

Efficient and High-Quality Depth Estimation via Pixel-Space Diffusion with Linear Attention

Bingde Liu^{1,2}, Wu Ran¹, Jinglei Zhang¹, Huanhuan Yuan¹, and Chao Ma^{1,*}

¹ MoE Key Lab of Artificial Intelligence, Institute of AI, Shanghai Jiao Tong University, Shanghai, China

² Zhiyuan College, Shanghai Jiao Tong University, Shanghai, China

{cylarus1024, bonjourlemonde, zhangjinglei168, yuanhuanhuan, chaoma}@sjtu.edu.cn

Code: <https://github.com/VISION-SJTU/Lapis>

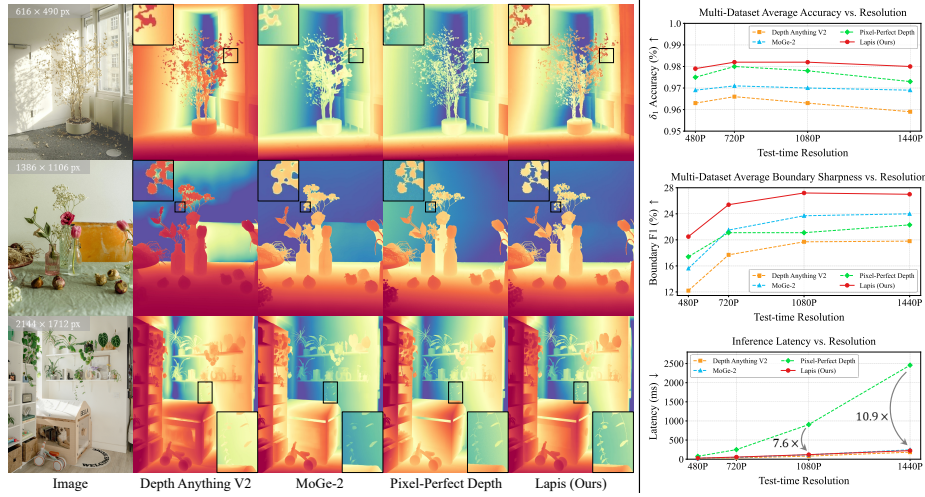


Fig. 1: Lapis: a one-step pixel-space diffusion framework with linear attention for efficient and high-quality monocular depth estimation. Lapis delivers precise geometry and intricate structures across resolutions, consistently outperforming SOTA models in accuracy and boundary sharpness with competitive latency.

Abstract. This work presents **Lapis**, a linear-attention-based pixel-space generative framework that achieves efficient and high-fidelity depth estimation with one-step diffusion. While generative frameworks have significantly advanced monocular depth estimation with superior detail fidelity, the $\mathcal{O}(N^2)$ complexity of standard attention and the multi-step denoising process introduce prohibitive computational costs when scaling them to high-resolution image applications. Although linear attention and one-step prediction are intuitively viable, directly applying them

* Corresponding author.

leads to poor structural consistency, detail loss, and noise. Lapis rectifies these limitations through a coarse-to-fine hierarchy. Specifically, a Patch-level Consistency Module restores structural coherence by integrating semantic and spatial priors. Subsequently, a Pixel-level Refinement Module recovers sharp geometric boundaries via skip-connection-based pixel correspondence. Furthermore, to mitigate sampling noise inherent in one-step diffusion, we leverage the manifold assumption and adopt a direct $\mathbf{x}$ -prediction strategy to target the clean data manifold. Extensive evaluations on multiple benchmarks demonstrate that Lapis consistently achieves state-of-the-art (SOTA) accuracy and boundary sharpness across various resolutions, reducing inference latency by up to $7.6\times$ at 1080P and $10.9\times$ at 1440P resolution compared to previous SOTA generative models.

Keywords: Monocular Depth Estimation · Diffusion Models · Linear Attention

1 Introduction

Monocular depth estimation is a cornerstone of computer vision, powering critical applications in real-world 3D/4D perception and reconstruction [28, 33, 36, 39, 66, 71, 79]. Driven by the demand for fine-grained geometric perception, the field has recently transitioned from discriminative regression [40, 53, 63, 64, 73, 74] to generative frameworks [18, 20, 25, 30, 31, 45, 69, 70]. Compared to their discriminative counterparts, generative models significantly advance fine-grained structural details. However, as applications increasingly demand higher resolutions to capture intricate, pixel-level geometry, as illustrated in Fig. 2, existing generative frameworks inevitably face a fundamental trade-off between fidelity and efficiency. On the one hand, most models [18, 20, 25, 30, 31, 45, 70] inherit latent-space U-Net [56] architectures from pre-trained image generators [55] for efficiency. Such a modeling paradigm is not only restricted by the irreversible information loss of the VAE [32] bottleneck, but also struggles to capture long-range dependencies necessary for global consistency at high resolutions. On the other hand, while recent works [69] employ pixel-space Diffusion Transformers (DiTs) [49] to eliminate the latent bottleneck and maintain long-range coherence, their $\mathcal{O}(N^2)$ complexity becomes computationally prohibitive at high resolutions. Moreover, the iterative multi-step sampling in both paradigms multiplies these overheads. Under these constraints, achieving a trajectory that harmonizes pixel-level precision with inference efficiency remains an open challenge.

In search of an architecture capable of capturing long-range consistency in pixel-space without quadratic complexity or multi-step overhead, an intuitively viable choice is to adopt linearized alternatives [29] for attention mechanisms [62] in DiTs with a one-step sampling strategy. Yet, such naive adaptation incurs significant precision loss compared to vanilla DiTs [49]. Specifically, we identify this degradation as stemming from two distinct scales: a patch-level structural misalignment that disrupts global consistency, and a pixel-level fidelity loss that

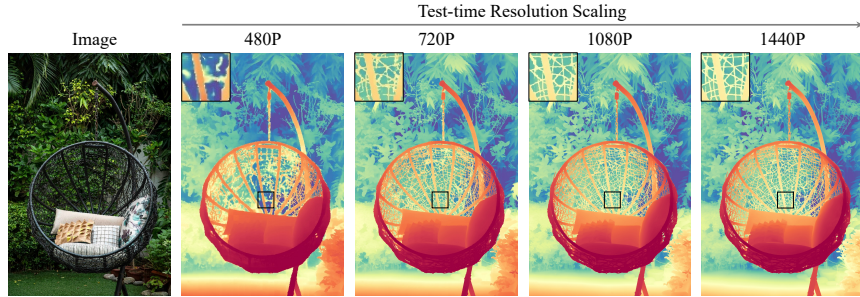


Fig. 2: Visualization of test-time resolution scaling using our model. A depth estimation model produces more fine-grained details as input resolution increases.

hampers fine-grained geometric recovery. Beyond these architectural weaknesses, the widely-adopted $\mathbf{v}$ -prediction trajectories [1, 41, 42] exhibit severe instability in pixel-space when restricted to a single forward pass, yielding noise-corrupted results without iterative refinement.

