# Pixel-Perfect Depth with Semantics-Prompted Diffusion Transformers

**Gangwei Xu**[1,2*] **Haotong Lin**[3*] **Hongcheng Luo**[2] **Xianqi Wang**[1] **Jingfeng Yao**[1]
**Lianghui Zhu**[1] **Yuechuan Pu**[2] **Cheng Chi**[2] **Haiyang Sun**[2†] **Bing Wang**[2]
**Guang Chen**[2] **Hangjun Ye**[2] **Sida Peng**[3] **Xin Yang**[1†✉]

[1]Huazhong University of Science and Technology [2]Xiaomi EV [3]Zhejiang University
https://pixel-perfect-depth.github.io

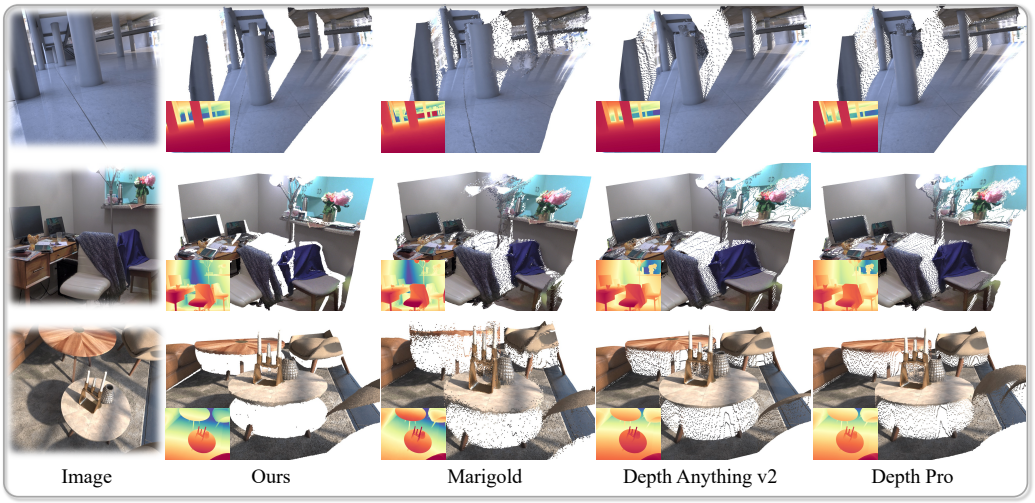


Figure 1: We present **Pixel-Perfect Depth**, a monocular depth estimation model with pixel-space diffusion transformers. Compared to existing discriminative [68, 4] and generative [29] models, its estimated depth maps can recover high-quality, flying-pixel-free point clouds.

## Abstract

This paper presents **Pixel-Perfect Depth**, a monocular depth estimation model based on pixel-space diffusion generation that recovers high-quality, flying-pixel-free point clouds from estimated depth maps. Current generative depth estimation models fine-tune Stable Diffusion and achieve impressive performance. However, they require a VAE to compress depth maps into latent space, which inevitably introduces *flying pixels* at edges and details. Our model addresses this challenge by directly performing diffusion generation in the pixel space, avoiding VAE-induced artifacts. To overcome the high complexity associated with pixel-space generation, we introduce two novel designs: 1) **Semantics-Prompted Diffusion Transformers (DiT)**, which incorporate semantic representations from vision foundation models into DiT to prompt the diffusion process, thereby preserving global semantic consistency while enhancing fine-grained visual details; and 2) **Cascade DiT Design** that progressively increases the number of tokens to further enhance efficiency and accuracy. Our model achieves the best performance among all published generative models across five benchmarks, and significantly outperforms all other models in edge-aware point cloud evaluation.

[*] Equal contribution, [†] Project leader, [✉] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).


Figure 2: **Qualitative comparisons**. GT(VAE) denotes the ground truth depth map after VAE reconstruction. Existing generative models [29] use a VAE to compress inputs into the latent space, inevitably introducing *flying pixels* at edges and details. In contrast, our model directly performs diffusion in pixel space, avoiding these issues. Depth maps are visualized on the point clouds.

## 1 Introduction

Monocular depth estimation (MDE) is a fundamental task with a wide range of downstream applications, such as 3D reconstruction, novel view synthesis, and robotic manipulation. Due to its significance, a large number of depth estimation models [29, 67, 68, 74] have emerged recently. These models achieve high-quality results in most zero-shot scenarios or regions, but suffer from *flying pixels* around object boundaries and fine details when converted into point clouds, as shown in Figure 1 and 4, which limits their practical applications in tasks such as free-viewpoint broadcast, robotic manipulation, and immersive content creation.

Current models suffer from the *flying pixels* problem due to different reasons. For discriminative models [68, 4, 74, 25], *flying pixels* mainly arise from their tendency to output an intermediate (*average*) depth value between the foreground and background at depth-discontinuous edges, in order to minimize regression loss. In contrast, generative models [29, 14, 18] bypass direct regression by modeling pixel-wise depth distributions, allowing them to preserve sharp edges and recover fine structures more faithfully. However, current generative depth models typically fine-tune Stable Diffusion [45] for depth estimation, which requires Variational Autoencoders (VAEs) to compress depth maps into a latent space. This compression inevitably leads to the loss of edge sharpness and structural fidelity, resulting in a significant number of *flying pixels*, as shown in Figure 2.

A trivial solution could be training a diffusion-based monocular depth model in pixel space, bypassing the use of VAEs. However, we find this highly challenging, due to the increased complexity and instability of modeling both global semantic consistency and fine-grained visual details, leading to extremely low-quality depth predictions (Table 2 and Figure 5). To further investigate this limitation, we examine prior studies on high-resolution image generation. Several works [24, 55, 80], through signal-to-noise ratio (SNR) analysis, have pointed out that adding noise with higher intensity is more likely to disrupt the global structures or low-frequency components of high-resolution images, thereby improving generation. This reveals that the primary difficulty in high-resolution pixel-space generation lies in effectively perceiving and modeling global image structures.

