Lotus-2: Advancing Geometric Dense Prediction with Powerful Image Generative Model

Jing He[1] Haodong Li[1*] Mingzhi Sheng[1*] Ying-Cong Chen[1,2†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
{jhe812, hli736, msheng}@connect.hkust-gz.edu.cn yingcongchen@ust.hk

Figure 1. **We present Lotus-2, a two-stage deterministic framework for monocular geometric dense prediction.** Our method leverages pre-trained generative model as a deterministic world prior to achieve **new state-of-the-art accuracy** while requiring **remarkably minimal data** (trained on only 0.66% of the samples used by MoGe-2 [38]). The decoupled, two-stage design ensures both *structurally correct* inference and *high-fidelity* detail refinement. This figure demonstrates Lotus-2's robust zero-shot generalization with sharp geometric details, especially in challenging cases like oil paintings and transparent objects.

Abstract

Recovering pixel-wise geometric properties from a single image is fundamentally ill-posed due to appearance ambiguity and non-injective mappings between 2D observations and 3D structures. While discriminative regression models achieve strong performance through large-scale supervision, their success is bounded by the scale, quality and diversity of available data and limited physical reasoning. Recent diffusion models exhibit powerful world priors *that encode geometry and semantics learned from massive image–text data, yet directly reusing their stochastic generative formulation is suboptimal for deterministic geometric inference: the former is optimized for diverse and high-fidelity image generation, whereas the latter requires stable and accurate predictions. In this work, we propose Lotus-2, a two-stage deterministic framework for stable, accurate and fine-grained geometric dense prediction, aiming to provide an optimal adaption protocol to fully exploit the pre-trained generative priors. Specifically, in the first stage, the core predictor employs a single-step deterministic formulation with a clean-data objective and a lightweight local continuity module (LCM) to generate globally coherent structures without grid artifacts. In the second stage, the detail sharpener performs a constrained multi-step rectified-flow refinement within the manifold defined by the core predictor, enhancing fine-grained geometry through noise-free deterministic flow matching. Using only 59K training samples—less than 1% of existing large-scale datasets—Lotus-2 establishes new state-of-the-art results in monocular depth estimation and highly competitive surface normal prediction. These results demonstrate that diffusion models can serve as deterministic world priors, enabling high-quality geometric reasoning beyond traditional discriminative and generative paradigms.*

1. Introduction

Geometric dense prediction aims to recover pixel-wise geometric or physical properties, such as depth, surface normal, or albedo, from a single image. This problem lies at the foundation of modern visual understanding and serves as a cornerstone for various downstream applications, including controllable image generation [15, 43], 3D/4D reconstruction [16, 17, 27], and autonomous driving [10, 23, 24]. The mapping from image appearance to underlying geometry is inherently ill-posed: a single image can correspond to multiple plausible 3D interpretations. Consequently, a model must infer a physically plausible and globally coherent structure beyond what is directly observable from appearance.

Traditional approaches have long attempted to solve this problem through either geometric reasoning or discriminative learning. Early multi-view geometry and photometric consistency methods rely on strong assumptions about scene structure, lighting, and reflectance, making them unsuitable for single-view and complex real-world scenarios. With the rise of deep learning, discriminative models [6, 7, 21, 37–40, 42] have become the dominant paradigm by directly regressing geometric quantities from single images. While such models have achieved remarkable progress through increasingly powerful architectures and large-scale training, their performance remains fundamentally constrained by the scale, quality and diversity of available data. Human perception leverages strong world priors to resolve the ambiguity of geometric dense prediction, however, discriminative models trained on limited data distributions lack such mechanisms. Consequently, they perform poorly in rare and challenging scenes, involving transparency, reflection, and low texture, where inference requires reasoning beyond observable appearance. Even recent large-scale efforts—such as MoGe [37, 38] and DepthAnything [39, 40], trained on millions of samples—still rely heavily on distributional coverage rather than true scene understanding from world modeling, see Fig. **??** for reference.

The emergence of diffusion models such as Stable Diffusion [31] and FLUX [2] has revealed a new paradigm for visual reasoning. Trained on billions of diverse image–text pairs (*e.g.*, LAION-5B [33]), these models exhibit remarkable capability in synthesizing geometrically coherent and physically consistent imagery across diverse scenes. This success suggests that diffusion backbones implicitly encode *world priors*—rich internal representations of geometry and semantics accumulated through large-scale generative training.

With this intuition, recent works have attempted to repurpose the world priors for dense prediction [8, 11, 18, 20, 22, 36, 44]. While these studies validate the promise of generative world priors, most of them directly adopt the original generative formulation of diffusion models without rethinking its suitability for dense prediction. For example, Marigold [18] fine-tunes Stable Diffusion by reformulating depth estimation as an image-conditioned depth generation problem. Although this design benefits from the pre-trained priors, it overlooks the fundamental difference between dense prediction and image generation: the former requires deterministic and accurate inference, whereas the latter optimizes for diverse and high-fidelity generation through stochastic multi-step sampling. This fundamental mismatch often results in inconsistent and inaccurate geometric structure. Post-processing (*e.g.*, test-time ensembling [8, 18]) doesn't solve it in a native manner, and needs repeated predictions and may produce blurry results.

Motivated by these limitations, we revisit the role of diffusion-based generative models in dense prediction and

propose a new perspective: their true value lies not in the generative mechanism itself, but in the *world priors* encoded within their pre-trained weights. Instead of treating diffusion as a stochastic generator, we view it as a structured world prior that can guide the inference towards deterministic and geometrically accurate dense prediction. Based on this insight, we introduce *Lotus-2*, a two-stage deterministic framework that decouples accurate global geometry prediction from meticulous detail sculpting, effectively combining the strengths of regression and generative expressiveness.

In the first stage, a *core predictor* extracts globally coherent and accurate geometry through a simple yet effective adaptation of the rectified-flow formulation in FLUX [2]. By systematically analyzing the key designs of stochastic generative formulation, including the stochasticity, multi-step sampling and parameterization type, we identify that a single-step deterministic formulation under a clean-data prediction yields much better stable and accurate results than the original stochastic multi-step residual-based design. This single-step predictor is further enhanced with a lightweight *local continuity module (LCM)*, which mitigates grid artifacts introduced by the non-parametric Pack–Unpack operations in FLUX while maintaining architectural compatibility and efficiency.

In the second stage, an *optional detail sharpener* performs a detail refinement through a deterministic multi-step rectified-flow process. It operates within the constrained manifold defined by the core predictor and learns the transition from the "accurate" to "accurate and fine-grained" annotation, progressively enriching geometric details while preserving global structure and accuracy. This design bridges the gap between regression and generative modeling: the former ensures structural stability and correctness, while the latter contributes fine-grained realism. Consequently, Lotus-2 effectively leverages the generative priors in a disciplined and interpretable manner, achieving both geometric consistency and high-frequency detail fidelity without sacrificing efficiency and stability.

In summary, our key contributions are:

- **Revisiting the role of diffusion models for dense prediction.** We reformulate diffusion-based generative models from stochastic image generators to structured world priors, emphasizing that their strength lies in the world modeling capability embedded within pre-trained weights rather than in the sampling trajectory itself.
- **A two-stage deterministic framework integrating the strengths of regression and generative refinement.** We propose *Lotus-2*, which decouples structure prediction and detail refinement: a *core predictor* performs single-step, clean-data regression for accurate and stable geometric estimation, while an optional *detail sharpener* applies multi-step rectified-flow refinement within the constrained manifold defined by the predictor.
- **A principled adaptation of the rectified-flow formulation.** Through systematic analysis of several key designs in the original stochastic generative formulation, including stochasticity, multi-step sampling, parameterization type and local continuity, we demonstrate that the single-step clean-data deterministic design achieves higher accuracy and better optimization stability than traditional formulation optimized for image generation.
- **State-of-the-art performance.** With only 59K training samples—merely 0.66% of the data used by MoGe [37, 38] and 0.09% of that used by DepthAnything [39, 40], Lotus-2 achieves new state-of-the-art results on monocular depth estimation and highly competitive results on normal estimation.

2. Preliminaries

Our Lotus-2 framework is founded on the mathematical formalism of rectified-flow and the architectural foundation of FLUX model. This section introduces the necessary technical background related to our methodology.

