DepthFM: Fast Generative Monocular Depth Estimation with Flow Matching

Ming Gui[1*], Johannes Schusterbauer[1*], Ulrich Prestel[1], Pingchuan Ma[1], Dmytro Kotovenko[1], Olga Grebenkova[1], Stefan Andreas Baumann[1], Vincent Tao Hu[1], Björn Ommer[1]

[1] CompVis @ LMU Munich, Munich Center for Machine Learning

Abstract

Current discriminative depth estimation methods often produce blurry artifacts, while generative approaches suffer from slow sampling due to curvatures in the noise-to-depth transport. Our method addresses these challenges by framing depth estimation as a direct transport between image and depth distributions. We are the first to explore flow matching in this field, and we demonstrate that its interpolation trajectories enhance both training and sampling efficiency while preserving high performance. While generative models typically require extensive training data, we mitigate this dependency by integrating external knowledge from a pre-trained image diffusion model, enabling effective transfer even across differing objectives. To further boost our model performance, we employ synthetic data and utilize image-depth pairs generated by a discriminative model on an in-the-wild image dataset. As a generative model, our model can reliably estimate depth confidence, which provides an additional advantage. Our approach achieves competitive zero-shot performance on standard benchmarks of complex natural scenes while improving sampling efficiency and only requiring minimal synthetic data for training.

Code — https://github.com/CompVis/depth-fm

1 Introduction

Monocular depth estimation is pivotal for 3D scene understanding due to its numerous applications, ranging from core vision tasks such as segmentation (He et al. 2021) and visual synthesis (Zhang, Rao, and Agrawala 2023) to application areas like robotics and autonomous driving (Cabon, Murray, and Humenberger 2020; Geiger et al. 2013). Despite the recent strides in this field, estimating realistic geometry from a single image remains challenging. State-of-the-art discriminative depth estimation models exhibit impressive overall performance but still lack fine-grained high-frequency details (Yang et al. 2024a,b). Current generative-based methods (Ke et al. 2024; Fu et al. 2024) address this issue by phrasing depth estimation as an image-conditional iterative denoising process using diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021). Although they can generate realistic and accurate depth maps, these methods suffer from extremely long inference times because of the integration over a highly curved ordinary differential equation (ODE) trajectory.

Flow Matching (FM) (Lipman et al. 2023; Liu, Gong, and Liu 2023; Albergo et al. 2023; Albergo and Vanden-Eijnden 2022; Neklyudov, Severo, and Makhzani 2022) is an attractive alternative paradigm. These methods emerged as a strong competitor to the currently prominent Diffusion Models (DM) (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2021) and offer enhanced flexibility in trajectory design and starting distribution. While diffusion models offer samples of great diversity, the curved diffusion trajectories through solution space entail high computational costs. Conversely, the straighter trajectories of flow matching result in much faster processing (Lipman et al. 2023; Lee, Kim, and Ye 2023). We hypothesize that these characteristics of flow matching are a much better fit for image-based depth estimation, in contrast to diffusion models. To further enhance training and inference efficiency, we use data-dependent couplings (Schusterbauer et al. 2024) for our model, which we call *DepthFM*. Unlike conventional diffusion models that start the generative process from noise and end with a depth map, our method directly models the trajectory from image to depth space.

However, there are several challenges in training efficient generative depth estimation models. First, the computational requirements for training generative models are extremely high (Zhang et al. 2024). Second, annotating depth is very difficult (Geiger, Lenz, and Urtasun 2012), making data efficiency a critical issue. To address these problems, we propose to seek external knowledge from a pre-trained *image diffusion model* and a pre-trained *discriminative depth estimation model*. We incorporate a strong image prior from unsupervised generative training together with a strong depth prior from a pre-trained discriminative model while preserving the advantages inherent to the generative approach. We augment our model with prior information by fine-tuning our approach from an image synthesis foundation model, specifically, SD2.1 (Rombach et al. 2022). We show the feasibility of transferring information between DM and FM by fine-tuning a flow matching model from a diffusion model prior. This provides our model with initial visual cues and significantly speeds up training. It also allows us to train

[*]These authors contributed equally.

Figure 1: We present *DepthFM*, a high-fidelity, fast, and flexible generative monocular depth estimation model.

exclusively on a small amount of *synthetic data* and still achieve robust generalization to real-world images.

In summary, to improve the sampling, training, and data efficiency in generative depth estimation, our contributions are as follows:

- We introduce DepthFM, a versatile and fast generative model for monocular depth estimation. We are the first to formulate monocular depth estimation as a direct transport problem, represented via flow matching. By utilizing more flexible trajectory and distribution choices compared to diffusion models, DepthFM enhances sampling efficiency, leading to more efficient depth estimation.
- To enhance training and data efficiency, we leverage external knowledge from both the generative and discriminative communities. We successfully transfer a robust image prior from a pre-trained diffusion model to a flow matching model, significantly reducing reliance on training data. Additionally, we demonstrate that a pre-trained, off-the-shelf discriminative model can further boost the performance of generative depth estimation.
- Ultimately, our findings show that flow matching is highly efficient and capable of synthesizing depth maps in a single inference step. Despite being trained solely on synthetic data, DepthFM delivers competitive or state-of-the-art performance on benchmark datasets and natural images. Additionally, our method demonstrates state-of-the-art performance in depth completion.