In this paper, we present **Pixel-Perfect Depth**, a framework for high-quality and flying-pixel-free monocular depth estimation using pixel-space diffusion transformers. Recognizing that the major difficulty in high-resolution pixel-space generation lies in perceiving and modeling global image structures. To address this challenge, we propose the Semantics-Prompted Diffusion Transformers (SP-DiT) that incorporate high-level semantic representations into the diffusion process to enhance the model's ability to preserve global structures and semantic coherence. Equipped with SP-DiT, our model can more effectively preserve global semantic consistency while generating fine-grained visual details in high-resolution pixel space. However, the semantic representations obtained from vision foundation models [38, 68, 58, 19] often do not align well with the internal representations of DiT, leading to training instability and convergence issues. To address this, we introduce a simple yet effective regularization technique for semantic representations, which ensures stable training and facilitates convergence to desirable solutions. As shown in Table 2 and Figure 5, the proposed SP-DiT significantly improves overall performance, achieving up to 78% improvement on the NYUv2 [52] AbsRel metric.

Furthermore, we introduce the Cascade DiT Design (Cas-DiT), an efficient architecture for diffusion transformers. We find that in diffusion transformers, the early blocks are primarily responsible

for capturing and generating global or low-frequency structures, while the later blocks focus on generating high-frequency details. Based on this insight, Cas-DiT adopts a progressive patch size strategy: larger patch sizes are used in the early DiT blocks to reduce the number of tokens and facilitate global image structure modeling; in the later DiT blocks, we increase the number of tokens, which is equivalent to using smaller patch sizes, allowing the model to focus on the generation of fine-grained spatial details. This coarse-to-fine cascaded design not only significantly reduces computational costs and improves efficiency, but also brings significant accuracy gains.

We highlight the main contributions of this paper below:

- We present **Pixel-Perfect Depth**, a monocular depth estimation model with pixel-space diffusion generation, capable of producing flying-pixel-free point clouds from estimated depth maps.
- We introduce **Semantics-Prompted DiT**, which integrates normalized semantic representations into the DiT to effectively preserve global semantic consistency while enhancing fine-grained visual details. This significantly boosts overall performance. We further propose a novel **Cascade DiT Design** to enhance the efficiency and accuracy of our model.
- Our model achieves the best performance across five benchmarks among all published generative depth estimation models.
- We introduce an edge-aware point cloud evaluation metric, which effectively assesses *flying pixels* at edges. Our model significantly outperforms previous models in this evaluation.

## 2 Related Work

### 2.1 Monocular Depth Estimation

Early monocular depth estimation methods relied primarily on manually designed features [46, 23]. The advent of neural networks revolutionized the field, though initial approaches [12, 11] struggled with cross-dataset generalization. To address this limitation, scale-invariant and relative loss [43] are introduced, enabling multi-dataset [31, 72, 7, 63, 61, 59, 57, 62, 44] training. Recent methods focus on improving the generalization ability [68, 4], depth consistency [66, 6, 26, 28], and metric scale [3, 32, 33, 74, 17, 75, 25, 40, 34] of depth estimation. These methods converge towards using transformer-based architectures [42]. Concurrent works [60, 65] explore point cloud representations to improve depth estimation performance. Several recent methods [27, 9, 49, 47, 48, 79] have attempted to use diffusion models for metric depth estimation. In contrast, our method focuses on relative depth and demonstrates improved generalization and fine-grained detail across a wide range of real-world scenes. Furthermore, our model significantly differs from these methods by introducing Semantics-Prompted DiT, which incorporates pretrained high-level semantic representations into the diffusion process, greatly enhancing performance.

More recently, [29] brought the new insight to the field by fine-tuning pretrained Stable Diffusion [45] for depth estimation, which demonstrated impressive zero-shot capabilities for relative depth. The following works [18, 16, 54, 78, 2] attempt to improve its performance and inference speed. However, they are all based on the latent diffusion model [45], which is trained in the latent space and requires a VAE to compress the depth map into a latent space. We focus on a pixel-space diffusion model that is trained directly in the pixel space without requiring any VAE.

### 2.2 Diffusion Generative Models

Diffusion generative models [20, 53, 39, 76, 70, 71] have demonstrated impressive results in image and video generation. Early approaches [20, 22, 21] such as DDPM [20] operate directly in the pixel space, enabling high-fidelity generation but incurring significant computational costs, especially at high resolutions. To address this limitation, Latent Diffusion Models perform the diffusion process in a lower-dimensional latent space obtained via a VAE, as popularized by Stable Diffusion [45]. This design significantly improves training and inference efficiency and has been widely adopted in recent works [13, 71, 76, 81, 30, 41, 69].

Diffusion models for monocular depth estimation typically follow a similar trend. For instance, Marigold [29] and its follow-ups [18, 16] fine-tune pretrained Stable Diffusion [45] models for depth

