FiffDepth: Feed-forward Transformation of Diffusion-Based Generators for Detailed Depth Estimation

Anonymous ICCV submission

Paper ID -

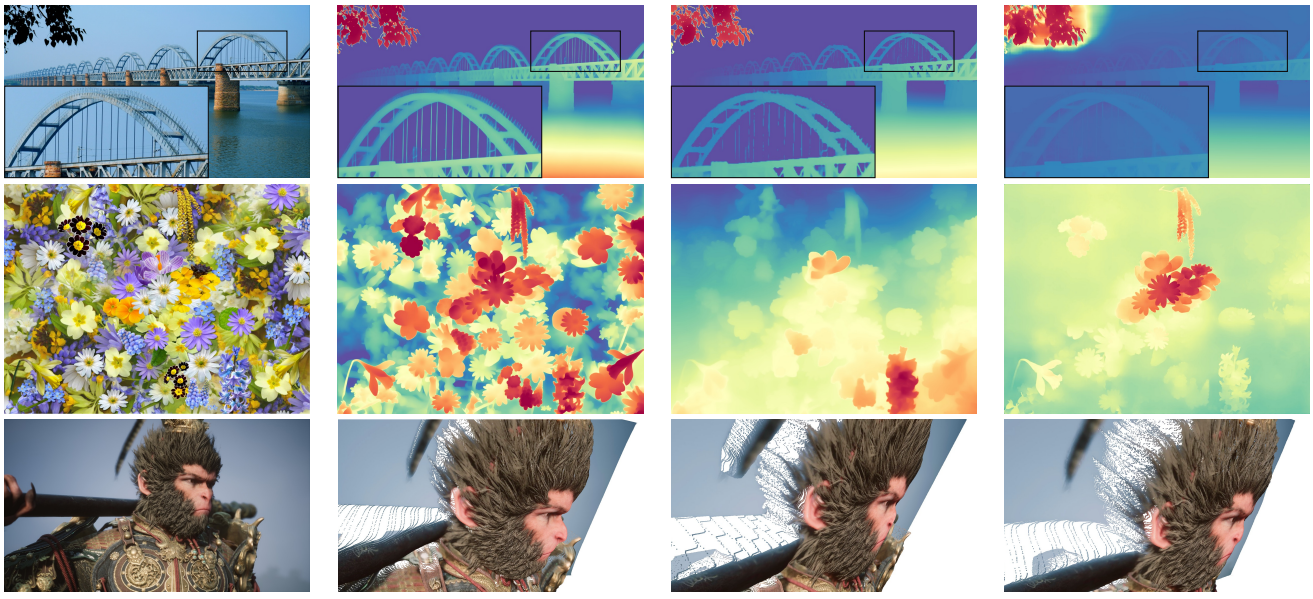

Figure 1. Compared to other methods, our model achieves more accurate details and better generalization in depth estimation. The final row shows the point cloud generated from the estimated depth results, and the corresponding depth map can be referenced in Figure 5.

Abstract

Monocular Depth Estimation (MDE) is a fundamental 3D vision problem with numerous applications such as 3D scene reconstruction, autonomous navigation, and AI content creation. However, robust and generalizable MDE remains challenging due to limited real-world labeled data and distribution gaps between synthetic datasets and real data. Existing methods often struggle with real-world test data with low efficiency, reduced accuracy, and lack of detail. To address these issues, we propose an efficient MDE approach named FiffDepth. The key feature of FiffDepth is its use of diffusion priors. It transforms diffusion-based image generators into a feed-forward architecture for detailed depth estimation. FiffDepth preserves key generative features and integrates the strong generalization capabilities of models like DINOv2. Through benchmark evaluations, we demonstrate that FiffDepth achieves exceptional accuracy, stability, and fine-grained detail, offering significant improvements in MDE performance against state-of-the-art MDE approaches.

1. Introduction

Monocular Depth Estimation (MDE) is a fundamental 3D vision problem with numerous applications in 3D scene reconstruction [44], autonomous navigation [45], and more recently in generating AI-based content [39]. MDE has made significant progress in the era of deep learning [19, 32, 33, 50]. Neural networks trained from paired datasets of im-

ages and pixel depth exhibit encouraging results and often outperform non-deep learning based counterparts that build on monocular depth cues. Despite significant progress, fundamental challenges remain in efficiency, accuracy, and generalization on diverse in-the-wild data. This is because 1) real depth datasets are usually noisy, and 2) while synthetic data can be used, there exist domain gaps between synthetic datasets and diverse in-the-wild data.

Specifically, current MDE research relies primarily on synthetic data due to its high-quality annotations and controlled environments. However, the scale and variety of synthetic datasets remain insufficient for comprehensive training. To address this, synthetic-to-real transfer techniques and the utilization of pre-trained models have emerged as viable solutions. Among pre-trained models, generative networks [35] preserve intricate image details more effectively than feed-forward networks (FFNs) like DINOv2 [29], thus holding greater promise for dense prediction models. However, generative models, while detail-rich, often fall short in synthetic-to-real transfer due to their limited generalization capabilities outside the training domain.

