D^{3}epth: Distilling Diffusion Models For Efficient Depth Estimation Through A Two-Stage Approach

Bo-Chih Chuang M11115209@MAIL.NTUST.EDU.TW
National Taiwan University of Science and Technology, Taipei, Taiwan

Wei-Tung Lin M11115116@MAIL.NTUST.EDU.TW
National Taiwan University of Science and Technology, Taipei, Taiwan

Shang-Fu Chen CHENSHANGFU@CMLAB.CSIE.NTU.EDU.TW
National Taiwan University, Taipei, Taiwan

Kai-Lung Hua HUA@MAIL.NTUST.EDU.TW
National Taiwan University of Science and Technology, Taipei, Taiwan

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Diffusion-based monocular depth estimation models demonstrate strong performance with limited supervision by leveraging pre-trained text-to-image models. However, their multi-step inference process and large model size create prohibitive computational overhead for practical applications. To retain the data efficiency of diffusion models while addressing their inference inefficiency, we propose a framework that enhances diffusion-based depth estimation through a two-stage training approach. The first stage distills implicit depth knowledge in the latent space by leveraging the rich representations from pre-trained diffusion models. The second stage refines explicit depth predictions in pixel space using Hybrid Depth Loss that combines Scale-Shift Invariant (SSI) loss for global structure preservation with Edge-aware Gradient Huber loss for fine-grained detail enhancement. Both components are adaptively weighted using a dynamic task weighting strategy, balancing structural consistency and boundary precision. Specifically, we demonstrate that our two-stage distillation approach yields D^{3}epth, an efficient variant that achieves state-of-the-art results while considerably reducing computational requirements. In parallel, our base model D^{2}epth, trained with enhanced pixel-space depth loss, also surpasses state-of-the-art performance across various benchmarks. Overall, these results deliver the accuracy benefits of diffusion-based methods at the efficiency level of traditional data-driven approaches.

Keywords: Monocular Depth Estimation; Knowledge Distillation; Diffusion Models

1. Introduction

Monocular depth estimation is a fundamental computer vision task that predicts per-pixel depth values from a single RGB image. The task is inherently challenging as it requires recovering 3D information from 2D observations, which is geometrically ill-posed without strong priors on scene structure and object properties Marr and Poggio (1979); Saxena et al. (2009); Eigen and Fergus (2015). Traditional approaches Yin et al. (2021); Ranftl et al. (2020, 2021); Yang et al. (2024a,b) to depth estimation rely on discriminative models trained on large datasets of RGB-depth pairs Geiger et al. (2013); Birkl et al. (2023); Yang et al. (2024a). Although these methods have achieved impressive performance by scaling up both training data and model capacity, they face significant limitations: they struggle

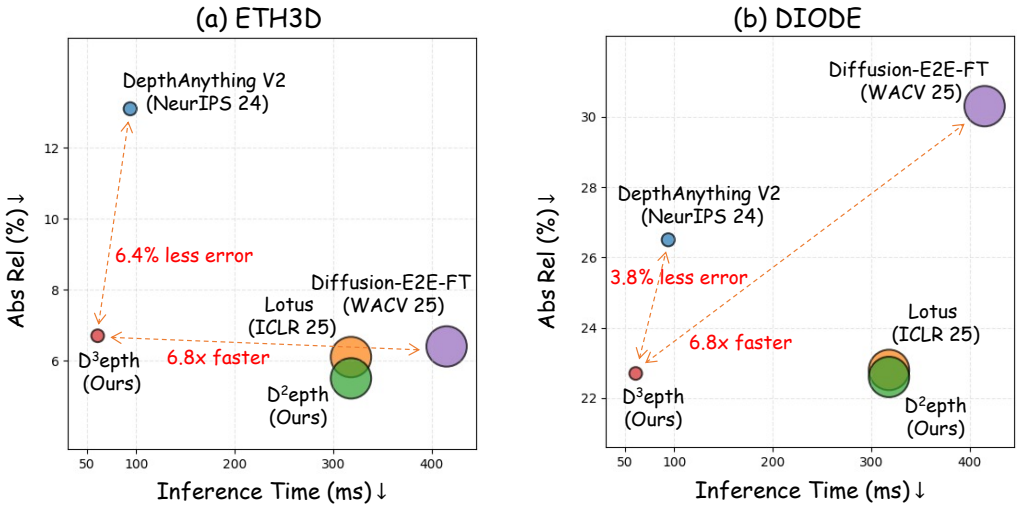

Figure 1: Accuracy vs. inference time comparison on (a)ETH3D and (b)DIODE dataset. Our distilled model (D^3epth) achieves competitive accuracy with considerably faster inference speed than SoTA methods (Diffusion-E2E-FT Garcia et al. (2024), Lotus He et al. (2024)) while outperforming data-driven methods (DepthAnything V2 Yang et al. (2024b)). Larger circles indicate higher parameter size.

with zero-shot generalization to novel scenes, domains, or imaging conditions not well-represented in their training data, and their reliance on massive datasets requires substantial computational resources for training that are often infeasible for real-world applications.

Recent advances in text-to-image diffusion models Ho et al. (2020); Rombach et al. (2022); Chen et al. (2023) have opened new possibilities for monocular depth estimation. These models, such as Stable Diffusion Rombach et al. (2022), trained on billions of internet images, learn rich visual priors that capture diverse scene structures, object shapes, and spatial relationships. Several works Ke et al. (2024); Fu et al. (2024); He et al. (2024) have successfully adapted these models for depth estimation, demonstrating impressive data efficiency by requiring approximately 1% of the training samples used in traditional data-driven methods. For instance, Marigold Ke et al. (2024) fine-tuned Stable Diffusion using only 74K synthetic samples to achieve competitive accuracy on real-world benchmarks. Lotus He et al. (2024) further improved performance and efficiency by removing the noise input and performing single-step inference from latent representations, achieving strong results with only 59K training samples. These works highlight the potential of diffusion models to reduce training costs considerably while maintaining strong generalization.

