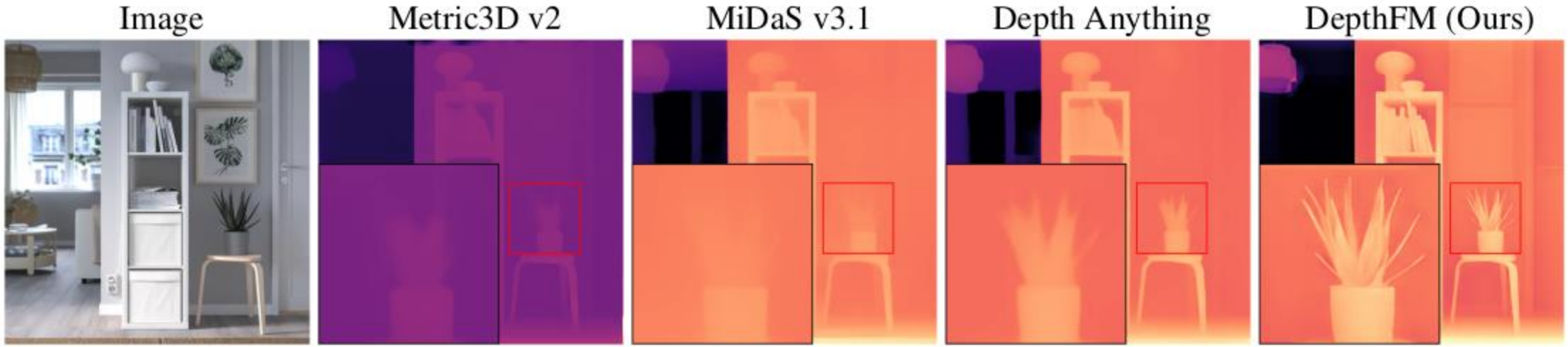
DEPTHFM: FAST GENERATIVE MONOCULAR DEPTH ESTIMATION WITH FLOW MATCHING

Authors: Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Tao Hu, Björn Ommer

INTRODUCTION AND MOTIVATION

- Monocular Depth Estimation: Importance for 3D scene understanding.
- Applications: Robotics, autonomous driving, visual synthesis.
- Challenges: Blurry artifacts (discriminative methods) and slow sampling (generative methods).
- DepthFM: A new approach using Flow Matching for faster and more accurate depth estimation.



RELATED WORK

Discriminative Methods (e.g., MiDaS, Depth Anything, Metric3D)

- Learn depth from labeled datasets but struggle with fine details

Generative Diffusion-based Methods (e.g., Marigold, GeoWizard)

- High visual fidelity but suffer from slow inference

Flow Matching (FM)

- Alternative to diffusion models with faster, straighter sampling trajectories
- Proposed solution: Direct image-to-depth transport

