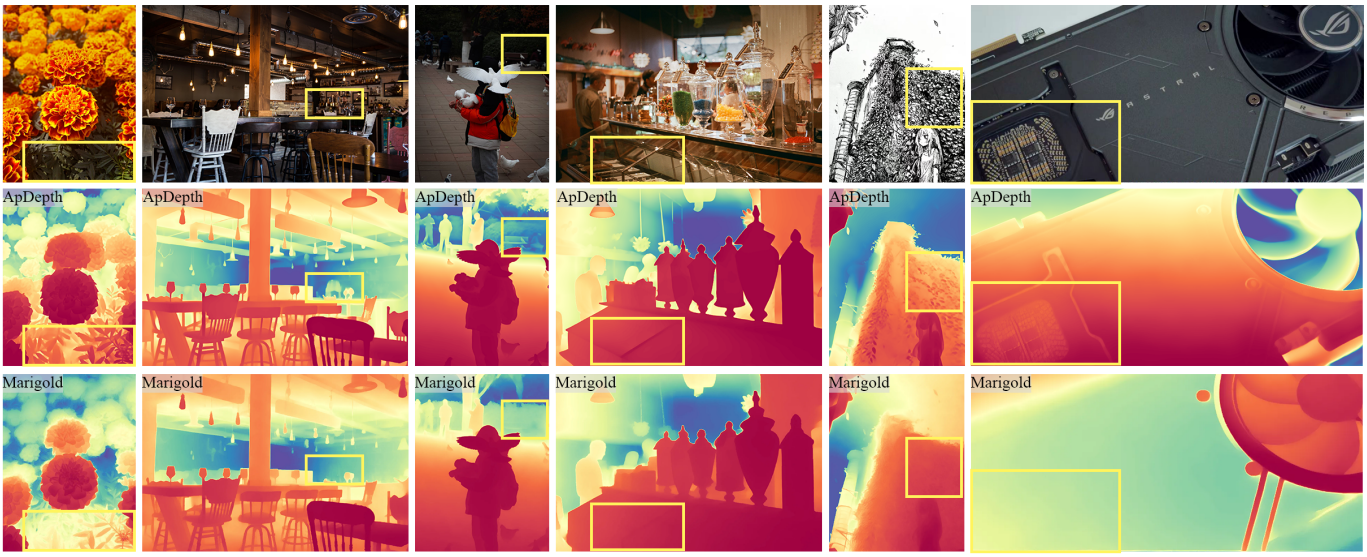
ApDepth: Aiming for Precise Monocular Depth Estimation Based on Diffusion Models

Jiawei Wang[1†], Jingxuan Wang[2†]

[1]*Shenyang Jianzhu University*
[2]*Beijing University of Chinese Medicine*
† Those authors contributed equally to this work.



Abstract

Monocular depth estimation (MDE) aims to recover per-pixel depth information from a single 2D image, and it plays a significant role in many fields such as autonomous driving, 3D reconstruction, robotics and so on. Significant progress has been made in MDE recently, and these advancements can be primarily categorized into two major methods: data-driven and model-driven. Data-driven method such as Depth Anything V2 have achieved promising results. Meanwhile, model-driven methods, mainly based on diffusion models, have shown great potential, yet they still offer vast room for further investigation. Existing diffusion-based methods face two major challenges: multi-step iterative inference incurs prohibitive runtime, while single-step deterministic inference often fails to preserve fine-grained details.

To address these limitations, we propose ApDepth, a novel diffusion-based framework for accurate and detailed monocular depth estimation with single-step inference. ApDepth ApDepth introduces a novel architecture that simultaneously leverages the knowledge a priori from both data-driven and model-driven approaches, while maintaining minimal resource overhead, effectively balancing inference efficiency and detail preservation. Extensive experiments demonstrate that ApDepth achieves competitive or superior performance across multiple benchmarks.

Codes are available at: https://haruko386.github.io/research/.