However, diffusion-based depth estimation methods face a critical limitation: *computational efficiency*, as shown in Table 2. Traditional diffusion models require multiple denoising steps during inference, making them substantially slower than data-driven approaches. For instance, Marigold requires 10–50 denoising steps and ensemble inference to achieve optimal performance, causing much longer inference time. Although Lotus improved by a single-step formulation and achieves significant speedup, its large model size still poses a challenge for real-time deployment. To overcome this, we adopt a knowledge distilla-

tion Hinton et al. (2015) strategy that transfers knowledge from a capable teacher model to a more efficient student, compressing large diffusion models into lightweight student networks, enabling efficient inference with comparable accuracy, as shown in various vision tasks including dense prediction He et al. (2025). However, distilling diffusion-based models poses unique challenges as student models are required to learn two objectives simultaneously: implicit depth latent representations guided by knowledge distillation loss and pixel-level depth predictions supervised by SSI loss Ranftl et al. (2020) or other pixel-wise losses such as MSE loss. Unlike traditional discriminative models that learn direct RGB-to-depth mappings, diffusion-based depth estimation models rely on rich visual knowledge encoded within the Variational Autoencoder (VAE) latent space, which stays frozen during training. This architectural design creates a fundamental disconnect between two different feature spaces: the semantically-rich latent space where implicit depth knowledge resides, and the pixel space where explicit depth supervision is applied. The student model must simultaneously acquire meaningful representations in the latent space to match the teacher's intermediate features while also producing accurate pixel-level depth maps that satisfy geometric constraints. These competing objectives operate in fundamentally different feature spaces with different optimization dynamics, leading to training instability and suboptimal convergence as the model struggles to balance semantic understanding with geometric precision.

In this work, we present the first study of knowledge distillation for diffusion-based depth estimation models and introduce D^{3}epth. Our approach aims to effectively transfer strong depth estimation capabilities from SD-based Rombach et al. (2022) models like Lotus He et al. (2024) into lightweight and efficient architectures, such as SDXS Song et al. (2024), through a two-stage approach. We further contribute a novel Hybrid Depth Loss that addresses the limitations of the conventional SSI loss by integrating Edge-aware Gradient Huber regularization, where the loss components are adaptively weighted through a dynamic strategy to preserve both global depth structure and fine-grained edge details. We also introduce base variant D^{2}epth which applies Hybrid Depth Loss and follows the Lotus He et al. (2024) training protocol on the SD architecture.

Comprehensive experiments on the ETH3D, DIODE, NYUv2, KITTI, and ScanNet datasets demonstrate that our approach achieves SoTA performance while maintaining computational efficiency compared to recent methods, including DepthAnything V2 Yang et al. (2024b), Lotus He et al. (2024), and Diffusion-E2E-FT Garcia et al. (2024) (see Fig. 1). These results underscore the efficacy of our two-stage training framework and Hybrid Depth Loss in improving accuracy and efficiency. Our contributions can be summarized as follows:

- We propose a novel two-stage distillation framework designed for diffusion-based depth estimation that addresses the fundamental limitations of vanilla knowledge distillation by separately transferring latent feature knowledge and depth-specific understanding.
- We introduce Hybrid Depth Loss combining Scale-Shift Invariant (SSI) loss with Edge Gradient Huber loss through dynamic weighting. It preserves global depth structure and fine-grained boundary details, outperforming existing training paradigms.
- We present comprehensive experiments showing that D^{2}epth achieves state-of-the-art performance across various datasets, while D^{3}epth delivers competitive accuracy with 70% parameter reduction and 5$\times$ inference speedup compared to state-of-the-art methods, enabling practical deployment in resource-constrained applications.

2. Related Work

2.1. Generative Models for Text-to-Image

Diffusion-based models have revolutionized the field of text-to-image (T2I) synthesis. Early works like DALL·E Ramesh et al. (2021), GLIDE Nichol et al. (2021), and Imagen Saharia et al. (2022) demonstrated the ability of diffusion processes to generate photorealistic images from natural language prompts. Stable Diffusion (SD) Rombach et al. (2022), in particular, leverages a latent UNet and is trained on the massive LAION-5B dataset Schuhmann et al. (2022). It balances fidelity and efficiency. Recent efforts have adapted these strong generative priors to downstream vision tasks. For instance, Marigold Ke et al. (2024) and GeoWizard Fu et al. (2024) explore using SD for dense predictions like depth and normals, revealing that pretrained T2I models encode transferable visual semantics. Our method follows this line by employing SDXS, a distilled and lightweight variant of SD, to enable real-time and efficient depth prediction.

2.2. Knowledge Distillation for Generative Models

To reduce the computational burden of large diffusion models, knowledge distillation (KD) has become a crucial strategy. Methods such as progressive distillation Salimans and Ho (2022) and feature-level supervision Song et al. (2024) have been proposed to compress high-capacity generative models into efficient student networks. For example, SDXS Song et al. (2024) distills both the UNet and text encoder components of Stable Diffusion, achieving fast, one-step inference with minimal degradation in visual quality. Building on these principles, our method adopts a two-stage distillation framework. In the first stage, latent-space KD transfers object-aware features and visual semantics from a high-capacity teacher (Lotus-D) to a compact student. In the second stage, we apply pixel-level supervision using a Hybrid Depth Loss that combines scale-and-shift invariant (SSI) with directional gradient losses. A dynamic weighting strategy balances global consistency and local detail, enabling the student to capture both global depth structure and fine-grained edge details.

2.3. Monocular Depth Estimation

Diffusion-based Approaches. Recent advancements in monocular depth estimation have explored the use of diffusion models for generating pixel-wise depth maps. Works like Marigold Ke et al. (2024) and GeoWizard Fu et al. (2024) repurpose pre-trained Stable Diffusion models for dense prediction tasks. These methods leverage the rich visual priors encoded in text-to-image diffusion models, but often rely on iterative inference, leading to high computational costs. Lotus-D He et al. (2024) improves efficiency by introducing a single-step stochastic inference mechanism tailored for dense prediction, showing strong zero-shot generalization from synthetic training. Our method builds on these ideas, employing a distilled student model with single-step inference for fast and accurate depth estimation, further enhanced by output and feature-level supervision.

Discriminative and Data-Driven Methods. Traditional approaches to monocular depth estimation rely on supervised learning with high-capacity CNNs or ViTs. Models like MiDaS Ranftl et al. (2020), DPT Ranftl et al. (2021), and DepthAnything V2 Yang et al. (2024b) are trained on diverse, large-scale datasets and exhibit strong generalization.