2.1. Rectified-Flow Formulation

The rectified-flow (RF) formulation [25, 26] provides a robust and deterministic framework for modeling the transformation between two arbitrary probability measures via an ordinary differential equation (ODE). Specifically, given a source distribution p_1 and a target distribution p_0, the ODE on time-step $t \in [0, 1]$ is defined as: $d\mathbf{z_t} = v(\mathbf{z_t}, t)dt$, which maps $\mathbf{z_1} \sim p_1$ to $\mathbf{z_0} \sim p_0$ under the velocity vector field $v(\mathbf{z_t}, t)$. Crucially, the core principle of RF is to transport samples along the straight-line path:

$$\mathbf{z_t} = t\mathbf{z_1} + (1 - t)\mathbf{z_0}, \tag{1}$$

thus the target vector field $\mathbf{v}$ is given by $\mathbf{v} = \frac{d\mathbf{z_t}}{dt} = \mathbf{z_1} - \mathbf{z_0}$. This straight-line mechanism fundamentally differs from the high-curvature paths of denoising diffusion models [13], which ensures a high efficiency and reduced error accumulation. For training, the velocity vector field $v(\mathbf{z_t}, t)$ is parameterized by a neural network f_θ, which is optimized by minimizing the distance to the target vector field $\mathbf{v}$. The loss function is thus defined as:

$$L_t = ||\mathbf{v} - f_\theta(\mathbf{z_t}, t)||^2 \tag{2}$$

$$= ||(\mathbf{z_1} - \mathbf{z_0}) - f_\theta(\mathbf{z_t}, t)||^2. \tag{3}$$

In practice, the expectation over the continuous time $t \in [0, 1]$ is approximated by randomly sampling a discrete time-step value from a pre-defined set at each training iteration. Given a total of T training time-steps, the pre-defined time-step set is:

$$\{t_i = \frac{i}{T} \mid i = 1, 2, \ldots, T\}. \tag{4}$$

During sampling (inference), the discrete Euler solver is used to iteratively generate the target sample ($t = 0$) from the source ($t = 1$). Formally, the iterative sampling process from current state $\mathbf{z}_{\mathbf{t}_{\text{curr}}}$ to next state $\mathbf{z}_{\mathbf{t}_{\text{next}}}$ is given by:

$$\mathbf{z}_{\mathbf{t}_{\text{next}}} = \mathbf{z}_{\mathbf{t}_{\text{curr}}} - \eta \cdot f_\theta(\mathbf{z}_{\mathbf{t}_{\text{curr}}}, t), \tag{5}$$

where $t_{\text{next}} < t_{\text{curr}}$ and η ($0 < \eta \leq 1$) denotes the step size, which is determined by the total number of inference time-steps T_{inf}.

2.2. Architectural Foundation of FLUX

We leverage the architecture and weights of FLUX[2], which utilizes a pre-trained variational autoencoder (VAE) to compress high-dimensional image data $\mathbf{x}$ into a compact latent space $\mathcal{Z}$. The VAE consists of an encoder E and a decoder D, where $E(\mathbf{x}) = \mathbf{z}^{\mathbf{x}}$ maps the image to a latent code, and $D(\mathbf{z}^{\mathbf{x}}) = \hat{\mathbf{x}}$ attempts to reconstruct the image from the latent code. The rectified-flow formulation of FLUX operates within this VAE latent space $\mathcal{Z}$.

In the specific task of image generation, the starting distribution p_1 is set to standard Gaussian noise in the latent space, *i.e.*, $\mathbf{z_1} \sim \mathcal{N}(0, I)$. The target distribution p_0 is the distribution of real, clean image latent, *i.e.*, $\mathbf{z_0} = E(\mathbf{x}) = \mathbf{z}^{\mathbf{x}}$. Based on this setup, the loss function in Eq. 2 is rewritten by:

$$L_t = ||(\epsilon - \mathbf{z}^{\mathbf{x}}) - f_\theta(\mathbf{z_t}, t)||^2. \tag{6}$$

Here, $\mathbf{z_t} = t\epsilon + (1-t)\mathbf{z}^{\mathbf{x}}$ is the linear interpolation between the noise and the target latent code. FLUX adopts the DiT (Diffusion Transformer) [28] architecture as its model f_θ.

The Pack-Unpack Operations in FLUX. To reduce computational overhead and memory usage, FLUX applies paired *Pack* and *Unpack* operations around the DiT model in the latent space. Pack is a parameter-free down-sampling procedure that rearranges the latent feature by grouping every non-overlapping 2×2 patch into the channel dimension,

$$\text{Pack} : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 4C}. \tag{7}$$

Conversely, Unpack restores the original resolution by inverting this rearrangement,

$$\text{Unpack} : \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 4C} \rightarrow \mathbb{R}^{H \times W \times C}. \tag{8}$$

This Pack-Unpack operation, while efficient, introduces a critical challenge: because it is parameter-free, it can introduce noticeable local pixel discontinuities ("grid-artifacts"). This issue is especially severe under the single-step formulation, degrading the overall quality and realism of the outputs.

3. Lotus-2

In this section, we present Lotus-2, a two-stage deterministic framework for stable, accurate and high-fidelity dense prediction, aiming to provide an optimal adaption protocol to effective and efficient leverage the pre-trained world priors of FLUX [2]. **We argue that directly inheriting the stochastic generative formulation—which is optimized for image synthesis—introduces instability and unnecessary complexity for deterministic geometric tasks.** The image synthesis aims at diverse and high-fidelity generation through stochastic multi-step sampling, while the dense prediction requires a deterministic and accurate inference. This fundamental misalignment results in high structural variance and significant prediction errors for dense prediction, thereby compromising overall accuracy. To better exploit the generative world priors, we propose a decoupled, two-stage adaption protocol. We first introduce the *Core Predictor* (Sec. 3.1) derived through a systematic analysis of the standard generative formulation, including its stochasticity (Sec. 3.1.1), multi-step iterative sampling (Sec. 3.1.2), parameterization type (Sec. 3.1.3), and local continuity (Sec. 3.1.4). This core predictor is dedicated solely to achieving highly-accurate and robust global geometry estimation. Subsequently, we address the challenge of fine-grained fidelity by proposing the *Detail Sharpener* (Sec. 3.2), which employs a constrained multi-step rectified-flow formulation designed only for meticulous detail sculpting within the established structural manifold. This decoupled, two-stage approach successfully achieves both structural accuracy and fine-grained fidelity, with its complete inference process detailed in Sec. 3.3.

3.1. Core Predictor: Robust and Accurate Geometric Prediction

3.1.1. Analysis-1: Stochastic *v.s.* Deterministic Formulation

Initial efforts to leverage diffusion priors for geometric dense prediction (*e.g.*, Marigold [18], GeoWizard [8]) inherit the model's original *stochastic generative formulation*. We term this approach as *Stochastic Direct Adaptation* (Stochastic-DA). In this setup, the process is framed as an image-conditioned geometric generation task: the model learns the flow from pure Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to the target geometry $\mathbf{z}^{\mathbf{y}}$ conditioned on the input image $\mathbf{z}^{\mathbf{x}}$ as illustrated in Fig. 2. Specifically, the latent feature at time t is defined as:

$$\mathbf{z_t} = t\epsilon + (1-t)\mathbf{z}^{\mathbf{y}}. \tag{9}$$

The neural network f_θ is trained to predict the velocity field $\mathbf{v} = \epsilon - \mathbf{z}^{\mathbf{y}}$ by incorporating the image latent $\mathbf{z}^{\mathbf{x}}$ as a conditional input (typically concatenated along the channel dimension of the input feature to the DiT backbone). The loss function for optimizing this stochastic generative formulation is given by:

$$L_t = ||(\epsilon - \mathbf{z}^{\mathbf{y}}) - f_\theta(\mathbf{z_t}, \mathbf{z}^{\mathbf{x}}, t)||^2. \tag{10}$$

