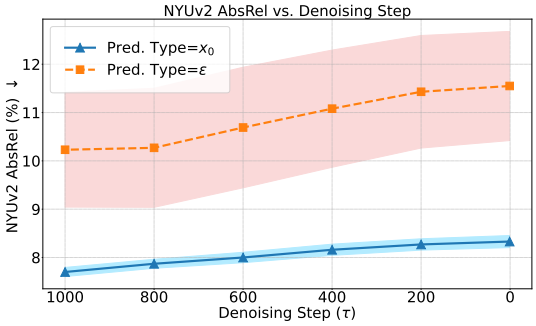

Figure 6: **Quantitative evaluation of the predicted depth maps $\hat{\mathbf{z}}_\tau^\mathbf{y}$ along the denoising process.** The experimental settings are same as Fig. 5. Six steps are selected for illustration. The banded regions around each line indicate the variance, wider areas representing larger variance.

We take the depth estimation as an example. During the inference process, we compute the predicted depth map $\hat{\mathbf{z}}_\tau^\mathbf{y}$ at each denoising step τ. As illustrated in Fig. 5, the depth maps predicted by ϵ-prediction significantly vary under different seeds while those predicted by x_0-prediction are more consistent. Although the large variance enhances diversity for image generation, it lead to unstable predictions in dense prediction tasks, potentially resulting in significant errors. For example in Fig. 5, the "dark gray cabinet" (highlighted in red circles) maybe wrongly considered as an "opened door" with significantly larger depth. While the predicted depth map *looks* more and more *plausible*, the error gradually propagates to the final prediction ($\tau = 1$) along the denoising process, indicating the persistent influence of the large variance. We further quantitatively measure the predicted depth maps by the absolute mean relative error (AbsRel) on NYUv2 dataset (Silberman et al., 2012). As shown in Fig. 6, ϵ-prediction exhibits higher error with much larger variance compared to x_0-prediction at the initial denoising steps ($\tau \to T$), and the prediction error propagates with a higher slope. In contrast, x_0-prediction, directly predicting $\hat{\mathbf{z}}_\tau^\mathbf{y}$

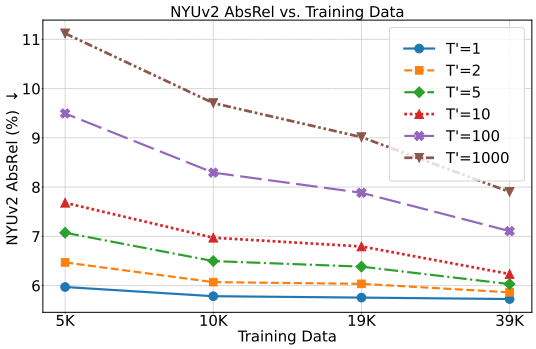

Figure 7: **Comparisons among various training time-steps and data scales** evaluated on NYUv2 in depth estimation. All models are fine-tuned on *Hypersim* using x_0-prediction. During inference, if $T' > 50$, the DDIM sampler is used with 50 denoising steps; otherwise, the number of denoising steps is equal to T'. The results demonstrate improved performance with decreased training time-steps. The single-step diffusion formulation ($T' = 1$) exhibits best performance across different data volumes.

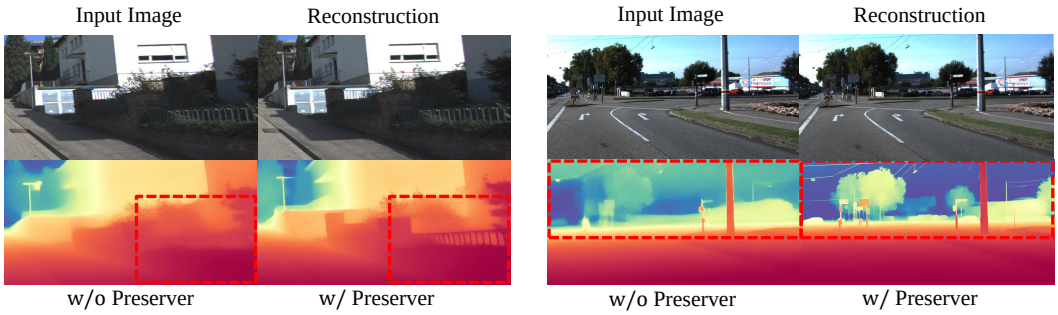

Figure 8: **Depth maps $w/$ and w/o the detail preserver and reconstruction outputs.** Fine-tuning the diffusion model for dense prediction tasks can potentially degrade its ability to generate highly detailed images, resulting in blurred predictions in regions with rich detail. To preserve these fine-grained details, we introduce a detail preserver that incorporates an additional reconstruction task, enhancing the model's capacity to produce more accurate dense annotations.