1 Introduction

Monocular depth estimation (MDE) is a fundamental task in computer vision, with applications major in autonomous driving, robotics, and augmented reality. The goal of MDE is to predict a dense depth map from a single RGB image, which is inherently an ill-posed problem due to the loss of 3D information during the projection from 3D to 2D. To address these issues, the model must possess a deep understanding of the current scene. Existing methods on Zero-shot monocular depth estimation can be broadly categorized into two main methods: data-driven and model-driven. Data-driven methods have achieved significant progress in recent years, leveraging large-scale annotated datasets to learn complex mappings from images to depth maps. Notable examples such as Depth Anything V2[21] and scaledepth[47] have demonstrated impressive performance on various benchmarks. However, these methods often struggle with data collection and their training time is usually very long.

In contrast, model-driven methods, particularly those based on diffusion models[39], have shown great potential in generating high-quality depth maps. For example, DiffusionDepth[5]is the first to reformulate the monocular depth estimation problem as a diffusion denoising approach. Marigold[32] presentes a fine-tuning protocol for Stable Diffusion[28], achieveing impressive results in both global structure and local details. However, the iterative denoising process results in low inference speed. GenPercept[41] proposes a deterministic single-step paragram, significantly reduced the time required for inference. Although these approaches have yielded promising results, they are largely constrained by extremely long inference times or poor local detail.

Stable Diffusion has two main components: image-to-latent (I2L) encoder-decoder and a denoising U-Net. I2L encoder-decoder is responsible for compressing the

input image into a latent representation and reconstructing the image from the latent space. The denoising U-Net is trained to iteratively refine noisy latent representations towards clean latent representations, conditioned on the input image and a time step. According to GenPercept[41], the primary perceptual priorknowledge of diffusion models is encapsulated within the U-Net of the diffusion model, while the I2L encoder-decoder primarily serves as a compression and reconstruction mechanism, exerting minimal influence on the outcome of image reconstruction. Hence, one of our improvement approaches is to provide U-Net with richer feature representations, enabling it to fully leverage the prior knowledge it has learned from large-scale data on the large-scale image and text dataset LAION-5B[31]

To address the issues raised above, we propose **ApDepth**, a novel diffusion-based framework for accurate and detailed monocular depth estimation with single-step inference. ApDepth modifies Stable Diffusion 2's stochastic multi-step generation by adopting a deterministic one-step perception approach, accelerating inference time from 24 seconds to 120 milliseconds for 640×480 images. To solve the problem of poor local detail inherent in single-step inference, **ApDepth** introduces a pretrained MDE Model, which is data-driven. In this way, our model can leverage not only the rich prior knowledge embedded in Stable Diffusion but also the complex, non-linear statistical correlations (texture-depth, semantic context) learned by the data-driven model. This provides the U-Net with richer feature representations, resulting in significantly improvement. We also introduced a two-stage training strategy that employs MSE and FFT loss functions to guide the diffusion model in learning more edge details. Briefly, our contributions are summarized as follows:

- We propose a deterministic single-step paragram, which can extremely reduce the time required for inference;
- We introduce a pretrained Data-Driven MDE Model to supply more abundant feature representations for U-Net, thereby enhancing the model's performance;
- We propose a two-stage training strategy to Enhance object edge details.

2 Related Work

2.1 Monocular Depth Estimation

Monocular Depth Estimation (MDE) aims to predict a depth value for each pixel from a single RGB image, representing a fundamental yet challenging task in computer vision. Since inferring 3D scene structure from a 2D image is an inherently ill-posed problem, learning-based approaches rely heavily on scene priors learned from large-scale datasets.