To address these multifaceted challenges, we propose three core innovations. First, we introduce the Patch-level Consistency Module (PCM) to rectify structural misalignment. The PCM leverages pre-trained semantic priors [48] to enforce global coherence while incorporating local convolutional biases to ensure seamless structural continuity across adjacent patches. Second, to mitigate pixel-level fidelity loss, we develop the Pixel-level Refinement Module (PRM), which employs skip-connection-based pixel correspondence to recover sharp boundaries and intricate structures. Third, we stabilize one-step inference by adopting direct pixel-space $\mathbf{x}$ -prediction over the volatile $\mathbf{v}$ -prediction. Grounded in the manifold assumption [37], which posits that natural data resides on a low-dimensional manifold, whereas noise-corrupted quantities do not, this strategy targets the clean data manifold directly, effectively suppressing the sampling instability inherent in one-step generation. Building on these components, we present **Lapis**, the first **linear-attention-based pixel-space** generative framework to enable consistent and high-fidelity depth reconstruction in a single forward pass while maintaining $\mathcal{O}(N)$ computational scaling in the diffusion process.

Experiments across 12 benchmarks demonstrate that Lapis achieves SOTA accuracy and boundary sharpness, consistently surpassing leading discriminative and generative baselines across varying resolutions with competitive efficiency, as shown in Fig. 1. For instance, at 1080P resolution, Lapis maintains a 5.1 average AbsRel (Tab. 2), a 10% relative improvement over the strongest competitor, while requiring only 13% of its latency (Tab. 4). These results underscore Lapis as an efficient and high-quality solution for high-resolution 3D perception.

Our main contributions are summarized as follows:

- We present Lapis, the first linearized pixel-space diffusion model for monocular depth estimation. It enables linear-time depth diffusion within a single forward pass, significantly outperforming generative alternatives in efficiency.

- We introduce a coarse-to-fine strategy to rectify the structural and detail degradation of linear attention. This includes a Patch-level Consistency Module that restores global structural coherence by integrating semantic and spatial priors, and a Pixel-level Refinement Module that recovers fine-grained geometric details via skip-connection-based pixel correspondence.
- We leverage the manifold assumption to address the unstable one-step pixel-space sampling via a direct $\mathbf{x}$ -prediction strategy. This approach effectively suppresses sampling noise and produces clean depth maps.
- Extensive evaluations show that Lapis achieves SOTA accuracy and boundary sharpness across multiple benchmarks while maintaining competitive latency, offering a holistic solution for high-resolution 3D perception.

2 Related Work

2.1 Generative Monocular Depth Estimation

As a cornerstone for 3D vision, monocular depth estimation (MDE) has evolved from early CNN-based regression on specific domains [13, 17, 35] toward open-domain expansion leveraging diverse datasets [53, 75], and finally to the architectural transformation afforded by large-scale Vision Transformers (ViTs) [2, 11, 12, 52]. Recent SOTA discriminative models, such as [40, 63, 64, 73, 74], utilize powerful pre-trained encoders [48] and massive training data to achieve remarkable zero-shot robustness. Despite their impressive accuracy and efficiency, these regression-based paradigms often produce over-smoothed predictions that discard fine-grained geometric details. To overcome this challenge, generative modeling works treat depth estimation as a conditional synthesis task. One prominent branch [18, 20, 24, 25, 30, 31, 45, 70] fine-tunes pre-trained latent-space image synthesis models [34, 55] to recover intricate structures during the diffusion process. However, these models remain inadequate for high-resolution depth estimation where high-fidelity geometric details are paramount. Their U-Net backbones [56] struggle with long-range consistency, while VAE latent downsampling [32] irreversibly discards fine-grained geometric details. To circumvent these constraints, recent research [69] has shifted toward pixel-space diffusion Transformers (DiTs) [49] to preserve complete spatial information for superior detail restoration. However, pixel-space diffusion predominantly relies on standard attention [62] with $\mathcal{O}(N^2)$ complexity, rendering it computationally prohibitive for high-resolution applications. Furthermore, iterative diffusion sampling compounds the latency overhead, posing significant challenges for efficient high-resolution deployment. By contrast, Lapis bridges the gap between geometric precision and inference efficiency via a linearized pixel-space DiT that enables high-fidelity, single-pass reconstruction. Unlike naive adaptations, our dual-level refinement and $\mathbf{x}$ -prediction strategy ensure that these efficiency gains enhance, rather than compromise, the restoration of intricate geometric structures.

2.2 Sampling Strategies for Generative Depth Estimation

Early generative MDE frameworks [18, 30] leverage pre-trained latent diffusion models (LDMs) [55] with ϵ -prediction [27]. Consequently, they inherit the heavy iterative sampling schedule (10–50 steps), leading to substantial computational latency. To accelerate inference, recent research has branched into two primary directions. On one hand, several approaches [20, 31, 69] employ flow-matching with $\mathbf{v}$ -prediction for straighter sampling trajectories, but still require at least 4 steps to sufficiently suppress noise artifacts. On the other hand, existing one-step $\mathbf{x}$ -prediction models [25, 45, 70] remain fundamentally constrained by the LDM paradigm. Specifically, the VAE-based latent space [32] is explicitly regularized toward an isotropic Gaussian distribution, inherently compromising the low-dimensional manifold of natural geometry [37]. As a result, performing one-step $\mathbf{x}$ -prediction within such a distorted space leads to poor results. In contrast to existing $\mathbf{v}$ -prediction or latent-space $\mathbf{x}$ -prediction models, Lapis performs $\mathbf{x}$ -prediction directly in pixel-space, thereby preserving the original geometric manifold and achieving SOTA accuracy within a single forward pass.

2.3 Linear Attention

To maintain global context while bypassing the quadratic scaling of standard self-attention [62], linear attention [29] utilizes kernel approximations to eliminate the Softmax operation, achieving $\mathcal{O}(N)$ complexity. Despite its widespread adoption in natural language processing [50, 62, 68] for long-sequence modeling, it remains disproportionately sparse in visual tasks. Existing explorations [6, 15, 16, 22, 23, 44] primarily refine the linear attention formulation for categorical tasks with limited data, such as image classification, object detection, instance segmentation, and semantic segmentation, which rely on coarse-grained semantic features and are inherently tolerant of feature fluctuations. In contrast, MDE is a dense regression task requiring high-precision numerical fidelity, making it far more sensitive to the representational decay often associated with linear approximations. More recently, generative models [7, 67] have adopted linear attention for visual synthesis tasks. But in these contexts, visual plausibility, *e.g.*, FID [26] and FVD [60], often takes precedence over metric accuracy. Lapis stands as the first application of linear attention for MDE. We demonstrate that linearized models can achieve high-quality 3D dense perception efficiently without compromising accuracy.