without any coefficients to amplify the prediction variance, yields more stable and correct dense predictions than ϵ-prediction. In conclusion, to mitigate the errors from large variance that adversely affect the performance of dense prediction, we replace the standard ϵ-prediction with the more tailored x_0-prediction.

4.2 Number of Time-Steps

Although x_0-prediction can improve the prediction quality, the multi-step diffusion formulation still leads to the propagation of predicted errors during the denoising process (Fig. 5, 6). Furthermore, utilizing multiple time-steps enhances the model's capacity, typically requiring large-scale training data to optimize and is beneficial for complex tasks such as image generation. However, for simpler tasks like dense prediction, where large-scale, high-quality training data is also scarce, employing multiple time-steps can make the model difficult to optimize. Additionally, training/inferring a multi-step diffusion model is slow and computation-intensive, hindering its practical application.

Therefore, to address these challenges, we propose fine-tuning the pre-trained diffusion model with fewer training time steps. Specifically, the original set of training time-steps is defined as $[1, T] = \{1, 2, 3, \dots, T\}$, where T denotes the total number of original training time-steps. We fine-tune the pre-trained diffusion model using a sub-sequence derived from this set. We define the length of this sub-sequence as T', where $T' \leqslant T$ and T is divisible by T'. This sub-sequence is obtained by evenly sampling the original set at intervals, defined as:

$$\{t_i = i \cdot k \mid i = 1, 2, \dots, T'\}, \tag{5}$$

where $k = T/T'$ is the sampling interval. During inference, the DDIM denoises the sample from noise to annotation using the same sub-sequence if $T' \leqslant 50$, otherwise we use 50 denoising steps.

As illustrated in Fig. 7, we conduct experiments by varying the number of time-steps T' under x_0-prediction. The results clearly show that the performance gradually improves as the number of time-steps is reduced, no matter the training data scales, culminating in the best result when reduced to only a single step. We further consider more strict scenarios with more limited training data to assess its impact on model optimization. As depicted in Fig. 7, these experiments reveal that the multi-step formulation is more sensitive to increases in training data scales compared with single-step. Notably, the single-step formulation consistently yields lower prediction errors and demonstrates greater stability. Although it is conceivable that multi-step and single-step formulations might achieve comparable performance with unlimited high-quality data, it's expensive and sometimes impractical in dense prediction.

Decreasing the number of denoising steps can reduce the optimization space of the diffusion model, leading to more effective and efficient adaption, as suggested by the above phenomenon. Therefore, for better adaptation performance under limited resource, we reduce the number of training time-steps of diffusion formulation to only one, and fixing the only time-step t to T. Additionally, the single-step formulation is much more computationally efficient. It also naturally prevents the harmful error propagation as discussed in Sec. 4.1, further enhancing the diffusion's adaptation performance in dense prediction.

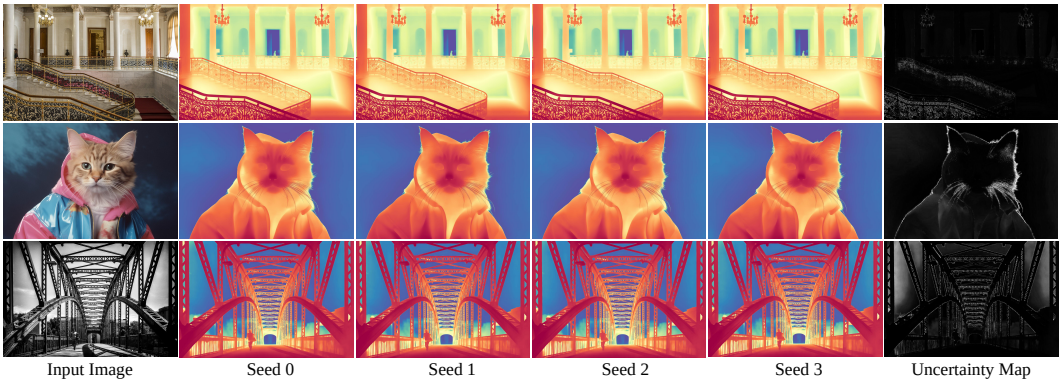

Figure 9: **Depth maps of multiple inferences and uncertainty maps.** Areas like the sky, object edges, and intricate details (*e.g.*, cat whiskers) typically exhibit high uncertainty.