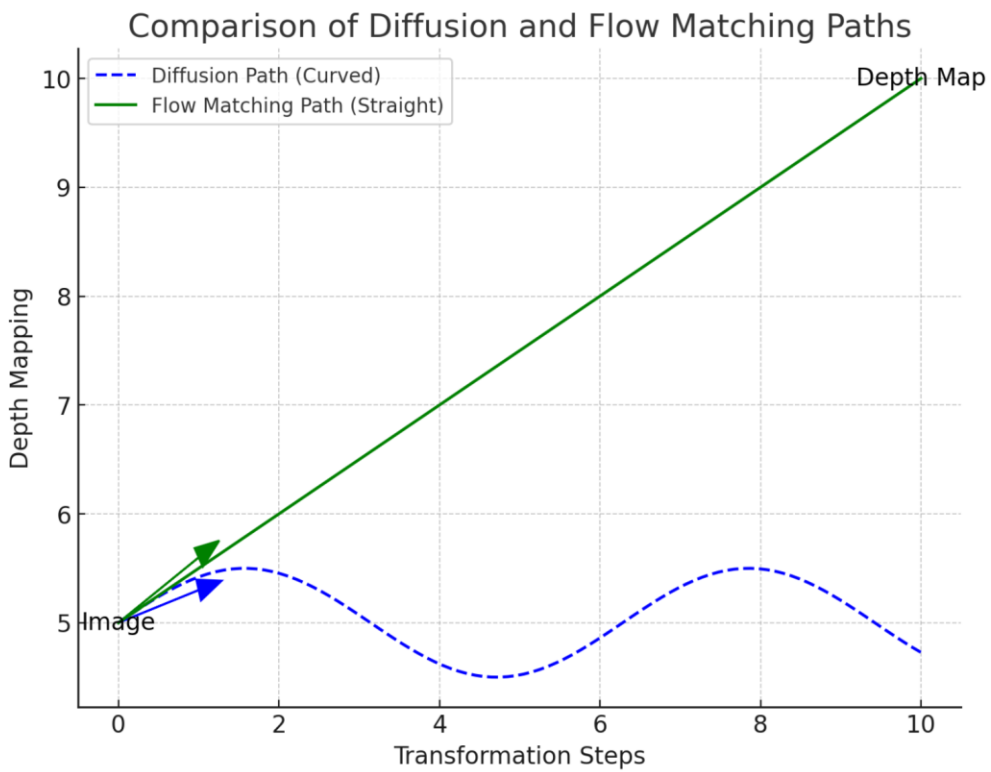
INNOVATIONS OF DEPTHFM

- Flow Matching for Depth Estimation: Direct transport from image to depth.
- Leveraging Pre-trained Models: Image prior (SD2.1) and discriminative depth prior (Metric3D v2).
- Data Efficiency: Combines generative and discriminative strengths for better performance with less data.

FLOW MATCHING APPROACH

- Flow Matching (FM) creates a direct transport between image and depth data.
- Unlike diffusion models that start from noise, FM starts directly from the image features.
- This approach results in a straight path from the image to the depth map, minimizing computational cost.
- Key Idea: Use a vector field to guide the transformation, making it efficient and accurate.
- Why It's Better: Faster inference, sharper depth maps, and fewer intermediate steps.
- Real-Life Analogy: Instead of taking a long, winding path, FM takes a straight road from start to finish.

FLOW MATCHING APPROACH



- **Diffusion Models (Blue Dashed Line):**

Start from pure noise and follow a curved, winding path to the depth map.
Computationally expensive and time-consuming.
Require multiple iterative steps for gradual denoising.

- **Flow Matching (Green Solid Line):**

Directly connects image features to depth map using a straight path.
Faster and more efficient as it avoids unnecessary intermediate steps.
Does not involve complex denoising processes.