The pioneering work of Eigen *et al.* [7] introduced a multi-scale network, demonstrating for the first time the feasibility of end-to-end depth regression using deep convolutional neural networks. Subsequent research has advanced the field primarily along the following fronts:

Network Architectures and Representation Learning. Researchers have explored various depth representations to improve regression accuracy. For instance, DORN [8] formulated depth estimation as an ordinal regression problem. AdaBins [2] introduced adaptive binning, which dynamically partitions the depth range into bins and combines classification with regression, significantly enhancing detail recovery. BTS [20] leveraged local planar guidance layers at multiple scales to exploit local contextual information. With the rise of Vision Transformers, DPT [25] successfully adapted the ViT architecture for dense prediction tasks by combining features from different transformer layers to capture both global and local information, challenging the dominance of CNN backbones. Works like PixelFormer [1] and NeWCRFs [46] further explored the potential of attention mechanisms and conditional random fields in capturing long-range dependencies and structured prediction.

Data Scaling and Generalization Capability. To enhance model generalization to unseen scenarios, *i.e.*, "in-the-wild" depth estimation, researchers have focused on integrating diverse datasets. MiDaS [26] learned a universal affine-invariant depth representation by training on a massive mixture of multiple datasets. Although its output lacks metric scale, it achieved a breakthrough in cross-dataset generalization. LeReS [45] proposed a strategy to recover metric scale by optimizing image-level shift and scale parameters. More recently, **Depth Anything** [43] pushed data scaling to a new level. It first leveraged the powerful visual prior from the DINOv2 [24] foundation model pre-trained on 142 million images, and was subsequently trained on a massive dataset comprising 62 million pseudo-labeled and 1.5 million real depth-annotated images, achieving remarkable zero-shot generalization performance. Its successor, Depth Anything V2 [21], further refined the training pipeline by completely removing real depth annotations and relying solely on synthetic and pseudo-labeled data, while maintaining impressive results.

Leveraging Privileged Information and Specific Settings. Another line of research attempts to utilize additional information to aid depth estimation. For example, the Metric3D series [44, 17] exploits camera intrinsics during both training and inference to recover metric depth, achieving high accuracy on specific benchmarks. However, the application of such methods is limited in "in-the-wild" images where camera information is unavailable.

Despite significant progress, most of the aforementioned methods follow a deterministic regression paradigm, directly learning a mapping from the input image to the output depth map. These approaches typically predict only the mode of the conditional distribution and struggle to capture the inherent ambiguities in depth estimation (*e.g.*, transparent objects, occlusions, motion blur). Our work differs from these methods in its fundamental paradigm, as we focus on exploring the potential of *generative models*, particularly diffusion models, for capturing the multi-modal distribution of depth estimates and enhancing generalization capability.

2.2 Diffusion Models

Diffusion models, as a class of generative models, have emerged as a powerful framework for data synthesis

and dense prediction tasks. Their core principle involves a two-step process: a *forward pass* that progressively injects noise into the data, and a *reverse pass* that learns to denoise, effectively reconstructing the data from noise.

The development of diffusion models can be traced through several key milestones. Initially inspired by non-equilibrium thermodynamics, the concept was formalized in [34]. A significant breakthrough came with Denoising Diffusion Probabilistic Models (DDPMs) by [16], which established a simple and stable training objective by predicting the injected noise. Subsequent works generalized this perspective through the lens of score-based generative modeling [35, 36] and stochastic differential equations [37]. The recent introduction of Latent Diffusion Models (LDMs) [28] further improved computational efficiency by conducting the diffusion process in a lower-dimensional latent space, enabling high-resolution image synthesis and facilitating their application to various conditional generation tasks, including monocular depth estimation.

Problem Formulation. Diffusion models learn a distribution $p(\mathbf{z}_0)$ by defining a forward process that gradually corrupts data $\mathbf{z}_0$ with Gaussian noise, and then learning a reverse process to recover it.

Forward Process. The forward process is a fixed Markov chain that adds noise over T steps:

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

where $\beta_t \in (0,1)$ is a variance schedule. A notable property is that we can sample $\mathbf{z}_t$ at any timestep t in closed form:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2}$$

where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Process. The reverse process is a parameterized Markov chain that starts from $p(\mathbf{z}_T) = \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I})$ and learns to denoise:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t, t)). \tag{3}$$

The goal is to match the true reverse distribution $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0)$.