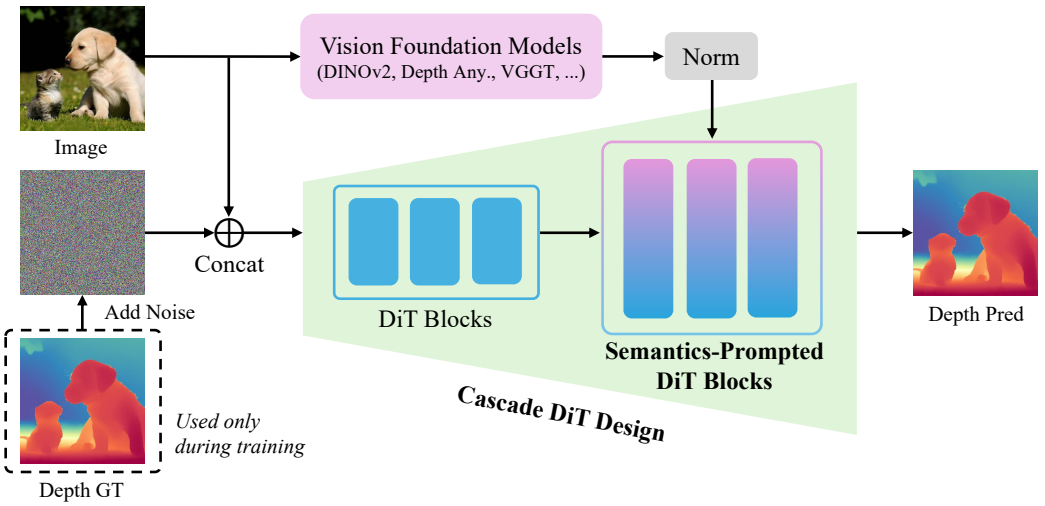


Figure 3: **Overview of Pixel-Perfect Depth.** Given an input image, we concatenate it with noise and feed it into the proposed Cascade DiT. Meanwhile, the image is also processed by a pretrained encoder from Vision Foundation Models to extract high-level semantics, forming our Semantics-Prompted DiT. We perform diffusion generation directly in pixel space without using any VAE.

prediction, benefiting from fast convergence and strong priors learned from large-scale datasets. However, the VAE's latent compression leads to *flying pixels* in the resulting point clouds. In contrast, pixel-space diffusion avoids such artifacts but remains computationally intensive and slow to converge at high resolutions. To address this, we propose Semantics-Prompted DiT and Cascade DiT Design, which enables efficient high-resolution depth estimation without latent compression.

## 3 Method

### 3.1 Pixel-Perfect Depth

Given an input image, our goal is to estimate a pixel-perfect depth map that is free of *flying pixels* when converted to point clouds. Existing models [29, 14, 18, 68, 4] often suffer from *flying pixels* due to their inherent modeling paradigms. Discriminative models tend to smooth object edges and blur fine details because of their mean-prediction bias, which results in noticeable *flying pixels* in the reconstructed point clouds. Generative models, in theory, can better capture the multi-modal depth distribution at object edges. However, current generative models typically fine-tune Stable Diffusion [45] for depth estimation, relying on its strong image priors. This requires compressing the depth map into a latent space via a VAE, inevitably causing *flying pixels*.

To unleash the potential of generative models for depth estimation, we propose **Pixel-Perfect Depth** that performs diffusion directly in the pixel space instead of the latent space. It allows us to directly model the pixel-wise distribution of depth, such as the discontinuities at object edges. However, training a generative diffusion model directly in the high-resolution pixel space (*e.g.*, 1024×768) is computationally demanding and hard to optimize. To overcome these challenges, we introduce Semantics-Prompted DiT and Cascaded DiT Design, detailed in the following sections.

### 3.2 Generative Formulation

We adopt Flow Matching [35, 36, 1] as the generative core of our depth estimation framework. Flow Matching learns a continuous transformation from Gaussian noise to data samples via a first-order Ordinary Differential Equation (ODE). In our case, we model the transformation from Gaussian noise to depth samples. Specifically, given clean depth samples $\mathbf{x}_0 \sim \mathcal{D}$ and Gaussian noise $\mathbf{x}_1 \sim \mathcal{N}(0, 1)$, we define an interpolated sample at continuous time $t \in [0, 1]$ as:

$$\mathbf{x}_t = t \cdot \mathbf{x}_1 + (1 - t) \cdot \mathbf{x}_0. \tag{1}$$

This defines a velocity field:

$$\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - \mathbf{x}_0, \tag{2}$$

which describes the direction from clean data to noise. Our model $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$ is trained to predict the velocity field, based on the current noisy sample $\mathbf{x}_t$, the time step $t$, and the input image $\mathbf{c}$. The training objective is the mean squared error (MSE) between the predicted and true velocity:

$$\mathcal{L}_{\text{velocity}(\theta)} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t}\left[\left\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}_t\right\|^2\right]. \tag{3}$$

At inference, we start from noise $\mathbf{x}_1$ and solve the ODE by discretizing the time interval $[0, 1]$ into steps $t_i$, iteratively updating the sample as follows:

$$\mathbf{x}_{t_{i-1}} = \mathbf{x}_{t_i} + \mathbf{v}_\theta(\mathbf{x}_{t_i}, t_i, \mathbf{c})(t_{i-1} - t_i), \tag{4}$$

where $t_i$ decreases from 1 to 0, gradually transforming the initial noise $\mathbf{x}_1$ into the depth sample $\mathbf{x}_0$.

### 3.3 Semantics-Prompted Diffusion Transformers

Our Semantics-Prompted DiT builds on Diffusion Transformers (DiT) [39] for its simplicity, scalability, and strong performance in generative modeling. Unlike previous depth estimation models such as Depth Anything v2 [68] and Marigold [29], our architecture is purely Transformer-based, without any convolutional layers. By integrating high-level semantic representations, our model preserves global semantic consistency while enhancing fine-grained visual details, without sacrificing the simplicity and scalability of DiT.

Specifically, given the interpolated noise sample $\mathbf{x}_t$ and the corresponding image $\mathbf{c}$, we first concatenate them into a single input: $\mathbf{a}_t = \mathbf{x}_t \oplus \mathbf{c}$, where the image $\mathbf{c}$ serves as a condition. Then, we directly feed $\mathbf{a}_t$ into the DiT. The first layer of DiT is a patchify operation, which converts the spatial input $\mathbf{a}_t$ into a 1D sequence of $T$ tokens (patches), each with a dimension of $D$, by linearly embedding each patch of size $p \times p$ from the input $\mathbf{a}_t$. Subsequently, the input tokens are processed by a sequence of Transformer blocks, called DiT blocks. After the final DiT block, each token is linearly projected into a $p \times p$ tensor, which is then reshaped back to the original spatial resolution to obtain the predicted velocity $\mathbf{v}$ (*i.e.*, $\mathbf{x}_1 - \mathbf{x}_0$), with a channel dimension of 1.