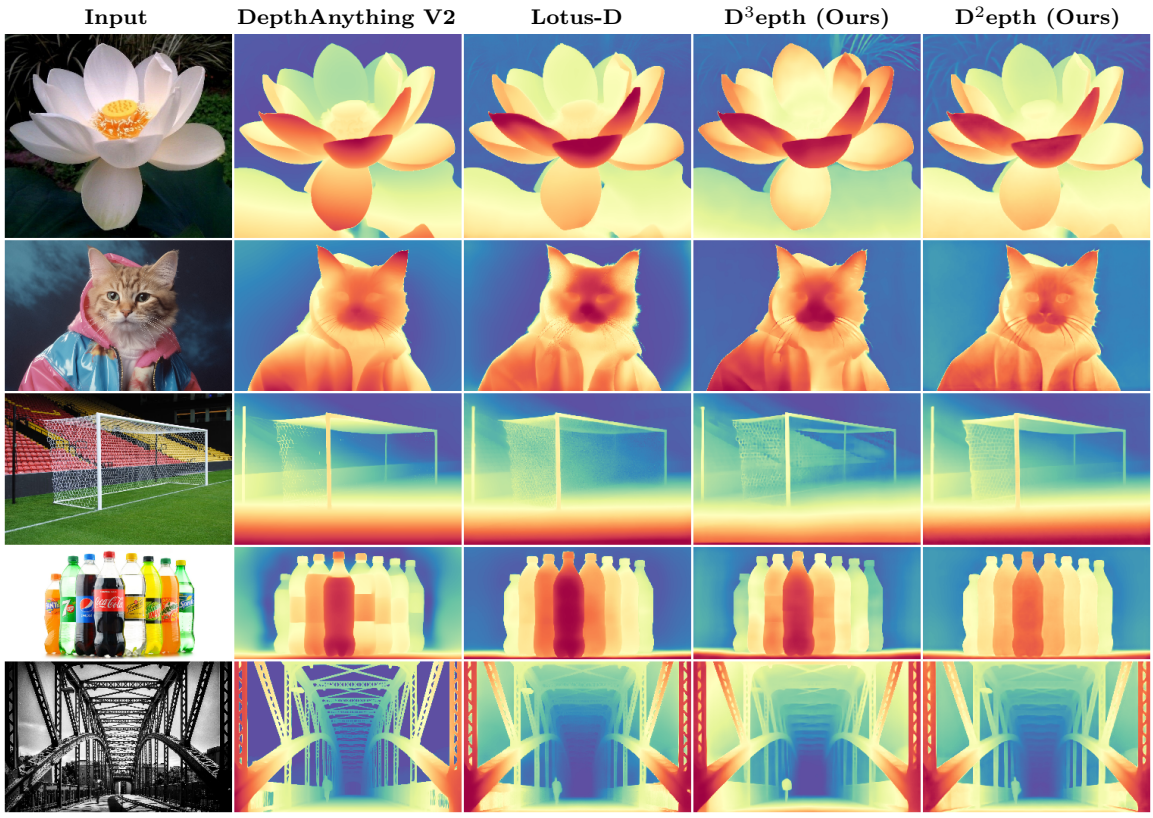

Figure 2: Qualitative results on in-the-wild examples. Our method preserves fine-grained structures more effectively in challenging scenes, including cat whiskers, background seats, and cloud patterns.

These discriminative models typically use pixel-wise regression losses and multi-scale fusion strategies to capture fine-grained geometry. While effective, they often require millions of annotated samples and lack the compactness needed for real-time deployment. In contrast, our work distills a diffusion-based backbone into a compact student, combining the strengths of diffusion- and data-driven approaches for practical depth estimation.

3. Preliminary

3.1. Discriminative Diffusion for Depth Estimation

Recent diffusion-based depth estimation methods leverage visual priors from large-scale text-to-image models Rombach et al. (2022). Lotus He et al. (2024) addresses the computational inefficiency of multi-step diffusion inference by reformulating depth estimation as a single-step prediction problem.

The approach trains a U-Net model f_θ with a dual-task objective:

$$\mathcal{L}_{\text{Latent}} = \|\boldsymbol{z}_d - f_\theta(\boldsymbol{z}_x, T, s_d)\|^2 + \|\boldsymbol{z}_x - f_\theta(\boldsymbol{z}_x, T, s_x)\|^2, \tag{1}$$

where $\boldsymbol{z}_x$ and $\boldsymbol{z}_d$ are latent encodings of input image and depth, T is the maximum timestep, and s_x, s_d are task-specific tokens for image reconstruction and depth prediction. During inference, the model performs deterministic prediction: $\hat{\boldsymbol{z}}_d = f_\theta(\boldsymbol{z}_x, T, s_d)$, followed by VAE decoding to obtain the final depth map.

3.2. Challenges in Diffusion-based Depth Distillation

While Lotus demonstrates impressive data efficiency and performance, directly applying either knowledge distillation or pixel-level supervision alone yields limited benefits, as shown in Table 4. Due to the intrinsic mismatch between latent and pixel spaces, joint optimization without proper decoupling introduces convergence issues, highlighting the need for a specialized two-stage framework.

Latent-Space Learning Complexity. Traditional knowledge distillation approaches that operate directly in the RGB-to-Depth mapping space fail to capture the nuanced latent representations that encode both visual understanding and depth-specific knowledge. Diffusion-based depth estimators like Lotus operate in the latent space of a pre-trained VAE, where the model learns to predict latent representations rather than direct pixel values. This latent knowledge encompasses two distinct components: (1) rich visual priors acquired from large-scale text-to-image pretraining, and (2) task-specific geometric understanding required for accurate depth estimation.

Convergence Failure in Naive Distillation. Monocular depth estimation is a discriminative task that relies on fine-grained, pixel-wise supervision for accurate geometric prediction. However, directly applying latent knowledge distillation from a generative teacher model such as Lotus-D He et al. (2024) requires the student to acquire semantically rich representations in latent space. These representations often conflict with the structural precision required by pixel-level objectives such as the scale-and-shift invariant loss Ranftl et al. (2020). The mismatch between semantic alignment and geometric accuracy introduces competing optimization goals, which in turn hinders convergence and affects training stability. These challenges necessitate a specialized distillation approach that can effectively transfer both the pre-trained visual priors and the task-specific depth estimation capabilities, while resolving convergence issues arising from joint latent and pixel space learning.