ARCHITECTURE

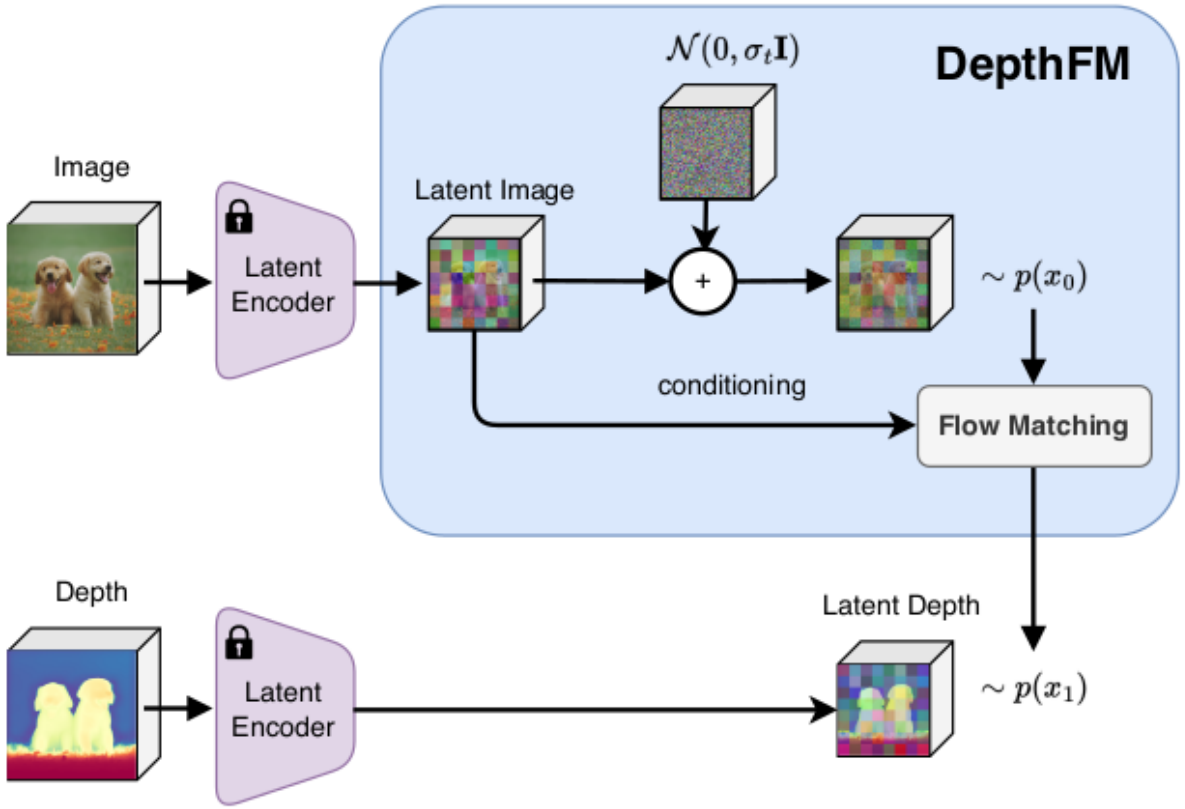


Figure 2: Overview of our training pipeline. We use flow matching to regress the vector field between the image latent x_0 and the corresponding depth latent x_1.

- Input: An image and its corresponding depth map.
- Latent Encoding: Converts both image and depth into latent representations.
- Noise Addition: Adds Gaussian noise to the latent image to improve robustness.

ARCHITECTURE

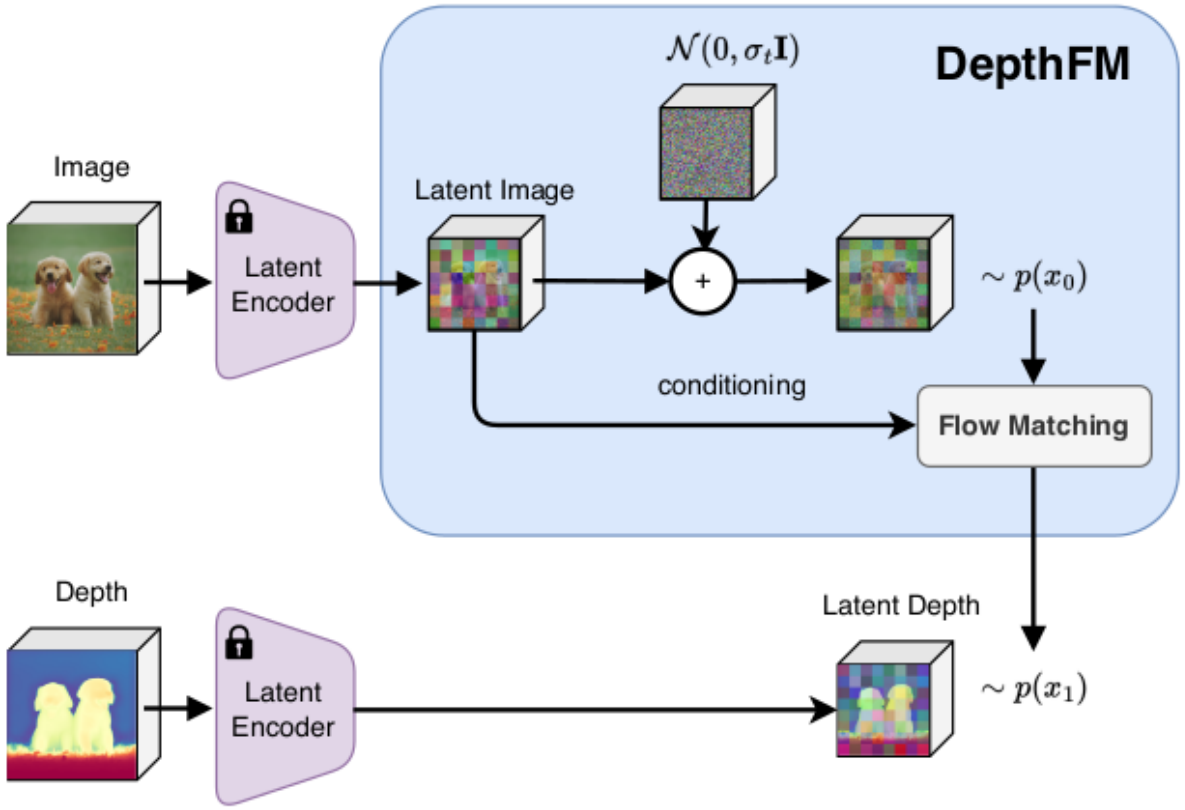


Figure 2: Overview of our training pipeline. We use flow matching to regress the vector field between the image latent x_0 and the corresponding depth latent x_1.