Previous studies [13, 19] have adopted pre-trained diffusion models, where the main idea is to directly finetuning pre-trained RGB image diffusion models into depth map diffusion generation models conditioned on images. However, this method may not be ideal, as dense prediction models require certainty over diversity. The introduction of any noise or uncertainty during the generation process by these methods is sub-optimal. In contrast, we observe that simply using the denoising diffusion module in a feedforward manner yields better and more stable results. This method capitalizes on extending the trajectories of image diffusion models into the depth domain, representing a significant advancement in both accuracy and efficiency for generative model-based depth estimation methods.

Specifically, we optimize diffusion trajectories for MDE tasks. To enable the diffusion model to better retain certain detailed generative features when fine-tuned into an MDE model, we preserve the original generative training trajectory while training the model for depth prediction, aiming to maintain the detailed features of the original generative model as much as possible. Furthermore, recognizing the limitations of fintuned diffusion models in maintaining robustness in diverse real-world images—with inaccuracies in predicted depth occurring mainly in the low-frequency components—we leverage the strengths of a DINOv2 [29] based model, which excels at predicting accurate low-frequency depth despite its reduced fine detail. To address this issue, we use the diffusion model itself to learn a filter that refines its inaccurate predictions, producing low-frequency outputs with a detail level similar to that of DINOv2's predictions, thereby matching DINOv2's results and optimizing the low-frequency component of our output. This approach allows us to incorporate a large amount of real image data for training without sacrificing detail preservation of generative models, while simultaneously leveraging DINOv2's strong generalization capabilities to enhance the overall robustness and precision of our MDE models.

In summary, the contributions of our work are as follows: 1) We propose an improved approach for transforming generative models into dense prediction models, specifically for depth prediction tasks, by leveraging diffusion model trajectories in a more stable, feed-forward manner; 2) We introduce a novel distillation method that transfers the robust generalization capabilities of models like DINOv2 to diffusion backbones; 3) Our method demonstrates higher stability, accuracy, and efficiency in depth estimation compared to other approaches based on generative models; 4) Compared to other FFN models, our approach achieves more detailed prediction results, marking a significant advancement in the field of MDE.

2. Related Works

2.1. Depth Estimation

Depth estimation has always been a widely researched topic, with numerous studies conducted in the past including single-image [1, 11, 12, 22, 26, 30, 48] and video depth estimation [8, 21, 41, 46, 52]. Early efforts in image depth estimation, such as DIW [6] and OASIS [7], focused on predicting relative (ordinal) depth. Subsequent approaches, including MegaDepth [25] and DiverseDepth [54], used extensive collections of photographs from the Internet to develop models that adapt to unseen data, while MiDaS [32] improved generalization by incorporating a diverse range of datasets during training. Recent advances, including DPT [33] and Omnidata [10], adopted transformer-based architectures to improve depth estimation performance.

However, due to limitations in models and data, these methods exhibit very limited generalization in various open-world scenes. Recently, some efforts [13, 19, 49, 50] in image depth estimation have made open-world depth estimation feasible by leveraging the power of vast amounts of unlabeled image data and the capabilities of pre-trained generative models. Although these methods have achieved remarkable progress in the field of depth estimation, challenges such as low efficiency, limited generalization, and insufficient detail preservation remain.

2.2. Diffusion Models as Representation Learner

The diffusion model training process strongly resembles that of denoising autoencoders (DAE) [2, 17, 43], as both are designed to recover clean images from noise-corrupted inputs. Recent studies have shown that semantic image representations learned from diffusion models can be effectively used for various downstream recognition tasks,