2 Related Works

Depth estimation can be broadly divided into discriminative and generative methods. Prominent examples of discriminative methods include MiDaS (Ranftl et al. 2020), Depth Anything (Yang et al. 2024a,b), and Metric3D (Yin et al. 2023; Hu et al. 2024a).

Recently, generative diffusion-based models have been explored for affine-invariant depth estimation (Ke et al. 2024; Fu et al. 2024). These models exploit the rich knowledge embedded in large-scale vision foundation models to produce high-quality depth maps. Despite their promising results, diffusion-based methods face significant challenges. First, they suffer from slow sampling times, even when using ODE approximations to solve the underlying SDEs. Second, the diffusion formulation relies on a Gaussian source distribution (Sohl-Dickstein et al. 2015; Song and Ermon 2019), which may not fully capture the natural relationship between images and their corresponding depth maps.

In contrast, flow matching-based models (Lipman et al. 2023; Liu, Gong, and Liu 2023; Albergo et al. 2023) have demonstrated promise across various tasks and offer faster sampling speeds (Liu, Gong, and Liu 2023). Additionally, optimal transport can be achieved even when source distribution deviates from a Gaussian distribution, which is advantageous for certain tasks (Tong et al. 2023a). Our work takes a first step in using flow matching for monocular depth estimation, aiming to reduce sampling costs by leveraging the inherent straight sampling trajectory (Lee, Kim, and Ye 2023). Furthermore, we seek to combine the strengths of both discriminative and generative depth estimation by designing a simple knowledge distillation pipeline and transferring knowledge from discriminative models to enhance generative depth estimation.

To the best of our knowledge, no prior work has explored the use of flow matching to facilitate the distribution transfer between images and depth maps, despite their intuitive proximity compared to noise and depth maps. In addition, we aim to improve training efficiency by incorporating insights from diffusion models, exploiting the similarity of their objectives (Lee, Kim, and Ye 2023). We discuss additional related work in the appendix.

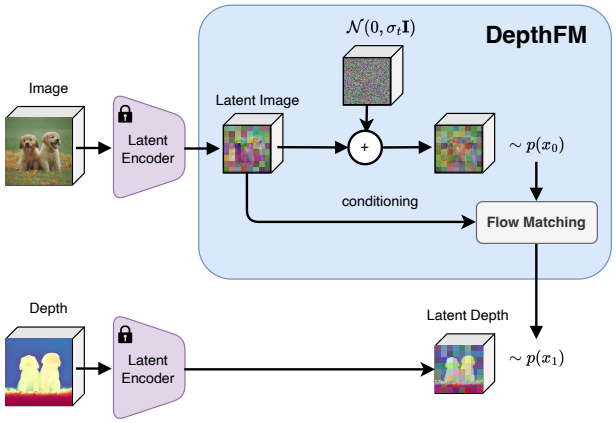

Figure 2: Overview of our training pipeline. We use flow matching to regress the vector field between the image latent x_0 and the corresponding depth latent x_1.

3 Methodology

Background: Flow Matching

Flow Matching (Lipman et al. 2023; Liu, Gong, and Liu 2023; Albergo et al. 2023; Neklyudov, Severo, and Makhzani 2022) belongs to the category of generative models designed to regress vector fields based on fixed conditional probability paths. Denote $\mathbb{R}^d$ as the data space with data points x and $u_t(x) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ the time-dependent vector field, which defines the ODE $dx = u_t(x)dt$. Let $\phi_t(x)$ represent the solution to this ODE with the initial condition $\phi_0(x) = x$.

The probability density path $p_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}_{>0}$ characterizes the probability distribution of x at timestep t with $\int p_t(x)dx = 1$. According to the continuity condition, the pushforward function $p_t = [\phi_t]_{\#}(p_0)$ then transports the probability density path p along u from timestep 0 to t.

Lipman et al. (2023) showed that we can efficiently train a neural network using the conditional flow matching objective, to regress conditional vector fields $u_t(x|x_1)$ by sampling $p_t(x|x_1)$:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t, p_{data}(x_1), x \sim p_t(x|x_1)} ||v_\theta(t,x) - u_t(x|x_1)||. \quad (1)$$

The most common approach to define $u_t(x|x_1)$ is as a straight path from x_0 to x_1 (Liu, Gong, and Liu 2023; Lipman et al. 2023). Assuming that p_0 is a standard Gaussian with $x_0 = \epsilon \sim \mathcal{N}(0, \mathbb{I})$, the intermediate interpolant is defined as $x_t = tx_1 + (1-t)\epsilon$, and a valid vector field that satisfies the probability path is given by $u_t(x|x_1) = \frac{x_1 - x}{1-t}$.