Unfortunately, performing diffusion directly in the pixel space leads to poor convergence and highly inaccurate depth predictions. As shown in Figure 5, the model struggles to model both global image structure and fine-grained details. To address this, we extract high-level semantic representations $\mathbf{e}$ as guidance from the input image $\mathbf{c}$ using a vision foundation model $f$, as follows:

$$\mathbf{e} = f(\mathbf{c}) \in \mathbb{R}^{T' \times D'}, \tag{5}$$

where $T'$ and $D'$ are the number of tokens and the embedding dimension of $f$, respectively. These high-level semantic representations are then incorporated into our DiT model, enabling it to more effectively preserve global semantic consistency while enhancing fine-grained visual details. However, we found that the magnitude of the obtained semantics $\mathbf{e}$ differs significantly from the magnitude of the tokens in our DiT model, which affects both the stability of the model's training and its performance. To address this, we normalize the semantic representation $\mathbf{e}$ along the feature dimension using L2 norm, as follows:

$$\hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}. \tag{6}$$

Subsequently, the normalized semantic representation is integrated into the tokens $\mathbf{z}$ of our DiT model via a multilayer perceptron (MLP) network $h_\phi$,

$$\mathbf{z}' = h_\phi(\mathbf{z} \oplus \mathcal{B}(\hat{\mathbf{e}})), \tag{7}$$

where $\mathcal{B}(\cdot)$ denotes the bilinear interpolation operator, which aligns the spatial resolution of the semantic representation $\hat{\mathbf{e}}$ with that of the DiT tokens. The resulting $\mathbf{z}'$ denotes the DiT tokens enhanced with semantics. After the fusion, the subsequent DiT blocks are prompted by semantics to effectively preserve global semantic consistency while enhancing fine-grained visual details in the high-resolution pixel space. We refer to these subsequent DiT blocks as Semantics-Prompted DiT.

In this work, we experiment with various pretrained vision foundation models, including DINOv2 [38], VGGT [58], MAE [19], and Depth Anything v2 [68]. All of them significantly boost performance and facilitate more stable and efficient training, as shown in Table 3. Note that we only utilize the encoder of each vision foundation model, *e.g.*, a 24-layer Vision Transformer encoder (ViT-L/14) for both DINOv2 [38] and Depth Anything v2 [68].

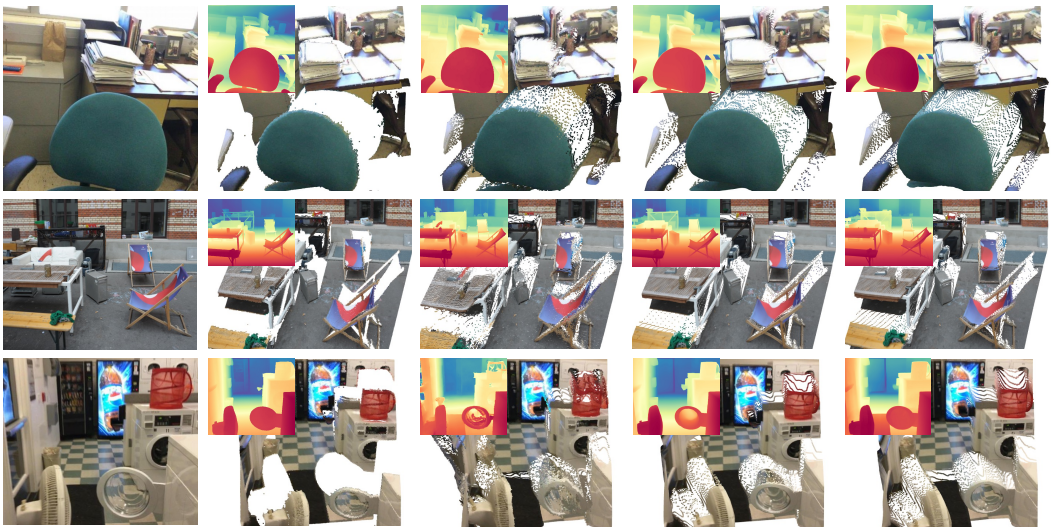


Figure 4: **Qualitative point cloud results in complex scenes.** Our model produces significantly fewer *flying pixels* compared to other depth estimation models [29, 68, 4], with depth maps overlaid on the point clouds for visualization.

### 3.4 Cascade DiT Design

Although the proposed Semantics-Prompted DiT significantly improves accuracy performance, performing diffusion directly in the pixel space remains computationally expensive. To address this issue, We propose a novel Cascaded DiT Design to reduce the computational burden of the model. We observe that in DiT architectures, the early blocks are primarily responsible for capturing global image structures and low-frequency information, while the later blocks focus on modeling fine-grained, high-frequency details.

To optimize the efficiency and effectiveness of this process, we adopt a large patch size in the early DiT blocks. This design significantly reduces the number of tokens that need to be processed, leading to lower computational cost. Additionally, it encourages the model to prioritize learning and modeling global image structures and low-frequency information, which also better aligns with the high-level semantic representations extracted from the input image. In the later DiT blocks, we increase the number of tokens, which is equivalent to using smaller patch sizes. This allows the model to better focus on fine-grained spatial details. The resulting coarse-to-fine cascaded design mirrors the hierarchical nature of visual perception and improves both the efficiency and accuracy of depth estimation.

Specifically, for our diffusion model with a total of $N$ DiT blocks, the first $N/2$ blocks constitute the coarse stage with a larger patch size, while the remaining $N/2$ blocks (*i.e.*, SP-DiT) form the fine stage using a smaller patch size.

### 3.5 Implementation Details

In this section, we provide essential information about the model architecture details, depth normalization, and training details.

**Model architecture details.** In our implementation, we use a total of $N = 24$ DiT blocks, each operating at a hidden dimension of $D = 1024$. The first 12 blocks are standard DiT blocks with a patch size of 16, corresponding to $(H/16) \times (W/16)$ tokens for an input of size $H \times W$. The remaining 12 blocks are our proposed Semantics-Prompted DiT blocks, which adopt a finer patch size of 8, resulting in $(H/8) \times (W/8)$ tokens. After the 12th block, we use an MLP network to increase the hidden dimension by a factor of 4, followed by reshaping to obtain more tokens.