3 Method

Lapis is designed as a high-efficiency, one-step generative framework for relative monocular depth estimation. As illustrated in Fig. 3, the model operates directly in pixel-space, taking a concatenated noisy depth map and a conditioning image as input to predict a clean depth map without the need for latent autoencoders. To circumvent the quadratic complexity inherent in standard DiTs [49], we replace all standard multi-head self-attention (MSA) layers with linearized

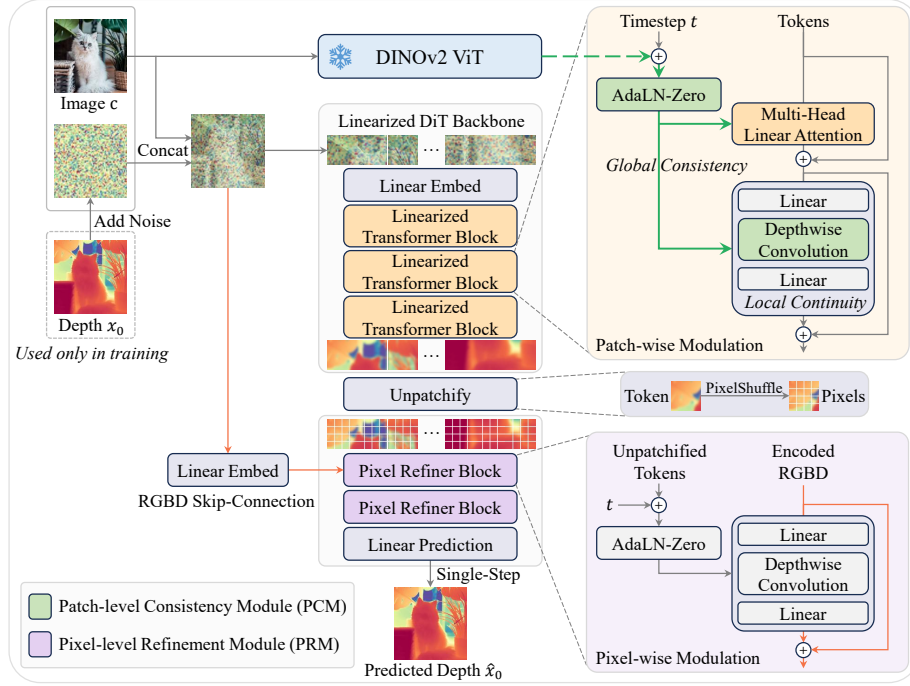


Fig. 3: Architecture overview. Lapis performs pixel-space $\mathbf{x}$ -prediction to enable high-quality one-step denoising. It first leverages a linearized DiT backbone to establish a coherent global depth layout, integrating semantic and spatial priors via our Patch-level Consistency Module. Subsequently, the Pixel-level Refinement Module restores fine-grained geometry via skip-connection-based pixel correspondence. The overall framework enables efficient and accurate depth estimation with pixel-level details.

attention kernels [29]. To bridge the representational fidelity gap often associated with linear approximations, we augment the transformer blocks with two synergistic components: the Patch-level Consistency Module (PCM) and the Pixel-level Refinement Module (PRM). While the PCM enforces spatial alignment across the token grid at the patch-level, the PRM restores high-frequency precision at the pixel-level, collectively harmonizing linear-time computational efficiency with high-fidelity depth estimation.

3.1 One-step Pixel-Space Diffusion

To eliminate the multi-step sampling latency inherent in iterative diffusion models, we adopt a native one-step framework. Guided by the low-dimensional manifold assumption [37], which posits that natural geometric data resides on a structured and continuous manifold, our model directly predicts the clean depth map in pixel-space. Specifically, we formulate the diffusion process following the flow-matching paradigm [1, 41, 42]. Given a clean depth sample $\mathbf{x}_0$, we sample

Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1)$, time step $t \sim [\tau, 1]$ with a small constant τ (*e.g.*, 0.05), and apply the linear schedule as:

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

Our model $x_\theta(\cdot)$ then predicts $\mathbf{x}_0$ with $\mathbf{x}_t$, t and corresponding image $\mathbf{c}$ as:

$$\hat{\mathbf{x}}_0 = x_\theta(\mathbf{x}_t, t, \mathbf{c}). \quad (2)$$

Training losses. Our model is optimized via a joint objective that combines velocity-based flow supervision and multi-scale geometric regularization. The velocity loss is defined in $\mathbf{v}$ -space following flow-matching frameworks [1, 41, 42], where the target velocity is given by:

$$\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt} = \varepsilon - \mathbf{x}_0. \quad (3)$$

We supervise the network by minimizing the Mean Squared Error (MSE) between the predicted velocity $\hat{\mathbf{v}}_t = (\mathbf{x}_t - \hat{\mathbf{x}}_0) / t$ and its ground truth:

$$\mathcal{L}_{\text{velocity}} = \mathbb{E}_{\mathbf{x}_0, \varepsilon, t} [\|\hat{\mathbf{v}}_t - \mathbf{v}_t\|_{l_2}^2]. \quad (4)$$

We further incorporate a multi-scale gradient matching loss directly in $\mathbf{x}$ -space following [53, 74] to enhance the sharpness of depth maps:

$$\mathcal{L}_{\text{grad}} = \mathbb{E}_{\mathbf{x}_0, \varepsilon, t} \left[\sum_{s \in S} w_s \|\nabla_s \mathbf{x}_0 - \nabla_s \hat{\mathbf{x}}_0\|_{l_1} \right], \quad (5)$$

where ∇_s denotes the spatial gradient operator at different resolution scales S and w_s is the scale-specific weight. The final loss is a weighted combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{velocity}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}, \quad (6)$$

where λ_{grad} is a hyperparameter balancing accuracy and boundary sharpness.