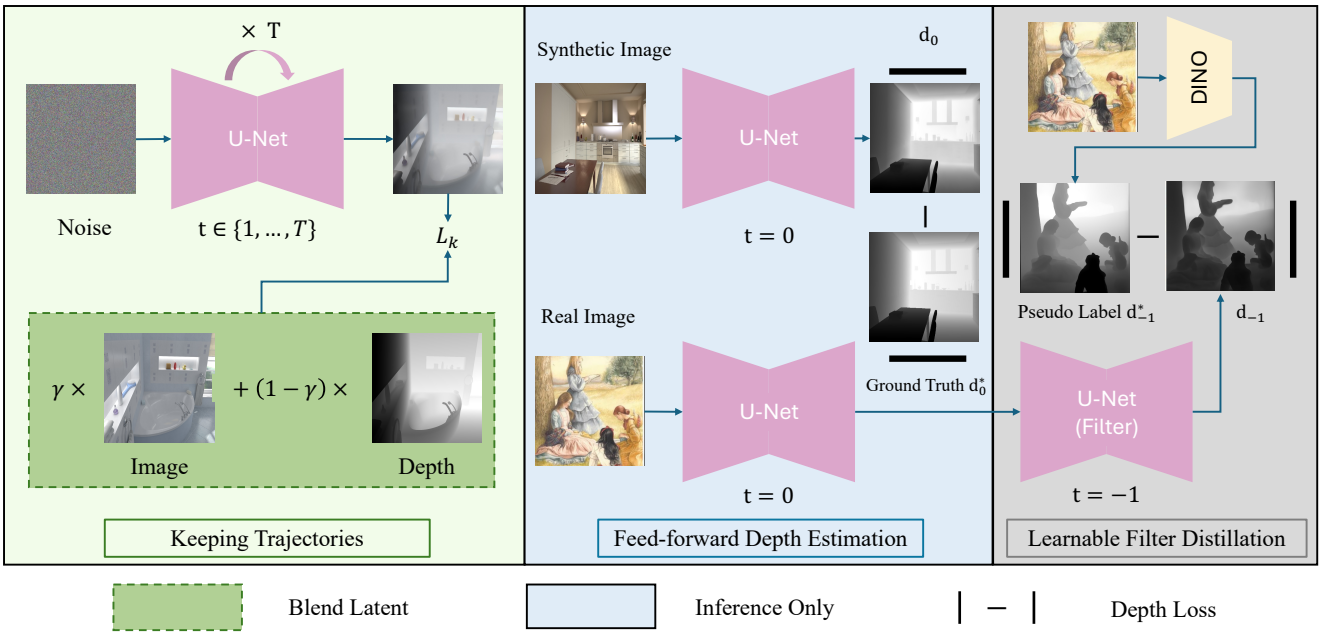

Figure 2. **Overview of the proposed method.** To simplify the representation, all the images we used above correspond to the respective latents. We transform the pre-trained diffusion model into a feed-forward approach for depth prediction, using only the result at $t = 0$ as the output during inference. During training, at $t = 0$, we use synthetic data to ensure detailed results, while at $t = -1$, we leverage pseudo-labels generated by DINOv2 for supervision.

such as correspondence [57], semantic segmentation [59], and keypoint detection [51]. Notably, the features extracted from diffusion models tend to preserve more intricate details, which has prompted the adoption of pre-trained diffusion models for dense prediction tasks. Recent studies, including Marigold [19] and GeoWizard [13], rely on the standard diffusion framework and pre-trained parameters to perform dense prediction tasks. Emerging approaches [16, 47, 53] attempt to bypass the stochastic phase of diffusion models by employing deterministic frameworks. However, these adaptations lack deeper exploration of the model's potential, often leading to suboptimal performance and the need for additional post-processing to refine results. Moreover, diffusion-based techniques generally exhibit limited generalization capabilities. For instance, BetterDepth [58] incorporates external depth priors as inputs but remains a stochastic framework and heavily depends on the quality of other models.

3. Method

3.1. Overview: Feed-forward Transformation of Diffusion Models

Fundamentally, generative models construct mappings between a latent space and the ambient data space. These mappings align closely with the needs of depth estimation and other visual recognition tasks, where precise mappings from image data to the corresponding labels are essential. In these tasks, the scarcity of labeled data often limits the precision of the trained models. In the context of learning from a collection of unlabeled data instances, advanced generative models, trained on massive datasets, are capable of learning robust mappings. They exhibit great promise for transferring knowledge to other visual prediction applications [51, 57, 59]. Our work builds on this approach, further exploring its application in depth estimation.

Specifically, we finetune a well-trained Stable Diffusion (SD) [35] model to construct our model. A diffusion process constructs multiple intermediate states by progressively adding noise to the data $\mathbf{x}_0$, defined as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\bar{\alpha}_t := \prod_{s=1}^{t} 1 - \beta_s$ with noise schedule $\{\beta_1, \ldots, \beta_T\}$. Then, diffusion generative model gradually learns the mapping between two distributions by denoising each step, denoted as $x_T \rightarrow x_{T-1} \rightarrow \cdots \rightarrow x_0$. The diffusion model is primarily a denoising model $\boldsymbol{\epsilon}_\theta$ that follows a loss function $\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(x_t, t)\|^2\right]$.

In depth estimation, the mapping from visual images to depth labels should ideally be deterministic. By leveraging a robust, pre-trained mapping within the generative model, there is no need to decompose the image-label mapping into multiple steps during training. Instead, we can extend the mapping trajectories of the existing diffusion model directly into the depth domain by adapting the learned diffu-

sion process to act as a deterministic one-step feed-forward network. Since our MDE model is built upon the diffusion trajectory for its extension, we set the time step input as $t = 0$ in this feed-forward step.

$$\mathbf{d}_0 = \hat{\boldsymbol{\epsilon}}_\theta \left(\mathbf{x}_0, t = 0\right). \tag{1}$$

Since we use diffusion model's parameters to construct our network, we also use $\hat{\boldsymbol{\epsilon}}_\theta$ to represent our network model here. The input $\mathbf{x}_0$ is the latent representation of RGB image, and $\mathbf{d}_0$ is the latent representation of "depth image," following Marigold's encoding approach. The depth latent $\mathbf{d}_0$ can be reconstructed into a depth map using the VAE of SD with negligible error. While there have been preliminary attempts [16, 47] to use similar approaches, they remain in a nascent stage and lack the precision, robustness, and richness in detail needed for effective depth estimation. In the following sections, we will delve into each step of our method, detailing how we adapt the diffusion model to enhance accuracy and detail in depth estimation tasks, such as the key technical contributions of preserving diffusion trajectories and improving synthetic-to-real robustness , as discussed in Section 3.2 and Section 3.3, respectively. The overall workflow of the method is illustrated in Figure 2.