4.3 Detail Preserver

Despite the effectiveness of the above designs, the model still struggles with processing detailed areas (Fig. 8, w/o Preserver). The original diffusion model excels at generating detailed images. However, when adapted to predict dense annotations, it can lose such detailed generation ability, due to unexpected catastrophic forgetting (Zhai et al., 2023; Du et al., 2024). This leads to challenges in predicting dense annotations in intricate regions.

To preserve the rich details of the input images, we introduce a novel regularization strategy called *Detail Preserver*. Inspired by previous works (Long et al., 2024; Fu et al., 2024), we utilize a task switcher $s \in \{s_x, s_y\}$, enabling the denoiser model f_θ to either generate annotation or reconstruct the input image. When activated by s_y, the model focuses on predicting annotation. Conversely, when s_x is selected, it reconstructs the input image. The switcher s is a one-dimensional vector encoded by the positional encoder and then added with the time embeddings of diffusion model, ensuring seamless domain switching without mutual interference. This dual capability enables the diffusion model to make detailed predictions and thus leading to better performance. Overall, the loss function L_t is:

$$L_t = ||\mathbf{z^x} - f_\theta(\mathbf{z_t^y}, \mathbf{z^x}, t, s_x)||^2 + ||\mathbf{z^y} - f_\theta(\mathbf{z_t^y}, \mathbf{z^x}, t, s_y)||^2, \tag{6}$$

where $t = T$ and thus $\mathbf{z_t^y}$ is a pure Gaussian noise.

4.4 Stochastic Nature of Diffusion Model

One major characteristic of generative models is their stochastic nature, which, in image generation, enables the production of diverse outputs. In perception tasks like dense prediction, this stochasticity has the potential to allow the model generating predictions with uncertainty maps. Specifically, for any input image, we can conduct multiple inferences using different initialization noises and aggregate these predictions to calculate its uncertainty map. Thanks to our systematic analysis and tailored fine-tuning protocol, our method effectively reduces excessive flickering (large variance), only allowing for more accurate uncertainty calculations in naturally uncertain areas, such as the sky, object edges, and fine details (*e.g.* cat whiskers), as shown in Fig. 9.

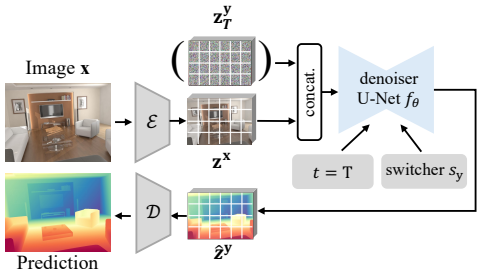

Figure 10: **Inference Pipeline of Lotus.** The noise $\mathbf{z_T^y}$ in bracket is used for **Lotus-G** and omitted for **Lotus-D**.

Most existing perception models are deterministic. To align with these, we can remove the noise input $\mathbf{z_t^y}$ and only input the encoded image features $\mathbf{z^x}$ to the U-Net denoiser. The model still performs well. In this paper, we finally present two versions of Lotus: **Lotus-G** (generative) with noise input and **Lotus-D** (discriminative) without noise input, catering to different needs.

4.5 Inference

The inference pipeline is illustrated in Fig. 10. We initialize the annotation map with standard Gaussian noise $\mathbf{z_T^y}$, and encode the input image into its latent code $\mathbf{z^x}$. The noise $\mathbf{z_T^y}$ and the image

$\mathbf{z}^{\mathbf{x}}$ are concatenated and fed into the denoiser U-Net model. In our single-step formulation, we set $t = T$ and the switcher to s_y. The denoiser U-Net model then predicts the latent code of the annotation map. The final annotation map is decoded from the predicted latent code via the VAE decoder. For deterministic prediction, we eliminate the Gaussian noise $\mathbf{z}_{\mathbf{T}}^{\mathbf{y}}$ and only feed the latent code of the input image into U-Net.