3.2 Linearized Transformer Architecture

Linear attention formulation. In standard self-attention [62] (for single head case), given query, key, and value matrices $Q, K, V \in \mathbb{R}^{N \times d}$, where N is the number of tokens and d is the feature dimension, the attention output is computed as:

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V, \quad (7)$$

which incurs a quadratic time complexity of $\mathcal{O}(N^2 d)$. This cost becomes prohibitive for high-resolution input where N is large (*e.g.*, 1024×768 pixels equal to 3072 tokens under patch size 16). Instead, linear attention [29] approximates the softmax similarity with a non-negative kernel function $\phi(\cdot)$. By leveraging

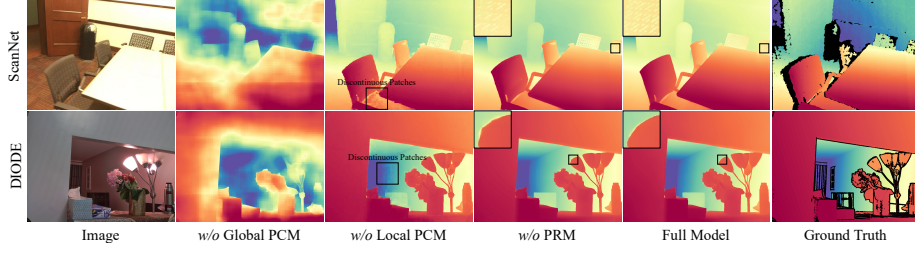


Fig. 4: Qualitative ablation of architectural components. We visually validate the necessity of each model component in Lapis. In comparison with the full model, removing the global PCM leads to severe global inconsistencies, omitting the local PCM introduces discontinuous patch artifacts, and the absence of the PRM results in a significant loss of sharp details.

the associativity of matrix multiplication, the computation can be reformulated as:

$$\text{Attn}_{\text{Linear}}(Q_i, K, V) = \frac{\phi(Q_i) \sum_{j=1}^N \phi(K_j)^T V_j}{\phi(Q_i) \sum_{j=1}^N \phi(K_j)^T} \text{ for each } i \in [1, N]. \quad (8)$$

With pre-computed global context terms $\sum_{j=1}^N \phi(K_j)^T V_j \in \mathbb{R}^{d \times d}$ and normalization factor $\sum_{j=1}^N \phi(K_j)^T \in \mathbb{R}^{d \times 1}$, Eq. (8) avoids the explicit construction of the $N \times N$ attention matrix and reduces the time complexity to $\mathcal{O}(Nd^2)$. We adopt the standard DiT [49] as our generative backbone for its structural simplicity and strong scaling properties. By substituting all MSA layers inside it with Eq. (8), we effectively reduce its complexity to $\mathcal{O}(N)$.

The fidelity gap in linearization. Despite its efficiency, we identify a dual-level fidelity gap when applying linear attention to dense depth estimation. First, the low-rank nature of linear kernels [15, 22] often leads to patch-level structural misalignment, where geometric coherence is weakened. Second, the iterative linear averaging inherent in these layers results in a significant loss of pixel-level detail. To resolve these issues, we introduce the Patch-level Consistency Module (PCM) and the Pixel-Level Refinement Module (PRM), as detailed below.

3.3 Patch-level Consistency Module

Global relational consistency. The core limitation of linearized attention lies in the sacrifice of the exponential softmax kernel, which provides a sharp attentional focalization in long-range modeling [15, 22]. We observe that forcing linear attention to simultaneously reconstruct global scene layouts and local geometric details is highly impractical. As illustrated in the second column of Fig. 4, the model struggles to preserve global geometry layout and subsequently blurs local

details. To alleviate this, we supervise macro-level global geometry with semantic priors, thereby exempting the linearized backbone from the complexities of global spatial coordination. Specifically, let $\mathbf{s} \in \mathbb{R}^{N \times D}$ denote the DiT tokens, where N and D represent the sequence length and embedding dimension, respectively. We align $\mathbf{s}$ with semantic descriptors $\mathbf{z} \in \mathbb{R}^{N \times D'}$, which are extracted from the input image $\mathbf{c}$ via a frozen DINOv2 [48] encoder. Here, D' corresponds to the feature dimension of the DINOv2 backbone. To align and refine high-level semantics, we apply a gating mechanism to $\mathbf{z}$ as follows:

$$\hat{\mathbf{z}}_i = \mathbf{W}_o (\mathbf{W}_v \bar{\mathbf{z}}_i \odot \sigma(\mathbf{W}_g \bar{\mathbf{z}}_i)), \quad (9)$$

where $\{\bar{\mathbf{z}}_i\}_{i=1}^N$ are tokens normalized from $\mathbf{z}$, $\sigma(\cdot)$ denotes the sigmoid function, $\odot$ represents the element-wise product, and $\mathbf{W}_v, \mathbf{W}_g, \mathbf{W}_o$ are learnable channel-wise projection matrices. The refined tokens $\hat{\mathbf{z}}$ are subsequently fused with the timestep embedding $\mathbf{t}$ to condition the feature flow in DiT blocks using an extended AdaLN-Zero [49] mechanism:

$$\text{AdaLN-Zero}(f, \mathbf{s}, \mathbf{t}, \hat{\mathbf{z}}) = \boldsymbol{\alpha} \odot f(\text{LayerNorm}(\mathbf{s}) \odot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta}), \quad (10)$$

where the modulation parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ are regressed from the joint embedding $\hat{\mathbf{z}} + \mathbf{t}$ via a multi-layer perceptron (MLP). We apply such a mechanism to both the linear attention layers and the feed-forward network (FFN) layers [62], which are represented by $f(\cdot)$. In this way, we effectively anchor the feature flow within a consistent global geometric layout without introducing additional parameters to the modulation process. These semantically-modulated blocks are interleaved in the linearized DiT backbone for computational efficiency.

Local spatial continuity. Despite the global structural rectification, linear attention still struggles to maintain local patch-wise continuity. As illustrated in the third column of Fig. 4, this deficiency manifests as localized depth discontinuities and perceptible grid artifacts. We address this structural fragmentation by incorporating a local inductive bias into the FFN layer via a residual 3×3 depth-wise convolution (DWC), a structure we term ConvFFN. Specifically, we formulate the feature transition in the ConvFFN as follows:

$$\mathbf{s}' = \text{SiLU}(\text{Linear}(\mathbf{s})), \quad \mathbf{s}_{\text{out}} = \text{Linear}(\text{SiLU}(\text{DWC}(\mathbf{s}')) + \mathbf{s}'), \quad (11)$$

where $\text{SiLU}(\cdot)$ denotes the SiLU activation function [14]. This local inductive bias enforces spatial correlation across adjacent tokens, effectively suppressing the aforementioned grid artifacts to ensure smooth geometric transitions.