3.2. Keeping Diffusion Trajectories

Since our approach leverages the trajectory of the diffusion model, it is crucial to prevent degradation of this trajectory during training. To achieve this, when fine-tuning the diffusion model to transition it into a feed-forward depth estimator, we simultaneously maintain the feed-forward step along with the preceding denoising training steps from the original diffusion model. While this trajectory was initially developed for image generation, directly applying it as-is does not facilitate an optimal transition to the depth domain. Thus, instead of predicting purely image-based latents, we modify the target latent to be a blend of image and depth representations.

$$\begin{aligned} \mathbf{b}_0 &= \gamma \mathbf{x}_0 + (1-\gamma)\mathbf{d}_0, \\ \mathbf{b}_t &= \sqrt{\bar{\alpha}_t}\mathbf{b}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t \in \{1, \ldots, T\}. \end{aligned} \tag{2}$$

Here, $\mathbf{b_0}$ represents the blended latent. γ controls the balance between the image and depth latents. In this process, we also use v-prediction re-parameterization approach [36] to define the training objective:

$$\begin{aligned} \mathbf{v}_t &= \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\mathbf{b}_0, \\ L_k &= \left\| \mathbf{v}_t - \hat{\boldsymbol{\epsilon}}_\theta \left(\mathbf{b}_t, t\right) \right\|_2^2, t \in \{1, \ldots, T\}. \end{aligned} \tag{3}$$

Intuitively, this approach forces the diffusion model to preserve the shared features between the image generation task and the depth estimation task, which are captured in the blended training target. Therefore, it allows the diffusion model to adapt more naturally to depth estimation while retaining essential generative features. Consequently, during fine-tuning, the model maintains features that enhance the accuracy and detail of depth predictions. This part is used exclusively during training. At inference time, our model functions as a fully deterministic framework.

3.3. Learnable Filter Distillation

Following previous methods, the above training process only uses synthetic data because it provides high-quality depth Ground Truth. However, this reliance limits both Marigold [19] and our approach, as SD-based MDE models trained solely on synthetic datasets often struggle to generalize well to in-the-wild data. Nevertheless, in this case, the predictions produced by the SD-based MDE model still retain the necessary details—precisely those details we expect to preserve in the final output. Therefore, enhancing the model's robustness essentially means improving the accuracy of the low-frequency components in the model's output on real image.

Inspired by prior work [50], we observe that DINOv2 [29], trained on synthetic data, can generalize effectively to real-world images. However, its depth estimates often lack the necessary fine details—in other words, it can accurately predict the low-frequency depth components for massive real images but misses the high-frequency details. This characteristic aligns well with our needs, so we attempt to leverage the abundance of labels generated by the DINOv2 model to enhance our model's robustness. A straightforward method, as used in Depth Anything v2 [50], is to use a fintuned DINOv2-G depth model to generate pseudo labels, expanding the training set. However, since DINOv2's depth predictions lack fine detail, directly incorporating these pseudo labels on the feed-forward output $\mathbf{d}_0$ risks ignoring the detailed features inherent to our model.

To address this, we propose learning a filter, denoted as $F(d_0)$ that processes our output to remove high-frequency details (e.g., fine details). This filtering produces results at a detail level similar to that of DINOv2, allowing us to focus supervision solely on the less accurate low-frequency components. The filter F is obtained through learning. In fact, we observed that the fine-tuned SD model itself already serves as an effective filter learner because the diffusion model is inherently designed to model the subtle differences across different time steps. Therefore, we directly apply an additional SD step on the feed-forward output $\mathbf{d}_0$ to simulate this filter. To maintain consistency with the inputs and concepts used in the diffusion model, we represent this filter $F : t_0 \rightarrow t_{-1}$ as the process that transforms the output from t_0 to t_{-1}:

$$\mathbf{d}_{-1} = \hat{\boldsymbol{\epsilon}}_\theta \left(\mathbf{d}_0, t = -1\right). \tag{4}$$

At this stage, we can use the labels predicted by DINOv2

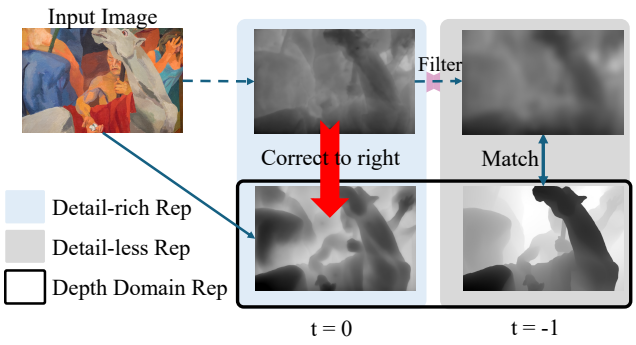

Figure 3. **Filter learning.** We use a learnable filter to map our results to detail levels similar to DINOv2's, matching its outputs and thereby transferring DINOv2's generalization capabilities to our model without compromising our inherent details.

to supervise $\mathbf{d}_{-1}$, allowing us to transfer DINOv2's robustness without interfering with the detailed features in $\mathbf{d}_0$. In this process, we use real-world image data for $\mathbf{x}_0$. Figure 3 illustrates the effects of Eq. (4).