- Flow Matching:
  - Learns a vector field to map latent image features to latent depth features.
  - Uses a direct transport approach to connect image and depth representations efficiently.
- Output: Decoded depth map with high accuracy and fidelity.

ARCHITECTURE

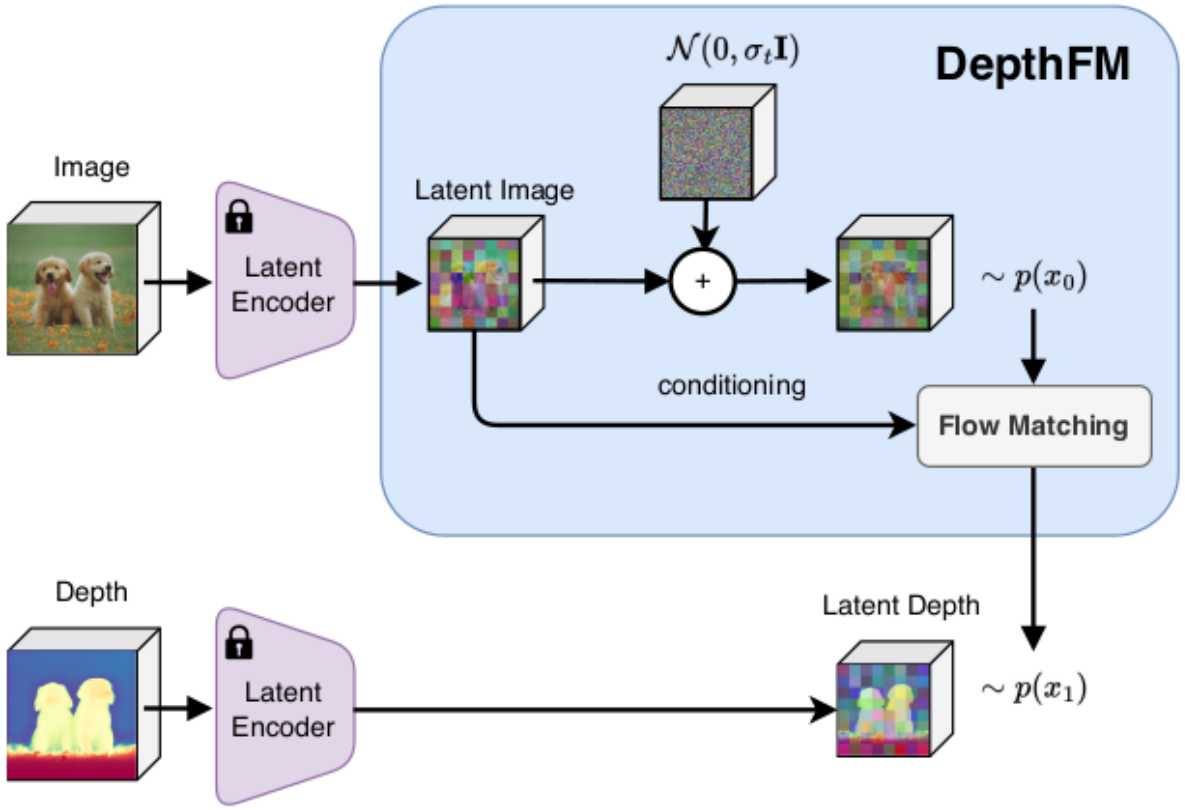


Figure 2: Overview of our training pipeline. We use flow matching to regress the vector field between the image latent x_0 and the corresponding depth latent x_1.

- Key Points:
  - Direct mapping from image to depth without intermediate denoising.
  - Efficient and fast due to straight path transformation.
  - Robust against input noise, thanks to noise augmentation.

- Diagram: DepthFM Training Pipeline (showing flow from image to depth via latent space).

