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

Jiawei Wang[a], Shuai Yuan[b,*], Mingbo Lei[a] and Yibo Chen[b]

[a]School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang, Liaoning 110168, China,
[b]School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, China,

ARTICLE INFO

Keywords:
Diffusion Model
Depth Estimation
Two-Stage training
Frequency-Domain

ABSTRACT

This paper presents **ApDepth**, a novel single-step diffusion framework for monocular depth estimation that achieves fast inference while preserving fine-grained edge details. Current diffusion-based models achieve impressive generalization through multi-step iterative denoising. However, they face a critical trade-off: multi-step inference incurs prohibitive runtime, whereas accelerating to a single-step deterministic generation inevitably leads to severe degradation of local structural boundaries. Our model addresses this challenge by integrating a pre-trained data-driven MDE model to provide strong semantic priors, maintaining robust global structures even under heavily compressed diffusion steps. To mitigate the inherent edge degradation, we introduce a coarse-to-fine two-stage training paradigm with two novel designs: 1) **Feature Alignment via Cosine Similarity**, which enriches the U-Net's structural understanding by explicitly aligning its intermediate features with an auxiliary encoder; and 2) an **Iteration-Based Loss Scheduling Strategy**, which initially employs spatial-domain constraints (latent MSE and pixel L_1 losses) to establish an accurate global metric scale, and subsequently transitions to a novel **Frequency-Domain (FFT) Loss** in the final training phase to explicitly recover sharp high-frequency edges. By effectively balancing inference efficiency and detail preservation, ApDepth significantly accelerates inference speed and 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 major applications 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 this issue, the model must possess a deep understanding of the current scene. Existing methods for zero-shot monocular depth estimation can be broadly categorized into two main groups: data-driven and model-driven approaches. 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 [1] and ScaleDepth [2] have demonstrated impressive performance on various benchmarks. However, these methods heavily

*Corresponding author
reidyuan@163.com (Shuai Yuan)
ORCID(s): 0009-0006-0793-6698 (Jiawei Wang)

rely on exhaustive data collection and often incur prohibitive computational costs during training.

In contrast, model-driven methods, particularly those based on diffusion models [3], have shown great potential in generating high-quality depth maps. For example, DiffusionDepth [4] pioneers the reformulation of monocular depth estimation as a diffusion denoising process. Marigold [5] presents a fine-tuning protocol for Stable Diffusion [6], achieving impressive results in both global structure and local details. DepthFM [7] adopts flow matching [8] during training to enhance generalization. However, the iterative denoising process results in a low inference speed. E2E-FT [9] and GenPercept [10] propose a deterministic single-step paradigm, significantly reducing the time required for inference. Although these approaches have yielded promising results, they are largely constrained by either extremely long inference times or poor local detail preservation.

To address these challenges, we propose **ApDepth**, a single-step diffusion framework for accurate and detailed monocular depth estimation. Specifically, ApDepth modifies the stochastic multi-step generation into a deterministic one-step perception approach. To ensure robust global structures despite the heavily compressed diffusion process, we integrate a pre-trained feed-forward MDE (M_{FFD}) model. By utilizing the M_{FFD} as an initial structural reference for the global depth context, we provide a stable semantic baseline that allows the diffusion model to focus efficiently on local detail refinement. This combined approach significantly accelerates inference time (e.g., from 12s to 120ms for 640×480 images) while maintaining high global accuracy.

However, compressing the iterative denoising process into a single step inevitably degrades local edge details. To mitigate this, we propose a coarse-to-fine two-stage training paradigm [11]. In the first stage, we introduce an auxiliary encoder and a cosine similarity loss to explicitly align the intermediate features of the denoising U-Net, thereby enriching the model's structural understanding. In the second stage, we decouple the optimization into two phases: establishing a solid global depth foundation via spatial-domain losses, followed by a fine-tuning phase using a novel frequency-domain loss (FFT loss). This explicit frequency-domain guidance effectively forces the model to reconstruct the sharp and accurate object edges that are typically lost during single-step generation. Extensive experiments on standard zero-shot monocular depth estimation benchmarks demonstrate the superiority of our approach. Compared to state-of-the-art diffusion-based and data-driven models, **ApDepth** establishes a new paradigm by achieving highly competitive—and in many cases superior—geometric accuracy and fine-grained edge preservation, all while operating at a mere fraction of the traditional inference cost.

Briefly, our main contributions are summarized as follows:

- **A Single-Step Diffusion Framework (ApDepth):** We propose a deterministic single-step paradigm for diffusion-based depth estimation. By seamlessly incorporating a pre-trained data-driven MDE model as a semantic prior, our framework significantly accelerates inference speed while ensuring robust global depth structures.
- **Enhanced Feature Alignment via Cosine Similarity:** Building upon existing two-stage training paradigms, we design an optimized feature alignment strategy in the first stage. By introducing a spatial-preserving Conv Adapter and a cosine similarity loss to explicitly align the intermediate U-Net features with an external encoder, we substantially enrich the model's structural understanding and mitigate texture overfitting.
- **Spatial- and Frequency-Domain Loss Strategy:** We refine the second training stage into a coarse-to-fine optimization process. It first employs latent MSE and pixel L_1 losses for global metric accuracy, followed by a transition to a novel frequency-domain (FFT) loss to explicitly supervise high-frequency components and recover sharp local 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.* [12] 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 [13] formulated depth estimation as an ordinal regression problem. AdaBins [14] introduced adaptive binning, which dynamically partitions the depth range into bins and combines classification with regression, significantly enhancing detail recovery. BTS [15] leveraged local planar guidance layers at multiple scales to exploit local contextual information. With the rise of Vision Transformers, DPT [16] 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 [17] and NeWCRFs [18] 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 [19] 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 [20] proposed a strategy to recover metric scale by optimizing image-level shift and scale parameters. More recently, **Depth Anything** [21] pushed data scaling to a new level. It first leveraged the powerful visual prior from the DINOv2 [22] 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 [1], 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 [23, 24] 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 highly competitive class of generative models, have emerged as a powerful framework for data synthesis and dense prediction tasks, demonstrating impressive performance across a wide range of applications, including image generation [6, 25], inpainting [26, 27], and super-resolution [28, 29]. Their core principle involves a two-step process: a *forward pass* that progressively transforms a data distribution (e.g., images) into a noise distribution, and a *reverse pass* that learns to denoise, effectively reconstructing high-quality 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 [30]. A significant breakthrough came with Denoising Diffusion Probabilistic Models (DDPMs) [31], 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 [32, 33] and stochastic differential equations [34].

However, optimizing and evaluating these early models in pixel space, combined with the iterative nature of the reverse process, leads to low inference speeds and high training costs. To mitigate these issues, various methods have been proposed, such as designing advanced sampling strategies [35, 36, 37] and adopting hierarchical approaches [38, 39], yet the overall computational cost remains high. The recent introduction of Latent Diffusion Models (LDMs) [6] addressed this bottleneck by operating in a compressed, lower-dimensional latent space utilizing an image-to-latent encoder-decoder. By pre-training on large-scale, high-quality datasets such as LAION-5B [40], LDMs learn powerful image priors that enable high-resolution image synthesis with affordable computation costs. In this work, we take a pre-trained latent diffusion model as our backbone to leverage the strong image priors learned by such models.

Formally, these 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 [31] 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.