3.4. Final Objective

For $\mathbf{d}_0$ and $\mathbf{d}_{-1}$, we follow the method of MiDaS [32] using MAE loss L_{MAE} and gradient matching loss L_{GM} as depth loss. However, unlike the original approach, we apply these losses in the latent space of Stable Diffusion instead:

$$L_{\text{MAE}}\left(\mathbf{d}, \mathbf{d}^*\right) = \frac{1}{M} \sum_{i=1}^{M} |\mathbf{d}_i - \mathbf{d}_i^*|, \tag{5}$$

where $\mathbf{d}$ represents the ground truth latent, and $\mathbf{d}^*$ is the model's predicted value. M denotes the number of pixels in the depth latent.

$$\mathcal{L}_{\text{GM}}\left(\mathbf{d}, \mathbf{d}^*\right) = \frac{1}{M} \sum_{i=1}^{M} \left(|\nabla_x R_i| + |\nabla_y R_i|\right), \tag{6}$$

where $R_i = \mathbf{d}_i - \mathbf{d}_i^*$. Hence, the final objective is expressed as a weighted sum of the losses L_{MAE}, L_{GM}, and L_k:

$$L_{\text{final}} = \sum_{t \in \{-1,0\}} \left(\lambda_{\text{MAE}} L_{\text{MAE}}\left(\mathbf{d}_t, \mathbf{d}_t^*\right) + \lambda_{\text{GM}} L_{\text{GM}}\left(\mathbf{d}_t, \mathbf{d}_t^*\right)\right) + \lambda_k L_k. \tag{7}$$

λ_{MAE} λ_{GM}, and λ_k are the weighting factors for their respective loss term. $\mathbf{d}_0^*$ represents the ground truth (GT) from the synthetic dataset, and $\mathbf{d}_{-1}^*$ represents the pseudo label generated by DINOv2 for real images.

4. Experiments

4.1. Implementation Details

During the training process, we preserve the diffusion trajectory while following the original DDPM noise scheduler [18] using 1000 diffusion steps. To better leverage pre-trained models, we use the Depth Anything V2-Large model as the DINO v2 model for supervision. This choice was made because DAv2-Giant has not yet released the weights, so we can only use other versions. Our training dataset consists of two parts. For training at $t = 0$ and during trajectory retention, we follow previous approaches and use two synthetic datasets, Hypersim [34] and Virtual KITTI [5], which cover both indoor and outdoor scenes, with a total of 74K images. For training at $t = -1$, we use real-world data from the LAION-Art dataset, a subset of LAION-5B [38] containing 8 million samples. However, we observed that training with only 0.2 million samples was sufficient. Synthetic data and real data each account for half of each batch. The parameters are set as follows: $\gamma = 0.5$, $\lambda_{\text{MAE}} = 1$, $\lambda_{\text{GM}} = 0.5$, $\lambda_k = 0.2$.

4.2. Comparison

Zero-shot affine-invariant depth. For the evaluation of affine-invariant depth, we use the same datasets and evaluation protocol as Marigold. These datasets include NYUv2 [40], ScanNet [9], KITTI [14], ETH3D [37], and DIODE [42]. We compared FiffDepth with 14 methods that produce affine-invariant depth maps/disparities, all claiming zero-shot generalization capabilities. These include the earlier methods [10, 32, 33, 54–56] , as well as the more recent ones [13, 15, 16, 19, 27, 47, 49, 50]. As shown in Table 1, FiffDepth achieves the best or state-of-the-art comparable results in most test scenarios. For visualization results, please refer to Figure 4. Our method not only accurately predicts the relative depth relationships but also excels in identifying and predicting depth for very fine objects. We also evaluate our method on the DA-2K introduced by Depth Anything v2. On this dataset, our method also performs comparably to Depth Anything v2.

Compared to methods like Depth Anything, which rely on massive training datasets, our model achieves comparable generalization to DAv2 while being trained on only a small amount of real data. To validate the generalization capability of our method, we present some test examples from special scenarios in Figure 5, including games, artworks, AI-generated content, and movies. Our method shows comparable generalization to DAv2 while preserving more details. In contrast, other methods' results are unsatisfactory in both generalization and detail preservation.

Zero-shot boundaries. To further demonstrate the accuracy of our method in predicting fine structures, we also employ the Zero-shot Boundaries Metric introduced in the

Method	Training Data	NYUv2		KITTI		ETH3D		ScanNet		DIODE-Full		DA-2K
		AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	Acc (%)
DiverseDepth	320K	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	79.3
MiDaS	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	80.6
LeReS	354K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	81.1
Omnidata v2	12.2M	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	76.8
HDN	300K	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	78.0	85.7
DPT	1.4M	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	**18.2**	75.8	83.2
Marigold	74K*	5.5	96.4	9.9	91.6	6.4	96.0	6.4	95.1	30.8	77.3	86.8
e2e-ft	74K*	5.2	96.6	9.6	91.9	6.4	95.9	5.8	96.2	30.2	77.9	83.6
DepthFM	74K*	6.5	95.6	8.3	93.4	7.8	95.9	6.8	94.9	24.5	74.1	85.8
GenPercept	74K*	5.6	96.0	13.0	84.2	7.0	95.6	6.2	96.1	30.7	77.6	85.1
Lotus-D	59K*	5.3	96.7	8.1	92.8	6.5	95.3	5.8	96.3	29.9	78.1	86.8
Lotus-G	59K*	5.4	96.6	8.5	87.7	**6.2**	96.1	6.0	96.0	29.4	78.5	86.2
GeoWizard	280K*	5.2	96.6	9.7	92.1	6.4	96.1	6.1	95.3	29.7	**79.2**	88.1
DepthAnything v1-L	62.6M*	**4.3**	**98.1**	7.6	**94.7**	12.7	88.2	**4.2**	**98.0**	27.7	75.9	88.5
DepthAnything v2-L	62.6M*	4.5	97.9	7.4	94.6	13.1	86.5	**4.2**	97.8	26.2	75.4	**97.1**
FiffDepth (Ours)	274K*	4.4	97.8	**7.3**	93.5	7.1	**97.2**	**4.2**	97.9	23.9	78.1	**97.1**