4. Proposed Method

4.1. Two-Stage Training Pipeline

We introduce a two-stage training pipeline that decouples representation learning from detail refinement (see Fig. 3). Stage 1 distills rich semantic knowledge from a large teacher model through latent-space supervision, while Stage 2 transitions to pixel-level optimization for fine-grained detail preservation.

Stage 1: Latent Representation Learning. We distill knowledge from the Lotus-D teacher model He et al. (2024) to our lightweight student through three complementary objectives: target supervision ($\mathcal{L}_{\text{Latent}}$), output distillation ($\mathcal{L}_{\text{LatentKD}}$), and intermediate feature alignment ($\mathcal{L}_{\text{FeatKD}}$).

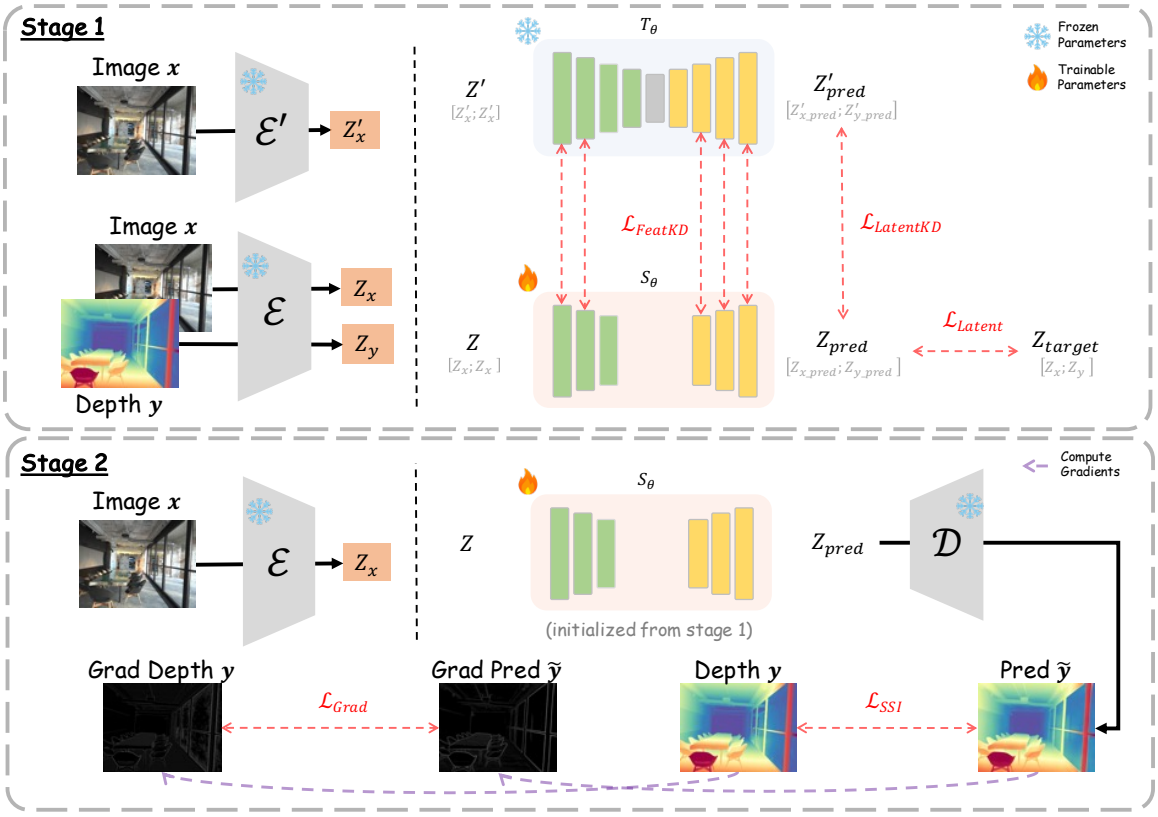

Figure 3: Overview of the D^{3}epth two-stage training pipeline. Stage 1 distills latent representations from the teacher (T_θ) to the student (S_θ) using latent-space distillation losses. Stage 2 finetunes the student with Hybrid Depth Loss in pixel space to preserve global structure and local details.

Target supervision directly trains the student to predict ground-truth latent depth representations, while output distillation enforces consistency between student and teacher predictions. Both terms employ Mean Square Error (MSE) loss, preserving Lotus-D's single-step inference process. Following BK-SDM Kim et al. (2023), we align intermediate UNet features based on spatial resolution matching. Identical spatial dimensions eliminate the need for projection layers, reducing computational overhead. Refer to Sec. 5.3 for details. Stage 1 objective combines all distillation terms as follows:

$$\mathcal{L}_{\text{stage\_1}} = \mathcal{L}_{\text{Latent}} + \lambda_{\text{LatentKD}} \cdot \mathcal{L}_{\text{LatentKD}} + \lambda_{\text{FeatKD}} \cdot \mathcal{L}_{\text{FeatKD}}, \tag{2}$$

Stage 2: Detail-Preserving Fine-Tuning. While latent supervision captures rich object-level details, it compromises global structural smoothness, resulting in artifacts such as depth discontinuities across homogeneous regions. This limitation stems from the mismatch between denoising objectives and pixel-level depth metrics Garcia et al. (2024). To address this, we perform end-to-end pixel-level fine-tuning initialized from Stage 1 weights.

This pipeline processes RGB input $\boldsymbol{x}$ through a frozen VAE encoder $\mathcal{E}$ to obtain $\boldsymbol{Z} = \mathcal{E}(\boldsymbol{x})$, predicts latent depth $\boldsymbol{z}_{\text{pred}} = S_\theta(\boldsymbol{Z}, t, s)$ via a UNet with fixed timestep $t = T$, and decodes to pixel space $\tilde{\boldsymbol{y}} = \mathcal{D}(\boldsymbol{z}_{\text{pred}})$ where s is the task-specific token for depth prediction. We fine-tune the model using Hybrid Depth Loss, detailed in Sec. 4.2, to better preserve global consistency while maintaining local detail.

