Diff2Flow: Training Flow Matching Models via Diffusion Model Alignment

Figure 9. Diff2Flow enables fast monocular depth estimation with high fidelity.

A. Implementation Details

Text-to-Image. For the text-to-image task, we fine-tune Stable Diffusion 2.1, aligning its $\mathbf{v}$-parameterization to Flow Matching. For the comparison, we fine-tune three models: the diffusion baseline, the diffusion model with Flow Matching loss, and our proposed Diff2Flow adaptation. Each model is trained for 20k iterations using a constant learning rate of 1×10^{-5} and a batch size of 64 on the LAION-Aesthetics dataset [53], which contains high-aesthetic-score images paired with synthetically generated captions. We evaluate all models on the COCO 2017 dataset [29] using ODE sampling. In our Low-Rank Adaptation (LoRA) setup, we set the rank of convolutions and attention layers to 20% of each layer's respective feature dimension. This configuration results in a model with 222M trainable parameters for the LoRA version, compared to the 866M parameters of the original Stable Diffusion 2.1 model.

Reflow. Adapting Stable Diffusion with our proposed method, allows us to perform rectification of sampling trajectories, as proposed in [32] for Flow Matching models. Rectification relies on pre-computed image-noise pairs, which we generate by sampling approximately 1.8M images with a classifier-free guidance scale of 7.5, using prompts from the LAION-Aesthetics dataset [53] and 40 sampling steps. We then perform 1-rectification training for 60k gradient updates on these image-noise pairs. We fix the LoRA rank to 64 across all convolutional, self-attention, and feedforward layers, resulting in a model with a total of 62M trainable parameters. We train this model with a batch size of 128 and a decaying learning rate schedule starting from 2×10^{-5}, and evaluate it on the COCO 2017 dataset [29].

Image-to-Depth. We follow the training paradigm of [14, 24] and use a mixture of Hypersim [45] and Virtual Kitti v2 [4] data. Similar to [14] we log-normalize the depth data, as we found it to make better use of the input data space. We evaluate zero-shot on five benchmark datasets: NYUv2 [40], DIODE [62], ScanNet [5], KITTI [13], and ETH3D [52]. We use the evaluation suite from [24] and align an ensemble of estimated depth maps to the ground truth depth with least squares fitting. We report the average relative difference between the ground-truth depth and the aligned predicted depth at each pixel

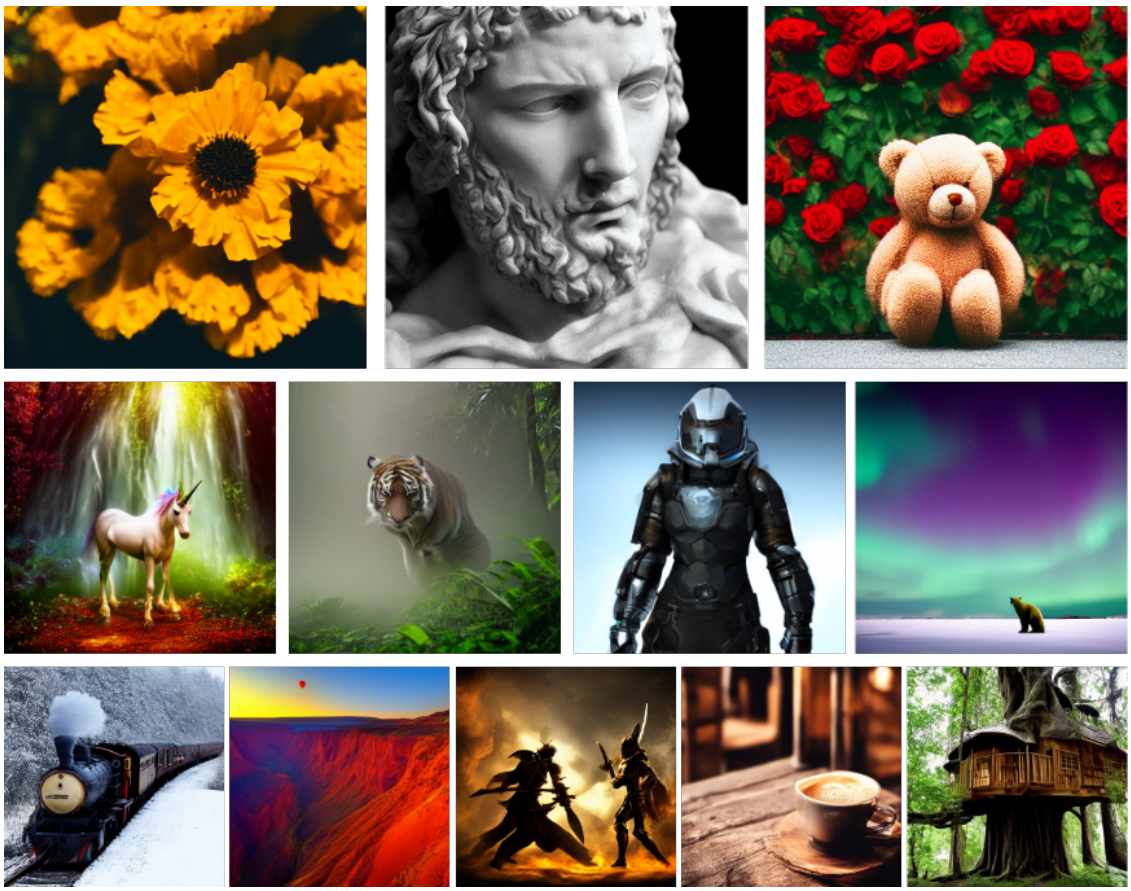

Figure 10. 4-step inference results of our Diff2Flow-reflow model, using Stable Diffusion 1.5 as the prior diffusion model