Training Objective. The model is trained by optimizing a variational bound on the negative log-likelihood. A simplified, reweighted objective [16] is:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right], \tag{4}$$

where $t \sim \mathcal{U}[1,T]$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{z}_t$ is computed via Eq. 2, and $\boldsymbol{\epsilon}_\theta$ is a neural network that predicts the noise.

Conditional Diffusion. For tasks like depth estimation, the process is conditioned on an input $\mathbf{x}$ (e.g., an RGB image). The noise prediction network then becomes $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{x})$, guiding the generation to be consistent with the conditioning signal. This formulation allows the model to learn the conditional distribution $p(\mathbf{z}_0|\mathbf{x})$, which is crucial for predictive tasks beyond unconditional generation.

2.3 Monocular Depth Estimation baesd on Diffusion Models

Many methods have been tried to apply diffusion models to monocular depth estimation. The DDP[18] first proposes an architecture to encode the image but decode a depth map. DiffusionDepth[5] perform multi-step denoising in the lattent space. DepthGen[29] employs a denoising diffusion model with a self-supervised plus supervised training strategy. Marigold[32] fine-tunes Stable Diffusion to achieve high-quality depth estimation through multi-step denoising. E2E-FT[10] and GenPercept[41] fine-tuned Marigold to make it an end-to-end model. Lotus[14] switched to predicting the target using exactly one step of denoising. DepthFM[12] adopted flow matching[22] to train model. DepthMaster[38] proposed a two stage training strategy: in the first stage, the encoder is trained to align features; in the second stage, the Fourier enhancement module is employed to enhance the model's feature representation. Despite these advances in applying diffusion models to MDE, no study has yet achieved both reasonable inference duration and satisfactory detail preservation.

In this work, we feed both the original image and a high-frequency image extracted via high-pass filtering into U-NET, thereby achieving further improvements in both performance and visual quality.

3 Method

We first analyzed the shortcomings of current model-driven monocular depth estimation methods in Section 3.1. Next, we introduce how the idea of using pre-trained models to assist training came about in3.2. Finally, we present our two-stage training strategy 3.3 and a new loss function in 3.4

3.1 Problem Formulation

The traditional approach to generation, often exemplified by the standard denoising-diffusion paradigm, is the **Stochastic Multi-Step Generation** process (Process a in the diagram). The input image $\mathbf{x}$ is first encoded into a latent representation $\mathbf{z}^{(x)}$ using a VAE Encoder. This latent is then transformed into a pure noise vector $\mathbf{z}_T^{(y)} = \epsilon$, representing the starting point of the reverse diffusion process in the target domain's latent space.

The core of this process is an iterative reverse denoising loop, repeated for a large number of steps, T. At each step $t \in \{T, T-1, \ldots, 1\}$, a U-Net model, ϵ_θ, predicts the noise component based on the current noisy latent $\mathbf{z}_t^{(y)}$ and the time step t, incrementally refining the latent representation towards a clean sample. This is typically governed by a stochastic sampling process:

$$\mathbf{z}_{t-1}^{(y)} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t^{(y)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t^{(y)}, t)\right) + \sigma_t\mathbf{w} \tag{5}$$

where α_t and $\bar{\alpha}_t$ are variance schedule constants, $\epsilon_\theta(\cdot)$ is the predicted noise, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ introduces stochasticity. After T steps, the final clean latent $\mathbf{z}_{1\to 0}^{(y)}$ is passed to the VAE Decoder to obtain the final output $\hat{\mathbf{y}}$. A significant drawback of this approach is its computational cost. The required repetition of the U-Net computation for T times ($\mathbf{z}_T^{(y)} \to \cdots \to \mathbf{z}_{t-1}^{(y)}$) means the inference time is **excessively long** for real-time applications, limiting its utility in latency-sensitive tasks.

In contrast to the multi-step approach, our proposed method adopts a **Deterministic Single-Step Perception** paradigm (Process c). This method fundamentally reformulates the task to better fit the model's structure while achieving high efficiency.