Table 1. Quantitative comparison with other affine-invariant depth estimators on several zero-shot benchmarks. We use AbsRel (absolute relative error: $|d^* - d| / d$) and δ_1 (percentage of $\max(d^*/d, d/d^*) < 1.25$). All metrics are reported as percentages; **bold numbers** are the best, underscored second best. Methods marked with an asterisk (*) utilize pre-trained models.

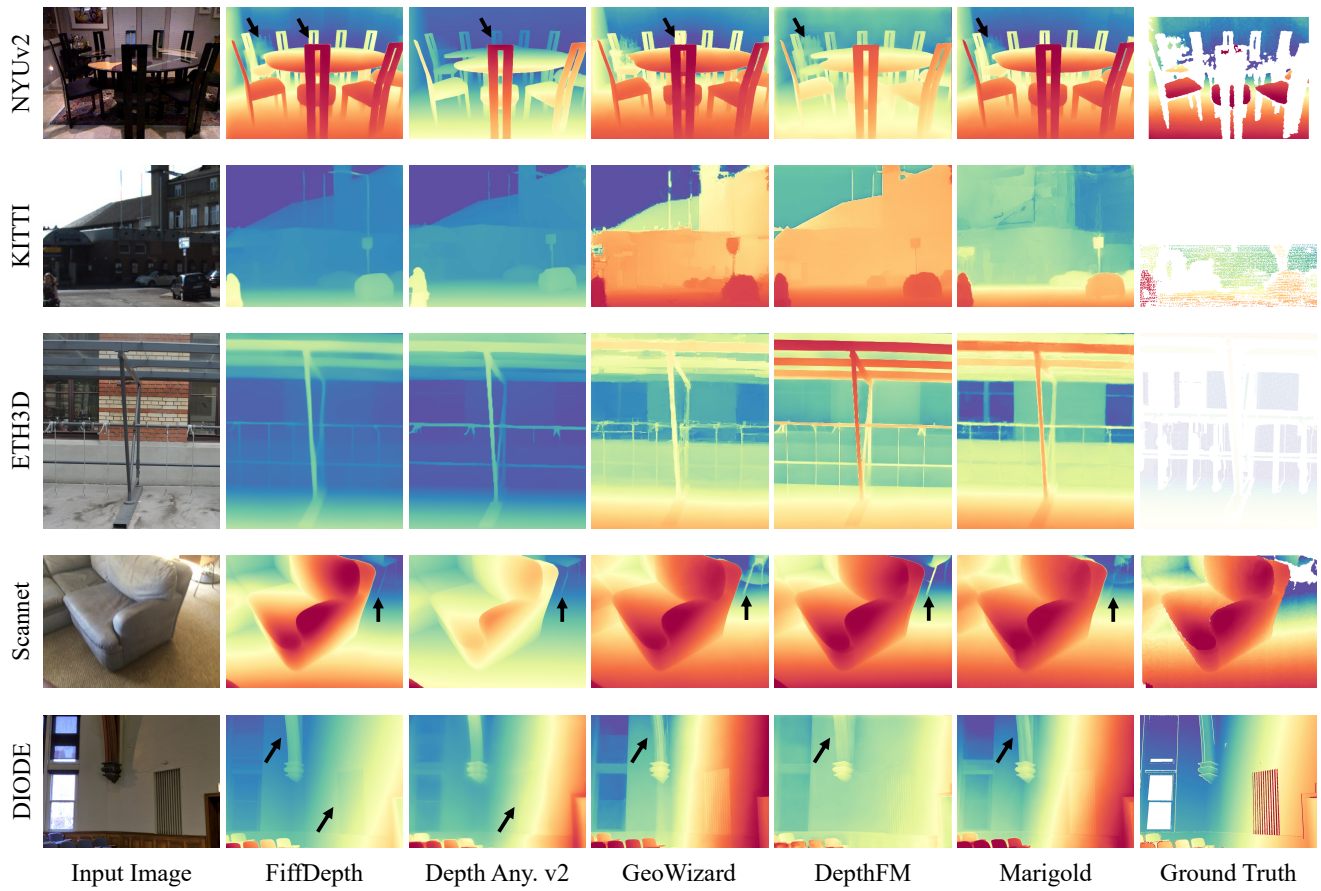

Figure 4. **Qualitative comparison across different datasets.** Our method is capable of predicting the depth of various fine objects, such as lampposts, railings, and chair legs.