Despite the robust generative priors established by this formulation, diffusion models often struggle with tasks requiring high-level semantic understanding. As recent studies indicate [41, 42, 43], the reconstruction objectives used to train denoising networks do not sufficiently address the elimination of irrelevant details. Consequently, these models tend to overfit low-level textures rather than capturing the true underlying structure. To overcome this limitation, approaches

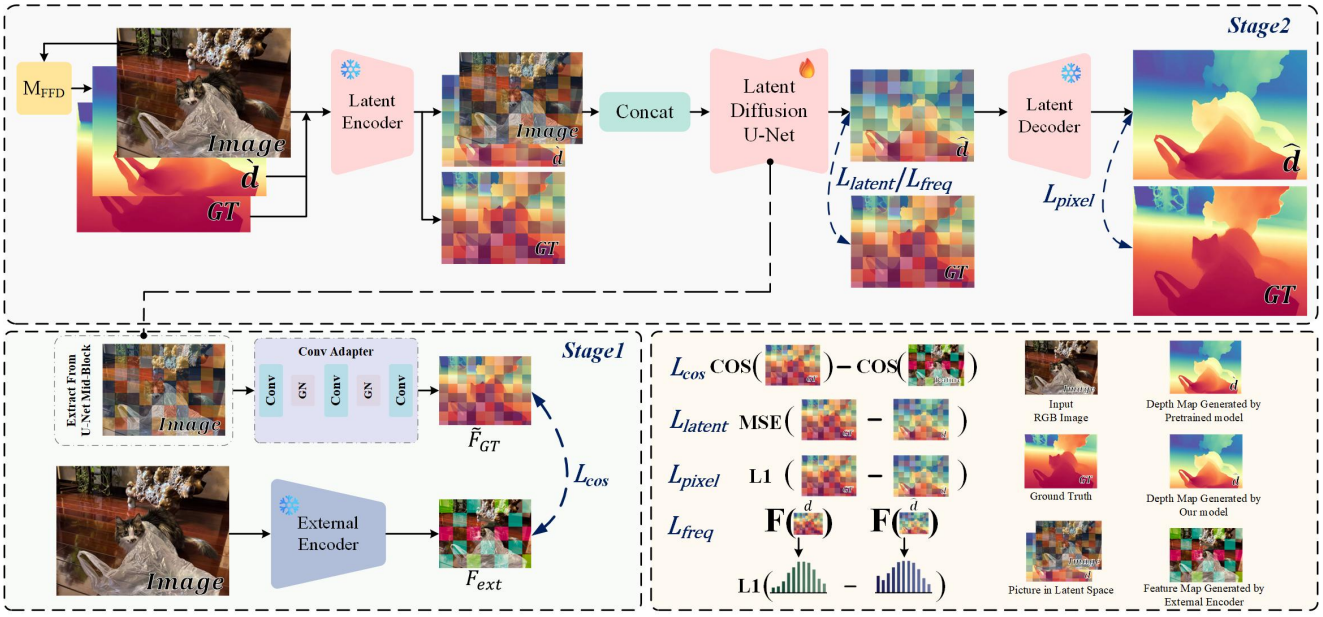


Figure 1: Overview of the ApDepth two-stage training framework. Stage 1 (Bottom Left): To establish a robust semantic foundation, the intermediate features of the U-Net are processed by a spatial-preserving Conv Adapter and explicitly aligned with representations from a frozen external encoder via a cosine similarity loss ($\mathcal{L}_{\text{cos}}$). **Stage 2 (Top):** For fine-grained depth estimation, the RGB image and a semantic prior generated by a pre-trained M_{FFD} model are encoded into the latent space and concatenated. The trainable Latent Diffusion U-Net predicts the depth latent, supervised by an iteration-based loss scheduling strategy. Initially, spatial constraints ($\mathcal{L}_{\text{latent}}$ and $\mathcal{L}_{\text{pixel}}$) enforce global metric accuracy. In the final phase, a frequency-domain loss ($\mathcal{L}_{\text{freq}}$) is introduced to reconstruct sharp, high-frequency edge details.

like REPA [43] have suggested that incorporating high-quality external representations can accelerate convergence in generation tasks. Building upon this insight, our work demonstrates that integrating external semantic information not only mitigates texture overfitting but also significantly enhances the model's generalization capabilities for discriminative tasks like monocular depth estimation.

2.3. Monocular Depth Estimation based on Diffusion Models

Recently, diffusion models have been extensively explored to formulate monocular depth estimation (MDE) as a generative denoising process. Early pioneering works, such as DDP [44] and DiffusionDepth [4], established the foundational architecture for encoding images and iteratively decoding depth maps in the latent space. To further enhance depth quality, DepthGen [45] introduced a combined self-supervised and supervised training strategy, while Marigold [5] successfully fine-tuned Stable Diffusion to achieve state-of-the-art geometric precision through multi-step latent denoising. Subsequent works, such as E2E-FT [9] and GenPercept [10], built upon Marigold to enable more efficient end-to-end training.

Despite their impressive accuracy, the iterative sampling process of these models incurs prohibitive inference latency, limiting their real-world applicability. To address this computational bottleneck, recent methods have shifted towards accelerated sampling. For instance, Lotus [46] streamlined the process by predicting the target depth using exactly one denoising step, and DepthFM [7] adopted flow matching [8] to achieve highly efficient generation. However, compressing the generative process into a single or very few steps inevitably leads to the degradation of fine-grained structures. To mitigate this loss of local details, DepthMaster [11] proposed a two-stage training strategy involving explicit feature alignment and a Fourier enhancement module.

Despite these significant advances, achieving an optimal balance between ultra-fast inference duration and satisfactory detail preservation remains an open challenge. In this work, we propose **ApDepth** to bridge this gap. By operating within a single-step denoising framework for maximum efficiency, we explicitly feed both the original RGB image and its high-frequency components—extracted via high-pass filtering—into the denoising U-Net. This explicit frequency-domain conditioning forces the network to perceive sharp structural boundaries, thereby achieving substantial improvements in both quantitative performance and visual edge quality.

3. Method

In this section, we present the architecture and training pipeline of ApDepth. We begin by revisiting the standard diffusion formulation and analyzing its limitations in Section 3.1. Next, we provide an overview of our proposed two-stage training strategy in Section 3.2. We then detail the first stage, which focuses on explicit feature alignment, in Section 3.3. Finally, Section 3.4 introduces the second stage, where

we incorporate a pre-trained feed-forward model and hybrid frequency-domain losses for fine-grained depth estimation.

3.1. Motivation and Problem Formulation

The traditional approach to conditional generation, often exemplified by the standard denoising-diffusion paradigm, is the **Stochastic Multi-Step Generation** process [31, 6]. In a standard conditional Latent Diffusion Model (LDM), the input RGB image $\mathbf{x}$ is first encoded into a latent condition $\mathbf{z}^{(x)}$ using a pre-trained VAE Encoder. Meanwhile, the reverse diffusion process starts from a pure noise vector $\mathbf{z}_T^{(y)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ sampled in the target domain's latent space (e.g., depth). The core of this process is an iterative reverse denoising loop, repeated for a large number of steps, T. At each step $t \in \{T, T-1, \dots, 1\}$, a U-Net model, ϵ_θ, predicts the noise component based on the current noisy latent $\mathbf{z}_t^{(y)}$, the timestep t, and the image condition $\mathbf{z}^{(x)}$. This incrementally refines the latent representation towards a clean sample, typically governed by a stochastic sampling process:

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

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 $\mathbf{z}_0^{(y)}$ is passed to the VAE Decoder to obtain the final depth output $\hat{\mathbf{y}}$. A significant drawback of this approach is its computational cost. The required repetition of the heavy U-Net computation for T times makes the inference time **computationally prohibitive** for real-time applications, severely limiting its utility in latency-sensitive tasks.