5 Experiments

5.1 Experimental Settings

Implementation details. We implement Lotus based on Stable Diffusion V2 (Rombach et al., 2022), without text conditioning. During training, we fix the time-step $t = 1000$. For depth estimation, we predict in disparity space, *i.e.*, $d = 1/d'$, where d represents the values in disparity space and d' denotes the true depth. For more details, please see the supplementary materials.

Training Datasets. Both depth and normal estimation are trained on two synthetic dataset covering indoor and outdoor scenes: ① *Hypersim* (Roberts et al., 2021) is a photorealistic synthetic dataset featuring 461 indoor scenes. We use the official training split, which contains approximately 54K samples. After filtering out incomplete samples, around 39K samples remain, all resized to 576×768 for training. ② *Virtual KITTI* (Cabon et al., 2020) is a synthetic street-scene dataset with five urban scenes under various imaging and weather conditions. We utilize four of these scenes for training, comprising about 20K samples. All samples are cropped to 352×1216, with the far plane at 80m.

Following Marigold (Ke et al., 2024), we probabilistically choose one of the two datasets and then draw samples from it for each batch (*Hypersim* 90% and *Virtual KITTI* 10%).

Evaluation Datasets and Metrics. ① For zero-shot affine-invariant depth estimation, we evaluate Lotus on NYUv2 (Silberman et al., 2012), ScanNet (Dai et al., 2017), KITTI (Geiger et al., 2013), and ETH3D (Schops et al., 2017) using absolute mean relative error (*AbsRel*), and also report $\delta 1$ and $\delta 2$ values. ② For surface normal prediction, we employ NYUv2, ScanNet, iBims-1 (Koch et al., 2018), and Sintel (Butler et al., 2012) datasets, reporting mean angular error (*m.*) as well as the percentage of pixels with an angular error below 11.25° and 30°. Please see supplementary materials for further details on the evaluation datasets and metrics.

5.2 Quantitative Comparisons

① For depth estimation (Tab. 1), Lotus-G demonstrates promising performance across all evaluation datasets, achieving the overall best rank compared to other generative baselines. Notice that we only require single step denoising process, also significantly boosting the inference speed as shown in Fig. 3. Lotus-D also performs well, achieving comparable results to DepthAnything series. It is worthy to notice that Lotus is trained on only 0.059M images compared to DepthAnything's 62.6M images. ② For normal estimation (Tab. 2), both Lotus-G and Lotus-D outperform all other generative and discriminative methods in terms of average ranking. Please see the supplementary materials for **Qualitative Comparisons**.

5.3 Ablation Study

As shown in Tab. 3, we conduct ablation studies to validate our designs. Starting with "Direct Adaptation", we incrementally test the effects of different components, such as parameterization types, the single-step diffusion process, and the detail preserver. Initially, we train the model using only the *Hypersim* dataset to establish a baseline. We then expand the training dataset using a mixture dataset strategy by including *Virtual KITTI*, aiming to enhance the model's generalization ability across different domains. The findings from these ablations validate the effectiveness of our proposed adaptation protocol, demonstrating that each design plays a vital role in optimizing the diffusion models for dense prediction tasks.

6 Conclusion and Future Work

In this paper, we introduce Lotus, a diffusion-based visual foundation model for dense prediction. Through systematic analysis and specifically tailored diffusion formulation, Lotus finds a way to better fit the rich visual prior from pre-trained diffusion models into dense prediction. Extensive experiments demonstrate that Lotus achieves promising performance on zero-shot depth and normal estimation with minimal training data, paving the way of various practical applications. Please see the supplementary materials for our discussion about **Applications** and **Future Work**.

Table 1: **Quantitative comparison on zero-shot affine-invariant depth estimation** between Lotus and SoTA methods. The upper section lists discriminative methods, the lower lists generative ones. The best and second best performances are highlighted. **Lotus-G** outperforms all others methods while **Lotus-D** is only slightly inferior to DepthAnything. Please note that DepthAnything is trained on 62.6M images while Lotus is only trained on 0.059M images. §indicates results revised by ourselves. ⋆denotes the method relies on pre-trained Stable Diffusion.