DATA EFFICIENCY TECHNIQUES

Leveraging External Knowledge

- Pre-trained diffusion model (Stable Diffusion 2.1) for image priors
- Pre-trained discriminative model for depth priors

Dual Knowledge Transfer

- Fine-tuning with diffusion models speeds up training
- Discriminative model improves depth accuracy with minimal data

USING PRE-TRAINED MODELS IN DEPTHFM

Pre-trained Image Diffusion Model (Image Prior)

Model Used: Stable Diffusion 2.1 (SD2.1)

Purpose: To provide a strong image prior for depth estimation.

- Leverages rich visual knowledge from large datasets.
- Captures fine-grained visual details efficiently.
- **How It's Used:**
  - Fine-tuned with the Flow Matching objective.
  - Transfers visual understanding to depth estimation.
- **Benefit:**
  - Speeds up training and enhances generalization.

PRE-TRAINED DISCRIMINATIVE DEPTH ESTIMATION MODEL (DEPTH PRIOR)

- **Model Used:** Metric3D v2
- **Purpose:** To provide a strong depth prior and enhance data efficiency.
- **Use:**
  - Provides accurate depth predictions from large annotated datasets.
  - Generates high-quality synthetic depth-image pairs.
- **How It's Used:**
  - Acts as a teacher model to generate synthetic data.
  - Fine-tunes DepthFM using generated pseudo-depth pairs.
- **Benefit:**
  - Increases robustness and accuracy with limited labeled data.

EXPERIMENTAL SETUP

Training Data

- Synthetic datasets: Hypersim, Virtual KITTI
- Discriminative samples from Unsplash dataset

Evaluation Datasets

- NYUv2, KITTI, ETH3D, ScanNet, DIODE

Metrics

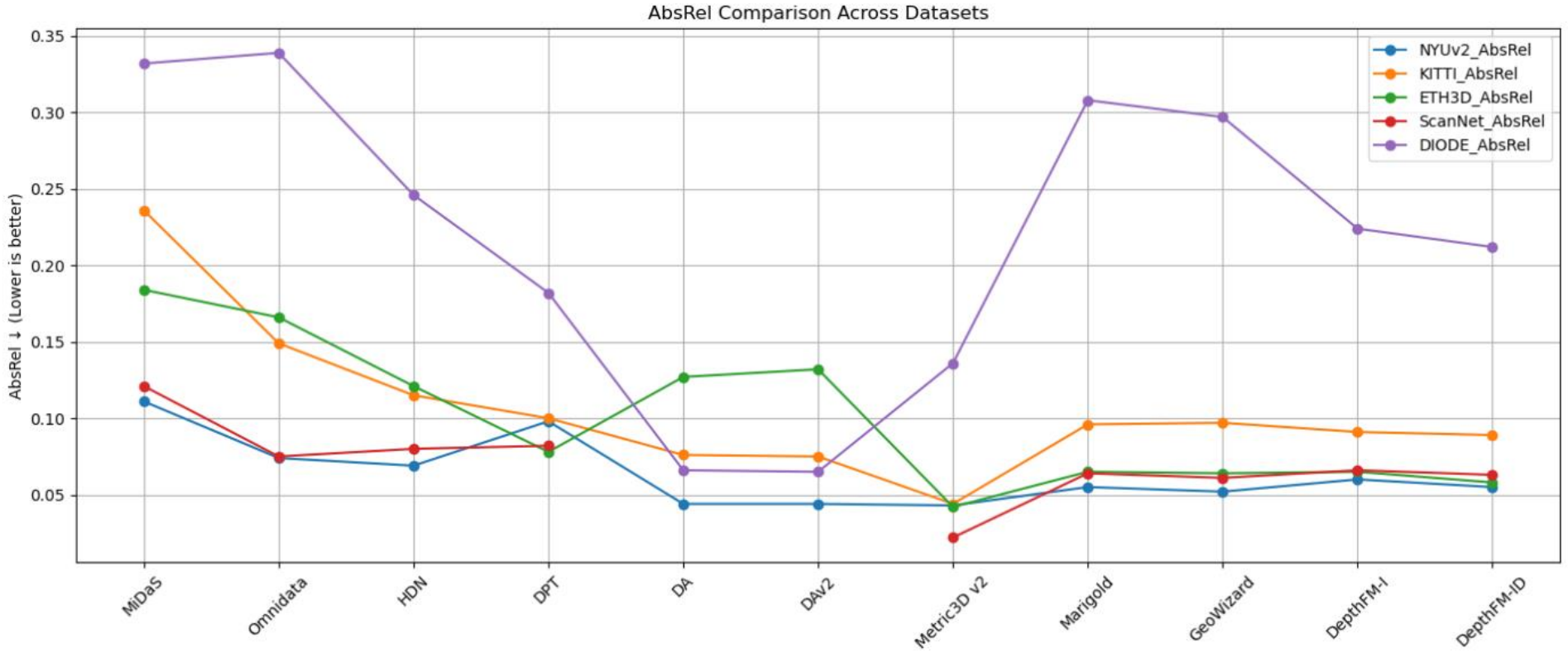
- Absolute Mean Relative Error (RelAbs)
- Accuracy $\delta 1$ (percentage of pixels satisfying depth accuracy threshold)