(AbsRel), as well as δ_1-Accuracy, which is the percentage of pixels for which the ratio between the aligned predicted depth and the ground-truth depth is below 1.25. Similar to Marigold [24], we train for 20k gradient updates with a batch size of 32 and a decaying learning rate schedule. For LoRA fine-tuning, we explore two variants: the first, "LoRA base" sets the rank to 20% of the respective feature dimension for all convolutional and attention layers, resulting in 222M trainable parameters. The second, a smaller LoRA model, fixes the rank to 64 across all convolutional, self-attention, and feedforward layers, resulting in 62M trainable parameters. We train our models on a resolution of 384×512. During evaluation, we resize the images to this size and subsequently resize our depth prediction to the ground truth resolution. We evaluate our models using an ensemble size of four and 10 sampling steps.

B. Qualitative Results

B.1. Reflow

In addition to the samples presented in Fig. 6, we provide further qualitative results in Fig. 10 and Fig. 11. By applying only the first rectified flow, Diff2Flow significantly reduces the number of diffusion generation steps while maintaining competitive performance compared to state-of-the-art flow matching approaches.

B.2. Image-to-Depth

In addition to the examples shown in Fig. 7, we present additional qualitative comparisons for our monocular depth estimation in Fig. 12. Our method consistently generates depth predictions with perceptually higher fidelity and finer details compared to the state-of-the-art models. Fig. 13 shows additional depth estimations of our Diff2Flow depth estimation model for in-the-wild images.

Figure 11. 2-step inference results of our Diff2Flow-reflow model, using Stable Diffusion 1.5 as the prior diffusion model