In contrast to the multi-step approach, our proposed method builds upon the paradigm of **Deterministic Single-Step Perception** [10, 46]. This method fundamentally reformulates the task to better fit the model's structure while achieving maximum efficiency. The input image $\mathbf{x}$ is encoded into its latent representation $\mathbf{z}^{(x)}$. Instead of an iterative denoising chain starting from pure noise, this paradigm repurposes the U-Net, which we denote as $\mathcal{F}_\theta$, to directly map the image latent to the target depth latent $\mathbf{z}_{\text{pred}}$ in a single, non-iterative feed-forward pass. The timestep is fixed (e.g., $t = 1$) to ensure a deterministic transformation:

$$\mathbf{z}_{\text{pred}} = \mathcal{F}_\theta(\mathbf{z}^{(x)}, 1). \quad (6)$$

The resulting predicted latent $\mathbf{z}_{\text{pred}}$ is then fed into the VAE Decoder to produce the final perception output $\hat{\mathbf{y}}$. This direct transformation bypasses the lengthy reverse diffusion chain, significantly accelerating the inference process.

However, while offering superior inference efficiency, the single-step nature of this transformation inevitably leads to a compromise in quality. By skipping the iterative refinement steps inherent to traditional diffusion models, the network struggles to eliminate irrelevant details and ambiguities, resulting in a noticeable **loss of high-frequency structural details** and sharp object edges in the final output

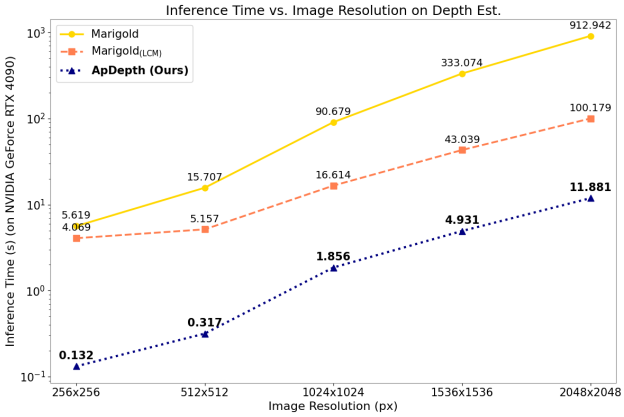


Figure 2: Inference time comparison in depth estimation between ApDepth and Marigold.

compared to the multi-step stochastic process. To address this limitation, we build upon the recently proposed two-stage training paradigm [11]. Rather than adopting it naively, we introduce crucial optimizations: we significantly enhance the feature alignment process in the first stage and propose a novel frequency-domain refinement for the second stage, which we detail in the following sections.

3.2. Overview of the Two-Stage Training Strategy

In the single-step deterministic paradigm, the U-Net is tasked with directly mapping an image latent to a depth latent in a single forward pass. However, empirical observations reveal a fundamental conflict during this learning process: the model is required to learn robust semantic understanding, accurate metric geometry, and high-frequency edge details simultaneously. Forcing the network to tackle all these disparate objectives concurrently often leads to severe learning conflicts, resulting in predictions that either overfit to superficial textures or suffer from heavily blurred structural edges.

Furthermore, capturing coarse geometric structure and fine-grained high-frequency details in a single pass presents a significant training challenge. If extreme high-frequency constraints are introduced before the global scale is properly established, the model tends to overfit to local gradients, causing training instability. Conversely, relying exclusively on spatial-domain losses often exhibits a well-known smoothing effect, failing to resolve sharp depth discontinuities.

Based on these observations, rather than forcing the network to solve this ill-posed problem simultaneously, we adopt and fundamentally optimize a two-stage training strategy [11], as illustrated in Figure 1. Our core motivation is to *divide and conquer* the complex mapping problem, explicitly decoupling semantic alignment from geometric depth refinement to fully leverage the potential of our proposed modules.

In the **first stage**, our goal is to train a model that can robustly generalize across diverse scenarios by focusing solely on establishing a strong semantic foundation. We achieve this by explicitly aligning the generative features of

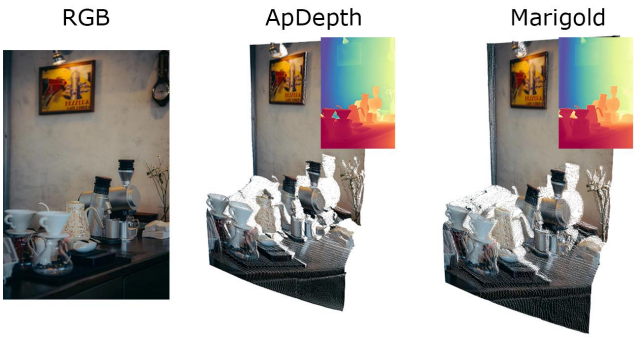


Figure 3: Compared to multi-step inference, our model not only achieves better results under single-step inference but also significantly reduces the number of flying pixels.

the U-Net with high-quality semantic representations from an external external encoder. This stage operates purely in the feature space without the interference of metric depth regression, preventing the network from prematurely overfitting to low-level texture details and ensuring the latent representations capture accurate global scene contexts.

Once robust semantic priors are firmly established, the **second stage** shifts the focus entirely to depth estimation and detail refinement. Guided by the explicit geometric priors from the pre-trained M_{FFD} model, we further decouple the depth learning process itself into a targeted curriculum. Initially, the model jointly enforces latent-space and pixel-level constraints to accurately capture the global structure and correct the metric scale. Subsequently, it transitions to a pure frequency-domain constraint, explicitly focusing on recovering missing fine-grained details and preserving sharp depth discontinuities. By isolating these learning objectives, our refined strategy effectively circumvents the gradient interference inherent in single-step paradigms, ensuring both global structural accuracy and exceptional local detail preservation.

3.3. Stage 1: Feature Alignment and Structure Learning

Since the denoising network in diffusion models is fundamentally trained on reconstruction tasks, it tends to prioritize texture details over structural integrity, which can lead to unrealistic depth predictions [11]. To prevent the U-Net from overfitting to low-level texture details, inspired by [11], we introduce a Feature Alignment module in the first stage, as illustrated in the bottom-left of Figure 1. This module leverages high-quality semantic representations from a pre-trained external encoder (DINOv2) [22].

Previous feature alignment methods typically utilize a standard Multi-Layer Perceptron (MLP) to project features between different latent spaces. However, MLPs operate point-wise and fail to explicitly model the spatial context inherent in dense feature maps. To address this, we propose a spatial-preserving **ConvFeatureAdapter**. This module strictly utilizes convolutional layers to maintain 2D spatial relationships while matching the channel dimensions of the external encoder. Given the intermediate U-Net feature $F_{\text{unet}} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$, the adapted feature $\tilde{F}_{\text{unet}}$ is computed as follows:

$$F_1 = \sigma(\text{GN}(\text{Conv}_{1\times1}(F_{\text{unet}}))) \tag{7}$$

$$F_2 = \sigma(\text{GN}(\text{Conv}_{3\times3}(F_1))) \tag{8}$$

$$\tilde{F}_{\text{unet}} = \text{Conv}_{1\times1}(F_2) \tag{9}$$

where $\text{Conv}_{k\times k}$ denotes a 2D convolution with kernel size k, GN represents Group Normalization, and σ is the SiLU activation function.This fully convolutional architecture provides two distinct advantages. First, the inclusion of the 3×3 convolution block explicitly perceives local neighborhood information, retaining the critical spatial geometric priors required for dense depth estimation. Second, we employ Group Normalization instead of standard Batch Normalization; this ensures stable and reliable statistical scaling even under the small batch sizes typically necessitated by the high memory requirements of training diffusion models.

