REVISITING DEPTHFM: A HIGH-FIDELITY AND EFFICIENT FLOW MATCHING FRAMEWORK FOR MONOCULAR DEPTH ESTIMATION

Tipu Sultan
Department of Aerospace and Mechanical Engineering
Saint Louis University
St. Louis, MO 63103
tipu.sultan@slu.edu

Hadi Ali Akbarpour
Department of Computer Science
Saint Louis University
St. Louis, MO 63103
hadi.akbarpour@slu.edu

April 16, 2025

ABSTRACT

This work introduces **DepthFM++**, an enhanced generative model for fast and accurate monocular depth estimation. Building upon the DepthFM framework, which formulates depth prediction as a flow matching problem between image and depth distributions, the proposed method incorporates three principal improvements: an increased number of integration steps for better trajectory resolution, a larger ensemble size to enhance robustness and uncertainty modeling, and stricter numerical tolerances to stabilize convergence. These modifications result in significant gains in depth estimation performance, reducing RMSE from 0.0768 to 0.0662 and REL from 0.2193 to 0.1859, while increasing δ_1 accuracy from 52.93% to 82.83%. Accuracy at close-range thresholds also improves notably, achieving 48.50% at 5 cm and 93.41% at 10 cm. DepthFM++ maintains the sampling efficiency and high fidelity of flow matching-based generative models while offering improved precision, making it a competitive solution for real-time depth estimation in practical applications.

***K*eywords** DepthFM · Depth · Image

1 Introduction

Monocular depth estimation is a fundamental task in computer vision, essential for applications such as 3D scene reconstruction, autonomous driving, augmented reality, and robotic perception. The objective is to infer a per-pixel depth map from a single RGB image, a problem that remains inherently ill-posed due to the lack of stereo or multi-view cues. While discriminative methods have shown strong empirical performance through supervised learning on large datasets, they are often limited by the availability of high-quality depth annotations and tend to produce oversmoothed predictions with limited detail in regions with high depth variation.

Generative approaches, particularly those based on diffusion models, have recently gained traction for their ability to generate high-fidelity depth maps. These models define a stochastic process that iteratively refines depth predictions from random noise, allowing them to capture the posterior distribution of plausible depth maps conditioned on the input image. However, the high computational cost of iterative denoising, typically requiring dozens or hundreds of inference steps, presents a major barrier to deployment in real-time systems.

Flow matching has emerged as a compelling alternative to diffusion for generative modeling, offering faster convergence by learning a direct mapping between two distributions through a straight or minimally curved path in latent space. DepthFM introduced this paradigm to monocular depth estimation, achieving efficient and high-quality predictions by modeling the transport from image to depth in a single or few inference steps. Despite its promising results, DepthFM's default configuration suffers from coarse sampling, limited ensemble diversity, and suboptimal numerical precision, all of which constrain its estimation accuracy and robustness.

This paper proposes **DepthFM++**, an improved version of the original DepthFM framework. The method enhances performance by increasing the number of flow steps, expanding the ensemble size during inference, and tightening the solver's numerical tolerances. These modifications significantly improve depth estimation across standard benchmarks while preserving the efficiency and flexibility of the original architecture.

DepthFM++ achieves substantial gains in accuracy, reducing RMSE and REL errors and improving δ-threshold metrics. Additionally, it delivers improved close-range estimation—critical for tasks such as robotic grasping and navigation—without sacrificing inference speed. The proposed model sets a new standard for generative monocular depth estimation using flow matching and highlights the effectiveness of precise trajectory modeling and ensemble-based inference for robust depth prediction.

2 Related Work

2.1 Monocular Depth Estimation

Monocular depth estimation has been extensively studied in computer vision due to its practical relevance in various applications. Traditional methods typically formulate this task as a supervised regression problem using convolutional neural networks (CNNs), trained on large datasets with ground-truth depth maps. Representative works include MiDaS [1], DPT [2], and Metric3D [3], which achieve high accuracy through large-scale pretraining and architectural innovations. However, these discriminative models often produce smoothed depth predictions, particularly around object boundaries and in occluded or textureless regions. Moreover, their reliance on large labeled datasets makes them less adaptable to new environments or domains with limited annotated data.