RESULTS

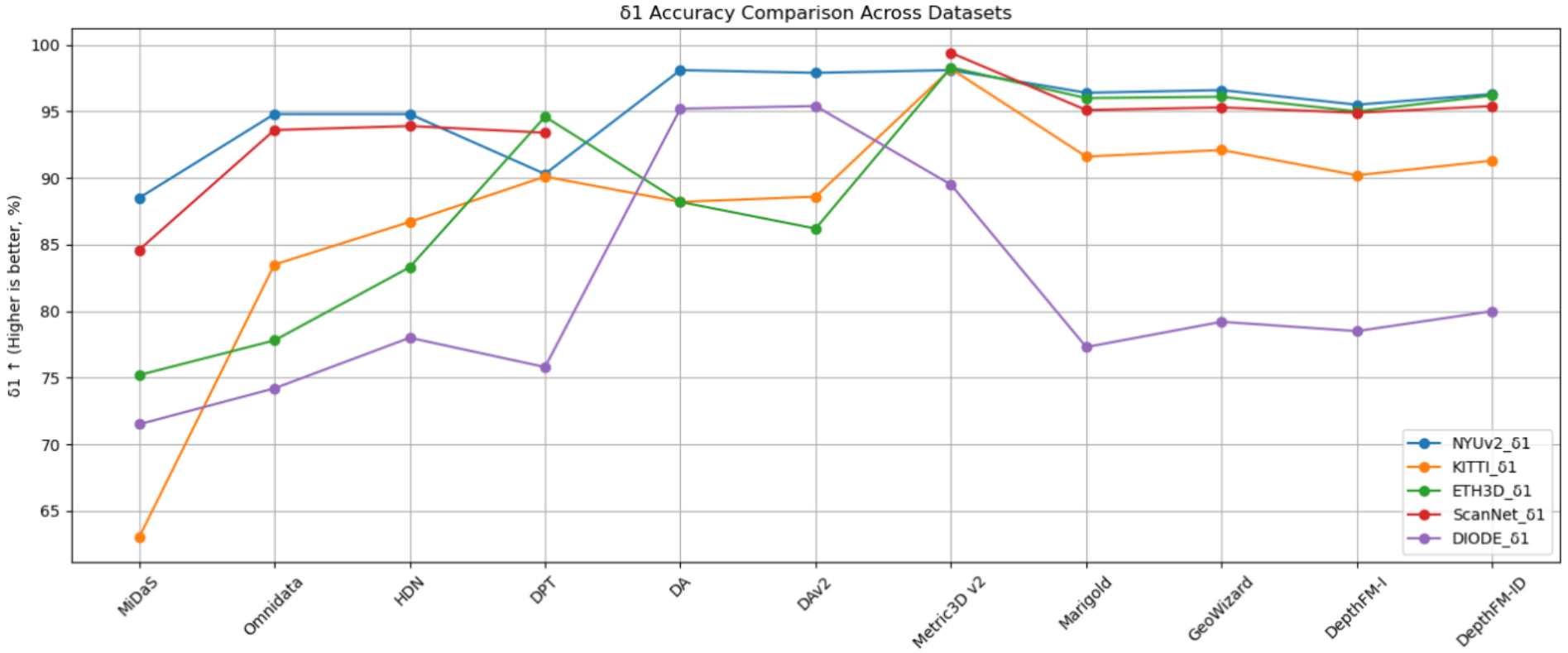
	Method	#Train samples		NYUv2		KITTI		ETH3D		ScanNet		DIODE	
		Real	Synthetic	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Discriminative	MiDaS (Ranftl et al. 2020)	2M	—	0.111	88.5	0.236	63.0	0.184	75.2	0.121	84.6	0.332	71.5
	Omnidata (Eftekhar et al. 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.