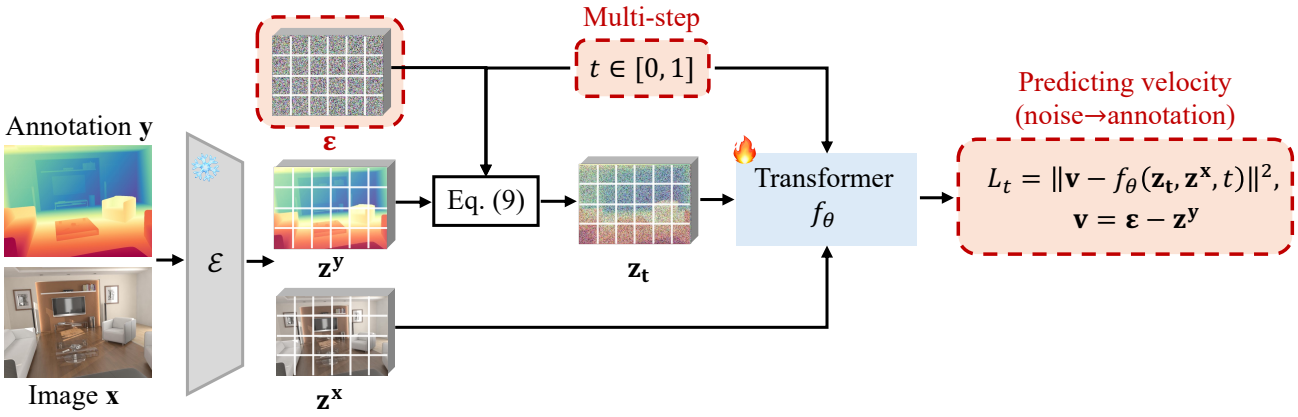


Figure 2. **Adaptation protocol of stochastic formulation (Stochastic-DA).** This framework models a conditional generative flow by estimating the velocity field from a random noise latent ϵ to the annotation latent $\mathbf{z^y}$, conditioned on the image latent $\mathbf{z^x}$. The target velocity vector is $\mathbf{v} = \epsilon - \mathbf{z^y}$. This inherent reliance on noise initialization inherently leads to non-deterministic variance in deterministic geometric prediction.

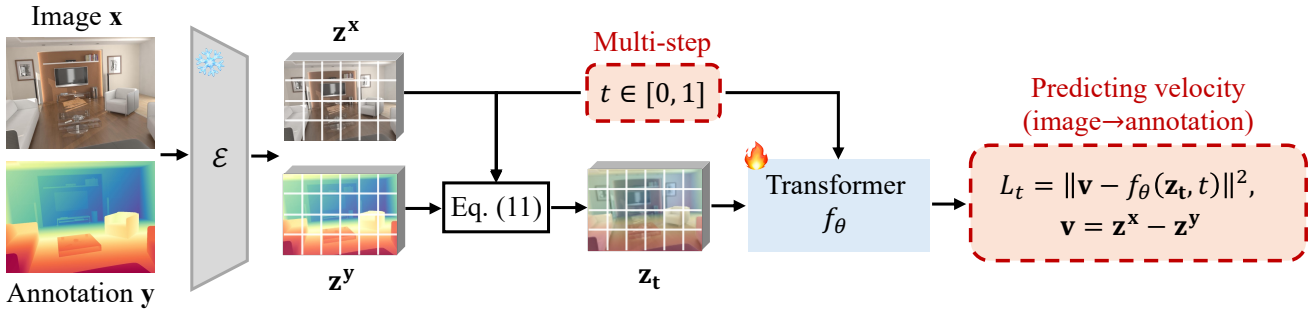


Figure 3. **Adaptation protocol of deterministic formulation (Deterministic-DA).** This architecture shifts the paradigm to a noise-free rectified-flow formulation. It directly estimates the velocity field from the source image latent $\mathbf{z^x}$ to the target annotation latent $\mathbf{z^y}$, where the target velocity vector is $\mathbf{v} = \mathbf{z^x} - \mathbf{z^y}$. This deterministic setup ensures the stability and structurally consistency for geometric dense prediction.

The core limitation of this approach is its inherent non-deterministic variance. Because the inference must begin with an initial sample of pure Gaussian noise, $\mathbf{z_1} = \epsilon \sim \mathcal{N}(0, I)$, different random initializations lead to diverse outputs, resulting in inconsistent geometric structures for the same input image, as illustrated in Fig. 4. This variance is beneficial for diverse image generation; however, it leads to physically implausible geometric structures for dense prediction, thus hindering accuracy. While ensemble averaging is commonly used to mitigate this variance, it inherently introduces prediction bias and also compromises overall accuracy by blending both correct and incorrect structural hypotheses.

To resolve this fundamental mismatch, we discard the stochastic conditional generative formulation and shift the paradigm to a purely deterministic flow matching between two distributions. We formulate the problem as learning a noise-free transformation between the image feature $\mathbf{z^x}$ and the geometric feature $\mathbf{z^y}$, directly utilizing the inherent determinism of the rectified-flow framework. We term this approach as *Deterministic Direct Adaptation* (Deterministic-DA) of the rectified-flow formulation. The architecture for this approach is illustrated in Fig. 3. Specifically, Deterministic-DA defines the two distributions as the image and annotation spaces, respectively: the source is the image latent $\mathbf{z^x}$ and the target is the annotation latent $\mathbf{z^y}$. The latent feature at time t is defined as:

$$\mathbf{z_t} = t\mathbf{z^x} + (1 - t)\mathbf{z^y}, \tag{11}$$

where the model f_θ is trained to predict the velocity $\mathbf{v} = \mathbf{z^x} - \mathbf{z^y}$. The training objective for this deterministic flow is:

$$L_t = ||(\mathbf{z^x} - \mathbf{z^y}) - f_\theta(\mathbf{z_t}, t)||^2. \tag{12}$$

This approach is inherently noise-free during both training

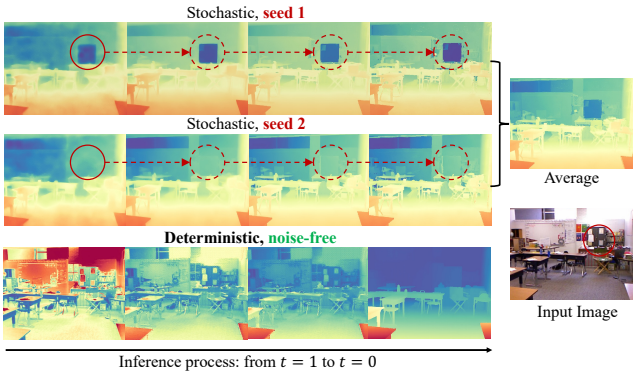


Figure 4. **Comparison between stochastic and deterministic formulation.** The figure visualizes the iterative inference process from $t = 1$ to $t = 0$. The stochastic formulation (Stochastic-DA) exhibits significant structural variance: distinct random noise initializations yield inconsistent geometric structures across the entire inference process (highlighted in red circles). While averaging is employed to mitigate the variance, the final prediction remains compromised by the blending of conflicting structural hypotheses. In contrast, the deterministic formulation (Deterministic-DA) ensures a noise-free and stable trajectory, preventing structural variance and improving geometric coherence and prediction accuracy.

and inference. As shown in the Fig. 4 and Tab. 3, this deterministic approach significantly improves structural consistency and prediction accuracy compared to its stochastic counterpart.

3.1.2. Analysis-2, Multi-Step Iterative Sampling

While the multi-step formulation enhances the capacity of generative models, it is optimized for high-fidelity image synthesis and demands large-scale training data. For dense geometric prediction, where high-quality supervision data is scarce, this inherited multi-step formulation is computationally intensive and makes the model difficult to optimize effectively. Furthermore, the prediction errors are accumulated during this multi-step iterative sampling, further compromising the accuracy. The iterative nature also hinders its practical application due to slow inference speeds.

To address these challenges, we propose fine-tuning the pre-trained rectified-flow model with fewer training time-steps. As illustrated in Fig. 6, we conduct experiments by gradually reducing the number of training time-steps T. This is achieved by modifying the value of T in Eq. 4 to define new, smaller time-step sets for training. The results clearly show that the performance gradually improves as the number of time-steps T is reduced, culminating in the best result when reduced to only a single step. Under a stricter setting with more limited training data, the multi-step formulation is more sensitive to variations in training data scale compared to the single-step formulation. The single-step formulation demonstrates greater stability and yields lower prediction errors. While it is plausible that, given unlimited high-quality data, both multi- and single-step formulations could reach comparable performance, such a setting is often costly and impractical for dense prediction tasks.