4.2. Hybrid Depth Loss

Scale-and-Shift Invariant Loss Standard MSE loss penalizes absolute depth deviations and exhibits sensitivity to global scale variations. We adopt scale-and-shift invariant (SSI) loss Ranftl et al. (2020) for affine-invariant supervision. Given predicted depth $\boldsymbol{d}$ and ground truth $\boldsymbol{d}^*$, we compute the aligned prediction $\hat{\boldsymbol{d}} = s\boldsymbol{d} + t$, where scale s and shift t minimize least-squares error. The SSI objective is:

$$\mathcal{L}_{\text{SSI}} = \frac{1}{HW} \sum_{i,j} \left| d^*_{i,j} - \hat{d}_{i,j} \right|, \tag{3}$$

which provides scale-and-shift invariant supervision and encourages consistency in relative depth predictions. Here, (i, j) denotes pixel coordinates, and H, W are the height and width of the image.

Edge-Aware Gradient Huber Loss SSI loss ensures global consistency but may blur local discontinuities. We introduce gradient-based supervision to preserve edge structures and fine geometric details. We compute 4-directional gradients (1×horizontal, 1×vertical, 2×diagonal) for ground truth $\mathbf{G}_{\text{GT}}$ and prediction $\mathbf{G}_{\text{pred}} \in \mathbb{R}^{H\times W\times 4}$. We also apply a modified Huber loss Huber (1992) with threshold δ to improve training stability:

$$\mathcal{L}_{\text{Grad}} = \begin{cases} \delta \cdot |\mathbf{G}_{\text{GT}} - \mathbf{G}_{\text{pred}}|, & \text{if } |\mathbf{G}_{\text{GT}} - \mathbf{G}_{\text{pred}}| \le \delta, \\ \frac{1}{2}(\mathbf{G}_{\text{GT}} - \mathbf{G}_{\text{pred}})^2 + \frac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \tag{4}$$

Dynamic Task Weighting Strategy To balance global consistency and local detail during optimization, we adopt a task weighting strategy inspired by Dynamic Weight Average (DWA) Liu et al. (2019), which adaptively adjusts each loss component's contribution based on its recent learning dynamics. Compared to gradient-based methods such as GradNorm Chen et al. (2018), this approach is simpler to implement, relying only on scalar loss values rather than gradient magnitudes. At training step t, the weight λ_k for each task k is updated as:

$$\lambda_k(t) := \frac{K \exp(w_k(t-1)/T)}{\sum_i \exp(w_i(t-1)/T)}, \quad w_k(t-1) = \frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)}, \tag{5}$$

Here, T controls the sharpness of the weighting distribution, and K is the number of tasks. Following DWA Liu et al. (2019), we fixed temperature parameter $T = 2$, and employ epoch-averaged losses to reduce variance, with w_k initialized as 1 for $t = 1, 2$. In our setting, λ_{SSI} and λ_{Grad} are derived from λ_k in Eq. (5), dynamically balancing $\mathcal{L}_{\text{SSI}}$ for scale-consistent structure and $\mathcal{L}_{\text{Grad}}$ for edge-aware detail preservation. The complete Stage 2 objective is:

$$\mathcal{L}_{\text{stage\_2}} = \lambda_{\text{SSI}} \cdot \mathcal{L}_{\text{SSI}} + \lambda_{\text{Grad}} \cdot \mathcal{L}_{\text{Grad}}, \tag{6}$$

Table 1: Quantitative comparison on zero-shot affine-invariant depth estimation between D^{3}epth and state-of-the-art methods. The upper section lists data-driven methods, while the lower section lists model-driven methods that rely on pre-trained Stable Diffusion. The **best** results are highlighted in bold, and the second-best results are underlined. * indicates numbers reported from Lotus He et al. (2024).

Method	Training Data↓	Params↓	ETH3D (Various)		DIODE (Various)		KITTI (Outdoor)		NYUv2 (Indoor)		ScanNet (Indoor)	
			AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Data-driven methods												
DiverseDepth (TPAMI '21)	320K	**70M**	22.8	69.4	37.6	63.1	19.0	70.4	11.7	87.5	10.9	88.2
MiDaS (TPAMI '22)	2M	344M	18.4	75.2	33.2	71.5	23.6	63.0	11.1	88.5	12.1	84.6
LeRes (CVPR '21)	354K	115M	17.1	77.7	27.1	76.6	14.9	78.4	9.0	91.6	9.1	91.7
Omnidata V1 (ICCV '21)	12.2M	123M	16.6	77.8	33.9	74.2	14.9	83.5	7.4	94.5	7.5	93.6
DPT (ICCV '21)	1.4M	344M	**7.8**	**94.6**	**18.2**	75.8	10.0	90.1	9.8	90.3	8.2	93.4
HDN (NIPS '22)	**300K**	-	12.1	83.3	24.6	**78.0**	11.5	86.7	6.9	94.8	8.0	93.9
DepthAnything V2 (NIPS '24)	62.6M	335M	13.1	86.5	26.5	73.4	**7.4**	94.6	4.5	97.9	**4.2**	97.8
DepthAnything (CVPR '24)	62.6M	335M	12.7	88.2	26.0	75.9	7.6	**94.7**	**4.3**	**98.1**	4.3	**98.1**
Model-driven methods												
GeoWizard* (ECCV '24)	280K	944M	6.6	95.8	33.5	72.3	14.4	82.0	5.6	96.3	6.4	95.0
Marigold (CVPR '24)	74K	949M	6.5	95.9	30.8	77.3	9.9	91.6	5.5	96.4	6.4	95.2
GenPercept* (ICLR '25)	74K	949M	7.0	95.6	35.7	75.6	13.0	84.2	5.6	96.0	6.2	96.1
Diffusion-E2E-FT (WACV '25)	74K	949M	6.4	95.9	30.3	**77.6**	9.6	92.1	5.4	96.5	5.8	96.5
Lotus-D (ICLR '25)	**59K**	951M	6.1	97.0	22.8	73.8	8.1	93.1	**5.1**	**97.2**	5.5	96.5
D^{2}epth (Ours)	**59K**	951M	**5.5**	**97.4**	**22.6**	74.1	**7.8**	**93.7**	**5.1**	97.1	**5.3**	**96.7**
D^{3}epth (Ours)	**59K**	**332M**	6.7	95.6	22.7	73.2	9.5	90.6	7.0	95.1	7.7	93.8

Table 2: Detailed efficiency and performance comparison of data-driven method DepthAnything V2 Yang et al. (2024b), model-driven method Diffusion-E2E-FT Garcia et al. (2024), Lotus-D He et al. (2024), D^{2}epth and D^{3}epth. Parameter size is shown in MB and inference times in milliseconds (ms). Inference times were measured with image size 512×512. The **best** results are highlighted in bold, and the second-best results are underlined.