Furthermore, rather than minimizing the Kullback-Leibler (KL) divergence or Mean Squared Error (MSE) between feature distributions, we optimize a **Cosine Similarity Loss**. In the highly-dimensional feature space of vision foundation models, absolute activation magnitudes can fluctuate significantly, causing L_1 or L_2 based losses to easily dominate the gradient and induce training instability. Cosine similarity circumvents this by strictly focusing on the angular direction—and thereby the pure semantic content—of the feature vectors.

We spatially interpolate the external DINOv2 features F_{ext} to match the resolution of the U-Net features, and apply L_2 normalization along the channel dimension for both representations:

$$\hat{F}_{\text{unet}} = \frac{\tilde{F}_{\text{unet}}}{\|\tilde{F}_{\text{unet}}\|_2}, \quad \hat{F}_{\text{ext}} = \frac{F_{\text{ext}}}{\|F_{\text{ext}}\|_2} \tag{10}$$

The feature alignment loss is then formulated to maximize the channel-wise cosine similarity, driving the angular distance between the network features and the semantic priors to zero:

$$\mathcal{L}_{\text{fa}} = 1 - \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} \hat{F}_{\text{unet}}^{(c,i,j)} \cdot \hat{F}_{\text{ext}}^{(c,i,j)} \tag{11}$$

This cosine-based constraint ensures a highly stable and structurally accurate transfer of semantic priors from the external model directly into the U-Net's latent representations.

3.4. Stage 2: Fine-grained Depth Estimation

In the second stage, our primary objective is to enhance the model's capability to estimate depth while preserving fine-grained structural details. To circumvent the limitations of pure single-step denoising—which often leads to over-smoothing—we propose to efficiently leverage the strengths of feed-forward models alongside our diffusion framework. Specifically, we introduce a pre-trained, data-driven feed-forward monocular depth estimation ($\mathrm{M_{FFD}}$) model—namely, Depth Anything V2 (DA2) [1]—to guide the generative process.

Feed-forward MDE models are known to achieve strong zero-shot generalization by training on massive and diverse datasets. As a robust data-driven model, DA2 provides basic scale-invariant geometric priors to ensure a stable global depth context [1]. By encoding the DA2 depth map into the latent space and concatenating it with the RGB image latent as the condition for our U-Net, we establish a reliable structural baseline. This explicit geometric guidance alleviates the burden on the diffusion model to infer global scene layouts entirely from scratch. Consequently, the U-Net can dedicate its generative capacity to rectifying local ambiguities and synthesizing fine-grained textures via iterative fine-tuning. Furthermore, because we treat the $\mathrm{M_{FFD}}$ as a geometric prior reservoir, our diffusion model requires significantly less training data to learn local refinement, thereby boosting overall training efficiency.

As illustrated in the overall training framework (Figure 1), to achieve accurate structure capture alongside sharp edge preservation, we employ a hybrid, iteration-based loss scheduling strategy, moving from coarse global alignment to fine structural refinement.

Phase 1: Global Structure and Metric Alignment. During the initial phase (the first 20,000 iterations of Stage 2), the model is supervised using a combination of Latent Mean Squared Error (MSE) and Pixel-level L_1 loss. The loss function is formulated as:

$$\mathcal{L}_{\text{phase1}} = \text{MSE}(\mathbf{z}_{\text{pred}}, \mathbf{z}_{\text{gt}}) + \mathcal{L}_1(\hat{\mathbf{y}}, \mathbf{y}_{\text{gt}}), \tag{12}$$

where $\mathbf{z}_{\text{pred}}$ and $\mathbf{z}_{\text{gt}}$ denote the predicted and ground-truth depth latents, respectively, while $\hat{\mathbf{y}}$ and $\mathbf{y}_{\text{gt}}$ represent the corresponding decoded depth prediction and the ground-truth depth map in the pixel space. This dual-space supervision offers distinct advantages. The Latent MSE loss ensures that the predicted features strictly adhere to the VAE's learned latent manifold, promoting stable gradients and rapid convergence of the global semantic structure. Simultaneously, the Pixel L_1 loss explicitly penalizes metric depth errors in the decoded image space. Since standard diffusion models often generate visually plausible but metrically inaccurate outputs, incorporating pixel-space supervision early on forces the model to align its generative priors with real-world geometric constraints.

Phase 2: High-Frequency Detail Refinement.

Once the global structure and metric scale are stabilized, standard pixel-wise losses (L_1/MSE) begin to exhibit a well-known smoothing effect, struggling to resolve sharp object edges. Therefore, for the final fine-tuning phase (iterations 20,001 to 21,000), we freeze the spatial losses and transition entirely to a **Frequency-Domain Refinement Loss**:

Figure 4: Illustration of our specially designed frequency-domain loss function, which effectively enhances high-frequency edge details in the predicted depth maps.

$$\mathcal{L}_{\text{phase2}} = \lambda_{\text{freq}} \cdot \mathcal{L}_{\text{freq}}. \tag{13}$$

This loss compares the magnitude spectra of the predicted and ground-truth depth maps in the Fourier domain. Given the predicted depth map $\hat{D}$ and the ground-truth depth map D, we first compute their two-dimensional discrete Fourier transforms:

$$\mathcal{F}_{\hat{D}} = \text{FFT2}(\hat{D}), \quad \mathcal{F}_D = \text{FFT2}(D), \tag{14}$$

and extract the centralized magnitude spectra:

$$M_{\hat{D}} = \left|\text{FFTShift}(\mathcal{F}_{\hat{D}})\right|, \quad M_D = \left|\text{FFTShift}(\mathcal{F}_D)\right|. \tag{15}$$

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

$$\mathcal{L}_{\text{base}} = \|M_{\hat{D}} - M_D\|_p, \quad p \in \{1, 2\}. \tag{16}$$

To explicitly penalize blurry edges, we design a radial high-pass weighting function $W(u, v)$ that increases the loss contribution for frequencies farther from the spectrum center:

$$W(u, v) = 1 + \lambda \cdot \frac{\sqrt{(u - u_c)^2 + (v - v_c)^2}}{\sqrt{u_c^2 + v_c^2} + \epsilon}, \tag{17}$$

where (u_c, v_c) denotes the spectrum center, λ controls the high-pass strength, and ϵ ensures numerical stability. The final frequency-enhanced loss is thus defined as:

$$\mathcal{L}_{\text{freq}} = \frac{1}{N} \sum_{u,v} W(u, v) \cdot |M_{\hat{D}}(u, v) - M_D(u, v)|^p. \tag{18}$$

By delaying the application of the frequency loss until the final training phase, we prevent the network from getting distracted by high-frequency noise during early feature alignment. Instead, the model focuses purely on recovering missing fine-grained details and preserving sharp depth discontinuities, resulting in highly accurate, edge-aware depth estimation.