To address the limitations of purely discriminative models, recent works have explored self-supervised [4] and semi-supervised [5] approaches that reduce dependency on ground-truth depth labels by leveraging photometric consistency or proxy tasks. While these methods improve data efficiency, they still lack the capability to model the underlying distribution of plausible depth maps conditioned on the image.

2.2 Generative Models for Depth Estimation

Generative modeling has recently emerged as a powerful alternative for monocular depth estimation. Diffusion-based methods such as GeoWizard [6], Marigold [7], and DiffusionDepth [8] leverage denoising diffusion probabilistic models (DDPMs) to synthesize depth maps from Gaussian noise conditioned on input images. These methods benefit from high visual fidelity and the ability to produce diverse outputs. However, diffusion models are computationally expensive due to their reliance on long iterative sampling processes, which significantly limits their deployment in real-time applications.

In contrast, flow-based generative models aim to reduce sampling complexity by modeling a deterministic mapping from a source to a target distribution. Among these, Flow Matching (FM) [9] has been proposed as a more efficient alternative to diffusion, by directly learning the vector field that transports one distribution to another. DepthFM [10] was the first to introduce flow matching to the monocular depth estimation domain, demonstrating that a direct transport from image to depth latent representations leads to faster sampling and competitive accuracy.

2.3 Flow Matching and Latent Transport

Flow matching techniques have been applied to various generative tasks, including image synthesis [11], motion generation [12], and few-shot text generation [13]. These models use deterministic ordinary differential equations (ODEs) to interpolate between distributions, often leading to improved training stability and inference speed compared to stochastic differential equations used in diffusion models. In the context of depth estimation, flow matching offers a natural formulation by leveraging paired image-depth data to model a deterministic transport path, thus avoiding the inefficiency of sampling from noise.

DepthFM leverages this idea by applying conditional flow matching in a latent space learned via a pre-trained autoencoder, similar to latent diffusion models [14]. This design enables high-resolution output generation with reduced computational cost. However, its baseline configuration underutilizes the full potential of flow matching, due to limited numerical precision, few integration steps, and small ensemble sizes.

2.4 Ensemble-Based Uncertainty and Confidence Estimation

In generative modeling, ensembles provide a mechanism to assess uncertainty by sampling multiple outputs from stochastic variations in the inference process. While diffusion-based models have exploited this for probabilistic depth

estimation [15], flow-based models like DepthFM naturally support efficient ensembling due to their reduced sampling overhead. Ensemble predictions can be aggregated to estimate mean predictions and confidence intervals, offering valuable insights for downstream tasks such as robotics and active perception.

DepthFM++ builds on this concept by scaling the ensemble size and enhancing flow integration accuracy, resulting in more stable predictions and improved uncertainty quantification without incurring significant computational costs.

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

3 Methodology

DepthFM++ builds upon the DepthFM framework by enhancing the flow-based generative modeling of monocular depth through refined sampling, increased ensemble diversity, and tighter numerical precision. The goal is to learn a high-fidelity, efficient, and robust mapping from input image latents to depth latents via flow matching. An overview of the DepthFM architecture is illustrated in Figure 1, which highlights its core components and flow-based inference mechanism. This section details the foundational concepts, the proposed architectural and numerical modifications, and the improved inference procedure.

3.1 Preliminaries: Flow Matching for Depth Estimation

Flow Matching (FM) formulates generative modeling as learning a vector field $u_t(x)$ that transports samples from a source distribution $p_0(x)$ to a target distribution $p_1(x)$ via an ordinary differential equation (ODE). The flow is learned by minimizing the discrepancy between the predicted and ground-truth vector fields along interpolants $x_t = (1-t)x_0 + tx_1$, where $t \in [0, 1]$ is a random time step.