The input image $\mathbf{x}$ is encoded into its latent representation $\mathbf{z}^{(x)}$. This image latent $\mathbf{z}^{(x)}$ is then directly passed through the adapted U-Net, ϵ_θ, in a single, non-iterative step, bypassing the lengthy reverse diffusion chain. The time step is fixed to $t = 1$ to ensure a direct, deterministic transformation from the image latent to the target latent representation, $\mathbf{z}_{\text{pred}}$.

$$\mathbf{z}_{\text{pred}} = \epsilon_\theta(\mathbf{z}^{(x)}, 1) \tag{6}$$

The resulting predicted latent $\mathbf{z}_{\text{pred}}$ is then fed into a VAE Decoder (or a Customized Decoder) to produce the final perception output $\hat{\mathbf{y}}$. This direct transformation significantly accelerates the inference process. While offering superior inference efficiency, the single-step nature of the transformation can lead to a compromise in quality. By skipping the iterative refinement steps inherent to diffusion models, this method may capture **fewer fine-grained details** in the resulting output $\hat{\mathbf{y}}$ compared to the multi-step stochastic process.

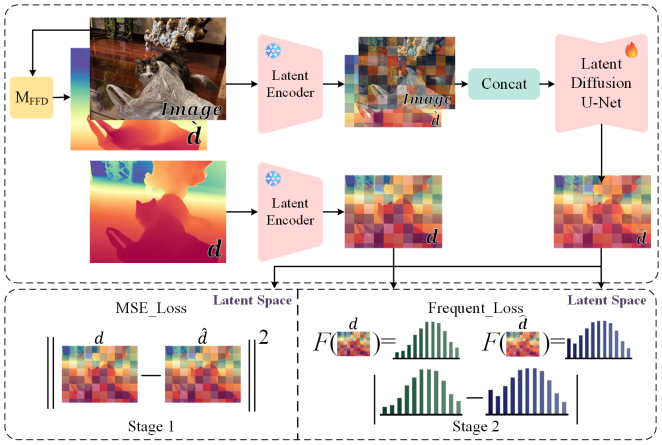


Figure 1: Overview of the ApDepth training framework. It first generates a depth map $\grave{d}$ from the input RGB image x using a feed-forward model. Subsequently, both $\grave{d}$ and x are concatenated and fed into a U-Net for depth estimation. Depending on the training stage, we employ different loss functions to guide the model. In Stage 1, we use MSE loss to guide the model in learning global features. In Stage 2, we utilize our proposed frequency loss to guide the model to learn edge details.

3.2 Pretrained Model assist

Training depth estimation models on synthetic and real-world datasets exhibits complementary characteristics. Synthetic data, withw accurate and complete depth labels, enable models to capture fine-grained spatial structures and learn detailed geometric relationships, yet they often suffer from a domain gap that limits generalization to natural scenes. Conversely, real-world datasets provide authentic image distributions and semantic diversity, fostering robust zero-shot generalization but at the cost of noisy and incomplete depth labels that lead to over-smoothed predictions. In diffusion-based depth estimation, the loss of detail caused by single-step denoising further aggravates this issue. As revealed by GenPercept [42], most of Stable Diffusion's representational knowledge resides in its U-Net architecture, implying that enriching the U-Net's semantic context is crucial for improving performance. To this end, we introduce a pretrained, data-driven monocular depth estimation model—specifically, Depth Anything V2 [21]—to guide the diffusion process. Trained on 63.5 million diverse real-world images, this model provides rich semantic priors and complex texture–depth correlations, enabling the U-Net to learn more meaningful geometric representations and achieve both fine-grained detail reconstruction and strong generalization in real-world scenes.