3.4 Pixel-Level Refinement Module

Although the PCM ensures token-level coherence, mapping discretized tokens back to the continuous pixel-space still entails a significant fidelity gap. As previously analyzed, the iterative averaging effect in deep linear-attention-based

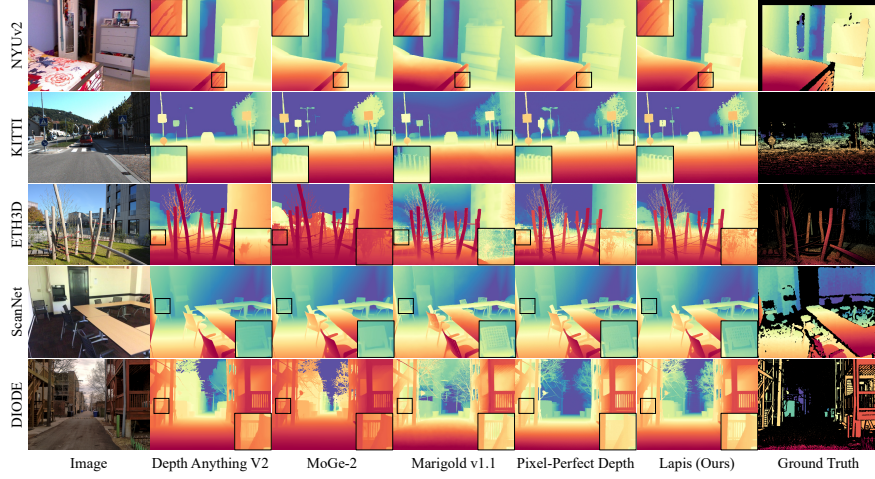


Fig. 5: Qualitative comparison of monocular depth estimation models. Our model excels at capturing thin structures while preserving the accurate global geometry.

architectures leads to cumulative signal dilution, where detailed structures progressively fade across layers. Such degradation is evident in the fourth column of Fig. 4, where a backbone without PRM fails to resolve sharp geometric boundaries. To counteract this, the Pixel-level Refinement Module (PRM) re-integrates pixel-level input cues via a long-range skip connection to effectively recover sharp geometric boundaries. Specifically, the original input y is concatenated with a learnable patch-wise positional encoding and transformed by an MLP to yield the PRM input $y' \in \mathbb{R}^{D_{\text{pixel}} \times H \times W}$, where D_{pixel} denotes the embedding dimension of the PRM. To bridge the gap between patch-level representation and pixel-level representation, latent tokens $\mathbf{s} \in \mathbb{R}^{N \times D}$ from the final DiT block are first reshaped into a spatial grid $\mathbb{R}^{D \times h \times w}$ and then projected and upsampled via PixelShuffle [59] to match the target pixel-level resolution, following the transformation chain as follows:

$$\mathbf{s} \xrightarrow{\text{MLP}} \mathbf{s}' \in \mathbb{R}^{(D_{\text{pixel}} \times p^2) \times h \times w} \xrightarrow{\text{PixelShuffle}} \mathbf{s}_{\text{pixel}} \in \mathbb{R}^{D_{\text{pixel}} \times H \times W}, \quad (12)$$

where p denotes the patch size, and $H = h \cdot p$ and $W = w \cdot p$ represent the height and width of the original input image. The PRM is architected as a sequence of Pixel Refiner blocks, each using an AdaLN-Zero-modulated ConvFFN layer to decode the pixel-level latents into depth values with pixel-level cues:

$$\text{AdaLN-Zero}(g, \mathbf{y}', \mathbf{s}_{\text{pixel}}) = \alpha \odot g(\text{LayerNorm}(\mathbf{y}') \odot (1 + \gamma) + \beta), \quad (13)$$

where $g(\cdot)$ denotes the ConvFFN layer, and the scaling and shifting parameters (γ, β, α) are regressed from the features $\mathbf{s}_{\text{pixel}}$ via an MLP. This design effectively recovers pixel-level details from the raw input pixels during the decoding stage. Finally, a standard linear prediction head in DiT [49] is used to project the refined pixel-level features into a high-fidelity depth map $\mathbf{d} \in \mathbb{R}^{H \times W}$.

Table 1: Zero-shot relative depth estimation on native resolution. Lapis surpasses both established discriminative models and recent generative baselines.

Type	Method	Training Data	NYUv2		KITTI		ETH3D		ScanNet		DIODE		Average	
			AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$
Discriminative	MiDaS [53]	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	–	–	–	–
	LeRes [76]	354K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	–	–	–	–
	OmniData [12]	12.2M	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	–	–	–	–
	DPT [52]	1.4M	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	–	–	–	–
	HDN [78]	300K	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	–	–	–	–
	Lotus-D [25]	59K	5.1	97.2	8.1	93.1	6.1	97.0	5.5	96.5	–	–	–	–
	Depth Anything [73] *	62.6M	4.5	97.9	8.7	93.6	6.6	98.1	4.4	98.0	11.1	94.8	7.0	96.5
	Depth Anything V2 [74] *	62.6M	4.5	97.9	7.9	93.6	7.4	97.7	4.2	97.8	10.0	94.9	6.8	96.4
	Depth Anything 3 [40] *	1.2M (Scenes)	4.1	97.4	7.0	95.6	4.2	97.7	4.1	97.6	7.1	94.6	5.3	96.6
	MoGe [63] *	9M	3.7	97.9	5.9	97.3	3.5	98.4	3.6	98.3	7.7	94.8	4.9	97.3
	MoGe-2 [64] *	8.9M	3.5	98.0	5.9	97.1	4.5	96.7	3.4	98.3	6.7	95.5	4.8	97.1
Generative	GeoWizard [18]	280K	5.6	96.3	14.4	82.0	6.6	95.8	6.4	95.0	–	–	–	–
	GenPercept [70]	74K	5.6	96.0	13.0	84.2	7.0	95.6	6.2	96.1	–	–	–	–
	Diffusion-E2E-FT [45]	74K	5.4	96.5	9.6	92.1	6.4	95.9	5.8	96.5	–	–	–	–
	Lotus-G [25]	59K	5.4	96.8	8.5	92.2	5.9	97.0	5.9	95.7	–	–	–	–
	Marigold v1.0 [30]	74K	5.5	96.4	9.9	91.6	6.5	95.9	6.4	95.2	11.1*	89.5*	7.9	93.7
	Marigold v1.1 [31]	74K	5.5	96.4	10.5	90.2	6.9	95.7	5.8	96.3	10.1*	91.0*	7.7	93.9
	Lotus-2 [24]	59K	4.1	97.6	6.7	94.5	4.6	98.1	4.2	97.6	8.4*	94.7*	5.6	96.5
	Pixel-Perfect Depth [69] *	125K	3.7	98.1	5.4	97.1	3.1	99.2	3.8	98.2	4.9	97.5	4.2	98.0
	Lapis (Ours) *	122K	3.8	98.2	5.1	97.7	3.2	99.2	3.8	98.2	4.8	97.9	4.1	98.2

4 Experiments

4.1 Implementation Details

Depth normalization. Given raw depth map $\mathbf{d}$, our model follows [69] to predict the normalized log-scale depth $\mathbf{x}_0$ as:

$$\mathbf{x}_0 = \frac{\hat{\mathbf{d}} - \hat{\mathbf{d}}_2}{\hat{\mathbf{d}}_{98} - \hat{\mathbf{d}}_2} - 0.5, \quad (14)$$

where $\hat{\mathbf{d}} = \log(\mathbf{d} + \epsilon)$ with the offset ϵ fixed at 0.01 to ensure numerical stability. $\hat{\mathbf{d}}_2$ and $\hat{\mathbf{d}}_{98}$ represent the 2nd and 98th percentiles computed independently for each depth map. We further rescale $\mathbf{x}_0$ to ensure unit variance, facilitating a more stable data distribution.