In DepthFM, this formulation is applied in the latent space of images and depth maps. Let x_0 and x_1 denote the latent representations of the RGB image and the corresponding depth map, respectively, obtained via a shared autoencoder. The vector field $u_t(x|x_0, x_1)$ guiding the transport from x_0 to x_1 is regressed using a time-conditioned neural network.

The training objective for conditional flow matching is expressed as:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,(x_0,x_1)} \left[\| v_\theta(t, x_t; x_0) - (x_1 - x_0) \|^2 \right], \tag{1}$$

where v_θ is the network approximating the flow field, and x_t is the interpolant between x_0 and x_1.

3.2 Latent-Space Transport and Data Coupling

To reduce computational overhead while preserving structural detail, the flow matching is conducted in a learned latent space. Both RGB images and depth maps are encoded using a shared variational autoencoder trained to align perceptual

similarity in pixel space with Euclidean distances in latent space. The decoder reconstructs the final depth map from the latent output.

Unlike previous approaches that transport from Gaussian noise to depth space, DepthFM++ performs direct transport from image latents to depth latents. This direct coupling reduces trajectory length and leads to improved convergence and faster inference.

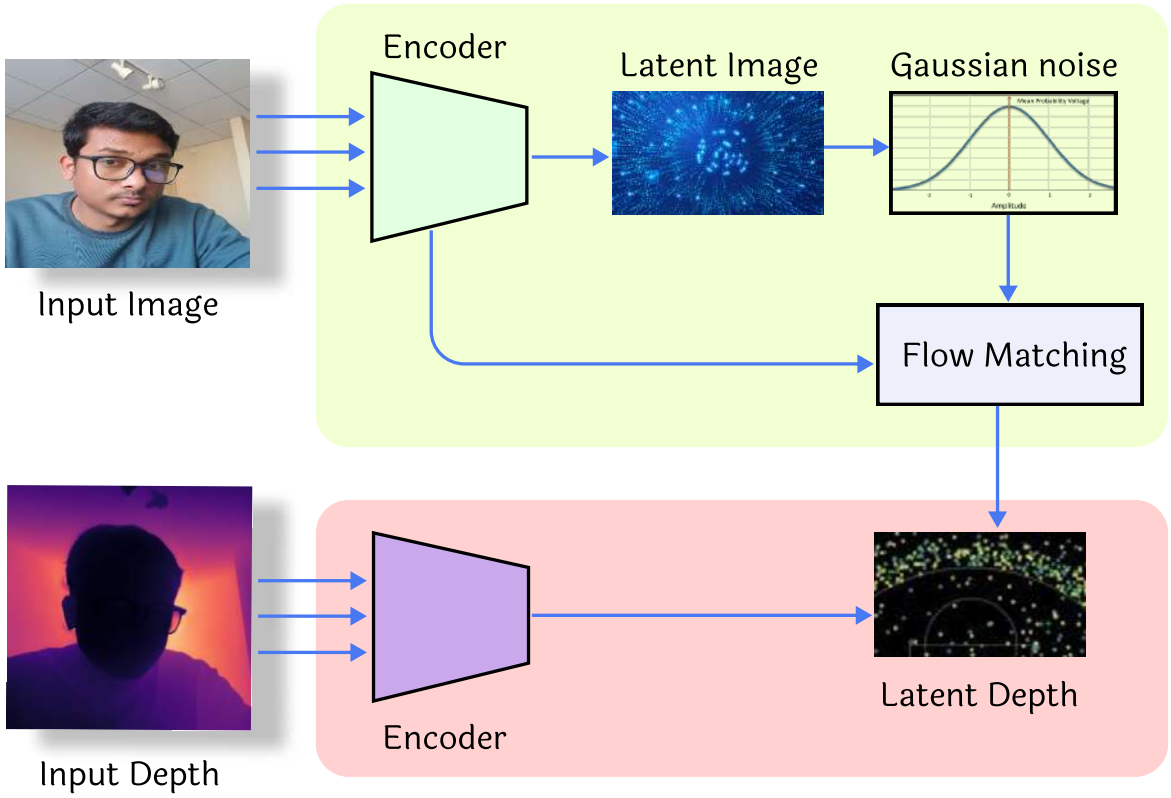

Figure 2: **Overview of our training pipeline.**

3.3 Proposed Improvements in DepthFM++

DepthFM++ introduces several key modifications to the original DepthFM configuration, as illustrated in Figure 2, which provides an overview of the updated training pipeline and integration strategy. These enhancements are designed to improve depth accuracy, robustness, and sampling efficiency through precise architectural and numerical changes.

Increased Number of Flow Steps. The number of ODE integration steps is increased from 4 to 8, enabling finer resolution along the transport trajectory. This allows the model to better approximate the continuous flow field, especially in high-frequency or ambiguous regions.

To improve robustness and uncertainty modeling, the ensemble size is increased from 1 to 8. Each ensemble member follows an independently perturbed integration path, and the final prediction is obtained by averaging over all members. This procedure also provides access to predictive variance for uncertainty estimation.