Method	Training Data	NYUv2 (Indoor)			KITTI (Outdoor)			ETH3D (Various)			ScanNet (Indoor)			Avg. Rank
		AbsRel↓	$\delta1$↑	$\delta2$↑	AbsRel↓	$\delta1$↑	$\delta2$↑	AbsRel↓	$\delta1$↑	$\delta2$↑	AbsRel↓	$\delta1$↑	$\delta2$↑	
DiverseDepth	320K	11.7	87.5	-	19.0	70.4	-	22.8	69.4	-	10.9	88.2	-	9.5
MiDaS	2M	11.1	88.5	-	23.6	63.0	-	18.4	75.2	-	12.1	84.6	-	9.5
LeRes	354K	9.0	91.6	-	14.9	78.4	-	17.1	77.7	-	9.1	91.7	-	7.6
Omnidata	12.2M	7.4	94.5	-	14.9	83.5	-	16.6	77.8	-	7.5	93.6	-	6.4
DPT	1.4M	9.8	90.3	-	10.0	90.1	-	7.8	94.6	-	8.2	93.4	-	5.5
HDN	300K	6.9	94.8	-	11.5	86.7	-	12.1	83.3	-	8.0	93.9	-	5.0
GenPercept⋆§	74K	5.6	96.0	99.2	13.0	84.2	-	7.0	95.6	98.8	6.2	96.1	99.1	3.8
DepthAnything V2	62.6M	4.5	97.9	99.3	7.4	94.6	98.6	13.1	86.5	-	4.2	97.8	99.3	2.9
Lotus-D (Ours)⋆	**59K**	**5.1**	**97.1**	**99.2**	**8.1**	**93.1**	**98.7**	**6.1**	**97.0**	**99.1**	**5.5**	**96.5**	**99.0**	**2.5**
DepthAnything	62.6M	4.3	98.1	99.6	7.6	94.7	99.2	12.7	88.2	-	4.3	98.1	99.6	2.0
GeoWizard⋆§	280K	5.6	96.3	99.1	14.4	82.0	96.6	6.6	95.8	98.4	6.4	95.0	98.4	3.3
Marigold$_{(LCM)}$§⋆	74K	6.1	95.8	99.0	9.8	91.8	98.7	6.8	95.6	99.0	6.9	94.6	98.6	3.2
Marigold⋆	74K	5.5	96.4	99.1	9.9	91.6	98.7	6.5	95.9	99.0	6.4	95.2	98.8	2.0
Lotus-G (Ours)⋆	**59K**	**5.4**	**96.8**	**99.2**	**8.9**	**92.0**	**98.4**	**6.3**	**96.6**	**99.1**	**6.1**	**95.6**	**98.8**	**1.2**

Table 2: **Quantitative comparison on zero-shot surface normal estimation** between Lotus and SoTA methods. Discriminative methods are shown in the upper section, generative methods in the lower. Both **Lotus-D** and **Lotus-G** outperform all other methods. ‡refers the Marigold normal model as detailed in this link. ⋆denotes the method relies on pre-trained Stable Diffusion.

Method	Training Data	NYUv2 (Indoor)			ScanNet (Indoor)			iBims-1 (Indoor)			Sintel (Outdoor)			Avg. Rank
		m.↓	11.25°↑	30°↑	m.↓	11.25°↑	30°↑	m.↓	11.25°↑	30°↑	m.↓	11.25°↑	30°↑	
OASIS	110K	29.2	23.8	60.7	32.8	15.4	52.6	32.6	23.5	57.4	43.1	7.0	35.7	7.0
Omnidata	12.2M	23.1	45.8	73.6	22.9	47.4	73.2	19.0	62.1	80.1	41.5	11.4	42.0	5.3
EESNU	2.5M	16.2	58.6	83.5	-	-	-	20.0	58.5	78.2	42.1	11.5	41.2	4.4
GenPercept⋆	74K	18.2	56.3	81.4	17.7	58.3	82.7	18.2	64.0	82.0	37.6	16.2	51.0	3.7
Omnidata V2	12.2M	17.2	55.5	83.0	16.2	60.2	84.7	18.2	63.9	81.1	40.5	14.7	43.5	3.6
DSINE	160K	16.4	59.6	83.5	16.2	61.0	84.4	17.1	67.4	82.3	34.9	21.5	52.7	1.8
Lotus-D (Ours)⋆	**59K**	**16.8**	**58.2**	**83.6**	**15.3**	**62.9**	**85.7**	**17.7**	**64.9**	**82.5**	**34.6**	**20.5**	**55.8**	**1.6**
Marigold‡⋆	74K	20.9	50.5	-	21.3	45.6	-	18.5	64.7	-	-	-	-	4.2
GeoWizard⋆	280K	18.9	50.7	81.5	17.4	53.8	83.5	19.3	63.0	80.3	40.3	12.3	43.5	3.2
StableNormal⋆	250K	18.6	53.5	81.7	17.1	57.4	84.1	18.2	65.0	82.4	36.7	14.1	50.7	2.0
Lotus-G (Ours)⋆	**59K**	**16.9**	**59.1**	**83.2**	**15.3**	**64.0**	**85.2**	**17.5**	**66.1**	**82.7**	**35.2**	**19.9**	**54.8**	**1.0**