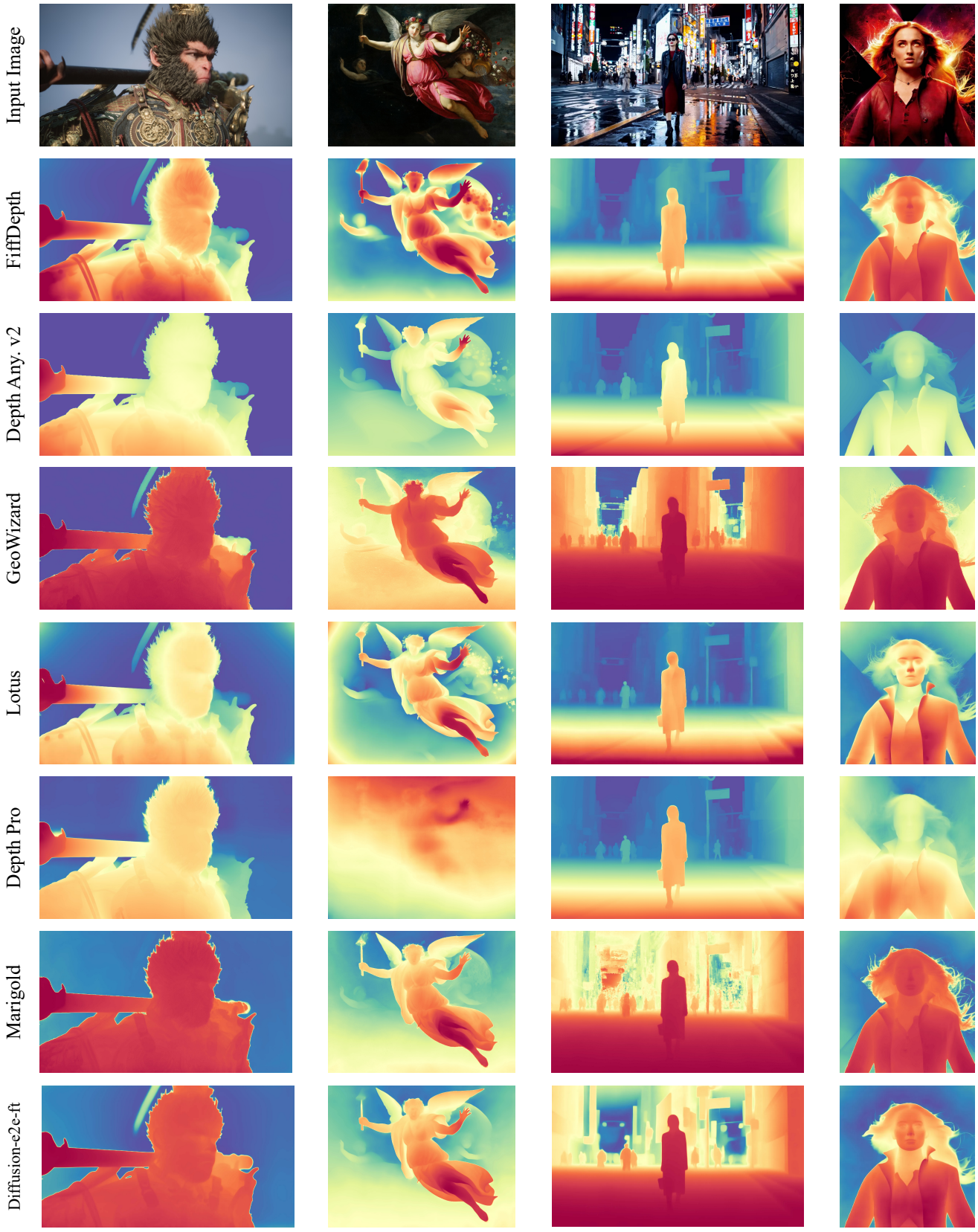

Figure 5. **Qualitative comparison on special scenarios.** In the special scenarios of games, artworks, AI-generated content, and movies, our method demonstrates strong generalization capability and the ability to predict detailed depth.

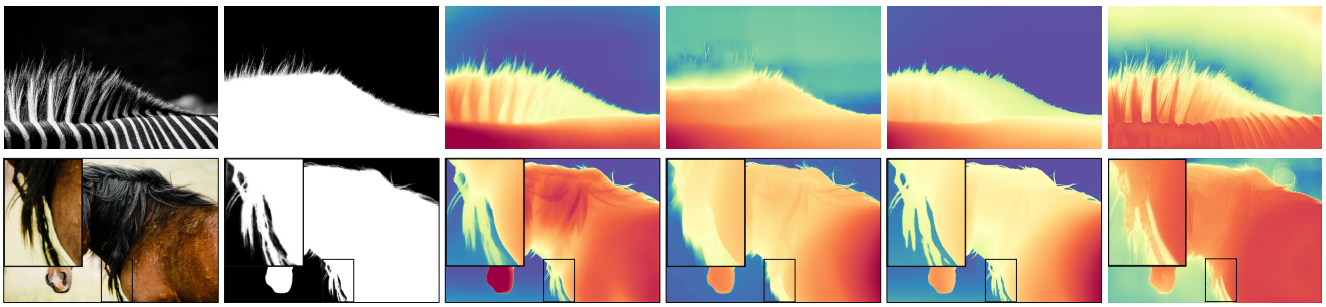

Figure 6. **Boundary visualization comparison.** These samples are from the AM-2k dataset.