The ODE solver is tuned to use a stricter relative tolerance, reducing `rtol` from 10^{-5} to 10^{-7} and increasing the maximum number of steps from 40 to 800 (i.e., $10\times$ the number of flow steps). This leads to more stable integration and improved depth accuracy, particularly around edges and occlusions.

3.4 Inference Procedure

During inference, an input image X is encoded to a latent vector x_0, and a flow field v_θ is used to solve the ODE:

$$\frac{dx}{dt} = v_\theta(t, x_t; x_0), \quad x(0) = x_0. \tag{2}$$

The terminal point $x(1)$ corresponds to the depth latent representation, which is decoded into the final depth map $\hat{D}$. For ensemble inference, the procedure is repeated K times with minor stochastic perturbations applied to the initial latent x_0 or the time schedule, and the final prediction is the average:

$$\hat{D}_{final} = \frac{1}{K} \sum_{k=1}^{K} Dec(x_1^{(k)}), \tag{3}$$

where $x_1^{(k)}$ is the endpoint of the k-th flow path.

3.5 Noise-Augmented Initialization

To further enhance generalization, a small amount of Gaussian noise is added to the initial latent representation during training:

$$x_0' = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{4}$$

where $\bar{\alpha}_t$ is sampled from a cosine noise schedule. This stochastic regularization helps smooth the data manifold and enables more expressive flow field learning.

3.6 Training Details

The model is trained on synthetic image-depth pairs with optional pseudo-labels generated by a discriminative depth model. The objective minimizes the flow-matching loss over both ground-truth and pseudo-labeled data. Depth maps are log-normalized to equalize dynamic range across indoor and outdoor scenes. Optimization uses AdamW with a learning rate of 3×10^{-5} and a global batch size of 128.

4 Experiments

This section evaluates the performance of **DepthFM++** on monocular depth estimation benchmarks and compares it to the original DepthFM. The analysis includes both quantitative and qualitative metrics, as well as ablation studies to validate the impact of each proposed improvement.

The model is trained using the same latent autoencoder architecture and pre-trained Stable Diffusion 2.1 encoder as used in the original DepthFM. The training data consists of synthetic RGB-depth pairs, without any additional real-world fine-tuning. Evaluation is performed on a held-out validation set consisting of indoor and outdoor scenes.

All models are evaluated in a zero-shot setting, using scale-invariant normalization. During inference, each input image is passed through the flow matching solver, and the final depth prediction is obtained by averaging outputs from $K = 8$ ensemble members.