Data Coupling in FM for Depth Estimation

The monocular depth estimation task involves training a function that uses image conditioning to generate depth. Let us denote the image in pixel space X_0 and the corresponding depth in pixel space X_1.

Flow in Latent Space In order to reduce the computational demands associated with training FM models for high-resolution depth estimation synthesis, we follow (Rombach et al. 2022; Dao et al. 2023; Schusterbauer et al. 2024; Hu et al. 2024b; Ke et al. 2024) and utilize an autoencoder model that provides a compressed latent space that aligns perceptually with the image pixel space. This approach also facilitates the direct inheritance of a robust model prior obtained from foundational LDMs such as Stable Diffusion. Similar to (Ke et al. 2024) we move both modalities (i.e., RGB images and depths) to the latent space. Without further mentioning, $x_0 = \texttt{Enc}(X_0)$, $x_1 = \texttt{Enc}(X_1)$ take place in the latent space. We can accordingly use the Decoder ($\texttt{Dec}$) to decode them back to image and depth space. Note that we employ a different normalization strategy compared to Marigold (Ke et al. 2024) in that we use log-scaled depth which ensures a more balanced capacity allocation for both indoor and outdoor scenes. Further details, including explanations and ablation studies, can be found in the appendix.

Direct Transport between Image and Depth Let (x_0, x_1) be ground truth image-depth feature pairs. In contrast to diffusion-based depth estimators (Ke et al. 2024; Fu et al. 2024), which map Gaussian noise ϵ with image conditioning to depth x_1 by $p(x_1|\epsilon; x_0)$, we formulate depth estimation as a direct distribution transport between the image feature x_0 and the depth feature x_1 by $p(x_1|x_0)$, which can be effectively solved using conditional flow matching (Tong et al. 2023b; Albergo et al. 2023). Specifically, we model the intrinsic transport relation between the image feature x_0 and the depth feature x_1. As shown in Table 1, using paired coupling between image and depth maps results in a far shorter transport path than mapping directly from Gaussian noise to depth maps. The transport between the image and the depth distribution can be defined as

$$x_t \sim p_t(x|(x_0, x_1)) = \mathcal{N}(x|tx_1 + (1-t)x_0, \sigma^2_{\min}\mathbb{I}), \quad (2)$$

$$u_t(x|(x_0, x_1)) = x_1 - x_0, \quad (3)$$

where $u_t(x|(x_0, x_1))$ is the vector field that transports x_0 along space to the marginal distribution $p_t(x|(x_0, x_1))$. To avoid singularity problems, we smooth both $p(x_0)$ and $p(x_1)$ with a minimum smoothing factor $\sigma_{\min}$ to obtain the corresponding data distributions $\mathcal{N}(x_0, \sigma^2_{\min})$ and $\mathcal{N}(x_1, \sigma^2_{\min})$.

Despite the different modalities and data manifolds of x_0 and x_1, the optimal transport condition between $p(x_0)$ and $p(x_1)$ is inherently satisfied due to image-to-depth pairs. This addresses the dynamic optimal transport problem in the transition for image-to-depth translation within the flow matching paradigm, ensuring more stable and faster training (Tong et al. 2023a). The loss thus takes the form:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1], (x_0, x_1) \sim \mathcal{D}^{GT}} ||v_\theta(t, x_t; \bar{x}) - (x_1 - x_0)||. \quad (4)$$

Here, $\bar{x}$ represents a clean copy of the image latent feature x_0 and serves as additional conditioning, $\mathcal{D}^{\text{GT}}$ refers to an image depth dataset. Notably, we improve the transport by injecting the image feature signal into the model, formulating it as $v_\theta(t, x_t; \bar{x})$ instead of $v_\theta(t, x_t)$.

Noise Augmentation Noise augmentation, first used in cascaded diffusion models (Ho et al. 2022), improves the

Coupling Pattern	EMD-$L_2 \downarrow$	EMD-$L_1 \downarrow$
Random	0.981	0.974
Paired (*Ours*)	**0.686**	**0.691**

Table 1: Paired (Image to Depth) Coupling is better than the Random Coupling. EMD = Earth Mover's Distance.

performance of super-resolution models by adding Gaussian noise to the low-resolution conditioning signal. We extend this technique and apply Gaussian noise augmentation to the terminal distribution at $t = 0$. We define the terminal data points x_0 as a weighted combination between the image feature $\bar{x}$ and the Gaussian noise ϵ: $x_0 := \sqrt{\bar{\alpha}_{t_s}}\bar{x} + \sqrt{1-\bar{\alpha}_{t_s}}\epsilon$, where $\bar{\alpha}_{t_s} \in [0,1]$ can be any predefined variance-preserving noise schedule (Ho, Jain, and Abbeel 2020; Song et al. 2021; Nichol and Dhariwal 2021), and t_s is a hyperparameter. We choose to augment the feature in a way that preserves the variance of the data point. We hypothesize that adding a small amount of Gaussian noise will smooth the base probability density and keep it well-defined over a larger manifold. In addition, adding this stochasticity allows us to use it in the sampling process for further applications, such as indicating confidence. In particular, unlike previous work on noise augmentation using noise-image pairs, we empirically show its feasibility on image-depth pairs, which has not been explored before.