Type	Method	Training Data	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
			AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Discriminative	DiverseDepth [47]	320K	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	-	-
	MiDaS [19]	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	-	-
	LeReS [20]	354K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	-	-
	Omnidata [48]	12.2M	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	-	-
	HDN [49]	300K	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	-	-
	DPT [16]	1.4M	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	-	-
	Depth Anything V2 [1]	63.5M	**4.3**	**98.0**	**8.0**	**94.6**	**6.2**	**98.0**	**4.3**	**98.1**	**6.6**	**95.2**
Generative	Marigold[50]	74K	5.5	96.4	9.9	91.6	6.5	96.0	6.4	95.1	10.0	90.7
	GeoWizard[51]	280K	5.2	96.6	9.7	92.1	6.4	96.1	6.1	95.3	12.0	89.8
	DepthFM[52]	74K	6.0	95.5	9.1	90.2	6.5	95.4	6.6	94.9	-	-
	GenPercept[53]	90K	5.6	96.0	9.9	90.4	6.2	95.8	5.6	96.5	-	-
	Lotus[46]	54K	5.3	96.7	8.5	92.8	5.9	97.0	5.9	95.7	9.8	92.4
	DepthMaster[11]	74K	5.0	97.2	**8.2**	93.7	**5.3**	**97.4**	5.5	96.7	-	-
	ApDepth(Ours)	74K	**4.5**	**97.3**	8.7	**93.7**	5.6	96.3	**4.5**	**97.3**	**8.3**	**93.2**

Table 1
Depth estimation performance comparison. We have obtained excellent results on indoor datasets, nonetheless, we still fall short on outdoor datasets. The best results in **bold** and the second best results are underlined.

4. Experiments

4.1. Implementation Details

Our architecture is built upon the pre-trained Stable Diffusion v2 [6] backbone. During both training and inference, the text-embedding cross-attention condition is disabled. To maximize inference efficiency, we follow the deterministic paradigm of GenPercept [10] to perform single-step inference, setting the prediction type to `sample` rather than the default `v-prediction`.

The two-stage training process is conducted on a single NVIDIA RTX 6000 Ada Generation GPU (48GB memory). Specifically, the first stage (feature alignment) is trained for 20,000 iterations, which takes approximately 7.7 days. The second stage (fine-grained depth estimation) is trained for a total of 21,000 iterations, requiring approximately 10 days. This second stage comprises 20,000 iterations for spatial alignment and 1,000 iterations for frequency-domain refinement. The batch size is set to 32 throughout the training. For evaluation, we test our model on a single NVIDIA GeForce RTX 4090 GPU (24GB memory), where evaluating all benchmark datasets takes approximately 30 minutes.

4.2. Dataset

Training Dataset. Our model is trained on Hypersim[54] and Virtual KITTI[55]. Hypersim is a large-scale synthetic indoor scene dataset comprising approximately 77,000 images, each featuring high-quality photorealistic rendering. We flollowed Marigold[5]'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.

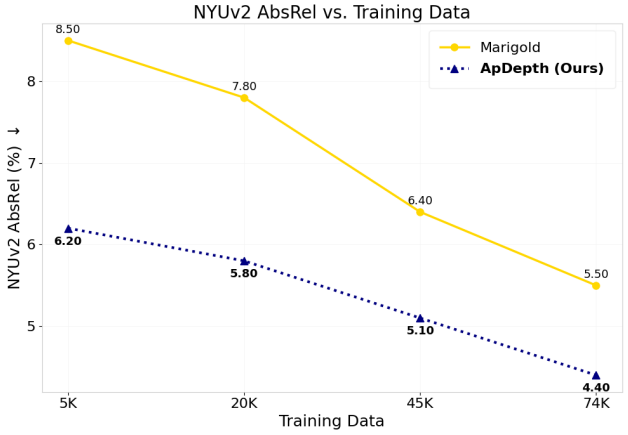


Figure 5: Quantitative comparison of AbsRel on the NYUv2 dataset across varying training data sizes. ApDepth consistently achieves lower prediction errors compared to Marigold under all data settings.

Evaluation Datasets. We evaluate our model's zero-shot performance on five real scene datasets. Indoor datasets NYU Depth V2[56] and Scannet[57]. 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[58], We follow the Eigen split[12], which consists of 652 images. Both indoor and outdoor datasets ETH3D[59] and DIODE[60]. 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** [19], 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:

$$\tilde{d} = a \cdot \hat{m} + s \cdot t \tag{19}$$

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** [19, 16, 20, 23] for assessing quality of depth estimation.

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

$$\text{AbsRel} = \frac{1}{M} \sum_{i=1}^{M} \frac{|\tilde{d}_i - d_i|}{d_i}, \tag{20}$$

   where M is the total number of pixels.
2. The second metric, δ_1 accuracy, measures the proportion of pixels satisfying:

$$\max\left(\frac{\tilde{d}_i}{d_i}, \frac{d_i}{\tilde{d}_i}\right) < 1.25. \tag{21}$$

Furthermore, for the evaluation on the DIODE dataset, we introduce a refined valid masking strategy. DIODE scenes typically contain highly complex structures with extreme depth discontinuities, making predictions at these high-frequency edges inherently ambiguous and prone to minor misalignment artifacts. To ensure a more reliable and objective evaluation of the model's core depth perception capabilities, we apply a Sobel filter to the ground truth depth maps to compute the horizontal and vertical gradients. We explicitly mask out high-frequency edge regions where the absolute gradient magnitude exceeds 0.3. The final valid mask used for our metric calculations on DIODE is the intersection of the dataset's original valid mask, our gradient-based edge mask, and a strict range boundary (0.6 to 350 meters, excluding invalid values like NaNs and Infs).

4.3. Ablation Studies

In this section, we validate the effectiveness of each key component in ApDepth through comprehensive ablation studies.

Ablation of the Pre-trained M_{FFD} Prior. We investigate the impact of the auxiliary M_{FFD} model's capacity on the overall performance of our ApDepth framework. Specifically, we select four scaled versions of Depth Anything V2 [1] (small, base, large, giant) as the pre-trained prior. By doing so, we aim to simulate the availability of various external MFFD models with different performance levels and computational constraints. As shown in Table 2, providing a depth map as a semantic prior offers a beneficial initial structural reference compared to the pure baseline. Predictably, as the capacity of the auxiliary model increases, the geometric accuracy improves, which naturally elevates the final depth estimation bounds. However, it is crucial to note that even when utilizing the lightest variant (DA2 small), our framework yields a stable and meaningful improvement over the baseline (e.g., reducing the AbsRel error from 10.7 to 9.7 on the KITTI dataset). This highlights the robustness of our core diffusion process and frequency-domain refinement, which can effectively reconstruct fine-grained details even guided by a highly constrained semantic prior. Finally, while larger M_{FFD} models introduce additional computational overhead, this scaling behavior offers users a flexible trade-off between peak geometric accuracy and inference efficiency.

Pretrained Model	NYUv2		KITTI	
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
base	5.3	96.5	10.7	89.3
DA2 small	5.2	96.7	9.7	92.1
DA2 base	5.1	96.8	9.2	92.6
DA2 large	4.7	97.1	9.1	93.2
DA2 giant	**4.5**	**97.1**	**8.7**	**93.7**

Table 2
Ablation of different pretrained MDE models used in our ApDepth framework. Incorporating the semantic prior consistently improves upon the baseline, with depth accuracy scaling predictably alongside the auxiliary model's capacity. The best results in **bold**.