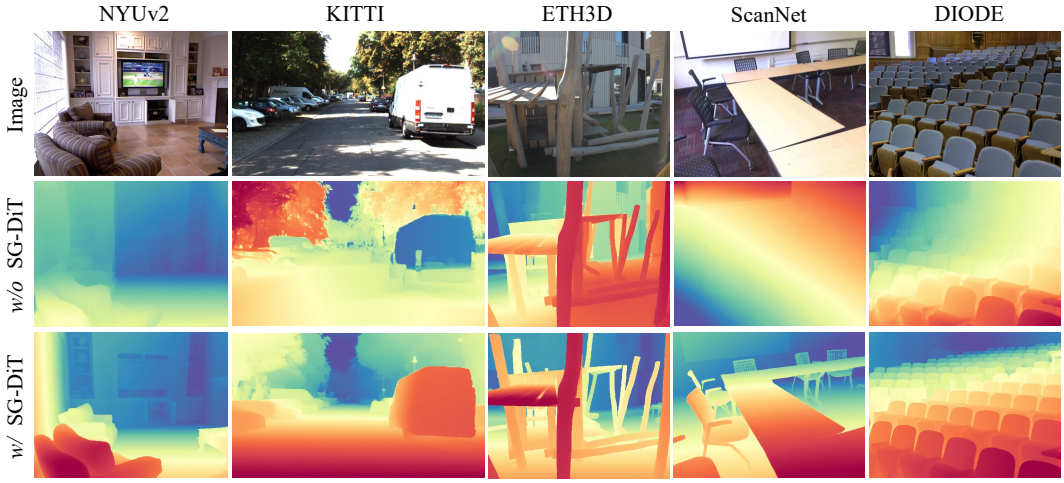


Figure 5: **Qualitative ablations for the proposed SP-DiT.** From top to bottom: input images from five benchmarks, results without SP-DiT, and results with SP-DiT. Without SP-DiT, the DiT model struggles with preserving global semantics and generating fine-grained visual details.

**Depth normalization.** The ground truth depth values are normalized to match the scale expected by the diffusion model. Before normalization, we convert the depth values into log scale to ensure a more balanced capacity allocation across both indoor and outdoor scenes. Specifically, we apply the transformation $\tilde{\mathbf{d}} = \log(\mathbf{d} + \epsilon)$, where $\tilde{\mathbf{d}}$ denotes the transformed depth, $\mathbf{d}$ is the original depth value, and $\epsilon$ is a small positive constant (*e.g.*, 1) to ensure numerical stability. We then normalize the log-scaled depth $\tilde{\mathbf{d}}$ using:

$$\hat{\mathbf{d}} = \frac{\tilde{\mathbf{d}} - d_{\text{min}}}{d_{\text{max}} - d_{\text{min}}} - 0.5, \tag{8}$$

where $d_{min}$ and $d_{max}$ are the 2% and 98% depth percentiles of each map, respectively.

**Training details.** We train two variants of the diffusion model at different resolutions: one at $512 \times 512$ and the other at $1024 \times 768$. Our training images are originally at a resolution of $1024 \times 768$. To train the $512 \times 512$ model, we resize the shorter side to 512 and then apply a random crop to obtain square inputs. We train all models for 800K steps using 8 NVIDIA GPUs, with a batch size of 4 per GPU. We train all models using the AdamW optimizer with a constant learning rate of $1 \times 10^{-4}$. We do not use any data augmentation during training. The training loss is the MSE loss between the predicted and true velocity, as shown in Equation 3.

# 4 Experiments

## 4.1 Experimental Setup

**Training datasets.** Our objective is to estimate pixel-perfect depth maps, which, when converted to point clouds, are free of *flying pixels* and geometric artifacts. To achieve this, it is essential to train on datasets with high-quality ground truth point clouds. We adopt Hypersim [44], a photorealistic synthetic dataset that offers accurate and clean 3D geometry. Specifically, we use the official training split of Hypersim, which contains approximately 54K samples, as our training data. The resolution of the dataset is $1024 \times 768$.

**Evaluation setup.** Following the majority of previous depth estimation models [29, 14, 18], we evaluate the zero-shot relative depth estimation performance on five real-world datasets: NYUv2 [52], KITTI [15], ETH3D [50], ScanNet [8], and DIODE [56], covering both indoor and outdoor scenes. To assess the quality of depth estimation, we adopt two widely-used evaluation metrics: Absolute Relative Error (AbsRel) and $\delta_1$ accuracy. To demonstrate that our model generates point clouds without *flying pixels*, we convert the estimated depth maps into 3D point clouds and evaluate them using the proposed edge-aware metric. For simplicity, all quantitative evaluations are conducted using the $512 \times 512$ model. We use the $1024 \times 768$ model for the qualitative results shown in Figure 1.

Table 1: **Zero-shot relative depth estimation.** Better: AbsRel ↓, $\delta_1$ ↑. **Bold** numbers are the best. Our model outperforms other generative models on five benchmarks.