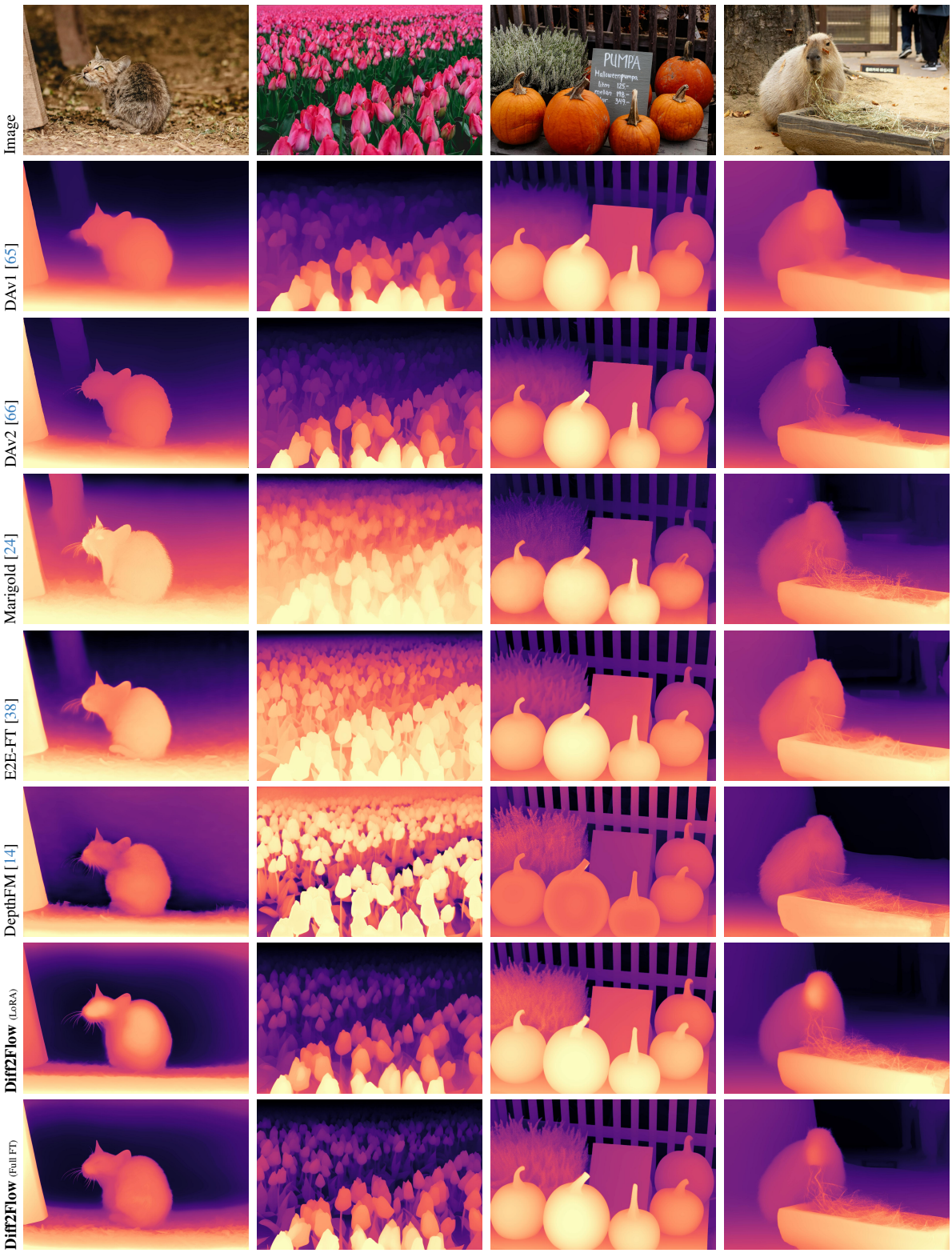

Figure 12. More qualitative results for monocular depth prediction compared to the state-of-the-art models (Part 1).

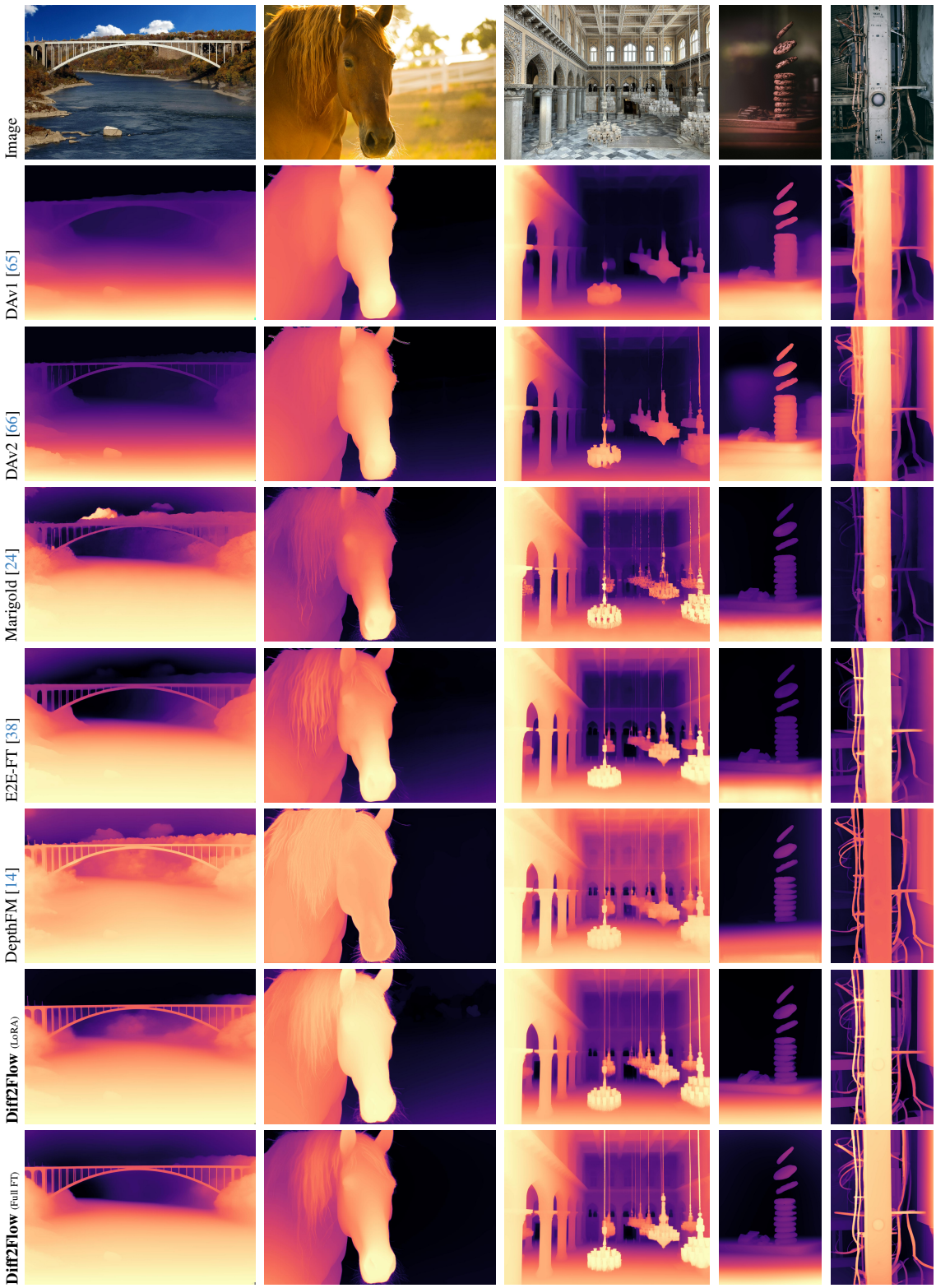

Figure 12. More qualitative results for monocular depth prediction compared to the state-of-the-art models (Part 2).

Figure 13. Qualitative results for monocular depth prediction.