Dual Knowledge Transfer for Better Training and Data Efficiency

Training depth estimation models faces significant challenges, including high computational demands during training (Zhang et al. 2024) and a lack of high-quality annotated depth data (Geiger, Lenz, and Urtasun 2012). To address these issues, we propose to leverage external knowledge from pre-trained image diffusion models and pre-trained discriminative depth estimation models and combine their strengths to improve training efficiency and performance.

Image Prior for Training Efficiency Intuitively, a visual synthesis model that generates sound images must also have some notion of the inherent depth of a scene. Similar to (Ke et al. 2024), we use a large pre-trained generative image model that has been trained on a large amount of data, and use this prior knowledge to infer depth. Diffusion models can be trained with different parameterizations, including the x, ϵ, and v parameterizations. A model parametrized with v regresses the *velocity* between samples from the two terminal distributions (Salimans and Ho 2022). In the context of FM, where the two terminal distribution samples are denoted as x_0 and x_1, the objective of the v parameterization can be mathematically formulated as $v = \alpha_t x_0 - \sigma_t x_1$, where α_t and σ_t is the fixed diffusion schedule. In comparison, our FM objective regresses a vector field of $\mathbf{v} = x_1 - x_0$. The similarity between the DM and FM objectives allows us to fine-tune the pre-trained diffusion model with the conditional flow matching loss and use the strong image prior to speed up convergence for generative depth estimation.

Depth Prior for Data Efficiency Generative depth estimation models (Ke et al. 2024; Fu et al. 2024) offer high visual fidelity but often lag behind discriminative models in quantitative metrics (Hu et al. 2024a; Yang et al. 2024a). While the high performance of discriminative methods is often due to training on a large amount of data, the higher fidelity of generative models is due to the ability to sample from the conditional posterior instead of the mean prediction of discriminative counterparts. To be data efficient while still producing high-fidelity depth maps, we integrate the strengths of both model types. Using a discriminative depth model as a teacher, we improve our generative model, increasing both robustness and performance.

In detail, let us denote the discriminative monocular depth estimation teacher model as T. We use this model to predict depth on an unlabeled in-the-wild image dataset $\hat{\mathcal{D}}^u$ and generate discriminative samples $\mathcal{D}^u$ in the form of $\mathcal{D}^u = \{(u_i, T(u_i)) | u_i \in \hat{\mathcal{D}}^u\}$. The loss $\mathcal{L}$ can be formulated as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(x_0,x_1)\sim\mathcal{D}} ||v_\theta((t,x_t);\bar{x}) - (x_1 - x_0)||, \tag{5}$$

where $\mathcal{D} = \mathcal{D}^{\text{GT}} \cup \texttt{part}(\mathcal{D}^u, k)$ and $\texttt{part}()$ is the partition function that randomly selects samples from $\mathcal{D}^u$ such that the number of selected samples corresponds to $|\texttt{part}(\mathcal{D}^u, k)| = k \cdot |\mathcal{D}^{\text{GT}}|$.

4 Experiments

Training and Evaluation Details

We train our depth estimation model on two synthetic datasets, Hypersim (Roberts et al. 2021) and Virtual KITTI (Cabon, Murray, and Humenberger 2020) to cover both indoor and outdoor scenes. Following (Ke et al. 2024) we take 54K training samples from Hypersim and 20K training samples from Virtual KITTI. We leverage Metric3D v2 (Hu et al. 2024a), as our teacher model. We apply this model on a subset of the general-purpose image dataset *Unsplash* (Unsplash 2023), to generate image-depth pairs.

We perform zero-shot evaluations on established real-world depth estimation benchmarks NYUv2 (Nathan Silberman and Fergus 2012), KITTI (Behley et al. 2019), ETH3D (Schops et al. 2017), ScanNet (Dai et al. 2017), and DIODE (Vasiljevic et al. 2019). Let d be the ground truth depth and $\hat{d}$ be the predicted depth. The evaluation metrics are the Absolute Mean Relative Error (**RelAbs**), calculated as $\frac{1}{M}\sum_{i=1}^{M} |d_i - \hat{d}_i|/d_i$ and **δ1 accuracy**, which measures the ratio of all pixels satisfying $\max(d_i/\hat{d}_i, \hat{d}_i/d_i) < 1.25$. Unless otherwise specified, we evaluate our model using an ensemble size of 10 and 4 Euler steps, and scale and shift our predictions to match the ground truth depth in log space.