| Type | Method | Training Data | NYUv2 | | KITTI | | ETH3D | | ScanNet | | DIODE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ |
| *Discriminative* | DiverseDepth[73] | 320K | 11.7 | 87.5 | 19.0 | 70.4 | 22.8 | 69.4 | 10.9 | 88.2 | - | - |
| | MiDaS[43] | 2M | 11.1 | 88.5 | 23.6 | 63.0 | 18.4 | 75.2 | 12.1 | 84.6 | - | - |
| | LeReS[75] | 354K | 9.0 | 91.6 | 14.9 | 78.4 | 17.1 | 77.7 | 9.1 | 91.7 | - | - |
| | Omnidata[10] | 12M | 7.4 | 94.5 | 14.9 | 83.5 | 16.6 | 77.8 | 7.5 | 93.6 | - | - |
| | HDN[77] | 300K | 6.9 | 94.8 | 11.5 | 86.7 | 12.1 | 83.3 | 8.0 | 93.9 | - | - |
| | DPT[42] | 1.2M | 9.8 | 90.3 | 10.0 | 90.1 | **7.8** | **94.6** | 8.2 | 93.4 | - | - |
| | DepthAny. v2[68] | 54K | 5.4 | 97.2 | 8.6 | 92.8 | 12.3 | 88.4 | - | - | 8.8 | 93.7 |
| | DepthAny. v2[68] | 62M | **4.5** | **97.9** | **7.4** | **94.6** | 13.1 | 86.5 | **6.5** | **97.2** | **6.6** | **95.2** |
| *Generative* | Marigold[29] | 74K | 5.5 | 96.4 | 9.9 | 91.6 | 6.5 | 96.0 | 6.4 | 95.1 | 10.0 | 90.7 |
| | GeoWizard[14] | 280K | 5.2 | 96.6 | 9.7 | 92.1 | 6.4 | 96.1 | 6.1 | 95.3 | 12.0 | 89.8 |
| | DepthFM[16] | 74K | 5.5 | 96.3 | 8.9 | 91.3 | 5.8 | 96.2 | 6.3 | 95.4 | - | - |
| | GenPercept[64] | 90K | 5.2 | 96.6 | 9.4 | 92.3 | 6.6 | 95.7 | 5.6 | 96.5 | - | - |
| | Lotus[18] | 54K | 5.4 | 96.8 | 8.5 | 92.2 | 5.9 | 97.0 | 5.9 | 95.7 | 9.8 | 92.4 |
| | Ours | 54K | **4.3** | **97.4** | **8.0** | **93.1** | **4.5** | **97.7** | **4.5** | **97.3** | **7.0** | **95.5** |

Table 2: **Ablation studies on five zero-shot benchmarks.** All metrics are presented in percentage terms, **bold** numbers are the best. Inference time was tested on an RTX 4090 GPU.

| Method | NYUv2 | | KITTI | | ETH3D | | ScanNet | | DIODE | | Time(s) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | |
| DiT (baseline) | 22.5 | 72.8 | 27.3 | 63.9 | 12.1 | 87.4 | 25.7 | 65.1 | 23.9 | 76.5 | 0.19 |
| SP-DiT | 4.8 | 96.7 | 8.6 | 92.2 | 4.6 | 97.5 | 6.2 | 94.8 | 8.2 | 94.1 | 0.20 |
| SP-DiT+Cas-DiT | **4.3** | **97.4** | **8.0** | **93.1** | **4.5** | **97.7** | **4.5** | **97.3** | **7.0** | **95.5** | **0.14** |

### 4.2 Ablations and Analysis

**Component-wise ablation analysis.** We adopt the DiT [39] model as our baseline and conduct ablations on our proposed modules. Quantitative results are shown in Table 2. Directly performing diffusion generation in high-resolution pixel space is highly challenging due to substantial computational costs and optimization difficulties, leading to significant performance degradation. As illustrated in Figure 5, the baseline model struggles with preserving global semantics and generating fine-grained visual details. In contrast, the proposed Semantics-Prompted DiT (SP-DiT) addresses these challenges, achieving significantly improved accuracy, for example, a 78% gain on the NYUv2 AbsRel metric. We further introduce a novel Cascaded DiT Design (Cas-DiT) that progressively increases the number of tokens. This coarse-to-fine design not only significantly improves efficiency, for example, reducing inference time by 30% on an RTX 4090 GPU, but also better models global context, leading to noticeable gains in accuracy.

**Ablations on vision foundation models (VFMs).** We evaluate the performance of SP-DiT using pretrained vision encoders from different VFMs, including MAE [19], DINOv2 [38], Depth Anything v2 [68], and VGGT [58], as illustrated in Table 3. All of them significantly boost performance.

### 4.3 Zero-Shot Relative Depth Estimation

To evaluate our model's zero-shot generalization, we compare it with recent depth estimation models [68, 4, 29, 18, 16] on five real-world benchmarks. As shown in Table 1, our model outperforms all other generative depth estimation models for all evaluation metrics. Unlike previous generative models, we do not rely on image priors from a pretrained Stable Diffusion [45] model. Instead, our diffusion model is trained from scratch and still achieves superior performance. Our model generalizes well to a wide range of real-world scenes, even when trained solely on synthetic depth datasets. Additionally, it outperforms discriminative models trained on similar amounts of training data. Unlike

Table 3: **Ablation studies on Vision Foundation Models (VFMs).** Note that we only utilize a pretrained encoder from these VFMs, such as a 24-layer ViT from DINOv2 or Depth Anything v2.

| VFM Type | NYUv2 | | KITTI | | ETH3D | | ScanNet | | DIODE | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ |
| w/o SP-DiT | 22.5 | 72.8 | 27.3 | 63.9 | 12.1 | 87.4 | 25.7 | 65.1 | 23.9 | 76.5 |
| SP-DiT (MAE [19]) | 6.4 | 95.0 | 14.4 | 84.9 | 7.3 | 94.8 | 7.7 | 92.5 | 11.6 | 91.3 |
| SP-DiT (DINOv2 [38]) | 4.8 | 96.4 | 9.3 | 91.2 | 5.6 | 96.2 | 5.1 | 96.9 | 9.2 | 93.5 |
| SP-DiT (VGGT [58]) | 4.7 | 96.7 | **7.6** | **94.1** | **4.1** | **97.8** | **3.8** | **98.0** | 7.8 | 94.9 |
| SP-DiT (DepthAny. v2 [68]) | **4.3** | **97.4** | 8.0 | 93.1 | 4.5 | 97.7 | 4.5 | 97.3 | **7.0** | **95.5** |

Table 4: **Edge-aware point cloud evaluation.** Our model achieves the best performance on the high-quality Hypersim test set. To further verify that VAE compression leads to *flying pixels*, we evaluate the ground truth depth maps after VAE reconstruction, denoted as GT(VAE).

| Method | Marigold[29] | GeoWizard[14] | DepthAny. v2[68] | DepthPro[4] | GT(VAE) | Ours |
|---|---|---|---|---|---|---|
| Chamfer Dist. ↓ | 0.17 | 0.16 | 0.18 | 0.14 | 0.12 | **0.08** |

previous models that use convolutional architectures, *e.g.*, denoising U-Net for generative models and DPT for discriminative models, our model is purely transformer-based, with no convolutional layers.