Reducing the number of training time steps T constrains the optimization space of rectified-flow formulation, thereby enabling more effective and efficient adaptation for geometric dense prediction. Motivated by this observation, we adopt the single-step formulation ($T = 1$, *i.e.*, $t = 1$ in Eq. 4). This single-step formulation further enhances computational efficiency.

3.1.3. Analysis-3, Parameterization Types

Under the single-step formulation derived above, the model degenerates into a regression task, which is trained to predict the velocity given the input image with a fixed time-step $t = 1$. The velocity $\mathbf{v} = \mathbf{z}^{\mathbf{x}} - \mathbf{z}^{\mathbf{y}}$ is the residual between the input image $\mathbf{z}^{\mathbf{x}}$ and its annotation $\mathbf{z}^{\mathbf{y}}$. We refer to this parameterization as *Residual Prediction*. During inference, the final prediction is obtained using the single-step Euler solver:

$$\hat{\mathbf{z}}^{\mathbf{y}} = \mathbf{z}^{\mathbf{x}} - f_\theta(\mathbf{z}^{\mathbf{x}}, t), \tag{13}$$

where $f_\theta(\mathbf{z}^{\mathbf{x}}, t)$ is the predicted residual.

However, such residual prediction is problematic for dense prediction tasks for two reasons: 172 Predicting $\mathbf{z}^{\mathbf{x}} - \mathbf{z}^{\mathbf{y}}$ requires the model to simultaneously learn image reconstruction and geometric estimation, which belong to substantially different distributions. This increases optimization difficulty and ultimately degrades accuracy; 173 The predicted residual is dominated by high-frequency appearance signals of the input image, such as textures, illumination, and color. Although the term $\mathbf{z}^{\mathbf{x}}$ in Eq. 13 attempts to remove these appearance components during inference, however, imperfect prediction makes this removal unreliable, and appearance interference inevitably leaks into the final result.

To overcome these limitations and better exploit pre-trained visual priors, we propose fine-tuning the model with *Clean-Data Prediction*, *i.e.*, directly predicting the clean annotation $\mathbf{z}^{\mathbf{y}}$. The clean-data prediction offers a simpler and more direct training objective, alleviates optimization difficulty, and eliminates appearance interference, thereby yielding superior performance.

As shown in Fig. 7, residual prediction produces predictions corrupted by image patterns (see red circles), whereas clean-data prediction yields accurate results without such interference. Consistently, Tab. 3 shows that clean-data prediction achieves significantly higher accuracy than the original residual prediction. Therefore, to mitigate appearance interference and improve prediction quality, we adopt clean-data prediction as the parameterization type.

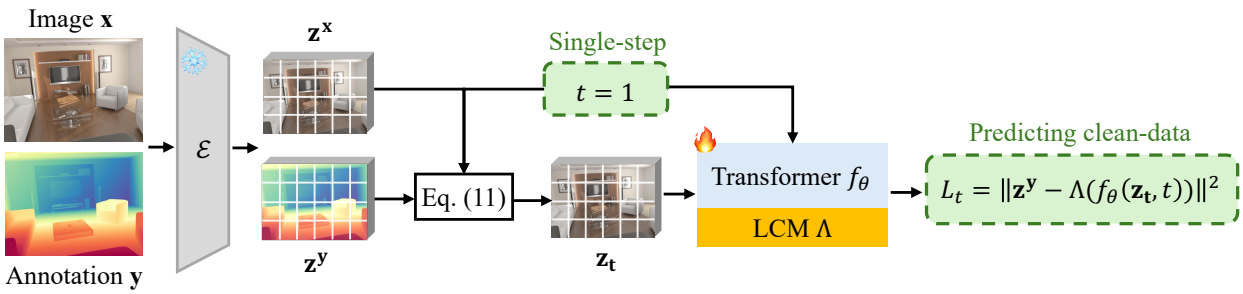


Figure 5. **Adaptation protocol of the core predictor in Lotus-2.** It adopts a single-step formulation ($t = 1$) with clean-data prediction to efficiently exploit the world priors of pre-trained FLUX model, where input latent $\mathbf{z_t}$ is equivalent to the image latent $\mathbf{z}^{\mathbf{x}}$, *i.e*, $\mathbf{z_t} = \mathbf{z_1} = \mathbf{z}^{\mathbf{x}}$ according to the Eq. 11. In addition, there is a pair of Pack-Unpack operations around the diffusion Transformer f_θ inherited from FLUX, a local continuity module (LCM) Λ is employed to mitigate grid artifacts caused by this Unpack operation.

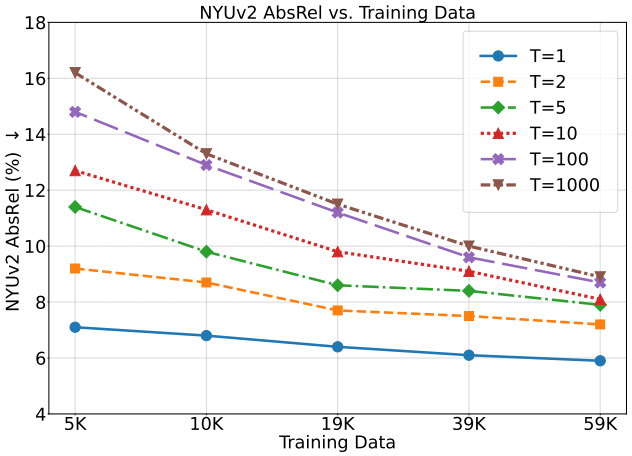


Figure 6. **Comparisons among various training time-steps and data scales** evaluated on NYUv2 in depth estimation. During inference, if the number of training time-steps $T > 50$, the inference time-steps are fixed at $T_{\text{inf}} = 50$; otherwise, $T_{\text{inf}} = T$. The results show that, when adapting the pre-trained rectified-flow model to dense prediction, reducing the number of training time-steps leads to improved performance. In particular, the single-step formulation ($T = 1$) achieves the best performance across all data scales.

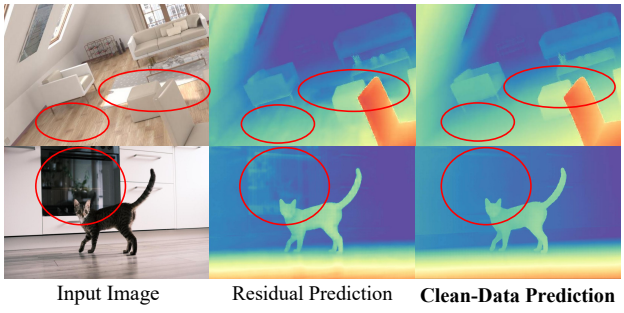


Figure 7. **Predictions under Different Model Parameterization Types.** Red circles highlight regions with obvious appearance artifacts when residual prediction is used. In contrast, clean-data prediction produces more accurate predictions without interference from image appearance.

3.1.4. Analysis-4, Local Continuity

The FLUX architecture employs non-parametric *Pack* and *Unpack* operations to reduce computational overhead in the latent space for image generation (Sec. 2.2). While efficient, the non-parametric nature of the Unpack operation, which rearranges feature channels back to spatial resolution after diffusion Transformer model, introduces spatial discontinuities at the boundaries of the 2×2 latent patches. This localized discontinuity, which lacks constraints on local spatial coherence, is detrimental to geometric fidelity in the final output (Fig. 8, "w/o LCM").

To address this issue without compromising efficiency, we propose the lightweight *Local Continuity Module* (LCM) after the Unpack operation of diffusion Transformer backbone, as shown in Fig. 5. LCM consists of two 3×3 convolutional layers with an intermediate GELU activation [12] to introduce nonlinearity, which is formally defined as:

$$\hat{\mathbf{z}}^{\mathbf{y}} = \Lambda\big(f_\theta(\mathbf{z_t}, t)\big), \quad \Lambda(h) = \phi_2 \circledast \gamma(\phi_1 \circledast h), \tag{14}$$

where $\Lambda(\cdot)$ denotes the LCM, $\circledast$ is the convolution operator, ϕ_1 and ϕ_2 are convolutional kernels, and $\gamma(\cdot)$ is the GELU activation.