Absolute Mean Relative Error (RelAbs)



Accuracy δ1 (percentage of pixels satisfying depth accuracy threshold)



RESULTS - COMPARISON WITH DISCRIMINATIVE MODELS

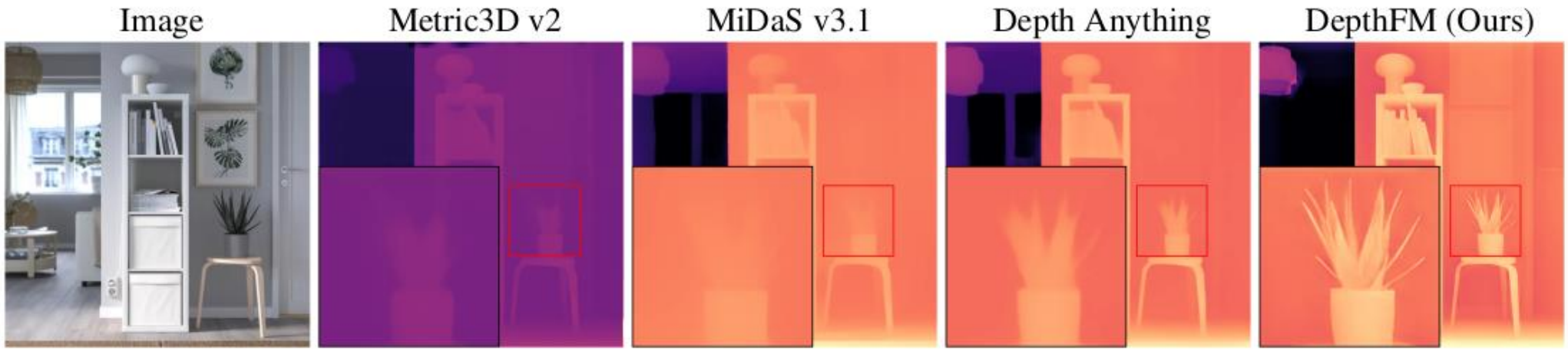


Figure 5: Qualitative comparison of our method against discriminative methods. Best viewed when zoomed in.

ABLATION STUDIES

Direct Image-to-Depth Transport	More effective than starting from Gaussian noise
Impact of Image and Depth Prior	Both contribute to improved accuracy and efficiency
Noise Augmentation Strategy	Improves robustness without degrading accuracy

ABLATION STUDIES

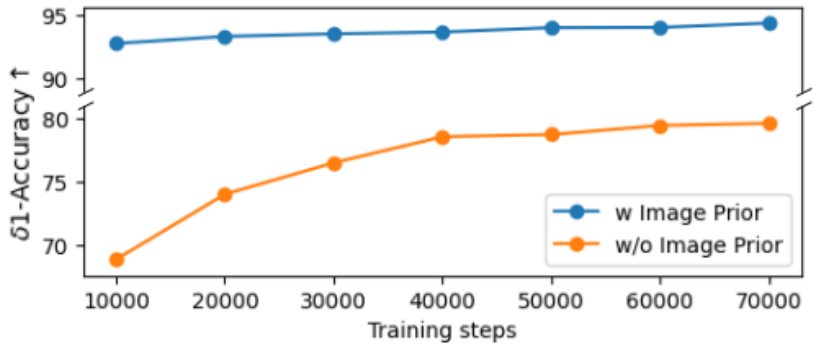


Figure 7: Image prior can boost the training efficiency and performance on NYUv2.

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.

These results clearly show that combining **image prior knowledge** with **direct transport** leads to faster convergence and better performance, making DepthFM more efficient and accurate than traditional methods.

CONCLUSION

DepthFM introduces Flow Matching for monocular depth estimation

Improves speed, efficiency, and accuracy over existing methods

Leverages external knowledge for minimal training data dependency

State-of-the-art zero-shot performance on multiple benchmarks

Future Work: Extend to real-time applications in robotics and AR/VR