Table 3: **Ablation studies** on the step-by-step design of our adaptation protocol for fitting pre-trained diffusion models into dense prediction. Here we show the results in monocular depth estimation.

Method	Training Data	NYUv2 (Indoor)			KITTI (Outdoor)			ETH3D (Various)			ScanNet (Indoor)		
		AbsRel↓	$\delta1$↑	$\delta2$↑	AbsRel↓	$\delta1$↑	$\delta2$↑	AbsRel↓	$\delta1$↑	$\delta2$↑	AbsRel↓	$\delta1$↑	$\delta2$↑
Direct Adaptation	39K	11.551	87.692	96.122	20.164	70.403	90.996	19.894	76.464	87.960	15.726	78.885	93.651
+ x_0-prediction	39K	8.332	92.769	97.941	17.008	74.969	93.611	11.075	87.952	94.978	10.212	89.130	97.181
+ Single Time-step	39K	5.587	96.272	99.113	13.262	83.210	97.237	7.586	94.143	97.678	6.262	95.394	98.791
+ Detail Preserver	39K	5.555	96.303	99.118	13.170	83.657	97.454	7.147	95.000	98.058	6.201	95.470	98.814
+ Mixture Dataset	59K	5.425	96.597	99.156	11.324	87.692	97.780	6.172	96.077	98.980	6.024	96.026	99.730
↪ − Noise Input	59K	5.334	96.729	99.198	9.334	92.813	98.795	6.846	95.290	98.899	5.982	96.287	99.087
+ Disparity Space (**Lotus-G**)	59K	5.398	96.844	99.204	8.932	91.977	98.352	6.337	96.655	99.070	6.042	95.559	98.857
↪ − Noise Input (**Lotus-D**)	59K	5.123	97.182	99.134	8.117	93.097	98.654	6.147	96.964	99.077	5.494	96.534	99.039

REFERENCES

Gilwon Bae and Andrew J Davison. Aleatoric uncertainty in monocular surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1472–1485, 2021.

Gilwon Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yaniv Benny and Lior Wolf. Dynamic dual-output diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11482–11491, 2022.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pp. 611–625. Springer, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 679–688, 2020.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Wenyu Du, Shuang Cheng, Tongxu Luo, Zihan Qiu, Zeyu Huang, Ka Chun Cheung, Reynold Cheng, and Jie Fu. Unlocking continual learning abilities in language models. *arXiv preprint arXiv:2406.17245*, 2024.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2002–2011, 2018.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jing He, Yiyi Zhou, Qi Zhang, Jun Peng, Yunhang Shen, Xiaoshuai Sun, Chao Chen, and Rongrong Ji. Pixelfolder: An efficient progressive pixel synthesis network for image generation. *arXiv preprint arXiv:2204.00833*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. doi: 10.1145/3641519.3657428.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18974, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 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.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7861–7871, 2024.

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

Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

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

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yunzhou Song, Jiahui Lei, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Track everything everywhere fast and robustly, 2024.

Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 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, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024b.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *arXiv preprint arXiv:2406.16864*, 2024.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2021a.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 204–213, 2021b.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3916–3925, 2022.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35: 14128–14139, 2022.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.