### 4.4 Edge-Aware Point Cloud Evaluation

Our objective is to estimate pixel-perfect depth maps that yield clean point clouds without *flying pixels*, which often occur at object edges due to inaccurate depth predictions in these regions. However, existing evaluation benchmarks and metrics often struggle to reflect *flying pixels* at object edges. For example, benchmarks like NYUv2 or KITTI usually lack edge annotations, while metrics such as AbsRel and $\delta_1$ are dominated by flat regions, making it difficult to assess depth accuracy at edges.

To address these limitations, we evaluate on the official test split of the Hypersim [44] dataset, which provides high-quality ground-truth point clouds and is not used during training. We further propose an edge-aware point cloud metric that quantifies depth accuracy at edges. Specifically, we extract edge masks from ground-truth depth maps using the Canny operator and compute the Chamfer Distance between predicted and ground-truth point clouds near these edges.

Quantitative results in Table 4 show that our method achieves the best performance. Discriminative models like Depth Pro [4] and Depth Anything v2 [68] tend to smooth edges, causing *flying pixels*. Generative models such as Marigold [29] rely on VAE compression, which blurs edges and details, causing artifacts in the reconstructed point clouds. To illustrate this, we encode and decode the ground-truth depth using a VAE (GT(VAE)), without any generative process. Table 4 and Figure 2 show that VAE compression introduces *flying pixels*, leading to a larger Chamfer Distance than ours.

## 5 Conclusion

We presented **Pixel-Perfect Depth**, a monocular depth estimation model that leverages pixel-space diffusion transformers to recover high-quality, flying-pixel-free point clouds. Unlike prior generative depth models that rely on latent-space diffusion with VAEs, our model performs diffusion directly in the pixel space, avoiding *flying pixels* caused by VAE compression. To tackle the complexity and optimization challenges of pixel-space diffusion, we introduce Semantics-Prompted DiT and Cascade DiT Design, which greatly boost performance. Our model significantly outperforms prior models in edge-aware point cloud evaluation.

**Limitations and future work.** This work has two known limitations. First, like most image-based diffusion models, it lacks temporal consistency when applied to video frames, resulting in a little flickering depth across frames. Second, its multi-step diffusion process leads to slower inference compared to discriminative models like Depth Anything v2 [68]. Future works can address these limitations by exploring video depth estimation methods [51, 26, 66, 6, 28] to improve temporal consistency and adopting DiT acceleration strategies [37, 82, 5] such as caching to speed up inference.

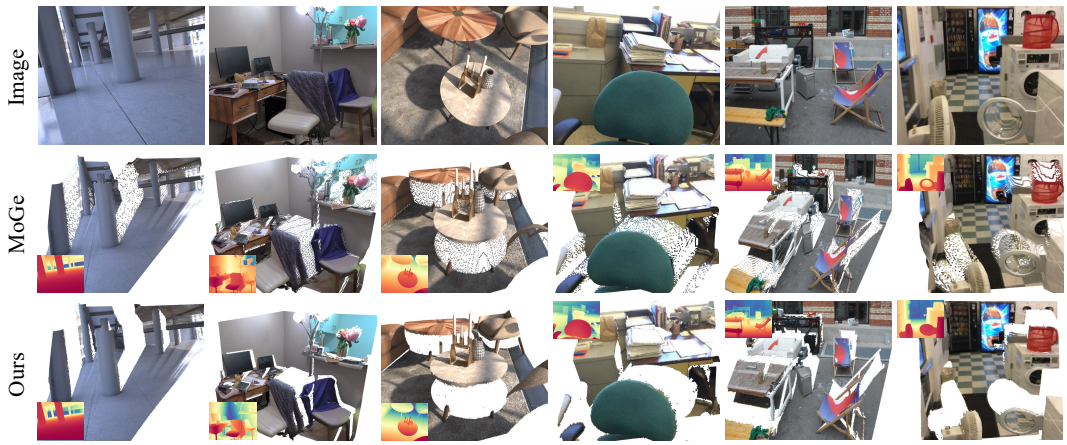


Figure 6: **Qualitative comparisons with MoGe [60].** We provide qualitative comparisons with the concurrent work MoGe [60]. Top: input images are taken from four test sets: Hypersim [44], DIODE [56], ScanNet [8], and ETH3D [50]. Middle: results of MoGe [60]. Bottom: our results. As a discriminative model, MoGe [60], like previous discriminative models [68, 4], also suffers from *flying pixels* at edges and details.

Table 5: **Quantitative comparisons with REPA [76].** Our model significantly outperforms REPA [76]. To ensure a fair comparison, the pretrained vision encoder used in both DiT+REPA and DiT+Ours is kept the same.

| Method | NYUv2 | | KITTI | | ETH3D | | ScanNet | | DIODE | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ | AbsRel↓ | $\delta_1$↑ |
| DiT (baseline) | 22.5 | 72.8 | 27.3 | 63.9 | 12.1 | 87.4 | 25.7 | 65.1 | 23.9 | 76.5 |
| DiT+REPA [76] | 17.6 | 78.0 | 23.4 | 70.6 | 9.1 | 91.2 | 20.1 | 74.3 | 14.6 | 86.9 |
| DiT+Ours | **4.3** | **97.4** | **8.0** | **93.1** | **4.5** | **97.7** | **4.5** | **97.3** | **7.0** | **95.5** |

## 6 More Qualitative Comparisons

We provide qualitative comparisons with the concurrent work MoGe [60], as shown in Figure 6. MoGe [60], as a discriminative model, suffers from *flying pixels* at edges and fine structures, a common issue observed in other discriminative models [68, 4]. Our model produces significantly fewer *flying pixels* compared to MoGe [60].