Zero-shot Depth Estimation

Main Quantitative Result Table 2 compares our model quantitatively with state-of-the-art depth estimation methods. Unlike other approaches that often require large training datasets, our method leverages the rich knowledge of a diffusion-based foundation model (*image prior*) and a discriminative teacher model (*depth prior*). This strategy reduces the computational burden while emphasizing our ap-

	Method	#Train samples		NYUv2		KITTI		ETH3D		ScanNet		DIODE	
		Real	Synthetic	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Discriminative	MiDaS (Ranftl et al. 2020)	2M	—	0.111	88.5	0.236	63.0	0.184	75.2	0.121	84.6	0.332	71.5
	Omnidata (Eftekhar et al. 2021)	11.9M	301K	0.074	94.5	0.149	83.5	0.166	77.8	0.075	93.6	0.339	74.2
	HDN (Zhang et al. 2022)	300K	—	0.069	94.8	0.115	86.7	0.121	83.3	0.080	93.9	0.246	78.0
	DPT (Ranftl, Bochkovskiy, and Koltun 2021)	1.2M	188K	0.098	90.3	0.100	90.1	0.078	94.6	0.082	93.4	0.182	75.8
	DA (Yang et al. 2024a)	1.5M	62M	0.043	98.1	0.076	94.7	0.127	88.2	—	—	0.066	95.2
	DAv2 (Yang et al. 2024b)	—	595K+62M	0.044	97.9	0.075	94.8	0.132	86.2	—	—	0.065	95.4
	Metric3D v2 (Hu et al. 2024a)	25M	91K	0.043	98.1	0.044	98.2	0.042	98.3	0.022†	99.4†	0.136	89.5
Generative	Marigold (Ke et al. 2024)	—	74K	0.055	96.4	0.099	91.6	0.065	96.0	0.064	95.1	0.308	77.3
	GeoWizard (Fu et al. 2024)	—	280K	0.052	96.6	0.097	92.1	0.064	96.1	0.061	95.3	0.297	79.2
	DepthFM-I	—	74K	0.060	95.5	0.091	90.2	0.065	95.4	0.066	94.9	0.224	78.5
	DepthFM-ID	—	74K+7.4K	0.055	96.3	0.089	91.3	0.058	96.2	0.063	95.4	0.212	80.0

Table 2: Quantitative comparison with affine-invariant depth estimators on *zero-shot* benchmarks. $\delta1$ is presented in percentage. Our method shows competitive performance across datasets. DepthFM-I and DepthFM-ID refer to our model trained with image prior and image-depth prior, respectively. DA stands for the Depth Anything model family. Some baselines are sourced from Marigold (Ke et al. 2024) and GeoWizard (Fu et al. 2024). State-of-the-art discriminative models, which heavily rely on *extensive* amounts of annotated training data, are listed in the upper part of the table. †: Models are trained with normals.

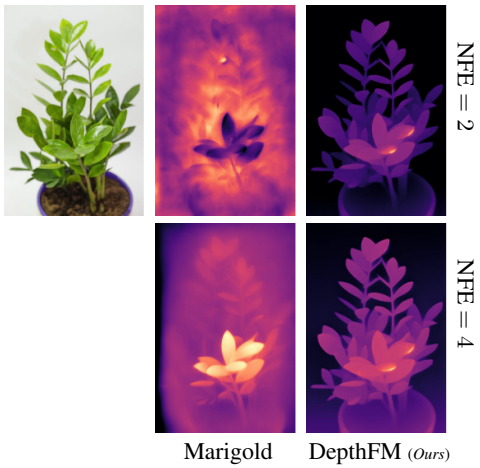

Figure 3: Zero-shot qualitative comparison with few inference steps. DepthFM can output realistic depth maps with just two inference steps.

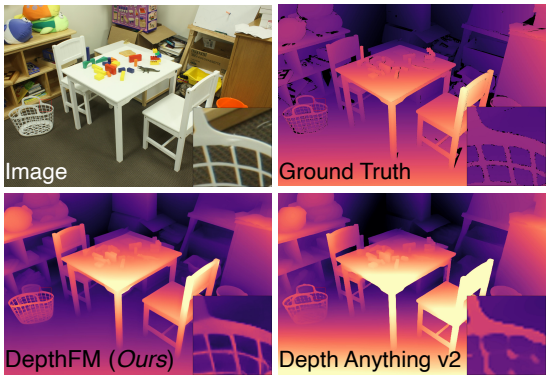

Figure 4: Depth predictions from high-resolution RGBD Middlebury-2014 dataset. Best viewed when zoomed in.

NFEs	1	2	4	10
Marigold	48.8	71.5	82.7	94.8
DepthFM (*Ours*)	**95.0**	**95.6**	**96.3**	**96.2**

Table 3: $\delta1$ evaluation on NYUv2 for different numbers of function evaluations (NFE). We fix the ensemble size to 10.

proach's adaptability and training efficiency. By training only on 74k synthetic samples and an additional 7.4k samples from a discriminative depth estimation method, our model demonstrates exceptional generalization and achieves high *zero-shot* depth estimation performance on both indoor and outdoor datasets.