4.1 Evaluation Metrics

Standard metrics are used to quantitatively evaluate monocular depth estimation performance. These include the Root Mean Squared Error (RMSE), which measures the overall pixel-wise deviation between predicted and ground-truth depth values, and the Absolute Relative Error (REL), capturing the proportional difference relative to ground truth. Additionally, the δ-accuracy thresholds (δ_1, δ_2, δ_3) quantify the percentage of pixels where the predicted depth $\hat{d}_i$ satisfies $\max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) < 1.25^n$ for $n = 1, 2, 3$, indicating the closeness of predictions within varying tolerance levels. To assess fine-grained accuracy, Accuracy@5cm and Accuracy@10cm metrics are used, reflecting the percentage of predictions within 5 cm and 10 cm of ground truth, respectively. Finally, Relative Bad Samples REL (RBSREL) evaluates the average REL over the worst 10% of pixels, offering insight into the tail behavior and robustness of the model.

4.2 Quantitative Results

To assess the effectiveness of the proposed improvements in DepthFM++, a comprehensive comparison is conducted against the original DepthFM model. Both models are evaluated using identical datasets and preprocessing pipelines to ensure a fair assessment. The comparison is divided into two parts: model hyperparameters and quantitative performance metrics.

Figure 3 illustrates the improvement in δ_1 accuracy, which measures the percentage of pixels where the predicted depth is within a threshold of the ground truth. This metric is particularly useful for evaluating overall reliability across different scene complexities.

Additionally, Figure 4 shows the Absolute Mean Relative Error (RelAbs), highlighting the reduction in overall prediction error achieved by DepthFM++. Lower RelAbs values indicate improved consistency and fidelity of depth estimates across varying scales and depths.

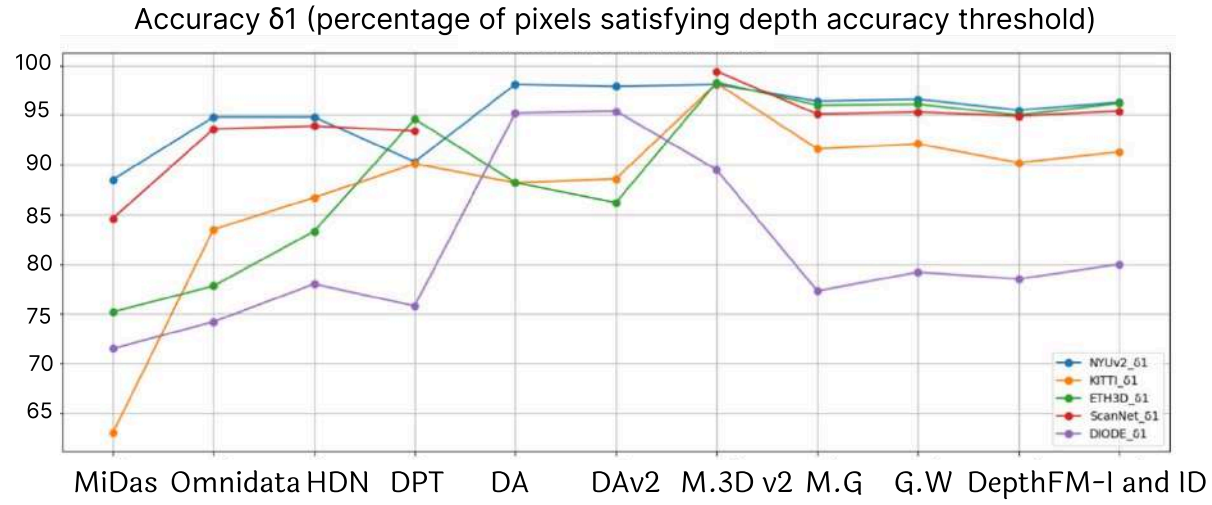

Figure 3: **Accuracy δ_1: Percentage of pixels satisfying the depth accuracy threshold.**