As shown in Fig .8, LCM effectively mitigates the local discontinuities introduced by Pack-Unpack, thereby eliminating grid artifacts. Furthermore, Tab. 3 demonstrates that LCM not only improves visual quality but also enhances prediction accuracy.

For comparison, we additionally evaluate a straightforward alternative: entirely removing the Pack-Unpack operations from FLUX architecture. While the removal of Pack-Unpack does eliminate grid artifacts (see the "w/o Pack-Unpack" cases in Fig. 8), this approach suffers from two

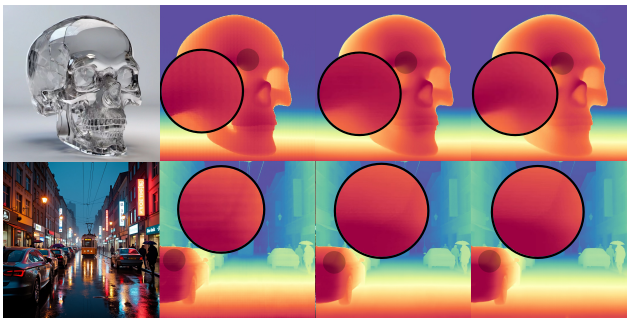


Figure 8. **The effects of different strategies for eliminating grid-like artifacts.** "w/o LCM" refers to only single-step formulation with clean-data prediction, which produces noticeable grid-like artifacts due to the discontinuity introduced by Pack-Unpack. Removing Pack-Unpack entirely alleviates this issue but compromises both accuracy and efficiency. In contrast, LCM effectively resolves the artifacts while improving accuracy and preserving model efficiency. (**Zoom in for clearer observation.**)

severe drawbacks: 172 since the input–output dimensionality of the diffusion Transformer changes, additional linear layers are required to align the dimensions, which causes the feature space shifts away from the pre-trained priors, degrading the prediction accuracy (see Tab. 3); 173 the absence of Pack-Unpack drastically compromises model efficiency, leading to much slower inference speed. Therefore, LCM offers an effective solution to the local discontinuity problem, while preserving the pre-trained priors and maintaining model efficiency.

3.1.5. Finalized Architecture and Objective

The final core predictor is built upon the foundational Deterministic-DA and integrates all derived components: the single-step formulation, the clean-data prediction, and the local continuity module (LCM), as shown in the Fig. 5. This comprehensive design transforms the instable and iterative generative flow into a highly efficient and structurally robust formulation, optimizing for deterministic geometric dense prediction. The overall training objective is defined as:

$$L_t = ||\mathbf{z}^{\mathbf{y}} - \Lambda(f_\theta(\mathbf{z_t}, t))||^2, \tag{15}$$

where $t = 1$ and the input latent $\mathbf{z_t} = \mathbf{z_1} = \mathbf{z}^{\mathbf{x}}$.

3.2. Detail Sharpener: High-Fidelity Geometric Refinement

The single-step core predictor excels at predicting accurate and globally coherent structure, but often produces predictions that are coarse and blurry in high-frequency detail areas, lacking fine-grained fidelity (see "w/o Sharpener" cases of Fig. 11). This limitation stems from the inherent difficulty of the single-step formulation in resolving high-frequency details. In contrast, multi-step flow (*e.g.*, Deterministic-DA) retains the complexity to model high-frequency dynamics and can produce sharper details; however, due to its optimization difficulty and the accumulation of high errors across multiple steps, it is prone to geometric hallucination (see the "Deterministic-DA" cases of Fig. 11), sacrificing structural correctness and overall accuracy.

To simultaneously achieve accuracy and fine-grained fidelity, we introduce the *Detail Sharpener*, a constrained multi-step rectified-flow model designed solely for geometric refinement within the manifold defined by the core predictor. Specifically, we first obtain a structurally correct but coarse prediction via the single-step core predictor, and then employ detail sharpener to progressively refine the high-frequency details. With this design, structural correctness is guaranteed by the core predictor, while the detail sharpener is solely responsible for enhancing sharpness.

As illustrated in Fig. 9, the detail sharpener is trained to learn a noise-free rectified-flow transformation from a coarse prediction $\mathbf{z}^{\mathbf{y_c}}$ to its high-fidelity ground-truth $\mathbf{z}^{\mathbf{y_f}}$. The flow is defined between the two known geometric states:

$$\mathbf{z_t} = t\mathbf{z}^{\mathbf{y_c}} + (1-t)\mathbf{z}^{\mathbf{y_f}}. \tag{16}$$

The model g_θ is fine-tuned from FLUX to predict the velocity $\mathbf{v} = \mathbf{z}^{\mathbf{y_c}} - \mathbf{z}^{\mathbf{y_f}}$. Thus, the training objective of detail sharpener is defined as:

$$L_t = ||(\mathbf{z}^{\mathbf{y_c}} - \mathbf{z}^{\mathbf{y_f}}) - g_\theta(\mathbf{z_t}, t)||^2. \tag{17}$$

We set the number of training steps $T' = 10$ to balance optimization and refinement. During inference, the number of inference steps T'_{inf} can be flexibly chosen up to T', depending on the desired level of sharpness.

As shown in Fig. 11, the detail sharpener noticeably enhances the sharpness while successfully avoiding the structural hallucinations observed in Deterministic-DA. Tab. 3 further confirms that incorporating the detail sharpener does not compromise geometric accuracy established by the core predictor.

3.3. Inference

Lotus-2 executes a two-stage deterministic inference pipeline, as illustrated in Fig. 10. The core predictor is dedicated to ensuring structural correctness and efficiency, while the detail sharpener is solely responsible for high-fidelity refinement. Rooted in our philosophy of deterministic modeling, both the core predictor and the detail sharpener are noise-free, guaranteeing structural consistency and stability for deterministic geometric dense prediction. The complete inference process proceeds as follows:

1. The input image $\mathbf{x}$ is first encoded into the VAE latent space using the encoder E, yielding the image latent $\mathbf{z}^{\mathbf{x}}$.

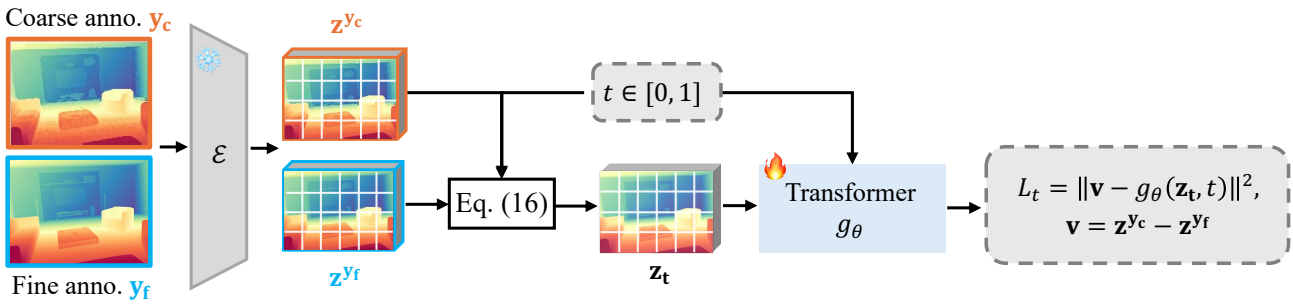


Figure 9. **The training pipeline of detail sharpener.** Starting from a structurally correct but coarse annotation predicted by the core predictor, the detail sharpener learns the transition from coarse to fine-grained annotation via a constrained multi-step rectified-flow within the manifold defined by the core predictor.