Method	Training Data↓	Params Total↓	Inference time↓	ETH3D (Various)		DIODE (Various)		KITTI (Outdoor)		NYUv2 (Indoor)		ScanNet (Indoor)	
				AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
DepthAnything V2 (NIPS '24)	62.6M	335M	94ms	13.1	86.5	26.5	73.4	**7.4**	**94.6**	**4.5**	**97.9**	**4.2**	**97.8**
Diffusion-E2E-FT (WACV '25)	74K	949M	415ms	6.4	95.9	30.3	**77.6**	9.6	92.1	5.4	96.5	5.8	96.5
Lotus-D (ICLR '25)	**59K**	950M	318ms	6.1	97.0	22.8	73.8	8.1	93.1	5.1	97.2	5.5	96.5
D^{2}epth (Ours)	**59K**	950M	318ms	**5.5**	**97.4**	**22.6**	74.1	7.8	93.7	5.1	97.1	5.3	96.7
D^{3}epth (Ours)	**59K**	**331M**	**61ms**	6.7	95.6	22.7	73.2	9.5	90.6	7.0	95.1	7.7	93.8

5. Experiment Results

5.1. Experimental Settings

Implementation details. D^{3}epth is implemented upon SDXS-0.9 Song et al. (2024), with text conditioning disabled. The depth maps are normalized to the range $[-1, 1]$ to match the input value range required by the VAE. A fixed time-step of $t = 1000$ is used throughout training. Depth is predicted in disparity space, *i.e.*, $d = 1/d'$, where d denotes the predicted disparity and d' is the ground truth depth. All experiments are conducted on NVIDIA RTX 3090 GPU. We adopt a two-stage training pipeline: the first stage performs latent-space knowledge distillation for 4,000 steps (∼50 hours) to transfer semantic features from the teacher model; the second stage fine-tunes the student at the pixel-level using the pretrained weights for 4,000 steps (∼14 hours) to enhance structural consistency and local

detail, totaling ∼64 hours of training. D^{2}epth is trained from scratch using Stable Diffusion v2 Rombach et al. (2022) and only performs the second-stage pixel-level fine-tuning with the latent model's weights for 4,000 steps, which takes around 65 hours.

Training Datasets. Following the training protocol established in Lotus (He et al., 2024), we adopt the same synthetic training setup that combines indoor and outdoor datasets: (i) *Hypersim* (Roberts et al., 2021), a photorealistic indoor dataset featuring 461 scenes, from which around 39K valid samples (resized to 576×768) are obtained after filtering the official 54K training split; and (ii) *Virtual KITTI* (Cabon et al., 2020), a synthetic urban dataset comprising five scenes under varied conditions, from which about 20K samples are selected (cropped to 352×1216 with a far plane of 80m). Each batch samples 90% from *Hypersim* and 10% from *Virtual KITTI*, following the probabilistic mixing strategy used in Marigold (Ke et al., 2024).

Evaluation Datasets and Metrics. We evaluate D^{3}epth for zero-shot affine-invariant depth estimation on five real-world datasets: NYUv2 (Silberman et al., 2012) and ScanNet (Dai et al., 2017), which provide indoor RGB-D data captured using Kinect sensors; KITTI (Geiger et al., 2013), containing outdoor driving scenes collected with vehicle-mounted cameras and LiDAR sensors; and ETH3D (Schops et al., 2017) and DIODE (Vasiljevic et al., 2019), which consist of both indoor and outdoor scenes derived from LiDAR scans. All five datasets are unseen during training and span diverse real-world environments.

We adopt the affine-invariant depth evaluation protocol from (Ranftl et al., 2020; Ke et al., 2024; Yang et al., 2024a,b), which aligns predicted depth maps to ground truth using least-squares fitting. Evaluation metrics consist of the *absolute mean relative error* (**AbsRel**), defined as $\frac{1}{M}\sum_{i=1}^{M}|a_i - d_i|/d_i$, where M denotes the total number of pixels, a_i is the predicted depth, and d_i is the ground truth depth. We also report $\delta 1$, which, along with AbsRel, serves as a complementary depth evaluation metric. It measures the proportion of pixels satisfying $\max(a_i/d_i, d_i/a_i) < 1.25$.

5.2. Quantitative and Qualitative Comparisons

As shown in Table 1, D^{2}epth consistently achieves top-tier performance across all evaluation datasets, performing better than existing model-driven baselines and attaining the best overall ranking. The lightweight variant D^{3}epth achieves a compelling balance between accuracy and efficiency, offering up to 5× faster inference speed and a 70% reduction in parameter count (comparable to that of DepthAnything), as summarized in Table 2. Notably, D^{3}epth surpasses the DepthAnything series on benchmarks such as ETH3D and DIODE, despite having a similar model size. As shown in Table 1, although DepthAnything achieves the best average performance, it performs worse than diffusion-based methods (including D^{3}epth) on Various scenes datasets, highlighting the generalization limitations of data-driven approaches under distribution shifts. In contrast, diffusion models benefit from strong generative priors, enabling more effective transfer with limited supervision. D^{3}epth further achieves competitive results with far fewer training samples, demonstrating potential for reducing annotation cost. Nonetheless, diffusion-based methods still face challenges in certain edge cases, and further refinement is needed to improve robustness.