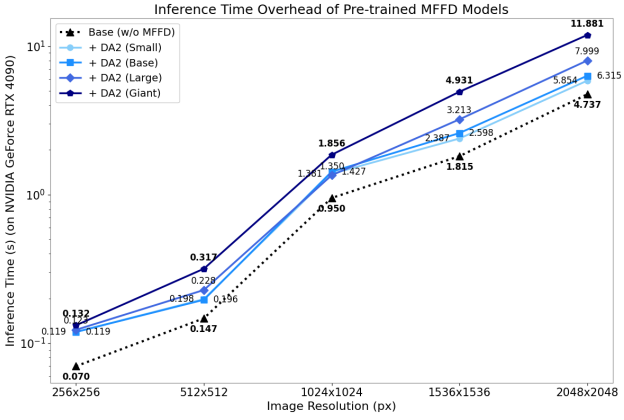


Figure 6: Inference time overhead comparison among varying scales of the pre-trained M_{FFD} model. As the capacity of the auxiliary M_{FFD} model increases from Small to Giant, the inference latency predictably rises across all image resolutions.

Adapter Type	NYUv2		KITTI	
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
MLP Adapter	4.7	96.7	8.9	93.3
Conv Adapter	**4.5**	**97.1**	**8.7**	**93.7**

Table 3
Ablation of different Adapter types used in Stage 1. We report the final depth estimation performance (after Stage 2) when employing different adapters for feature alignment in Stage 1. By explicitly maintaining 2D geometric context, our proposed Conv Adapter provides a superior structural foundation, leading to better final depth accuracy. The best results in **bold**.

Setting	NYUv2		KITTI	
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
$\lambda = 0.6$	**4.5**	**97.1**	**8.7**	**93.7**
$\lambda = 1.0$	4.5	97.1	8.8	93.7
$\lambda = 1.6$	4.6	97.0	8.8	93.5
$\lambda = 2.0$	4.6	97.0	8.8	93.6

Table 4
Ablation of the high-pass filter weight (λ) in the frequency-domain loss. A relatively small weight ($\lambda = 0.6$) yields the optimal balance, effectively enhancing edge sharpness without amplifying high-frequency noise to disrupt structural stability. The best results are in **bold**.

External Encoder	NYUv2		KITTI	
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Baseline (w/o Ext.)	4.5	96.8	8.8	93.5
DINOv2 [22]	**4.5**	**97.1**	**8.7**	**93.7**
OpenCLIP [61]	4.7	96.8	8.7	93.7
SAM [62]	4.6	97.0	8.8	93.5

Table 5
Ablation of different external encoders for Feature Alignment. DINOv2 provides the most effective patch-level semantic priors, consistently improving upon the baseline. The best results are in **bold**.

Ablation of Different Adapters. We investigate how different adapter architectures used in the Stage 1 Feature Alignment module impact the final depth estimation performance after the complete two-stage training. As shown in Table 3, compared to the standard Multi-Layer Perceptron (MLP) adapter, our proposed spatial-preserving Conv Adapter establishes a more robust semantic foundation, yielding consistent improvements in the final output across both indoor and outdoor scenarios. Specifically, the Conv Adapter consistently reduces the absolute relative error and improves the threshold accuracy on both the NYUv2 and KITTI datasets.

Ablation of High-Pass Filter Weight. We explore the influence of different high-pass filter weighting factors (λ) in our proposed frequency-domain loss. As shown in Table 4, while introducing frequency-domain supervision is crucial for capturing fine-grained details, assigning an excessively large weight to the high-pass filter gradually degrades the overall depth estimation performance. This performance drop suggests that an overly aggressive high-pass penalty may amplify high-frequency noise, which consequently disrupts the established geometric structure. We find that setting λ to a relatively small value (specifically, $\lambda = 0.6$) yields the optimal balance, effectively enhancing edge sharpness without compromising global structural stability.

Ablation of Different External Encoders. We investigate the impact of utilizing various pre-trained vision foundation models (DINOv2, OpenCLIP, and SAM) to provide semantic priors in our Stage 1 Feature Alignment module. As shown in Table 5, introducing external representations does not universally benefit the depth estimation process. For instance, OpenCLIP and SAM occasionally degrade the Absolute Relative error (AbsRel) on the indoor NYUv2 dataset compared to the baseline. This indicates that while OpenCLIP excels at image-level semantics and SAM at segmenting instance boundaries, their feature spaces may lack the continuous, fine-grained spatial geometric representations required for dense depth regression. In contrast, DINOv2 consistently delivers the most superior and balanced performance across both indoor and outdoor scene types. We attribute this to DINOv2's self-supervised patch-level training architecture, which extracts robust, localized semantic features that align most effectively with the generative features of our U-Net, providing the optimal structural guidance needed to prevent texture overfitting.

Hypersim	Virtual KITTI	NYUv2		KITTI	
		AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
✗	✓	10.3	90.9	10.4	90.3
✓	✗	4.7	96.6	9.2	92.1
✓	✓	**4.5**	**97.1**	**8.4**	**93.1**

Table 6
Ablation of training datasets. Hypersim delivers strong results; Virtual KITTI improves outdoor performance.

5. Conclusion

We presented **ApDepth**, a monocular depth estimation model that leverages a deterministic single-step diffusion framework to achieve highly efficient inference while preserving fine-grained edge details. Unlike prior diffusion models that struggle with either prohibitive multi-step latency or severe edge degradation in accelerated generation, our model utilizes a pre-trained data-driven MDE model to provide a stable structural baseline in a single forward pass. To tackle the inherent loss of high-frequency details in single-step inference, we introduce a coarse-to-fine two-stage training paradigm. This strategy features an enhanced Feature Alignment via Cosine Similarity, and a Spatial- and Frequency-Domain Loss strategy to explicitly recover sharp depth edges. Benefiting from these designs, ApDepth effectively balances computational efficiency and structural accuracy, achieving highly competitive, and often superior, performance across multiple benchmarks.

References

[1] X. Li, R. Deng, Y. Fan, P. Chen, S. Chen, Z. Liu, L. Han, S. Zhu, H. Sun, Y. Lu, Q. Li, Depth anything v2, arXiv preprint arXiv:2404.14442 (2024).