## 7 Additional Discussion with REPA

We provide an additional discussion on the recent image generation method REPA [76]. REPA [76] aligns intermediate tokens in diffusion models with pretrained vision encoder, significantly improving training efficiency and generation quality for image generation tasks. We compare our method with REPA [76], and the quantitative evaluation results are presented in Table 5. DiT+REPA refers to training the DiT model with REPA's representation alignment, while DiT+Ours denotes training the DiT model using our Semantics-Prompted DiT. For a fair comparison, the pretrained vision encoder used in both DiT+REPA and DiT+Ours is kept the same. Experimental results show that our Semantics-Prompted DiT significantly outperforms REPA [76]. We attribute our model's superiority over REPA to two factors. First, during training, REPA's implicit alignment of DiT tokens with the pretrained vision encoder is suboptimal, making it difficult for DiT to effectively leverage semantic guidance from the pretrained vision encoder. In contrast, our Semantics-Prompted DiT directly integrates semantic cues, resulting in more effective guidance. Second, at inference, REPA cannot leverage the pretrained vision encoder to provide semantic guidance, whereas our method effectively incorporates high-level semantics into the Semantics-Prompted DiT during inference to guide the diffusion process.

Image VAE-d4 (SD2) VAE-d16 (SD3.5) Ours

Figure 7: **Validation of flying pixels in different types of VAEs.** We present further qualitative comparisons showing that increasing the latent dimension in VAEs fails to eliminate *flying pixels*. VAE-d4 (SD2) denotes the reconstruction of ground truth depth maps using the VAE from Stable Diffusion 2, with a latent dimension of 4, which is also used in Marigold. VAE-d16 (SD3.5) uses the VAE from Stable Diffusion 3.5, which has a latent dimension of 16.

## 8 Analysis of Flying Pixels in Different Types of VAEs

To better understand the emergence of *flying pixels* in VAE-based reconstructions, we analyze VAEs with different latent dimensions (*i.e.*, channel) by using them to reconstruct ground truth depth maps. Figure 7 shows that both VAE variants exhibit *flying pixels* at object edges and details, revealing a common weakness of VAE reconstructions in preserving precise geometric structures. VAE-d4 (SD2) denotes the reconstruction of ground truth depth maps using the VAE from Stable Diffusion 2, with a latent dimension of 4, which is also used in Marigold [29]. VAE-d16 (SD3.5) uses the VAE from Stable Diffusion 3.5, which has a latent dimension of 16.

## References


[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[2] Yunpeng Bai and Qixing Huang. Fiffdepth: Feed-forward transformation of diffusion-based generators for detailed depth estimation. *arXiv preprint arXiv:2412.00671*, 2024.

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, 2023.

[4] Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

[5] Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. $\backslash\delta$-dit: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*, 2024.

[6] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025.

[7] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. *arXiv*, 2021.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017.

[9] Yiquan Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In *ECCV*, pages 432–449. Springer, 2024.

[10] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021.

[11] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015.

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014.

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[14] 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 *ECCV*, pages 241–258. Springer, 2025.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012.

[16] 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 *AAAI*, volume 39, pages 3203–3211, 2025.

[17] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, pages 9233–9243, 2023.

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

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020.

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(47):1–33, 2022.

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[23] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *IJCV*, 75:151–172, 2007.

[24] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, pages 13213–13232. PMLR, 2023.

[25] 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. *TPAMI*, 2024.

[26] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.

[27] 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 *ICCV*, pages 21741–21752, 2023.

[28] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. *arXiv preprint arXiv:2411.19189*, 2024.

[29] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024.

[30] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[31] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[32] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *CVPR*, pages 10016–10025, 2024.

[33] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. In *ECCV*, pages 250–267. Springer, 2024.

[34] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. *arXiv preprint arXiv:2412.14015*, 2024.

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

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

[37] Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-cache: Accelerating diffusion transformer via layer caching. *NIPS*, 37:133282–133304, 2024.

[38] 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.

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

[40] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021.

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020.

[44] 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 *ICCV*, 2021.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[46] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 31(5):824–840, 2008.

[47] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *NIPS*, 36:39443–39469, 2023.

[48] Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J Fleet. Zero-shot metric depth with a field-of-view conditioned diffusion model. *arXiv preprint arXiv:2312.13252*, 2023.

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

[50] 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 *CVPR*, pages 3260–3269, 2017.

[51] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024.

[52] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

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

[54] 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.

[55] Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*, 2023.

[56] 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.

[57] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *3DV*, 2019.

[58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025.

[59] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *ICME*, 2021.

[60] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024.

[61] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020.

[62] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018.

[63] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 2020.

[64] 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.

[65] Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. *arXiv preprint arXiv:2504.01016*, 2025.

[66] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv*, 2024.

[67] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024.

[68] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NIPS*, 37:21875–21911, 2024.

[69] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[70] Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *NIPS*, 37:56166–56189, 2024.

[71] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025.

[72] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020.

[73] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020.

[74] 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 *CVPR*, pages 9043–9053, 2023.

[75] 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 *CVPR*, pages 204–213, 2021.

[76] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[77] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *NIPS*, 35:14128–14139, 2022.

[78] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. *arXiv preprint arXiv:2407.17952*, 2024.

[79] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, pages 5729–5739, 2023.

[80] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *ECCV*, pages 1–22. Springer, 2024.

[81] Lianghui Zhu, Zilong Huang, Bencheng Liao, Jun Hao Liew, Hanshu Yan, Jiashi Feng, and Xinggang Wang. Dig: Scalable and efficient diffusion models with gated linear attention. *arXiv preprint arXiv:2405.18428*, 2024.

[82] Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating diffusion transformers with token-wise feature caching. *arXiv preprint arXiv:2410.05317*, 2024.