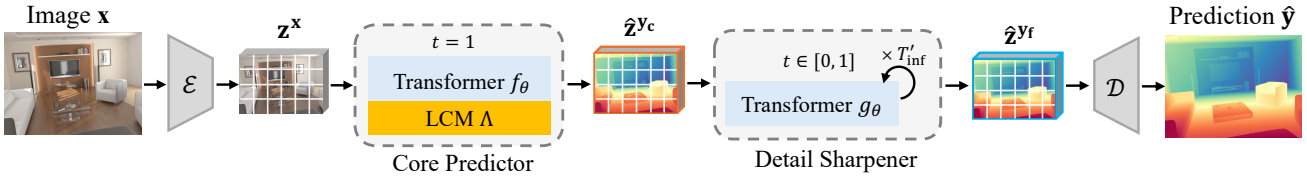


Figure 10. **The inference pipeline of Lotus-2.** It is a decoupled, two-stage deterministic pipeline that bridges the regression and geometric refinement. First, the core predictor produces stable and structurally consistent prediction via single-step regression. The detail sharpener then employs a constrained multi-step rectified-flow formulation to iteratively refinement without any stochastic noise. The refinement uses $T'_{\text{inf}} \leq 10$ steps, adjustable based on the desired level of sharpness. This design ensures both structural consistency and fine-grained fidelity in minimal steps.

2. The image latent $\mathbf{z}^{\mathbf{x}}$ is passed through the core predictor to generate the accurate but coarse prediction $\hat{\mathbf{z}}^{\mathbf{y_c}}$. This step guarantees global structural correctness and is performed with maximum efficiency (1 step).
3. The coarse prediction $\hat{\mathbf{z}}^{\mathbf{y_c}}$ is then fed into the detail sharpener to obtain the sharp and high-fidelity result $\hat{\mathbf{z}}^{\mathbf{y_f}}$. This iterative refinement is achieved by the discrete Euler solver (Eq. 5). Note that this refinement is optional based on the desired level of sharpness.
4. The final refined latent $\hat{\mathbf{z}}^{\mathbf{y_f}}$ is decoded back to the pixel space using the VAE decoder D to produce the final geometric prediction $\hat{\mathbf{y}}$.

4. Experiments

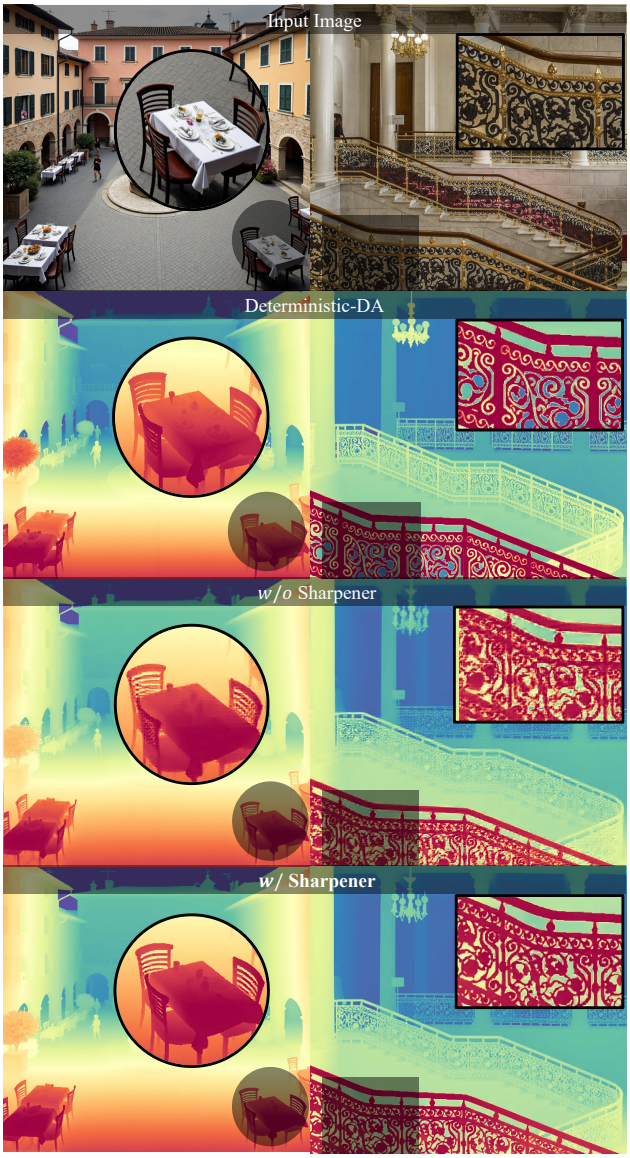
In this section, we systematically validate the design principles of Lotus-2: leveraging pre-trained generative priors as a stable and deterministic flow for structurally correct and high-fidelity geometric dense prediction. We first detail the experimental setup, then present a quantitative comparison against state-of-the-art methods, followed by comprehensive ablation studies validating our methodological contributions.

4.1. Experimental Settings

4.1.1. Implementation Details

We implement the proposed Lotus-2, which includes both the core predictor and the detail sharpener, by fine-tuning the pre-trained FLUX model [2] without utilizing the text conditioning. Our design adapts the rectified-flow formulation by setting the core predictor to a single-step formulation ($T = 1$, $t = 1$) with clean-data prediction and the detail sharpener to a constrained multi-step rectified-flow formulation ($T' = 10$, with time-steps defined by Eq. 4) within the manifold defined by the core predictor. For optimization, we use the Adam optimizer with a learning rate of 1×10^{-4}. All models are trained on 8 NVIDIA H100 GPUs (80G) with a total batch size of 64. To adapt the large-scale pre-trained architecture, we employ the parameter-efficient method LoRA [14], using a rank of 128 for depth estimation and 256 for normal estimation. For depth estimation, we operate in the disparity space, *i.e.*, $d = 1/d'$, where d' is the true depth. During inference, the core predictor directly predicts the coarse but structurally correct prediction in single inference step, while the detail sharpener utilizes the Euler sampler with $T'_{\text{inf}} = T' = 10$ steps for refinement.

Figure 11. **Comparisons in Detail Sharpness.** "w/o Sharpener" denotes predictions directly obtained by the core predictor, which suffer from blurry and coarse details. The "$w/$ Sharpener" cases demonstrate that the detail sharpener noticeably enhances the sharpness of fine-grained structures, particularly along boundaries, while avoiding the geometric hallucinations observed in Deterministic-DA, such as the misaligned chair backrest and stair railing.



4.1.2. Training Datasets

A core demonstration of this work is the ability to achieve SoTA performance using extremely limited supervised data. Both depth and normal estimation tasks are trained solely on a small collection of synthetic data, totaling approximately **59K samples**—a fraction of the millions used by large-scale discriminative models.

- *Hypersim* [30]: A photorealistic synthetic dataset of 461 indoor scenes. We utilize the official training split, retaining approximately 39K samples after filtering. All samples are resized to 576×768.
- *Virtual KITTI* (VKITTI) [4]: A synthetic street-scene dataset covering five urban scenes. We use four scenes, comprising about 20K samples, cropped to 352×1216.

To train the detail sharpener, we implement the methodology described in Sec. 3.2: we first generate coarse predictions ($\mathbf{y_c}$) on Hypersim and VKITTI using the trained core predictor, and then train the detail sharpener on the flow defined between these coarse predictions and the ground truth ($\mathbf{y_f}$).

4.1.3. Evaluation Datasets

We evaluate the generalization capability of Lotus-2 across five real-world datasets for depth estimation and four for surface normal prediction, none of which were seen during training.

- *Affine-Invariant Depth Estimation*: We evaluate across diverse indoor (NYUv2 [34], ScanNet [5]), outdoor (KITTI [9]), and high-resolution mixed scenes (ETH3D [32], DIODE [35]).
- *Surface Normal Prediction*: We use NYUv2, ScanNet, and iBims-1 [19] for real indoor evaluation, and Sintel [3] for highly dynamic synthetic outdoor scenes.

4.1.4. Metrics

We employ widely accepted metrics for both affine-invariant depth estimation and surface normal prediction.

- *Affine-Invariant Depth Estimation*: Following standard protocols [18, 29], we firstly align predictions to ground truth via least-squares fitting before evaluation. We report two primary metrics: the *absolute mean relative error* (AbsRel), defined as $\frac{1}{M}\sum_{i=1}^{M}|a_i - d_i|/d_i$ (lower is better); and the $\delta 1$ value, which is the proportion of pixels satisfying $\text{Max}(a_i/d_i, d_i/a_i) < 1.25$ (higher is better).
- *Surface Normal Prediction*: Following [1, 41], we measure the *mean angular error* (lower is better) and the percentage of pixels with an angular error below 11.25° (higher is better).

For overall comparison, we report the *Avg. Rank*, which is the average ranking of each method across all datasets and metrics. A lower Avg. Rank signifies superior overall performance.