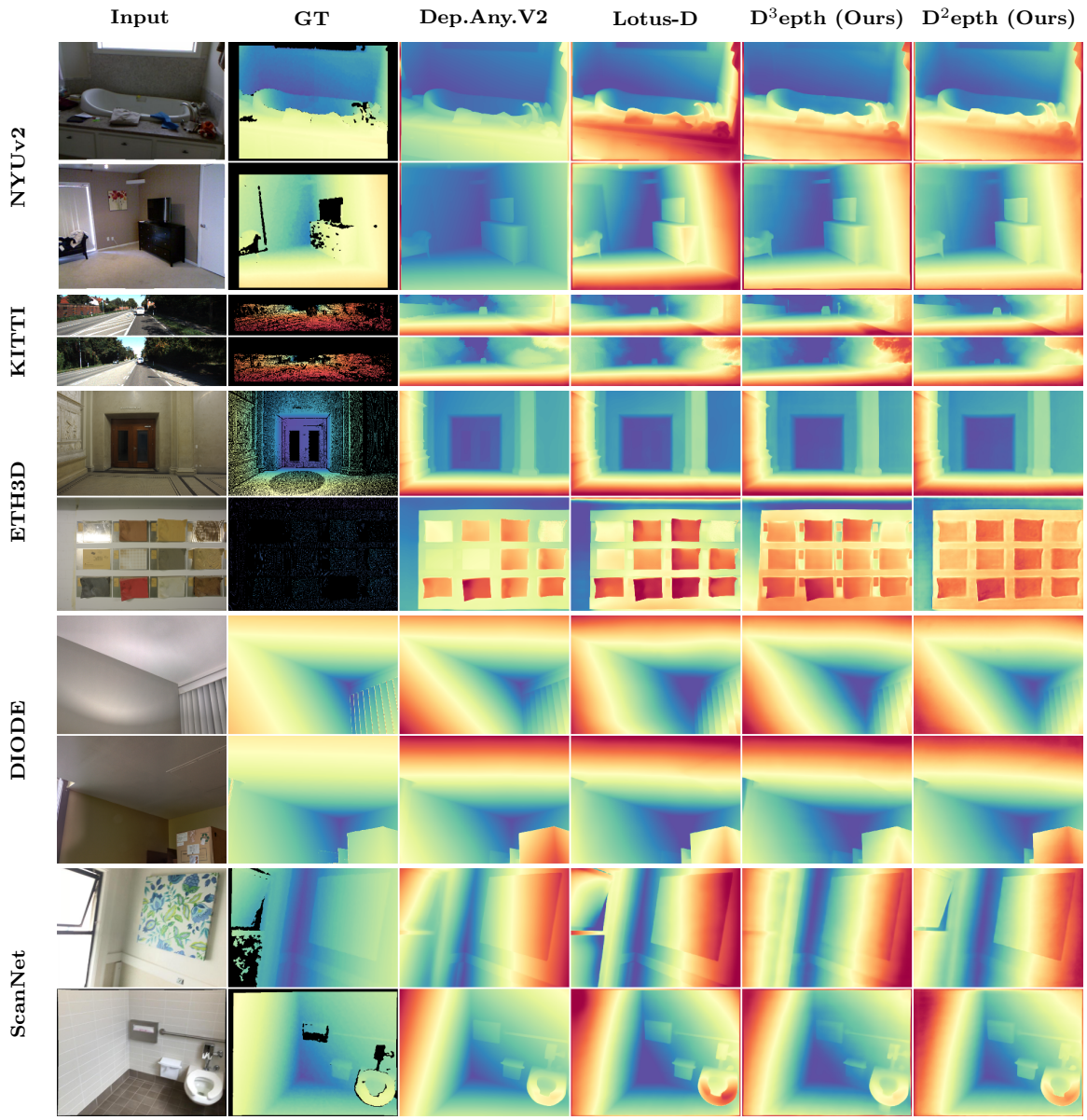

Figure 4: Qualitative comparison with other state-of-the-art methods across five datasets.

5.3. Ablation Study

Latent Representation Distillation. To evaluate the effectiveness of latent feature alignment, we apply distillation supervision to progressively deeper blocks within the U-Net. As shown in Table 3, Baseline serves as the baseline without any feature-level alignment. The teacher model comprises 4 encoder layers $(d0', d1', d2', d3')$, 4 decoder layers $(u0', u1', u2', u3')$, and a bottleneck layer (m'), while the student consists of 3 encoder layers $(d0, d1, d2)$ and 3 decoder layers $(u0, u1, u2)$ without a bottleneck. We study three KD configurations by aligning only blocks with matching feature map sizes. Shallow KD aligns

Table 3: Ablation study on latent representation distillation across different U-Net blocks. We compare the impact of applying knowledge distillation (KD) to progressively deeper encoder (di) and decoder (ui) layers in the student model. Baseline denotes that our model is directly trained with target supervision ($\mathcal{L}_{\text{Latent}}$) only, while Full KD includes all matching blocks between teacher and student. The **best** results are highlighted in bold, and the second-best results are underlined.

Layer	$d0$	$d1$	$u0$	$u1$	$u2$	ETH3D (Various) AbsRel↓	ETH3D (Various) $\delta1$↑	DIODE (Various) AbsRel↓	DIODE (Various) $\delta1$↑	KITTI (Outdoor) AbsRel↓	KITTI (Outdoor) $\delta1$↑	NYUv2 (Indoor) AbsRel↓	NYUv2 (Indoor) $\delta1$↑	ScanNet (Indoor) AbsRel↓	ScanNet (Indoor) $\delta1$↑
Baseline						8.6	94.1	24.2	**72.6**	9.7	90.0	7.8	94.3	8.6	92.4
Shallow KD	✓				✓	8.3	93.9	24.5	72.4	9.9	89.9	7.9	94.0	8.6	92.5
Intermediate KD	✓	✓		✓	✓	8.3	94.4	24.2	72.5	9.8	90.0	7.6	94.5	8.4	92.9
Full KD	✓	✓	✓	✓	✓	**7.8**	**94.9**	**24.0**	**72.6**	**9.6**	**90.1**	**7.5**	**94.6**	**8.2**	**93.1**

Table 4: Ablation study on key components of our distillation framework. We incrementally introduce KD, L_{SSI}, L_{Grad}, and two-stage training. Each component offers complementary improvements. The **best** results are highlighted in bold, and the second-best results are underlined.

Method	KD	$\boldsymbol{L_{SSI}}$	$\boldsymbol{L_{Grad}}$	Two Stage	ETH3D (Various) AbsRel↓	ETH3D (Various) $\delta1$↑	DIODE (Various) AbsRel↓	DIODE (Various) $\delta1$↑	KITTI (Outdoor) AbsRel↓	KITTI (Outdoor) $\delta1$↑	NYUv2 (Indoor) AbsRel↓	NYUv2 (Indoor) $\delta1$↑	ScanNet (Indoor) AbsRel↓	ScanNet (Indoor) $\delta1$↑
Baseline					8.6	94.1	24.2	72.6	9.7	90.0	7.8	94.3	8.6	92.4
Vanilla KD	✓				7.8	94.9	24.0	72.6	9.6	90.1	7.5	94.6	8.2	93.1
SSI Loss		✓			7.9	94.9	23.6	72.4	**9.4**	90.7	**7.0**	**95.1**	7.8	93.6
Hybrid Depth Loss		✓	✓		7.4	95.1	23.4	72.6	**9.4**	**90.9**	7.1	94.8	7.8	93.6
Two-Stage	✓	✓	✓	✓	**6.7**	**95.6**	**22.7**	**73.2**	9.5	90.6	**7.0**	**95.1**	**7.7**	**93.8**

the outermost layers ($d0' \leftrightarrow d0$, $u3' \leftrightarrow u2$), yielding modest gains. Intermediate KD adds deeper blocks ($d1' \leftrightarrow d1$, $u2' \leftrightarrow u1$), showing further improvements. Full KD includes another pair ($u1' \leftrightarrow u0$), consistently achieving the best results. These findings underscore the benefit of deep, structured guidance in latent distillation.