3.3 Two Stages training strats

Previous model-driven approaches, like Marigold[32] mostly employed MSE as the loss function. Within the latent space, the model performs MSE loss calculations and, through multi-step inference via the diffusion model, can predict detailed depth maps. However, single-step inference often results in depth maps with blurred edges and a lack of fine details. Inspired by DepthMaster[38], we introduce a two-stage training strategy. In the first stage, we utilize the MSE loss to guide the model in learning the overall structure of the depth map. In the second stage, we incorporate a Latent Frequency Loss (detailed in Section 3.4) to enhance the model's ability to capture fine-grained details and preserve depth discontinuities. By combining these two training stages, our model can effectively learn both the global geometry and local details of the depth map, leading to improved performance in monocular depth estimation.

3.4 Latent Frequency Loss

To enhance the model's ability to recover fine-grained structural details and preserve depth discontinuities, we introduce a frequency-domain based loss function, termed **Latent Frequency Loss**. Inspired by the success of frequency-domain analysis in generative and reconstruction models [6, 23], this loss compares the magnitude spectra of the predicted and ground-truth depth maps in the Fourier domain, thus constraining the model to learn both low-frequency global geometry and high-frequency structural details.

Given the predicted depth map $\hat{D}$ and the ground-truth depth map D, we first compute their two-dimensional discrete Fourier transforms:

$$\mathcal{F}_{\hat{D}} = \text{FFT2}(\hat{D}), \quad \mathcal{F}_D = \text{FFT2}(D), \tag{7}$$

and extract the centralized magnitude spectra:

$$M_{\hat{D}} = \left|\text{FFTShift}(\mathcal{F}_{\hat{D}})\right|, \quad M_D = \left|\text{FFTShift}(\mathcal{F}_D)\right|. \tag{8}$$

The base frequency loss is defined as the L_p difference between these magnitude spectra:

$$\mathcal{L}_{\text{base}} = \|M_{\hat{D}} - M_D\|_p, \quad p \in \{1, 2\}. \tag{9}$$

To emphasize high-frequency regions, we design a radial high-pass weighting function $W(u, v)$ that increases

the loss contribution for frequencies farther from the spectrum center:

$$W(u,v) = 1 + \lambda \cdot \frac{\sqrt{(u-u_c)^2 + (v-v_c)^2}}{\sqrt{u_c^2 + v_c^2} + \epsilon}, \quad (10)$$

where (u_c, v_c) denotes the spectrum center, λ controls the high-pass strength, and ϵ ensures numerical stability. This frequency weighting encourages the model to preserve sharp edges and fine details.

The final frequency-enhanced loss is thus defined as:

$$\mathcal{L}_{\text{freq}} = \frac{1}{N} \sum_{u,v} W(u,v) \cdot |M_{\hat{D}}(u,v) - M_D(u,v)|^p. \quad (11)$$

By incorporating frequency-domain constraints during the multi-step denoising process of diffusion-based models, the proposed Latent Frequency Loss enforces stability and enhances edge awareness, leading to more accurate and detail-preserving monocular depth estimation.

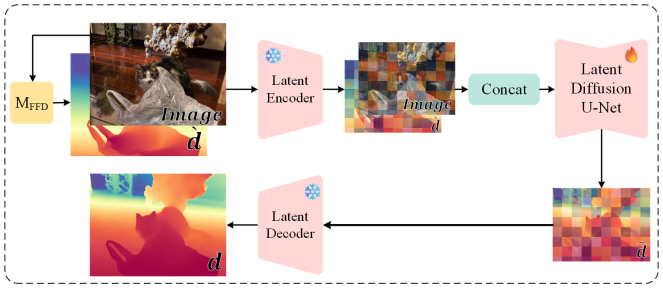


Figure 2: Overview of the ApDepth inference framework. It uses the original image x and the depth map $\grave{d}$ generated by the FFD model as input. After encoding into latent space, it is fed into U-Net and do a single-step inference. Finally, the depth map within the generated latent space is decoded to generate depth map d.

4 Experiments

4.1 Implementation

Our model is based on Stable Diffusion 2.1[28], but text-embedding is diaabled during both training and inference. We flollowed GenPercept[41] to do a single step inference and setprediction type = 'sample' instead of 'v-prediction'. Training our method takes 24K iterations for stage one and another 1k iterations for stage two with a bath size of 32. We train our model on a single NVIDIA GeForece RTX 4090 GPU, which has 24GB memory. It takes about 6 days to train our model.