Training datasets. We adopt a total of about 122K training samples from five synthetic datasets: Hypersim [54] (53.8K), TartanAir [65] (28.2K), VKITTI2 [5] (21.3K), UrbanSyn [21] (7.5K), and MatrixCity [38] (11.6K) to train our model. All training datasets are directly resized to 1024×768 resolution and simply concatenated together without weighted sampling.

Model architecture details. Our model comprises 520M parameters, maintaining consistency with the DiT-Large [49] configuration in hidden dimensions and the number of blocks while transitioning to a fully linearized backbone. The kernel feature map $\phi(\cdot)$ in Eq. (8) is implemented as $\phi(x) = \text{elu}(x) + 1$ [8]. Following [69], we employ the DINOv2-Large [48] encoder fine-tuned by MoGe-2 [64] as the frozen semantic encoder to provide robust structural priors. For the PRM, we use $M = 3$ blocks and an embedding dimension of $D_{\text{pixel}} = 32$.

Table 2: Zero-shot relative depth estimation on 1080P and 1440P. Notably, Lapis demonstrates the exceptional resolution-robustness of our dual-level PCM and PRM modeling and largely outperforms existing baselines in accuracy metrics.

Res.	Method	Middlebury 1920 × 1080		Booster 4112 × 3008		Cityscapes 2048 × 1024		DrivingStereo 1762 × 800		ETH3D 6048 × 4032		Average --	
		AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
1080P	Depth Anything V2 [74]	6.6	96.8	3.3	99.7	10.1	89.4	6.4	96.4	5.7	98.1	6.4	96.1
	Depth Anything 3 [40]	7.6	93.4	2.4	99.8	18.9	75.3	5.7	96.8	5.1	97.5	7.9	92.6
	MoGe-2 [64]	7.5	93.8	2.0	99.9	11.8	89.2	5.8	96.5	4.5	96.9	6.3	95.3
	Marigold v1.1 [31]	13.8	81.4	8.6	94.3	24.0	60.0	7.1	95.9	11.5	87.2	13.0	83.8
	Lotus-2 [24]	14.5	76.2	8.0	91.3	31.8	58.1	6.1	96.6	6.6	97.5	13.4	83.9
	Pixel-Perfect Depth [69]	6.3	96.5	2.9	99.8	9.3	93.5	6.1	96.8	4.1	99.1	5.7	97.1
	Lapis (Ours)	5.6	96.8	2.1	99.9	8.7	92.5	5.6	97.0	3.3	99.3	5.1	97.1
1440P	Depth Anything V2 [74]	6.9	96.5	3.4	99.6	10.6	88.7	6.4	96.5	5.7	97.9	6.6	95.8
	Depth Anything 3 [40]	7.9	93.4	2.9	99.9	20.4	71.4	5.6	96.9	6.2	96.6	8.6	91.6
	MoGe-2 [64]	7.7	93.5	2.0	99.9	12.0	88.9	5.8	96.5	4.6	96.8	6.4	95.1
	Marigold v1.1 [31]	17.1	73.9	10.4	91.2	40.8	42.1	7.3	95.3	15.8	78.6	18.3	76.2
	Lotus-2 [24]	16.6	71.1	10.0	88.9	65.1	47.5	6.2	96.6	8.4	96.3	21.3	80.1
	Pixel-Perfect Depth [69]	7.2	95.3	3.9	99.8	10.0	92.9	6.5	96.7	5.1	98.5	6.5	96.6
	Lapis (Ours)	5.8	96.4	2.1	99.9	9.1	91.8	5.7	97.0	3.4	99.2	5.2	96.9

Training details. We optimize the model using the total loss defined in Eq. (6), with the gradient loss weight set to $\lambda_{\text{grad}} = 1.5$. Training is conducted using the AdamW optimizer [43] with a constant learning rate of 1×10^{-4} and no weight decay. The model is trained for 400K steps on 4 NVIDIA RTX Pro 6000 Blackwell GPUs, with a per-GPU batch size of 8. An Exponential Moving Average (EMA) with a decay rate of 0.9999 is applied to the model weights during training, and all evaluations are performed using the EMA weights. Both training and inference are conducted in bfloat16 mixed precision.

4.2 Zero-shot Relative Depth Evaluation

Native resolution. We evaluate zero-shot relative depth estimation on five mainstream unseen real-world datasets: NYUv2 [47] (indoor), KITTI [19] (outdoor), ETH3D [58] (various), ScanNet [10] (indoor), and DIODE [61] (various). We adopt two widely accepted metrics: Absolute Relative Error (AbsRel) and δ_1 accuracy.³ Each model is evaluated at its training resolution (usually lower than 720P) to assess its peak accuracy. As summarized in Tab. 1, Lapis achieves the best performance with an averaged AbsRel of 4.1% and a δ_1 accuracy of 98.2%. Besides, as demonstrated in Fig. 5, our model excels in resolving thin structures and intricate details while maintaining a coherent global geometric layout.

Test-time resolution scaling at 1080P and 1440P. Benefiting from linearized attention and one-step diffusion, Lapis enables efficient scaling to high resolutions for enhanced geometric details. To validate that such a capability does not come at the cost of degraded global accuracy, we evaluate zero-shot relative depth at 1080P and 1440P test-time resolution on five datasets with high-resolution annotations: Middlebury [57] (indoor), Booster [51, 77] (indoor),

³ *indicates results re-evaluated using the evaluation protocol of Marigold [30, 31].

Table 3: Boundary sharpness across different resolutions. We report the F1-score ($\uparrow$) for depth boundary accuracy. Lapis consistently outperforms all existing methods across various resolutions, while achieving near-monotonic improvements in boundary sharpness as the test-time resolution scales up.

Method	Hypersim F1 $\uparrow$				Sintel F1 $\uparrow$				Spring F1 $\uparrow$				Average F1 $\uparrow$			
	480P	720P	1080P	1440P	480P	720P	1080P	1440P	480P	720P	1080P	1440P	480P	720P	1080P	1440P
Depth Anything V2 [74]	13.1	20.9	24.5	24.1	17.3	23.0	23.7	23.6	6.1	9.2	11.0	11.7	12.2	17.7	19.7	19.8
Depth Anything 3 [40]	12.5	19.3	23.1	24.4	15.8	20.9	23.7	24.6	5.8	8.7	10.2	10.3	11.4	16.3	19.0	19.8
MoGe-2 [64]	19.5	28.4	31.7	32.1	21.4	27.8	29.2	28.9	5.8	8.4	10.3	10.9	15.6	21.5	23.7	24.0
Marigold v1.1 [31]	20.9	22.2	8.9	7.7	17.5	18.8	6.1	5.6	5.0	5.6	2.9	2.6	14.5	15.6	6.0	5.3
Lotus-2 [24]	17.1	21.9	22.6	22.0	14.5	18.3	19.2	18.5	4.7	7.5	9.3	9.4	12.1	15.9	17.0	16.6
Pixel-Perfect Depth [69]	26.5	31.2	32.7	30.6	19.1	21.5	21.0	20.8	6.5	10.6	12.0	13.2	17.4	21.1	21.1	22.3
Lapis (Ours)	27.4	32.2	34.8	33.6	26.2	30.9	31.4	31.7	8.0	13.0	15.5	15.8	20.5	25.4	27.2	27.0

Table 4: Runtime statistics across different resolutions. We report latency (ms), throughput (Hz), and peak GPU memory (GB). Lapis achieves throughput competitive with leading discriminative models while consistently outperforming existing generative frameworks by a significant margin.