4.2. Comparison with State-of-the-Art

We benchmark Lotus-2 against recent state-of-the-art methods in both affine-invariant monocular depth estimation and surface normal prediction, including both large-scale discriminative models (*e.g.*, DepthAnything [39, 40], MoGe [37, 38]) and generative prior adaptation methods (*e.g.*, Marigold [18], GeoWizard [8]).

Table 1. **Quantitative comparison on zero-shot affine-invariant depth estimation** between Lotus-2 and SoTA methods. The best and second best performances are highlighted. §indicates results re-evaluated by ourselves using the evaluation protocol of Marigold [18]. ⋆denotes the method relies on pre-trained text-to-image generative models. Ours Lotus-2 achieves the best overall performance than all other methods.

2*Method	Training Data↓	NYUv2 (Indoor) AbsRel↓	NYUv2 (Indoor) $\delta1$↑	KITTI (Outdoor) AbsRel↓	KITTI (Outdoor) $\delta1$↑	ETH3D (Various) AbsRel↓	ETH3D (Various) $\delta1$↑	ScanNet (Indoor) AbsRel↓	ScanNet (Indoor) $\delta1$↑	DIODE (Various) AbsRel↓	DIODE (Various) $\delta1$↑	Avg. Rank
DiverseDepth	320K	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	19.5
MiDaS	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	18.7
LeRes	354K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	15.7
Omnidata	12.2M	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	15.4
DPT	1.4M	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	18.2	75.8	12.5
GeoWizard⋆§	280K	5.6	96.3	14.4	82.0	6.6	95.8	6.4	95.0	33.5	72.3	12.4
HDN	300K	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	78.0	12.2
GenPercept⋆§	74K	5.6	96.0	13.0	84.2	7.0	95.6	6.2	96.1	35.7	75.6	11.5
Marigold$_{(LCM)}$⋆§	74K	6.1	95.8	9.8	91.8	6.8	95.6	6.9	94.6	30.7	77.5	10.5
MoGe-2§	8.9M	3.6	98	11.8	89.2	16.6	81.5	3.5	98.2	39.3	70.0	10.4
Marigold⋆	74K	5.5	96.4	9.9	91.6	6.5	95.9	6.4	95.2	30.8	77.3	9.2
DICEPTION⋆	500K	7.2	93.9	7.5	94.5	5.3	96.7	7.5	93.8	24.3	74.1	9.2
DepthAnything V2	62.6M	4.5	97.9	7.4	94.6	13.1	86.5	4.2	97.8	26.5	73.4	7.3
Diffusion-E2E-FT⋆	74K	5.4	96.5	9.6	92.1	6.4	95.9	5.8	96.5	30.3	77.6	7.1
Lotus-G⋆	59K	5.4	96.8	8.5	92.2	5.9	97.0	5.9	95.7	22.9	72.9	7.1
DepthFM-ID⋆	81.4K	5.5	96.3	8.9	91.3	5.8	96.2	6.3	95.4	21.2	80.0	6.9
MoGe§	9M	3.6	97.9	7.3	95.2	8.4	93.0	3.5	98.4	36.3	71.2	6.9
DepthAnything	62.6M	4.3	98.1	7.6	94.7	12.7	88.2	4.3	98.1	26.0	75.9	6.2
Lotus-D⋆	59K	5.1	97.2	8.1	93.1	6.1	97.0	5.5	96.5	22.8	73.8	6.0
Lotus-2⋆	**59K**	**4.1**	**97.6**	**6.7**	**94.5**	**4.6**	**98.1**	**4.2**	**97.6**	**22.1**	**75.2**	**3.6**

Table 2. **Quantitative comparison on zero-shot surface normal estimation** between Lotus-2 and SoTA methods. ‡refers the Marigold normal model as detailed in this link. §indicates results re-evaluated by us using the evaluation protocol of DSINE [1]. Our Lotus-2 demonstrates highly competitive quantitative performance, crucially delivering the robust and fine-grained qualitative results as highlighted in Fig. **??**.

2*Method	Training Data↓	NYUv2 (Indoor) mean↓	NYUv2 (Indoor) 11.25°↑	ScanNet (Indoor) mean↓	ScanNet (Indoor) 11.25°↑	iBims-1 (Indoor) mean↓	iBims-1 (Indoor) 11.25°↑	Sintel (Outdoor) mean↓	Sintel (Outdoor) 11.25°↑	Avg. Rank
OASIS	110K	29.2	23.8	32.8	15.4	32.6	23.5	43.1	7.0	13.5
Omnidata	12.2M	23.1	45.8	22.9	47.4	19.0	62.1	41.5	11.4	11.9
GeoWizard⋆§	280K	18.9	50.7	17.4	53.8	19.3	63.0	40.3	12.3	10.4
StableNormal⋆§	250K	18.6	53.5	17.1	57.4	18.2	65.0	36.7	14.1	8.4
GenPercept⋆§	74K	18.2	56.3	17.7	58.3	18.2	64.0	37.6	16.2	8.3
EESNU	2.5M	16.2	58.6	-	-	20.0	58.5	42.1	11.5	7.3
Omnidata V2	12.2M	17.2	55.5	16.2	60.2	18.2	63.9	40.5	14.7	8.1
Marigold⋆‡	74K	20.9	50.5	21.3	45.6	18.5	64.7	-	-	8.1
Lotus-G⋆	59K	16.5	59.4	15.1	63.9	17.2	66.2	33.6	21.0	5.4
DSINE	160K	16.4	59.6	16.2	61.0	17.1	67.4	34.9	21.5	4.9
Diffusion-E2E-FT⋆§	74K	16.5	60.4	14.7	66.1	16.1	69.7	33.5	22.3	3.4
Lotus-D⋆	59K	16.2	59.8	14.7	64.0	17.1	66.4	32.3	22.4	3.4
Lotus-2⋆	**59K**	**16.9**	**59.0**	**14.2**	**66.8**	**15.4**	**70.4**	**30.3**	**27.6**	**2.9**
MoGe-2§	8.9M	14.7	62.3	12.8	68.4	14.7	70.4	29.3	24.8	1.1

4.2.1. Affine-Invariant Depth Estimation

As presented in Tab. 1, Lotus-2 establishes a new state-of-the-art in affine-invariant monocular depth estimation across the five real-world datasets. Notably, Lotus-2 achieves the best Avg. Rank despite being trained on only 59K samples. This result decisively validates the power of leveraging large-scale generative models as deterministic world priors, allowing Lotus-2 to surpass massive data-trained discriminative methods.

4.2.2. Surface Normal Prediction

For surface normal prediction, Lotus-2 demonstrates highly competitive performance (Tab. 2), showcasing the effectiveness of our deterministic adaptation in capturing complex geometry. Crucially, as highlighted in Fig. **??**, our deterministic adaption of world priors ensures robust and structurally correct geometric prediction, enabling strong generalization even in challenging or rare scenes. This robust foundation, coupled with our noise-free multi-step refinement (detail sharpener), proves highly effective at capturing the high-frequency surface detail required for local geometry, significantly outperforming other SoTA approaches.

4.3. Ablation Studies

4.3.1. Ablation on the Core Predictor

The core predictor is the structural foundation of Lotus-2. We systematically validate its design in Tab. 3 by incrementally incorporating the core contributions, showing consistent performance superiority across all four evaluation datasets.

We begin by validating the necessity of the deterministic formulation. Moving from the stochastic generative formulation (Stochastic-DA) to the noise-free deterministic formulation (Deterministic-DA) yields an immediate improvement in accuracy. This validates our core hypothesis that deterministic geometric prediction requires a stable flow (Sec. 3.1.1). Next, adopting the single-step formulation ($T = 1$) also provides a significant performance increase, confirming the single-step mechanism is the optimal strategy for efficiently leveraging pre-trained world priors under limited data (Sec. 3.1.2). Following this, switching to clean-data prediction from residual prediction consistently achieves higher structural accuracy. This confirms that its value lies in both eliminating high-frequency appearance interference (Fig. 7) and providing a more direct and effective optimization target (Sec. 3.1.3). Finally, we validate the local continuity module (LCM). This lightweight module successfully eliminates grid artifacts (Fig. 8) and provides the final accuracy boost. This contrasts with the "w/o Pack-Unpack" alternative, which compromises efficiency and degrades performance due to feature space misalignment (Sec. 3.1.4).