4.2 Dataset

Training Dataset. Our model is trained on Hypersim[27] and Virtual KITTI[3]. Hypersim is a large-scale synthetic indoor scene dataset comprising approximately 77,000 images, each featuring high-quality photorealistic rendering. We flollowed Marigold[32]'s official split with around 54K samples is used with the training resolution of 480×640. Virtual KITTI is a synthetic autonomous driving scene dataset that virtually reconstructs and extends the classic KITTI dataset. It contains approximately 6 scene sequences comprising 21,260 annotated frames, and provides diverse weather and lighting conditions (e.g., sunny, rainy, foggy days, and different time periods). We take around 20K samples for training with the resolution of 1216×352 andset the far plane to 80 meters. The two datasets are mixed with a ratio of 9:1 during training.

Evaluation Datasets. We evaluate our model's zero-shot performance on five real scene datasets. Indoor datasets NYU Depth V2[33] and Scannet[4]. We use the official test split with 654 images for NYUv2 and the split proposed by Marigold with 800 images for ScanNet. Outdoor street-scene dataset KITTI[11], We follow the Eigen split[7], which consists of 652 images. Both indoor and outdoor datasets ETH3D[30] and DIODE[40]. We use the Marigold's split to evaluate on 454 samples from ETH3D and 771 samples from DIODE.

Evaluation protocol. Following the protocol of **affine-invariant depth evaluation** [26], we first align the estimated predicted depth map $\hat{m}$ to the ground truth depth d with the **least squares fitting**. This step gives us the absolute aligned depth map $\tilde{d}$ as:

$$\tilde{d} = a \cdot \hat{m} + s \cdot t$$

where a is the scaling factor and t is the shifting bias. Both a and t are derived in the same units as the ground truth depth map d.

Next, we apply two widely recognized **metric-based error metrics** [26, 25, 45, 44] for assessing quality of depth estimation.

1. The first is **Absolute Mean Relative Error (AbsRel)**, calculated as:

$$\text{AbsRel} = \frac{1}{M} \sum_{i=1}^{M} \frac{\left|\tilde{d}_i - d_i\right|}{d_i},$$

where M is the total number of pixels.

2. The second metric, δ_1 accuracy, measures the proportion of pixels satisfying:

$$\max\left(\frac{\tilde{d}_i}{d_i}, \frac{d_i}{\tilde{d}_i}\right) < 1.25.$$

4.3 Ablation Studies

For ablation studies, we select the NYUv2[33] dataset, consisting of 785 samples, and the KITTI[11] dataset, following the KITTI Eigen[7] training split, to validate the effectiveness of each component in our proposed ApDepth framework.

Pretrained Model. We investigate the impact of different pretrained data-driven MDE models on the performance of our ApDepth framework. We select three version of models of Depth Anything V2[21] (small, base, large) and Distill Any Depth[15] (base) as the pretrained MDE models to assist the U-Net in learning richer feature representations. As show in Table 4.3, using the small version of Depth Anything V2 already brings noticeable improvements over the baseline model without any pretrained model.

Method	Training Data	NYUv2		KITTI		ETH3D		ScanNet		DIODE		Avg. Rank
		AbsRel↓	δ1↑	AbsRel↓	δ1↑	AbsRel↓	δ1↑	AbsRel↓	δ1↑	AbsRel↓	δ1↑	
Marigold[19]	74K	5.5	96.4	9.9	91.6	6.5	96.0	6.4	95.1	30.8	77.3	5.40
GeoWizard[9]	280K	5.2	96.6	9.7	92.1	6.4	96.1	6.1	95.3	29.7	**79.2**	3.70
DepthFM[13]	74K	6.0	95.5	9.1	90.2	6.5	95.4	6.6	94.9	22.4	78.5	5.30
GenPercept[42]	74K	5.6	96.0	9.9	90.4	6.2	95.8	6.2*	96.1*	35.7	75.6	5.45
Lotus[14]	59K	5.3	96.7	9.3	92.8	6.8	95.3	6.0	96.3	22.8	73.8	4.40
DepthMaster[38]	74K	5.0	97.2	**8.2**	**93.7**	**5.3**	**97.4**	5.5	96.7	**21.5**	77.6	1.70
ApDepth(Ours)	74K	**4.7**	**97.3**	9.1	93.2	5.7	96.3	**4.8**	**97.2**	29.8	78.3	2.05