Method	480P			720P			1080P			1440P		
	Lat $\downarrow$	TP $\uparrow$	Mem $\downarrow$	Lat $\downarrow$	TP $\uparrow$	Mem $\downarrow$	Lat $\downarrow$	TP $\uparrow$	Mem $\downarrow$	Lat $\downarrow$	TP $\uparrow$	Mem $\downarrow$
Depth Anything V2 [74]	18.6	53.7	1.3	33.9	29.5	1.5	80.6	12.4	1.7	183.5	5.5	2.1
Depth Anything 3 [40]	14.9	67.0	1.3	58.1	17.2	1.3	141.2	7.1	1.3	207.6	4.8	1.4
MoGe-2 [64]	24.1	41.5	1.3	49.3	20.3	1.5	115.3	8.7	1.7	245.5	4.1	2.1
Marigold v1.1 [31]	119.1	8.4	6.0	328.1	3.0	8.0	758.1	1.3	11.7	1479.0	0.7	16.9
Lotus-2 [24]	1258.6	0.8	37.5	3095.3	0.3	39.5	6952.7	0.1	43.3	13747.0	0.1	48.5
Pixel-Perfect Depth [69]	75.3	13.3	3.2	250.2	4.0	3.3	905.5	1.1	3.5	2457.0	0.4	3.8
Lapis (Ours)	27.5	36.4	3.3	55.4	18.0	3.4	119.0	8.4	3.6	224.7	4.5	3.9

Cityscapes [9] (outdoor), DrivingStereo [72] (outdoor), and ETH3D [58] (various). As reported in Tab. 2, Lapis maintains remarkable accuracy metrics, outperforming current SOTA models by approximately 10% and 19% in AbsRel, respectively. This demonstrates that our architecture not only scales efficiently but also preserves global structural integrity under significant resolution shifts.

4.3 Boundary Sharpness Evaluation

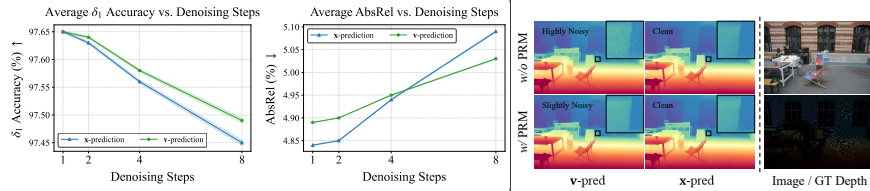
To quantify the details of the estimated depth maps, we evaluate the boundary sharpness F1-score proposed by [3] across three high-resolution synthetic benchmarks: Hypersim [54] (indoor), Sintel [4] (various), and Spring [46] (outdoor). Unlike global error metrics, this evaluation specifically measures the model’s ability to resolve sharp occlusion boundaries and fine-grained structural transitions. As demonstrated in Tab. 3, Lapis consistently outperforms both discriminative and generative baselines in boundary sharpness across all test resolutions, validating the efficacy of our PRM in recovering high-frequency spatial details.

4.4 Runtime Analysis

We evaluate the runtime performance of Lapis against competitive methods on a single NVIDIA RTX Pro 6000 Blackwell GPU with 1 sample per batch. As

Table 5: Ablation study of Lapis components. We evaluate the contribution of each module across five datasets.

Components	NYUv2		KITTI		ETH3D		ScanNet		DIODE		Average	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Full Model	4.3	98.0	5.7	96.9	4.0	98.6	4.5	97.7	5.6	97.0	4.8	97.7
— w/o Global PCM	19.6	68.2	20.2	69.9	19.1	71.4	19.1	69.1	21.2	71.9	19.8	70.1
— w/o Local PCM	4.5	97.7	6.4	96.5	4.3	98.5	4.5	97.7	5.6	97.0	5.1	97.5
— w/o PRM	4.5	97.8	5.7	96.9	4.0	98.6	4.8	97.6	5.5	97.1	4.9	97.6

**Fig. 6: Quantitative comparison under various denoising steps and qualitative comparison for the impact of the PRM on one-step generation.** A one-step x-prediction strategy is optimal in both efficiency and geometric fidelity.

summarized in Tab. 4, Lapis significantly outperforms all generative baselines in terms of inference latency. By leveraging a one-step diffusion formulation and linearized attention, our model achieves a processing speed that is highly competitive with purely regression-based architectures.

4.5 Ablation Study

To facilitate efficient analysis, we follow [40, 64] to adopt a lightweight configuration based on the DiT-Base [49] backbone. For the PRM, we reduce the complexity to $M = 2$ blocks with an embedding dimension of $D_{\text{pixel}} = 16$.

Effectiveness of components. We conduct a component-wise ablation to validate the synergy of the PCM and the PRM, as reported in Tab. 5 and Fig. 4. First, removing Global PCM causes a sharp accuracy drop, confirming the necessity of semantic supervision for global coherence. Second, Local PCM is critical for inter-patch continuity. Its absence increases AbsRel ($4.8 \rightarrow 5.1$) and introduces grid artifacts. Finally, while PRM has a marginal impact on accuracy metrics, it is essential for restoring high-frequency details otherwise diluted in the deep linearized backbone to ensure detailed geometry.

Prediction targets and diffusion steps. We compare x-prediction with v-prediction across varying sampling steps (1, 2, 4, 8). As shown in Fig. 6 (left), both strategies peak at one-step sampling, with x-prediction slightly outperforming in AbsRel. We further analyze pixel-level continuity. As visualized in

Fig. 6 (right), while PRM effectively suppresses noise in the $\mathbf{v}$ -prediction path, $\mathbf{x}$ -prediction consistently yields more spatially coherent depth maps. Consequently, a one-step $\mathbf{x}$ -prediction strategy is identified as our optimal choice, as it inherently aligns with the clean data manifold and maximizes both efficiency and geometric fidelity.