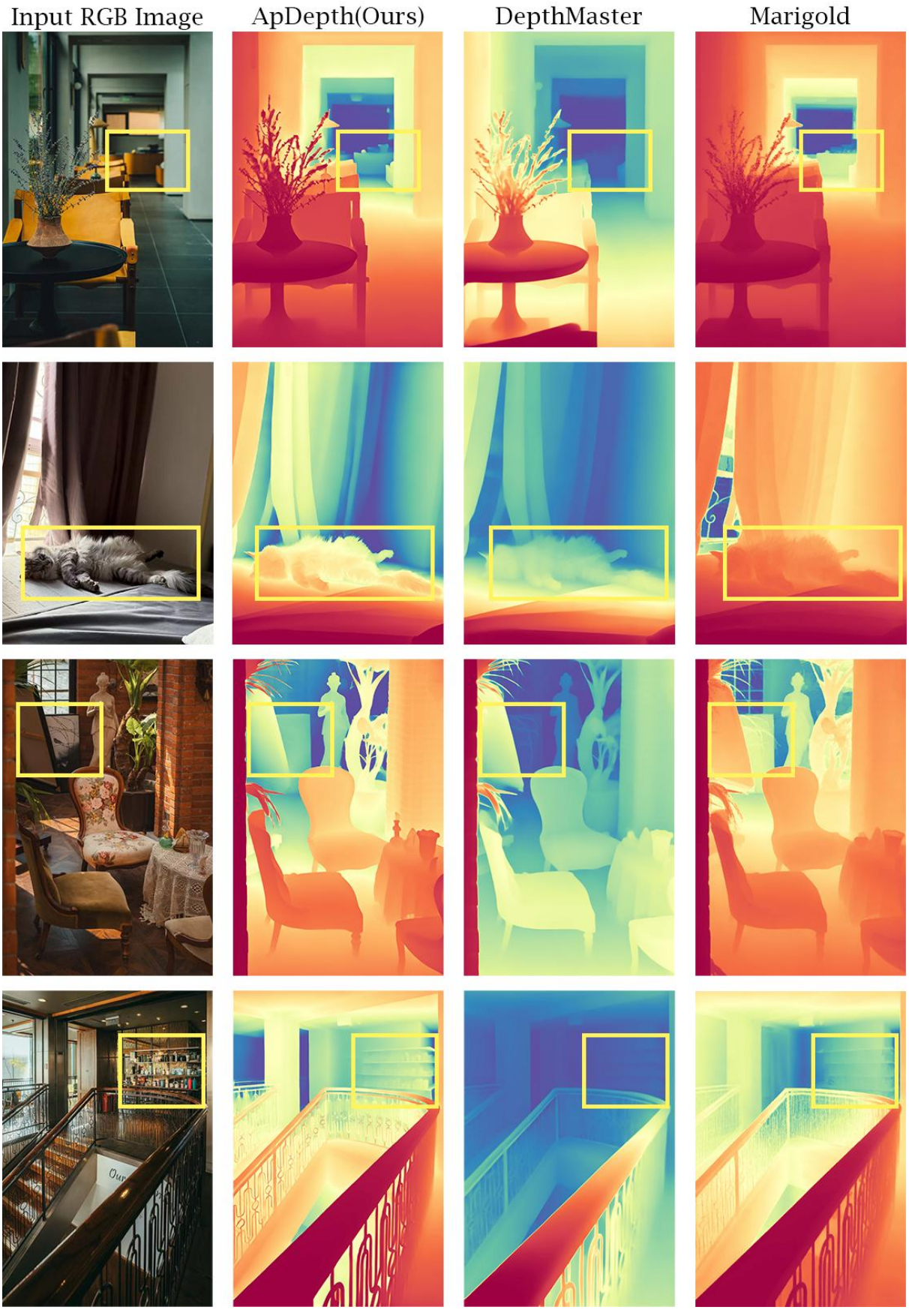
[2] R. Zhu, C. Wang, Z. Song, L. Liu, T. Zhang, Y. Zhang, Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation, arXiv preprint arXiv:2407.08187 (2024).

[3] F. Tosi, P. Z. Ramirez, M. Poggi, Diffusion models for monocular depth estimation: Overcoming challenging conditions, in: European Conference on Computer Vision, Springer, pp. 236–257.

[4] Y. Duan, X. Guo, Z. Zhu, Diffusiondepth: Diffusion denoising approach for monocular depth estimation, in: European Conference on Computer Vision, Springer, pp. 432–449.
[5] S. Shi, L. V. Gool, J. Luiten, Marigold: Repurposing diffusion models for monocular depth estimation, arXiv preprint arXiv:2404.09015 (2024).
[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.
[7] M. Gui, J. Schusterbauer, U. Prestel, P. Ma, D. Kotovenko, O. Grebenkova, S. A. Baumann, V. T. Hu, B. Ommer, Depthfm: Fast generative monocular depth estimation with flow matching, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 3203–3211.
[8] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, M. Le, Flow matching for generative modeling, arXiv preprint arXiv:2210.02747 (2022).
[9] G. M. Garcia, K. Abou Zeid, C. Schmidt, D. De Geus, A. Hermans, B. Leibe, Fine-tuning image-conditional diffusion models is easier than you think, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 753–762.
[10] G. Xu, Y. Ge, M. Liu, C. Fan, K. Xie, Z. Zhao, H. Chen, C. Shen, What matters when repurposing diffusion models for general dense perception tasks?, arXiv preprint arXiv:2403.06090 (2024).
[11] Z. Song, Z. Wang, B. Li, H. Zhang, R. Zhu, L. Liu, P.-T. Jiang, T. Zhang, Depthmaster: Taming diffusion models for monocular depth estimation, arXiv preprint arXiv:2501.02576 (2025).
[12] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, Advances in neural information processing systems 27 (2014).
[13] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2002–2011.
[14] S. F. Bhat, I. Alhashim, P. Wonka, Adabins: Depth estimation using adaptive bins, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4009–4018.
[15] J. H. Lee, M.-K. Han, D. W. Ko, I. H. Suh, From big to small: Multi-scale local planar guidance for monocular depth estimation, arXiv preprint arXiv:1907.10326 (2019).
[16] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 12179–12188.
[17] A. Agarwal, C. Arora, Attention attention everywhere: Monocular depth prediction with skip attention, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5861–5870.
[18] W. Yuan, X. Gu, Z. Dai, S. Zhu, P. Tan, New crfs: Neural window fully-connected crfs for monocular depth estimation, arXiv preprint arXiv:2203.01502 (2022).
[19] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, IEEE transactions on pattern analysis and machine intelligence 44 (2020) 1623–1637.
[20] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, C. Shen, Learning to recover 3d scene shape from a single image, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 204–213.
[21] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, H. Zhao, Depth anything: Unleashing the power of large-scale unlabeled data, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10371–10381.
[22] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
[23] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, C. Shen, Metric3d: Towards zero-shot metric 3d prediction from a single image, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9043–9053.
[24] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, S. 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).
[25] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al., Scaling rectified flow transformers for high-resolution image synthesis, in: Forty-first International Conference on Machine Learning.
[26] S. Xie, Z. Zhang, Z. Lin, T. Hinz, K. Zhang, Smartbrush: Text and shape guided object inpainting with diffusion model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22428–22437.
[27] H. Manukyan, A. Sargsyan, B. Atanyan, Z. Wang, S. Navasardyan, H. Shi, Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models, arXiv preprint arXiv:2312.14091 (2023).
[28] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, Y. Chen, Srdiff: Single image super-resolution with diffusion probabilistic models, Neurocomputing 479 (2022) 47–59.
[29] J. Wang, Z. Yue, S. Zhou, K. C. Chan, C. C. Loy, Exploiting diffusion prior for real-world image super-resolution, International Journal of Computer Vision (2024) 1–21.
[30] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International conference on machine learning, pmlr, pp. 2256–2265.
[31] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in neural information processing systems 33 (2020) 6840–6851.
[32] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, Advances in neural information processing systems 32 (2019).
[33] Y. Song, S. Ermon, Improved techniques for training score-based generative models, Advances in neural information processing systems 33 (2020) 12438–12448.
[34] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, arXiv preprint arXiv:2011.13456 (2020).
[35] Z. Kong, W. Ping, On fast sampling of diffusion probabilistic models, arXiv preprint arXiv:2106.00132 (2021).
[36] R. San-Roman, E. Nachmani, L. Wolf, Noise estimation for generative diffusion models, arXiv preprint arXiv:2104.02600 (2021).
[37] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).
[38] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, T. Salimans, Cascaded diffusion models for high fidelity image generation, Journal of Machine Learning Research 23 (2022) 1–33.
[39] A. Vahdat, K. Kreis, J. Kautz, Score-based generative modeling in latent space, Advances in neural information processing systems 34 (2021) 11287–11302.
[40] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, Advances in neural information processing systems 35 (2022) 25278–25294.
[41] Y. LeCun, A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27, Open Review 62 (2022) 1–62.
[42] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15619–15629.
[43] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, S. Xie, Representation alignment for generation: Training diffusion transformers is