Component Analysis. To assess the contribution of each component in our proposed distillation pipeline, we conduct a comprehensive ablation study, as reported in Table 4 (see also Fig. 5). Starting from Baseline, we sequentially add latent distillation (KD), scale-shift invariant loss (L_{SSI}), and edge-aware gradient loss (L_{Grad}). Each component contributes complementary benefits: KD enhances object-level detail alignment, L_{SSI} reinforces global structural consistency, and L_{Grad} sharpens local boundaries. Finally, incorporating our two-stage training strategy further boosts performance, with the full model (Two-Stage) achieving the lowest AbsRel and highest $\delta1$ across most datasets. This validates the effectiveness of our modular design and the synergistic impact of each component.

6. Conclusion and Future Work

To address the high computational cost associated with diffusion-based monocular depth estimation, we propose a two-stage knowledge distillation framework that explicitly decouples the learning of latent visual priors from pixel-level depth supervision. In the first stage, semantic-rich representations are distilled from a generative teacher into a lightweight stu-

dent model within the latent space. In the second stage, we further refine the student using a Hybrid Depth Loss that integrates Scale-Shift Invariant loss and Edge-aware Huber regularization with dynamic weighting to preserve both global structure and fine-grained detail. This design allows the student model D^{3}epth to maintain competitive accuracy while achieving a 5× speedup and 70% parameter reduction. Furthermore, our base model D^{2}epth, trained solely with the hybrid loss and without distillation, achieves state-of-the-art results across five standard benchmarks, underscoring the effectiveness of our approach in enhancing both efficiency and accuracy for real-world deployment.

In the future, our framework offers potential for broader applications. The first stage serves as a general distillation process from the teacher model, while the second stage requires task-specific loss design. This indicates that by tailoring supervision in the second stage, our framework can be extended to other tasks such as surface normal estimation. Moreover, we have also tested our method on dynamic video sequences, which reveal limited temporal consistency compared to video-based methods like VideoDepthAnything (Chen et al.). Addressing temporal modeling remains an important direction for future research.

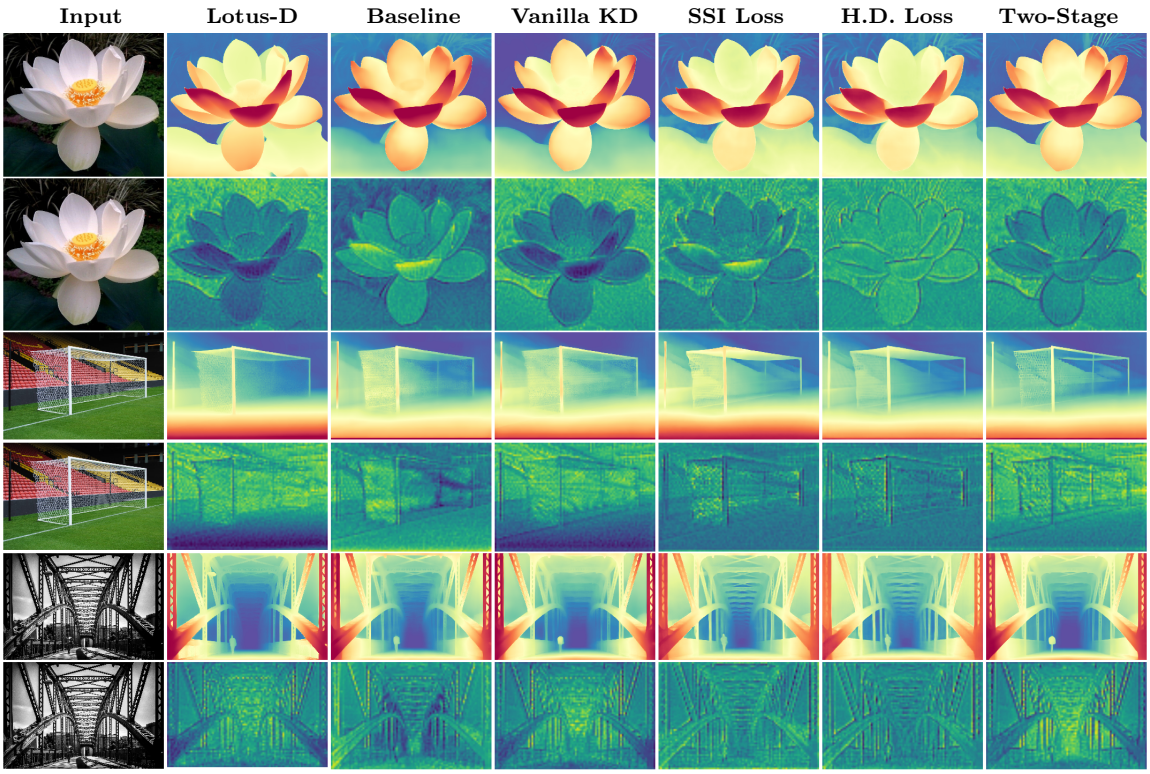

Figure 5: Feature map visualization on in-the-wild samples.

References

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

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.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

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. *arXiv preprint arXiv:2403.12013*, 2024.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

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 preprint arXiv:2409.18124*, 2024.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: Architecturally compressed stable diffusion for efficient text-to-image generation. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

David Marr and Tomaso Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1156):301–328, 1979.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

Yuda Song, Zehao Sun, and Xuanwu Yin. Sdxs: Real-time one-step latent diffusion models with image conditions. *arXiv preprint arXiv:2403.16627*, 2024.

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.

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*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024b.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2021.