5 Conclusion

In this paper, we presented Lapis, a pixel-space generative model based on linear attention for efficient and high-quality depth estimation, especially under high-resolution inputs. By decoupling patch-level context modeling from pixel-level refinement and leveraging a direct $\mathbf{x}$ -prediction strategy, we effectively resolve the degradations in the naively linearized backbone and suppress the noise in one-step pixel-space diffusion. Extensive experiments demonstrate that Lapis achieves SOTA zero-shot performance while significantly reducing inference latency. Our approach provides a scalable and computationally efficient foundation for real-world 3D perception, bridging the gap between generative quality and discriminative speed.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62322113, 62376156), as well as the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G15).

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571 (2023)
2. Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
3. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: International Conference on Learning Representations (2025)
4. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) Proceedings of the European Conference on Computer Vision. pp. 611–625. Part IV, LNCS 7577, Springer-Verlag (Oct 2012)
5. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
6. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: Multi-scale linear attention for high-resolution dense prediction. arXiv preprint arXiv:2205.14756 (2024)
7. Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., et al.: Sana-video: Efficient video generation with block linear diffusion transformer. arXiv preprint arXiv:2509.24695 (2025)

8. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2016)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scan-net: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2021)
12. Eftekhari, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10786–10796 (2021)
13. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems* **27** (2014)
14. Elfving, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks* **107**, 3–11 (2018)
15. Fan, Q., Huang, H., Ai, Y., He, R.: Rectifying magnitude neglect in linear attention. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 21505–21514 (2025)
16. Fan, Q., Huang, H., He, R.: Breaking the low-rank dilemma of linear attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25271–25280 (2025)
17. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)
18. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: Proceedings of the European Conference on Computer Vision (2024)
19. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *International Journal of Robotics Research* (2013)
20. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfin: Fast generative monocular depth estimation with flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3203–3211 (2025)
21. Gómez, J.L., Silva, M., Seoane, A., Borràs, A., Noriega, M., Ros, G., Iglesias-Guitián, J.A., López, A.M.: All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *Neurocomputing* **637**, 130038 (2025)
22. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention. In: Proceedings of the IEEE International Conference on Computer Vision (2023)
23. Han, D., Wang, Z., Xia, Z., Han, Y., Pu, Y., Ge, C., Song, J., Song, S., Zheng, B., Huang, G.: Demystify mamba in vision: A linear attention perspective. *Advances in Neural Information Processing Systems* **37**, 127181–127203 (2024)

24. He, J., Li, H., Sheng, M., Chen, Y.C.: Lotus-2: Advancing geometric dense prediction with powerful image generative model. arXiv preprint arXiv:2512.01030 (2025)
25. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In: International Conference on Learning Representations. vol. 2025, pp. 89454–89467 (2025)
26. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017)
27. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
28. Huang, J., Miao, S., Yang, B., Ma, Y., Liao, Y.: Vivid4d: Improving 4d reconstruction from monocular video by video inpainting. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 12592–12604 (2025)
29. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: *International Conference on Machine Learning*. pp. 5156–5165. PMLR (2020)
30. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
31. Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
32. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2022)
33. Kumar, R., Vats, V.: Few-shot novel view synthesis using depth aware 3d gaussian splatting. In: *Proceedings of the European Conference on Computer Vision*. pp. 1–13. Springer (2024)
34. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024), accessed: June 30, 2026
35. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
36. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20775–20785 (June 2024)
37. Li, T., He, K.: Back to basics: Let denoising generative models denoise. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 36115–36125 (2026)
38. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 3205–3215 (2023)
39. Li, Z., Yu, Z., Austin, D., Fang, M., Lan, S., Kautz, J., Alvarez, J.M.: FB-OCC: 3D occupancy prediction based on forward-backward view transformation. arXiv preprint arXiv:2307.01492 (2023)

40. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
41. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2023)
42. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
43. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2019)
44. Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., Xu, C., Xiang, T., Zhang, L.: Soft: Softmax-free transformer with linear complexity. arXiv preprint arXiv:2110.11945 (2022)
45. Martin Garcia, G., Knaebel, K., Schmidt, C., de Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2025)
46. Mehl, L., Schmalfluss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
47. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: Proceedings of the European Conference on Computer Vision (2012)
48. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
49. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4195–4205 (2023)
50. Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., Zhong, Y.: cosformer: Rethinking softmax in attention. In: International Conference on Learning Representations (2022)
51. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Stefano, L.D.: Booster: A benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(1), 85–102 (2024)
52. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. Proceedings of the IEEE International Conference on Computer Vision (2021)
53. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(3) (2022)
54. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)

56. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
57. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German Conference on Pattern Recognition. pp. 31–42. Springer (2014)
58. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
59. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1874–1883 (2016)
60. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation (2019), <https://openreview.net/forum?id=rylgEULtdN>, accessed: June 30, 2026
61. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. arXiv preprint arXiv:1908.00463 (2019)
62. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
63. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5261–5271 (2025)
64. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 35928–35959. Curran Associates, Inc. (2025)
65. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4909–4916. IEEE (2020)
66. Wu, Y., Zheng, W., Zuo, S., Huang, Y., Zhou, J., Lu, J.: Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 26360–26370 (October 2025)
67. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. arXiv preprint arXiv:2410.10629 (2024)
68. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nystr"omformer: A nystr"om-based algorithm for approximating self-attention. arXiv preprint arXiv:2102.03902 (2021)
69. Xu, G., Lin, H., Luo, H., Wang, X., YAO, J., Zhu, L., Pu, Y., Chi_, C., Sun, H., Wang, B., Chen, G., Ye, H., Peng, S., Yang, X.: Pixel-perfect depth with semantics-prompted diffusion transformers. In: Belgrave, D., Zhang, C., Lin, H., Pascanu,

- R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) *Advances in Neural Information Processing Systems*. vol. 38, pp. 174731–174755. Curran Associates, Inc. (2025)
70. Xu, G., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C., et al.: What matters when repurposing diffusion models for general dense perception tasks? In: *International Conference on Learning Representations*. vol. 2025, pp. 6786–6799 (2025)
 71. Xu, J., Deng, K., Fan, Z., Wang, S., Xie, J., Yang, J.: Ad-gs: Object-aware b-spline gaussian splatting for self-supervised autonomous driving. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 24770–24779 (October 2025)
 72. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019)
 73. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
 74. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. *arXiv preprint arXiv:2406.09414* (2024)
 75. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: *Proceedings of the IEEE International Conference on Computer Vision* (2019)
 76. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021)
 77. Zama Ramirez, P., Tosi, F., Poggi, M., Salti, S., Di Stefano, L., Mattoccia, S.: Open challenges in deep stereo: the booster dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022)
 78. Zhang, C., Yin, W., Wang, Z., Yu, G., Fu, B., Shen, C.: Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems* (2022)
 79. Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., et al.: Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems* **38**, 24195–24228 (2026)