Comparison against Generative Models Our DepthFM model achieves remarkable sampling efficiency without sacrificing performance. To highlight its inference speed, we quantitatively compare it to Marigold (Ke et al. 2024), a representative diffusion-based generative model. Both models share the same foundational image synthesis model (SD2.1) and network architecture and only differ in the respective training objective and starting distribution (noise versus image). Table 3 shows that DepthFM consistently outperforms Marigold in the low number of function evaluations (NFE) regime, achieving superior results even with *one* NFE, compared to Marigold's performance with *four* NFEs. This is further supported qualitatively in Figure 3, where DepthFM delivers high-quality results with minimal sampling steps, while Marigold requires more steps to produce reasonable results. These results demonstrate the efficiency and effectiveness of DepthFM for fast, high-quality depth estimation.

Comparison against Discriminative Models Despite recent advances in discriminative depth estimation, blurred lines at the edges of objects remain a common problem. Generative methods overcome this problem and pro-

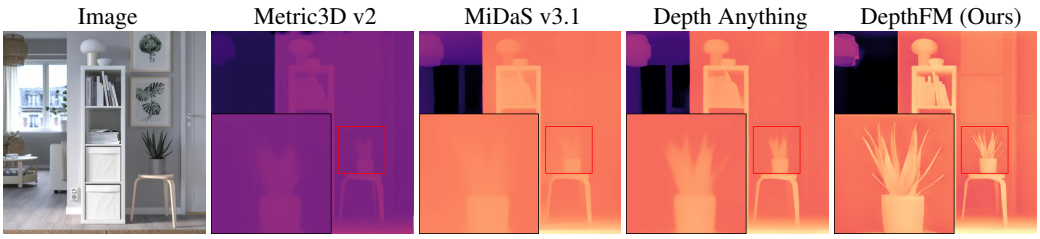

Figure 5: Qualitative comparison of our method against discriminative methods. Best viewed when zoomed in.

duce sharper depth predictions by allowing sampling from the image-conditional posterior distribution of valid depth maps. In contrast, discriminative models are prone to mode averaging and lack detail, as illustrated in Figure 4 & Figure 5. To further quantify fidelity, we use the method introduced by (Hu et al. 2019). On the high-resolution Middlebury-2014 dataset (Scharstein et al. 2014), we predict depth maps, extract edges using a Sobel filter, and then measure edge precision and recall. Higher precision indicates sharper and more precise edges, while higher recall reflects the accuracy of the predicted edges. Table 4 shows that our method significantly outperforms state-of-the-art discriminative depth estimation models in terms of fidelity, achieving higher precision and recall. We further explain the metrics and additional qualitative results in the appendix.

Another unique feature of our DepthFM model is its inherent ability to provide ensembles of depth predictions, due to the stochastic nature within the generative training paradigm. In addition to improving overall performance, ensembling also provides a robust method for quantifying confidence or uncalibrated uncertainty. We estimate the confidence of a prediction by calculating the standard deviation of predictions across ensemble members. A higher standard deviation implies that the model's predictions are less consistent and more sensitive to the stochasticity present in our model. Figure 6 shows an example image, the corresponding depth estimate, and the uncalibrated uncertainty. The ensemble members show noticeable differences, especially in the high frequency regions. Given the drastic depth contrasts within these regions, the variance along the edges again highlights the ability of our model to sample from a reasonable posterior distribution. We evaluate the correlation between uncertainty and depth prediction accuracy on the Middlebury-2014 dataset (Scharstein et al. 2014). Using a threshold to identify regions of high and low uncertainty, we find that the L1 loss is, on average, 3.21 times higher in high-uncertainty regions, indicating a strong correlation between uncertainty and error. Notably, uncertainty does not impact the fidelity of individual ensemble members. Regardless of variations, ensemble members and their mean consistently produce high-fidelity predictions, as shown in Figure 4 and Table 4. Additional visualizations are provided in the appendix.

Depth Completion

An important task related to depth estimation is depth completion. Due to hardware limitations of depth sensors, only

Method	EP (%) ↑	ER (%) ↑
Depth Anything (Yang et al. 2024a)	29.32	29.80
Depth Anything v2 (Yang et al. 2024b)	31.67	40.25
DepthFM *(Ours)*	**33.54**	**49.31**

Table 4: We measure zero-shot edge precision (EP) and recall (ER) on the high-resolution RGBD Middlebury-2014 dataset. Our method excels at high detail.

Method	RMSE ↓
NLSPN (Park et al. 2020)	0.092
DSN (de Queiroz Mendes et al. 2021)	0.102
Struct-MDC (Jeon et al. 2022)	0.245
CompletionFormer (Zhang et al. 2023)	0.090
DepthFM *(Ours)*	**0.077**

Table 5: Comparison of *Depth Completion* on NYUv2.

Figure 6: Uncertainty in absolute depth produced by the stochasticity in our model. *Left*: Original image. *Middle*: Mean depth prediction over ensemble members. *Right*: Standard deviation over ensemble members.

NFEs	1	10
noise → depth	92.4	92.6
image → depth *(Ours)*	**94.6**	**95.5**

Table 6: $\delta 1$ accuracy for different starting distributions on NYUv2. Direct transport is better than starting from noise.