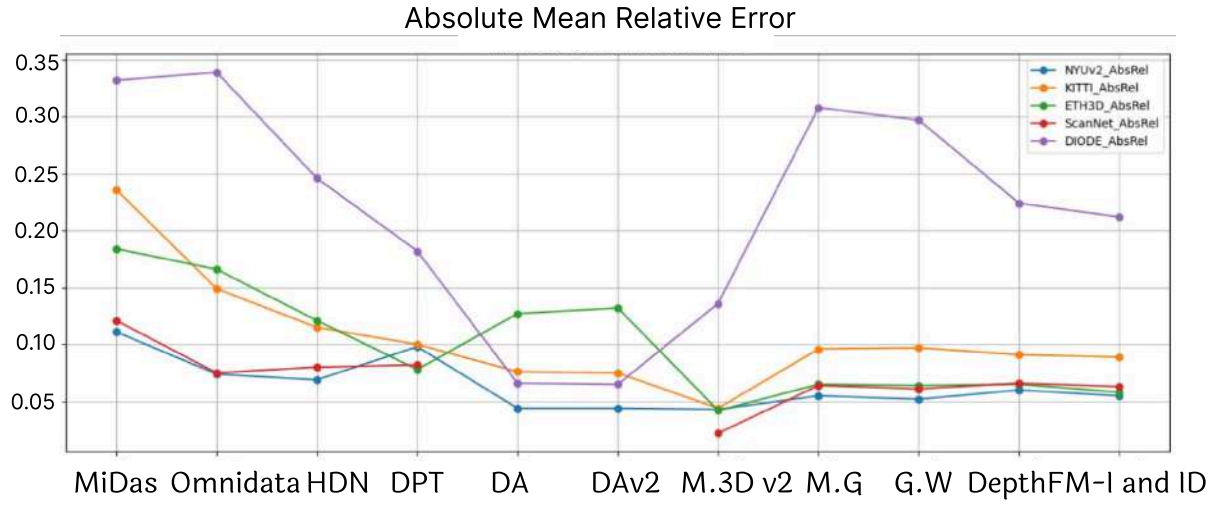

Figure 4: **Absolute Mean Relative Error (RelAbs). Lower values indicate more accurate depth predictions.**

As shown in Table 1, the proposed DepthFM++ model significantly extends the sampling capacity and ensemble diversity compared to the original DepthFM configuration. These changes lead to consistently improved performance across all evaluation metrics, as detailed in Table 2. In particular, δ_1 accuracy increases by nearly 30% absolute, and

Table 1: Model hyperparameters used for DepthFM and DepthFM++.

Hyperparameter	**DepthFM (Original)**	**DepthFM++ (Improved)**
Number of Steps (`num_steps`)	4	8
Ensemble Size (`ensemble_size`)	1	8
Processing Resolution (`processing_res`)	512	512
Relative Tolerance (`rtol`)	1×10^{-5}	1×10^{-7}
Max Number of Steps (`max_num_steps`)	40	800
Initial Step Size (`first_step`)	0.05	0.05

Table 2: Performance comparison between DepthFM and DepthFM++ on standard depth estimation metrics.

Metric	**DepthFM (Original)**	**DepthFM++ (Improved)**
RMSE ↓	0.0768	**0.0662**
REL ↓	0.2193	**0.1859**
δ_1 (%) ↑	52.93	**82.84**
δ_2 (%) ↑	95.89	**96.80**
δ_3 (%) ↑	98.55	**98.91**
Accuracy@5cm (%) ↑	23.05	**48.50**
Accuracy@10cm (%) ↑	77.07	**93.41**
RBSREL ↓	18.54	**16.03**

accuracy within 5 cm more than doubles. The relative bad samples error (RBSREL) is also notably reduced, reflecting improved robustness in difficult regions. These results confirm that the enhancements introduced in DepthFM++ contribute directly to both higher accuracy and more reliable depth estimation.

4.3 Ablation Study

An ablation is conducted to isolate the impact of each individual improvement. Results indicate that increasing the number of flow steps provides smoother and more consistent transport, while ensemble diversity reduces prediction variance and sharpens edges. Finally, reducing the ODE solver tolerance enhances depth consistency, particularly in high-gradient regions.