Method	Sintel F1↑	Spring F1↑	iBims F1↑	AM R↑	P3M R↑	DIS R↑
DepthAnything v2	0.228	0.056	0.111	0.107	0.131	0.056
Depth Pro	0.409	0.079	0.176	0.173	0.168	0.077
FiffDepth (Ours)	**0.423**	**0.086**	**0.189**	**0.176**	**0.179**	**0.091**

Table 2. **Zero-shot boundary accuracy.** We provide the F1 score for datasets containing ground-truth depth and boundary recall (R) for those with matting or segmentation labels.

Method	Marigold	Marigold (LCM)	GeoWizard	DepthFM	DepthAnything v2-L	Depth Pro	Ours
Time (s)	103	1.7	19	0.39	0.026	0.23	0.092

Table 3. **Running time comparison.** We performe inference on 100 512×512 images using these methods and report the average time.

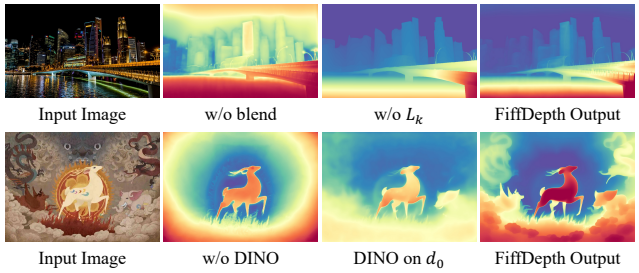

Figure 7. **Ablation studies.** The generalization capability and depth details of the method are affected when some essential components are missing.

recent work Depth Pro [3] to evaluate boundary sharpness. Following Depth Pro, we compute the depth average boundary F1 score for datasets with ground truth and the boundary recall (R) for datasets with matting or segmentation annotations. The former datasets include Sintel [4], Spring [28], and iBims [20], while the latter include AM-2k [24], P3M-10k [23], and DIS-5k [31]. For details on the boundaries metric and its computation, please refer to the Depth Pro paper for further details. Quantitative comparisons in Table 2 demonstrate that our method surpasses Depth Pro and other approaches in boundary prediction. Additionally, the visual results in Figure 6 further validate that our method predicts more accurate boundaries. Please refer to the **supplementary materials** for the results of other methods.

Running time. Generative MDE models adopt the diffusion paradigm, and the resulting instability necessitates test-time assembly, leading to a lengthy inference time. In contrast, our feed-forward approach provides significant efficiency advantages. We evaluate the average inference time for a 512×512 image on an NVIDIA Titan RTX GPU. As shown in Table 3, our method significantly outperforms other generative approaches in terms of efficiency and achieves performance comparable to DAv2. We tested these methods using their default settings.

Ablation studies. We conduct ablation studies to validate components of our method. Keeping the diffusion trajectory but predicting purely image latents affects the relative depth relationships between objects (1st row, 1st result in Fig. 7). Without keeping trajectory, some details are lost (1st row, 2nd result in Fig. 7). Omitting DINO supervision impacts the model's generalization ability (2nd row, 1st result in Fig. 7). Using DINO supervision at d_0 also reduces details (2nd row, 2nd result in Fig. 7). Please refer to **supplementary materials** for **quantitative** ablation studies.

5. Conclusion

In this work, we transform diffusion models into stable, feed-forward depth estimators, achieving significant improvements in accuracy and efficiency over generative model-based methods. By combining the detail preservation of generative models with the robust generalization of FFN models like DINOv2, our hybrid approach bridges the synthetic-to-real gap, enhancing stability, predictability, and resolution in MDE for diverse real-world scenarios.

References

[1] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11746–11752. IEEE, 2021. 2

[2] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *Advances in neural information processing systems*, 26, 2013. 2

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 8

[4] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 8

[5] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 5

[6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 2

[7] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 679–688, 2020. 2

[8] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7063–7072, 2019. 2

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5

[10] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 2, 5

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 2

[13] 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. 2, 3, 5

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5

[15] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 5

[16] 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. 3, 4, 5

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[19] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 5

[20] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 8

[21] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 2

[22] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[23] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021. 8

[24] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 8

[25] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

[26] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, 2023. 2

[27] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 5

[28] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. 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*, pages 4981–4991, 2023. 8

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4

[30] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022. 2

[31] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 8

[32] 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. 1, 2, 5

[33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1, 2, 5

[34] 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. 5

[35] 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. 2, 3

[36] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 4

[37] 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. 5

[38] 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. 5

[39] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 1

[40] 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. 5

[41] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 2

[42] 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. 5

[43] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[44] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 1

[45] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. 1

[46] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9466–9476, 2023. 2

[47] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024. 3, 4, 5

[48] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279, 2021. 2

[49] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 5

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 2, 4, 5

[51] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 3

[52] Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. Mamo: Leveraging memory and attention for monocular video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8754–8764, 2023. 2

[53] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *arXiv preprint arXiv:2406.16864*, 2024. 3

[54] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 2, 5

[55] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

[56] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35:14128–14139, 2022. 5

[57] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[58] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. *arXiv preprint arXiv:2407.17952*, 2024. 3

[59] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3