$\delta 1$ ↑	Scratch	LoRA	Finetune
NYUv2	80.0	75.6	**95.5**
DIODE	68.8	70.0	**78.5**

Table 7: Image pre-training and high model capacity are required for high accuracy.

Image Prior	Depth Prior	$\delta 1 \uparrow$
✗	✗	80.0
✓	✗	95.5
✓	✓	**96.3**

Table 8: Including Image and Depth Prior improves zero-shot performance on NYUv2.

a partial depth map is usually available. Therefore, the task is to fill in the rest of the missing depth values with suitable depth estimates. Following previous conventions (Park et al. 2020; de Queiroz Mendes et al. 2021), we fine-tune our DepthFM to complete depth maps where only 2% of the ground truth pixels are available and evaluate it using the root mean square error (RMSE). Table 5 shows that with minimal fine-tuning, DepthFM can achieve state-of-the-art depth completion results on the NYUv2 dataset (Nathan Silberman and Fergus 2012). We provide training details and additional depth completion results in the appendix.

Ablation Studies

Image-depth Coupling We compare DepthFM to a naïve Flow Matching (FM) baseline. While naïve FM also uses an optimal transport-based objective to regress vector fields, it starts from Gaussian noise with $p(x_0) \sim \mathcal{N}(0, \mathbb{I})$. In contrast, our method starts directly from the latent code of the input image. Both models have access to the image as conditioning information over the entire ODE trajectory. The results in Table 6 demonstrate that starting directly from the latent image representation yields significantly higher accuracy, especially in the low NFE regime.

Knowledge Transfer from Image Prior We investigate the importance of the image prior and evaluate its impact on performance. We provide metrics for training the same architecture with identical training settings from scratch, from an image prior, and using adapters (LoRA (Hu et al. 2021)) in Table 7. For LoRA, we use rank 8 and keep the rest of the training details the same. Our observations indicate that adapters significantly limit the fine-tuning process, as they do not provide enough modeling capacity to transfer the image prior to depth estimation. Training from scratch without fine-tuning does not achieve nearly as good a performance, despite our best efforts to optimize it. Therefore, we conclude that a strong image prior and sufficient modeling capacity is essential to provide important visual cues for depth inference. In Figure 7, we further plot the $\delta 1$-accuracy versus training steps on NYUv2. Compared with the default baseline, we demonstrate that we can also achieve better training efficiency and performance with the extra image prior.

Knowledge Transfer from Depth Prior Building upon the Image Prior, we further explore the impact of the depth prior. Combining these two knowledge sources leads to significant performance improvements, as demonstrated in Table 8. For the optimal mixing coefficient k, as referred to in Equation (5), we ablate it in the appendix. We identify an optimal mix where $k = 10\%$ discriminative samples combined with synthetic samples yields the best tradeoff. We find that training on more discriminative samples results in blurrier results indicating a trade-off between accuracy and fidelity.

	0.1	0.2	0.4	0.6	0.8
NYUv2 $\delta 1(\uparrow)$	93.7	94.4	**95.5**	95.5	95.4

Table 9: Noise augmentations level t_s.

Data and Training Efficiency Our method can achieve a better trade-off between training data and performance by using the external depth and the image prior. Note that our approach differs significantly from Depth Anything (Yang et al. 2024a) in two important ways. First, while Depth Anything uses 62M discriminative samples, we achieve optimal results with only 7.4K discriminative samples. Second, Depth Anything requires strong augmentation schemes to take advantage of the pseudo-labeled dataset, while our approach does not require such augmentations.

Noise Augmentation Following the notation from variance-preserving diffusion models, we apply noise to the image samples according to the cosine schedule proposed by (Nichol and Dhariwal 2021). Through empirical analysis in Table 9, we determine that a noise augmentation level of $t_s = 0.4$ is optimal.

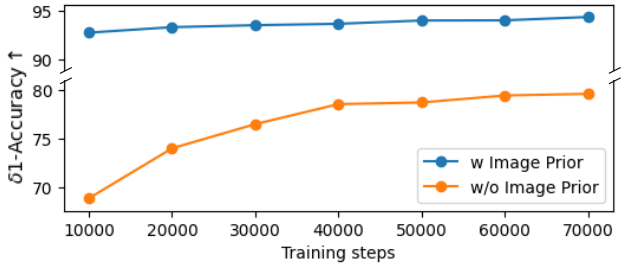

Figure 7: Image prior can boost the training efficiency and performance on NYUv2.

5 Conclusion

We present DepthFM, a flow matching approach to generative monocular depth estimation that improves sampling, data fidelity, training, and data efficiency. First, we improve sampling efficiency by learning a direct transport between image and depth, rather than denoising the Gaussian distribution in depth maps, making our approach faster than current diffusion-based solutions. Second, DepthFM provides high-fidelity depth maps without the common artifacts of discriminative depth estimation methods. Third, we improve training efficiency by using a pre-trained image diffusion model as a prior, providing valuable visual cues to aid depth inference. In addition, we improve data efficiency by using a combination of synthetic data and effectively integrating a discriminative depth prior.