easier than you think, arXiv preprint arXiv:2410.06940 (2024).
[44] Y. Ji, Z. Chen, E. Xie, L. Hong, X. Liu, Z. Liu, T. Lu, Z. Li, P. Luo, Ddp: Diffusion model for dense visual prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21741–21752.
[45] S. Saxena, A. Kar, M. Norouzi, D. J. Fleet, Monocular depth estimation using diffusion models, arXiv preprint arXiv:2302.14816 (2023).
[46] J. He, H. Li, W. Yin, Y. Liang, L. Li, K. Zhou, H. Zhang, B. Liu, Y.-C. Chen, Lotus: Diffusion-based visual foundation model for high-quality dense prediction, arXiv preprint arXiv:2409.18124 (2024).
[47] W. Yin, X. Wang, C. Shen, Y. Liu, Z. Tian, S. Xu, C. Sun, D. Renyin, Diversedepth: Affine-invariant depth prediction using diverse data, arXiv preprint arXiv:2002.00569 (2020).
[48] A. Eftekhar, A. Sax, J. Malik, A. Zamir, Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10786–10796.
[49] C. Zhang, W. Yin, B. Wang, G. Yu, B. Fu, C. Shen, Hierarchical normalization for robust monocular depth estimation, Advances in Neural Information Processing Systems 35 (2022) 14128–14139.
[50] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, K. Schindler, Repurposing diffusion-based image generators for monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9492–9502.
[51] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, X. Long, Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image, in: European Conference on Computer Vision, Springer, pp. 241–258.
[52] M. Gui, J. Schusterbauer, U. Prestel, P. Ma, D. Kotovenko, O. Grebenkova, S. A. Baumann, V. T. Hu, B. Ommer, Depthfm: Fast monocular depth estimation with flow matching, 2024.
[53] G. Xu, Y. Ge, M. Liu, C. Fan, K. Xie, Z. Zhao, H. Chen, C. Shen, What matters when repurposing diffusion models for general dense perception tasks?, arXiv preprint arXiv:2403.06090 (2024).
[54] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, J. M. Susskind, Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10912–10922.
[55] Y. Cabon, N. Murray, M. Humenberger, Virtual kitti 2, arXiv preprint arXiv:2001.10773 (2020).
[56] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European conference on computer vision, Springer, pp. 746–760.
[57] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5828–5839.
[58] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, The international journal of robotics research 32 (2013) 1231–1237.
[59] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3260–3269.
[60] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, et al., Diode: A dense indoor and outdoor depth dataset, arXiv preprint arXiv:1908.00463 (2019).
[61] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, J. Jitsev, Reproducible scaling laws for contrastive language-image learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2818–2829.
[62] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015–4026.


Input RGB Image
ApDepth(Ours)
DepthMaster
Marigold


Input RGB Image
ApDepth(Ours)
DepthMaster
Marigold

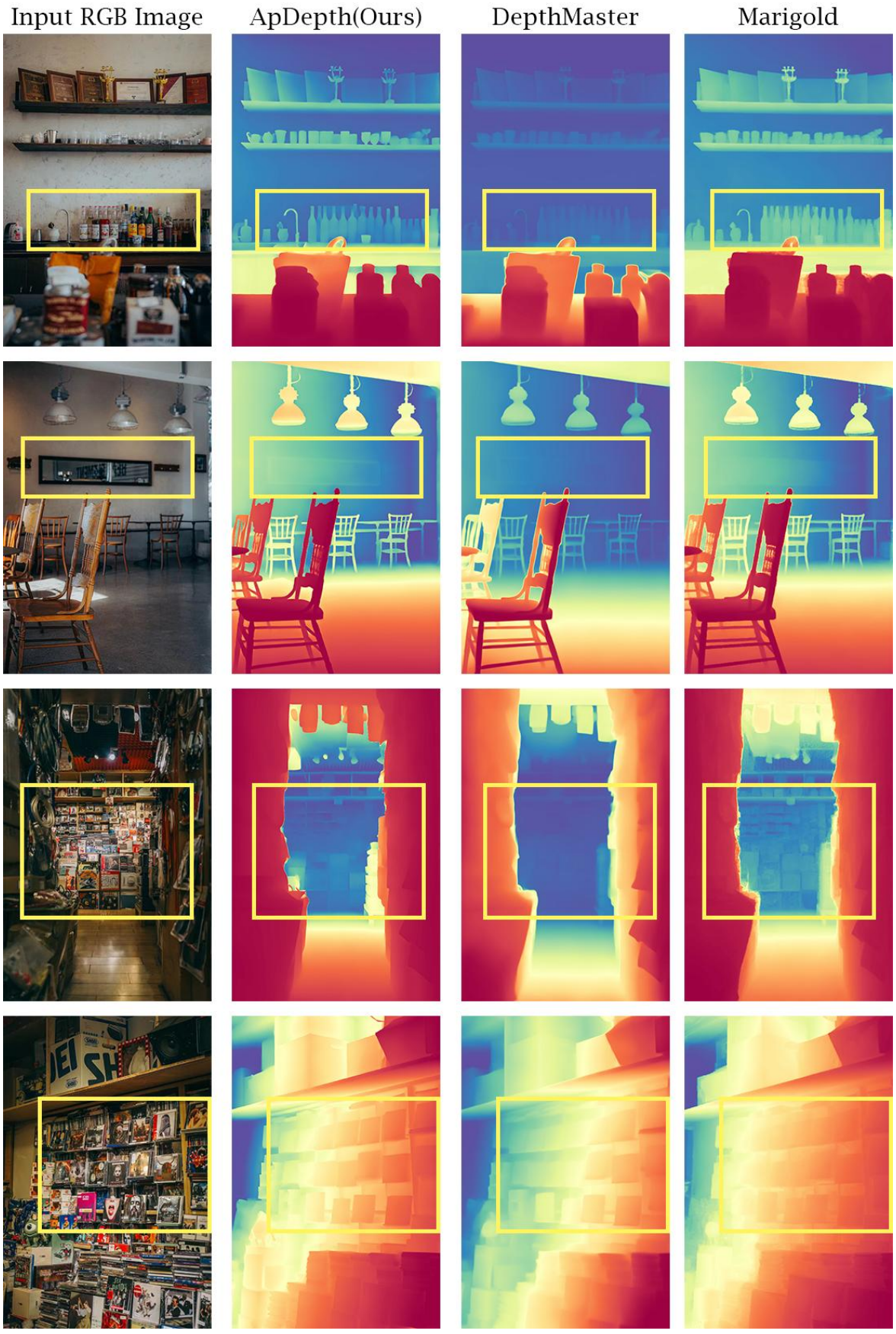


Figure 7: Comparison with Different Model-Driven Approaches on InDoor Images


Input RGB Image
ApDepth(Ours)
DepthMaster
Marigold
停
避让行人

Figure 8: Comparison with Different Model-Driven Approaches on Outdoor Images