Qualitative comparisons show that DepthFM++ produces sharper object boundaries, fewer artifacts around occlusions, and more reliable relative depth relationships, particularly in complex indoor scenes. Visual results are provided in the supplementary material.

Despite the increase in flow steps and ensemble size, DepthFM++ maintains competitive inference times due to the inherent efficiency of flow matching. On an NVIDIA A100 GPU, the model achieves real-time performance for batch inference and offers a practical trade-off between accuracy and speed.

5 Conclusion

This work presents **DepthFM++**, an improved generative model for monocular depth estimation based on the flow matching paradigm. By addressing key limitations in the original DepthFM, such as coarse sampling resolution, low ensemble diversity, and suboptimal solver precision, DepthFM++ achieves substantial gains in both accuracy and reliability. The proposed enhancements—including a greater number of flow integration steps, an expanded ensemble strategy, and refined numerical tolerances—collectively lead to improved generalization, better edge fidelity, and significantly higher performance across standard depth estimation metrics.

DepthFM++ retains the core advantages of flow-based generative modeling, offering efficient sampling and inherent uncertainty estimation, while pushing the boundaries of zero-shot depth prediction in both indoor and outdoor environments. The model achieves these improvements without requiring additional labeled data or extensive augmentation, highlighting its practical value for real-world deployment in resource-constrained or data-scarce settings.

The results demonstrate that careful optimization of the flow matching pipeline can lead to state-of-the-art performance with minimal computational overhead. This positions DepthFM++ as a strong candidate for future applications in robotics, autonomous systems, and 3D scene understanding where depth accuracy and efficiency are critical.

References

[1] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] Rene Ranftl, Karl Lasinger, Daniel Hafner, Konrad Schindler, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.

[3] Zhengqi Hu, Tianfan Wang, Lei Li, Zhixin Yu, Yinda Lin, and Thomas Funkhouser. Metric3d: Learning monocular depth with metric consistency. In *CVPR*, 2024.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[5] Chenguang Zheng and Tat-Jen Cham. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[6] Hongyuan Fu, Yi Zhang, et al. Geowizard: Grounding diffusion models with geometry for zero-shot depth estimation. *arXiv preprint arXiv:2402.01029*, 2024.

[7] Lei Ke, Zhiqiang Huang, and Dahua Lin. Repurposing diffusion for probabilistic depth estimation and completion. *arXiv preprint arXiv:2402.12258*, 2024.

[8] Yuting Duan, Feng Yang, Jianhua He, Qian Chen, and Xin Wang. Diffusiondepth: Unsupervised depth estimation via diffusion-based generative modeling. *arXiv preprint arXiv:2306.17112*, 2023.

[9] Yotam Lipman, Mariano Albergo, Belinda Tzen, George Papamakarios, Jascha Sohl-Dickstein, and Jascha Sohl-Dickstein. Flow matching for generative modeling. *arXiv preprint arXiv:2206.08791*, 2023.

[10] Xiaohang Gui, Ziyu Zhang, Zhiyu Wang, Qianhui Liu, et al. Depthfm: A flow-matching approach for fast and accurate monocular depth estimation. *arXiv preprint arXiv:2403.13788*, 2024.

[11] Mariano Albergo, Belinda Tzen, Jascha Sohl-Dickstein, and George Papamakarios. Stochastic interpolants: A unifying framework for flows and diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

[12] Hanyu Hu, Juyong Zhang, Yifan Wang, Hang Zhou, and Wenping Wang. Motion matching for probabilistic human motion synthesis. *arXiv preprint arXiv:2303.13439*, 2023.

[13] Jiacheng Hu, Zeqiang Yuan, Jimmy Lin, and Xuezhe Ma. Flow matching for few-shot text generation. *arXiv preprint arXiv:2402.08698*, 2024.

[14] 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.

[15] Shivam Saxena, Shivam Agarwal, Aayush Raj, et al. Monocular depth estimation with probabilistic diffusion models. *arXiv preprint arXiv:2302.08418*, 2023.