Acknowledgements

This project has been supported by the German Federal Ministry for Economic Affairs and Climate Action within the project "NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung", the German Research Foundation (DFG) project 421703927, Bayer AG, and the bidt project KLIMA-MEMES. The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS at JSC and the HPC resources supplied by the Erlangen National High Performance Computing Center (NHR@FAU funded by DFG).

References

Albergo, M. S.; Goldstein, M.; Boffi, N. M.; Ranganath, R.; and Vanden-Eijnden, E. 2023. Stochastic interpolants with data-dependent couplings. *arXiv*.

Albergo, M. S.; and Vanden-Eijnden, E. 2022. Building normalizing flows with stochastic interpolants. *arXiv*.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*.

Cabon, Y.; Murray, N.; and Humenberger, M. 2020. Virtual KITTI 2.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *CVPR*.

Dao, Q.; Phung, H.; Nguyen, B.; and Tran, A. 2023. Flow matching in latent space. *arXiv*.

de Queiroz Mendes, R.; Ribeiro, E. G.; dos Santos Rosa, N.; and Grassi, V. 2021. On deep learning techniques to boost monocular depth estimation for autonomous navigation. *Robotics and Autonomous Systems*.

Eftekhar, A.; Sax, A.; Malik, J.; and Zamir, A. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *CVPR*, 10786–10796.

Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. *arXiv*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 3354–3361. IEEE.

He, L.; Lu, J.; Wang, G.; Song, S.; and Zhou, J. 2021. SOSD-Net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *NeurIPS*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *JMLR*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv*.

Hu, J.; Ozay, M.; Zhang, Y.; and Okatani, T. 2019. Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps With Accurate Object Boundaries. *WACV*, 1043–1051.

Hu, M.; Yin, W.; Zhang, C.; Cai, Z.; Long, X.; Chen, H.; Wang, K.; Yu, G.; Shen, C.; and Shen, S. 2024a. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation. *arXiv*.

Hu, V. T.; Zhang, D. W.; Mettes, P.; Tang, M.; Zhao, D.; and Snoek, C. G. 2024b. Latent Space Editing in Transformer-based Flow Matching. In *AAAI*.

Jeon, J.; Lim, H.; Seo, D.-U.; and Myung, H. 2022. Struct-MDC: Mesh-Refined Unsupervised Depth Completion Leveraging Structural Regularities from Visual SLAM.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *CVPR*.

Lee, S.; Kim, B.; and Ye, J. C. 2023. Minimizing Trajectory Curvature of ODE-based Generative Models. *ICML*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow matching for generative modeling. *ICLR*.

Liu, X.; Gong, C.; and Liu, Q. 2023. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*.

Nathan Silberman, P. K., Derek Hoiem; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

Neklyudov, K.; Severo, D.; and Makhzani, A. 2022. Action Matching: A Variational Method for Learning Stochastic Dynamics from Samples. *arXiv*.

Nichol, A.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. *arXiv*.

Park, J.; Joo, K.; Hu, Z.; Liu, C.-K.; and Kweon, I. S. 2020. Non-Local Spatial Propagation Network for Depth Completion. In *ECCV*.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision Transformers for Dense Prediction. In *ICCV*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *TPAMI*.

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv*.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 31–42. Springer.

Schops, T.; Schonberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 3260–3269.

Schusterbauer, J.; Gui, M.; Ma, P.; Stracke, N.; Baumann, S. A.; and Ommer, B. 2024. Boosting Latent Diffusion with Flow Matching. *ECCV*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *arXiv*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Tong, A.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Fatras, K.; Wolf, G.; and Bengio, Y. 2023a. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*.

Tong, A.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Fatras, K.; Wolf, G.; and Bengio, Y. 2023b. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*.

Unsplash. 2023. Unsplash | https://unsplash.com/data.

Vasiljevic, I.; Kolkin, N.; Zhang, S.; Luo, R.; Wang, H.; Dai, F. Z.; Daniele, A. F.; Mostajabi, M.; Basart, S.; Walter, M. R.; and Shakhnarovich, G. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv*.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth Anything V2. *arXiv*.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *ICCV*.

Zhang, C.; Yin, W.; Wang, B.; Yu, G.; Fu, B.; and Shen, C. 2022. Hierarchical normalization for robust monocular depth estimation. *NeurIPS*, 35: 14128–14139.

Zhang, H.; Lu, Y.; Alkhouri, I.; Ravishankar, S.; Song, D.; and Qu, Q. 2024. Improving Training Efficiency of Diffusion Models via Multi-Stage Framework and Tailored Multi-Decoder Architecture. In *CVPR*, 7372–7381.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, Y.; Guo, X.; Poggi, M.; Zhu, Z.; Huang, G.; and Mattoccia, S. 2023. Completionformer: Depth completion with convolutions and vision transformers. In *CVPR*, 18527–18536.