4.3.2. Ablation on the Detail Sharpener

The detail sharpener is responsible for high-fidelity refinement via a constrained multi-step flow. This ablation validates the contribution of the detail sharpener to high geometric fidelity in both qualitative and quantitative manner and a spectral analysis.

As qualitatively demonstrated in Fig. 11, the detail sharpener achieves noticeable refinement in high-frequency areas. Quantitatively, the final line item in Tab. 3 shows that the multi-step flow of the detail sharpener maintains the near-optimal accuracy achieved by the core predictor. This preservation of accuracy confirms that the detail sharpener successfully operates on a decoupled objective—enhancing local fidelity—without compromising the structural accuracy established by the core predictor, thus validating the success of our two-stage design.

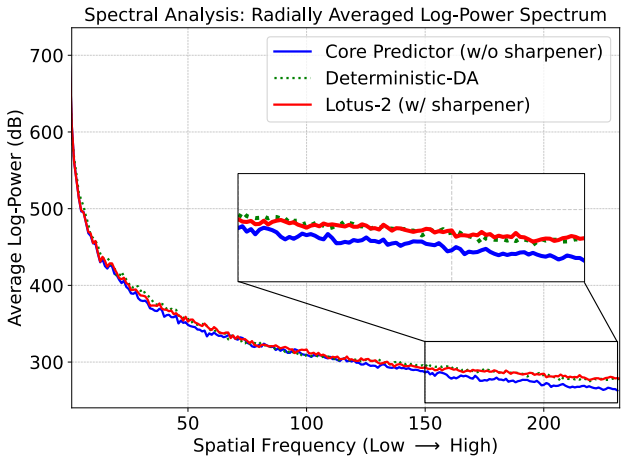


Figure 12. **Spectral analysis of high-fidelity refinement.** This plot compares the average log-power (y-axis) across spatial frequencies (x-axis) on NYUv2 dataset to validate the contribution of detail sharpener. The decay of the core predictor (w/o sharpener) curve confirms its coarse nature, while the Lotus-2 ($w/$ sharpener) curve shows recovery of high-frequency power.

To rigorously quantify the contribution of the detail sharpener to fine-grained fidelity, specifically its effect on high-frequency detail areas, we conduct a spectral analysis using the 1D radially averaged power spectrum as illustrated in Fig. 12. The results show that the prediction from the core predictor exhibits a clear decay in power at high frequencies, confirming its output is structurally correct but coarse. In contrast, both the Deterministic-DA and the our Lotus-2 retain significantly more high-frequency power, indicating successful detail refinement. This provides quantitative, signal-level evidence that the detail sharpener is essential for high-fidelity geometric prediction.

5. Conclusion

In this work, we addressed the fundamental challenge of geometric dense prediction—the task's ill-posed nature—by proposing a critical shift in how large-scale generative models are leveraged. We established the principle that for deterministic geometric inference, the power of diffusion backbones lies not in their stochastic sampling process but in their implicitly embedded deterministic world priors. Directly reusing the original stochastic generative flow proves suboptimal, leading to structural variance and unacceptable inconsistency in geometric outputs.

Table 3. **Ablation studies** of the proposed Lotus-2. The second portion of the table contains the key components of the *core predictor*, sequentially demonstrating the performance gains conferred by each design. The final row validates the *detail sharpener*. The shaded row (w/o Pack-Unpack) is included as an auxiliary ablation to validate the effect of the local continuity module (LCM). The results below are evaluated in monocular depth estimation across four datasets.

2*Method	NYUv2 (Indoor)		KITTI (Outdoor)		ETH3D (Various)		ScanNet (Indoor)	
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Stochastic-DA	8.261	93.468	13.196	78.204	17.384	77.842	9.373	91.569
Deterministic-DA	7.812	94.262	10.212	89.900	10.766	94.762	8.488	92.897
+ Single-Step Formulation	5.910	96.939	8.833	92.088	5.858	96.952	7.121	96.331
+ Clean-Data Prediction	4.384	97.627	6.843	94.325	4.980	97.552	4.446	97.529
+ Local Continuity Module	4.128	97.608	6.576	94.682	4.625	98.004	4.174	97.575
(w/o Pack-Unpack)	4.817	97.383	6.966	94.203	5.728	97.252	4.723	97.168
+ Detail Sharpener	4.122	97.623	6.767	94.492	4.643	98.101	4.188	97.597

To fully exploit these priors in a disciplined and stable manner, we introduced Lotus-2, a novel two-stage deterministic framework that decouples the inference process into two specialized, noise-free rectified-flow mappings.

The first stage, the *core predictor*, is implemented for maximum structural accuracy and efficiency. Through systematic ablation, we validated the necessity of our derived design choices: the deterministic shift, the highly efficient single-step formulation ($T = 1$), and the clean-data prediction objective, which together transform the complex generative flow into a robust geometric regressor. The lightweight local continuity module (LCM) further ensures fidelity by suppressing architectural artifacts without compromising efficiency.

The second stage, the *detail sharpener*, solves the final limitation of single-step regression—coarse high-frequency details. It performs a constrained multi-step refinement within the geometry manifold established by the core predictor. This process is inherently noise-free and is optimized to selectively enhance high-fidelity geometry without compromising the established global structural correctness, successfully validating the benefits of our decoupled design.

The experimental results decisively confirm our core hypothesis. By training on only 59K synthetic samples—less than 1% of existing large-scale datasets—Lotus-2 achieved new state-of-the-art performance in monocular depth estimation and demonstrated highly competitive results in surface normal prediction. This unprecedented data efficiency, combined with high inference stability and fine-grained fidelity, validates the efficacy of our deterministic adaptation protocol.

Ultimately, this work demonstrates that the vast knowledge accumulated by generative diffusion models can be repurposed to enable efficient, accurate, and physically consistent geometric reasoning, setting a new paradigm for structured prediction tasks beyond traditional discriminative and generative methods. This finding opens promising avenues for future research into extracting and utilizing structured knowledge from foundational generative models.

References

[1] Gilwon Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 10, 11

[2] BFL.ai. Bfl.ai announces the flux.1 suite of models, 2024. 2, 3, 4, 9

[3] 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*, pages 611–625. Springer, 2012. 10

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 10

[5] 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*, pages 5828–5839, 2017. 10

[6] 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*, pages 10786–10796, 2021. 2

[7] 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. 2

[8] 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 esti-

mation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 2, 4, 10

[9] 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. 10

[10] Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024. 2

[11] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 2

[12] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 9

[15] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2

[16] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 2

[17] Lutao Jiang, Jiantao Lin, Kanghao Chen, Wenhang Ge, Xin Yang, Yifan Jiang, Yuanhuiyi Lyu, Xu Zheng, Yinchuan Li, and Yingcong Chen. Dimer: Disentangled mesh reconstruction model. *arXiv preprint arXiv:2504.17670*, 2025. 2

[18] 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*, pages 9492–9502, 2024. 2, 4, 10, 11

[19] 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*, pages 0–0, 2018. 10

[20] 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*, pages 7861–7871, 2024. 2

[21] Haodong Li, Hao Lu, and Ying-Cong Chen. Bi-tta: Bidirectional test-time adapter for remote physiological measurement. In *European Conference on Computer Vision*, pages 356–374. Springer, 2024. 2

[22] Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. Da ^{2}: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025. 2

[23] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 2

[24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[26] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[27] 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*, pages 9970–9980, 2024. 2

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4

[29] 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. 10

[30] 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*, pages 10912–10922, 2021. 10

[31] 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*, pages 10684–10695, 2022. 2

[32] 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*, pages 3260–3269, 2017. 10

[33] 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. 2

[34] 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*, pages 746–760. Springer, 2012. 10

[35] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 10

[36] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025. 2

[37] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 2, 3, 10

[38] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 1, 2, 3, 10

[39] 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*, pages 10371–10381, 2024. 2, 3, 10

[40] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2, 3, 10

[41] 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. 10

[42] 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*, pages 3916–3925, 2022. 2

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[44] Canyu Zhao, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. *arXiv preprint*, 2025. 2