Table 1: Depth estimation performance comparison. Training Data is the total number of samples used (Real + Synthetic).

Pretrained Model	NYUv2		KITTI	
	AbsRel↓	δ1↑	AbsRel↓	δ1↑
base	5.3	96.5	10.7	89.3
DA2 small	5.2	96.7	9.7	92.1
DA2 base	-	-	-	-
DA2 large	**-**	**-**	**-**	**-**

5 Conclusion

In this work, we propose ApDepth, a novel diffusion-based framework for monocular depth estimation. By adopting a deterministic single-step perception paradigm, we significantly accelerate inference time compared to traditional multi-step methods. By incorporating a pretrained Data-Driven MDE Model, we effectively supply the U-Net with richer feature representations , enhancing the model's ability to learn complex statistical correlations. Additionally, the two-stage training strategy, which utilizes a Latent Frequency Loss , enhances fine-grained detail preservation and edge details by operating in the frequency domain. Benefiting from this design, ApDepth effectively balances inference efficiency and detail preservation. Extensive experiments validate the effectiveness of our approach, which achieves competitive or superior performance across multiple benchmarks.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023.

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021.

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

[4] 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.

[5] Yiquan Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In *European Conference on Computer Vision*, pages 432–449. Springer, 2024.

[6] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.

[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.

[8] 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*, pages 2002–2011, 2018.

[9] 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. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025.

[10] 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. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 753–762. IEEE, 2025.

[11] 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.

[12] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3203–3211, 2025.

[13] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching, 2024.

[14] 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.

[15] Xiankang He, Dongyan Guo, Hongji Li, Ruibo Li, Ying Cui, and Chi Zhang. Distill any depth: Distillation creates a stronger monocular depth estimator. *arXiv preprint arXiv: 2502.19204*, 2025.

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

[17] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[18] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21741–21752, 2023.

[19] 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.

[20] 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.

[21] Xuecheng Li, Renze Deng, Yulin Fan, Peng Chen, Siyuan Chen, Zhangjun Liu, Li Han, Shuai Zhu, Hao Sun, Yiyi Lu, and Qizhou Li. Depth anything v2. *arXiv preprint arXiv:2404.14442*, 2024.

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

[23] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. *Advances in neural information processing systems*, 31, 2018.

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

[26] 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*, 44:1623–1637, 2020.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] Simon Shi, Luc Van Gool, and Jonathon Luiten. Marigold: Repurposing diffusion models for monocular depth estimation. *arXiv preprint arXiv:2404.09015*, 2024.

[33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[36] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[38] Ziyang Song, Zerong Wang, Bo Li, Hao Zhang, Ruijie Zhu, Li Liu, Peng-Tao Jiang, and Tianzhu Zhang. Depthmaster: Taming diffusion models for monocular depth estimation. *arXiv preprint arXiv:2501.02576*, 2025.

[39] Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *European Conference on Computer Vision*, pages 236–257. Springer, 2024.

[40] 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.

[41] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024.

[42] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024.

[43] 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.

[44] 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, pages 9043–9053, 2023.

[45] 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*, pages 204–213, 2021.

[46] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022.

[47] Ruijie Zhu, Chuxin Wang, Ziyang Song, Li Liu, Tianzhu Zhang, and Yongdong Zhang. Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation. *arXiv preprint arXiv:2407.08187*, 2024.