MetricGold

Ansh Shah
IIIT Hyderabad
anshshahresearch@gmail.com

K. Madhava Krishna
IIIT Hyderabad

Code: https://github.com/AnshShah3009/MetricGold

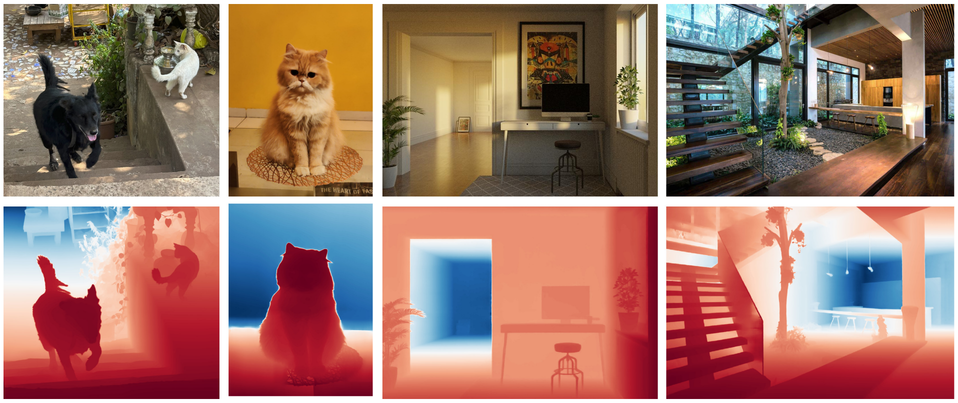

Figure 1: We present MetricGold, a diffusion model and associated fine-tuning protocol for monocular metric depth estimation. Its core principle is to leverage the rich visual knowledge stored in modern generative image models. Our model, derived from Stable Diffusion and fine-tuned with photorealistic synthetic data, can zero-shot transfer to unseen datasets, offering state of-the-art monocular depth estimation results.

Abstract

Monocular metric depth estimation poses a long-standing ill-posed challenge in computer vision, requiring an understanding of scene geometry and scaling to recover three-dimensional depth from a single image. While deep learning has driven significant advancements in this field, resulting in improved relative monocular depth estimators, these models often struggle with unfamiliar content and layouts due to their reliance on training data, limiting zero-shot generalization to new domains. This issue is further exacerbated when predicting scale-ergodic metric depth. To address these challenges, we introduce MetricGold, a method that leverages the extensive priors of generative diffusion models to enhance metric depth estimation. Inspired by MariGold [5] and employing techniques such as latent diffusion, log-scaled metric depth, and advanced loss functions, MetricGold can be fine-tuned on a single RTX 4090 GPU in just two days using photorealistic synthetic training data from HyperSIM, VirtualKitti [2], and TartanAir [13]. Our approach achieves state-of-the-art performance across various datasets, yielding sharper and more qualitatively superior depth estimations by mitigating sensor noise inherent in real-world data collection.

Loosing interest in a problem statement is sign of a curious and busy mind, At the same time putting effort to showcase your work so that people can replicate and understand it is our duty towards all people.

1 Introduction

Monocular metric depth estimation seeks to transform a single photographic image into a metric depth map, meaning it regresses a depth value for every pixel. This task is essential whenever understanding the 3D structure of a scene is required, but direct range or stereo measurements are unavailable. However, recovering the 3D structure from a 2D image is a geometrically ill-posed problem, as much of the depth information is lost in the projection process. Solving this issue requires prior knowledge, such as typical object shapes, sizes, common scene layouts, and occlusion patterns. In other words, monocular metric depth estimation implicitly demands an understanding of both scene geometry and scale. The emergence of deep learning has led to significant improvements in this field. Depth estimation is now treated as a neural image-to-image translation problem, typically learned in a supervised or semi-supervised manner using collections of paired, aligned RGB images and depth maps. Early approaches in this domain were limited by their training data, often focusing on specific environments like indoor scenes [10] or driving scenarios [2].

More recently, there has been a growing interest in developing general-purpose models, such as Metric Depth V2 [15] for metric depth and Marigold and Depth Anything V1 and V2 [5, 6, 14] for relative depth, which can be applied to a wide range of scenes without retraining or fine-tuned to specific applications with minimal data. These models often build on the strategy first introduced by MiDaS [9], which achieves generalization by training high-capacity models on data sourced from diverse RGB-D datasets across multiple domains. The latest advancements include a shift from convolutional encoder-decoder architectures [9] to larger, more powerful vision transformers [8], and incorporating surrogate tasks [3] to enhance the model's understanding of the visual world, leading to better depth predictions. For example, Depth Anything builds on this concept by training a DPT-style transformer for relative depth prediction over DINOV2 embeddings [7], while Depth Anything V2 introduces training on photo-realistic simulation datasets, such as [13], [10], and [2], instead of sensor-based datasets. They argue that sensor data introduces biases from sensor-specific corruptions, which the model might learn, potentially leading to suboptimal performance. Their results demonstrate that models trained on photo-realistic datasets produce sharper and more accurate depth predictions. Additionally, visual depth cues depend not only on the scene content but also on the (often unknown) camera intrinsics [16]. As shown in [14], relative depth can be predicted with strong generalization and accuracy, pushing the frontier of metric depth estimation even further.

The intuition behind our work is the following: Modern image diffusion models have been trained on internet-scale image collections specifically to generate high-quality images across a wide array of domains [1, 11, 12]. If the cornerstone of monocular depth estimation is indeed a comprehensive, encyclopedic representation of the visual world, then it should be possible to derive a broadly applicable depth estimator from a pretrained image diffusion model. In this paper, we set out to explore this option and develop Marigold, a latent diffusion model (LDM) based on Stable Diffusion [11], along with a fine-tuning protocol to adapt the model for depth estimation. The key to unlocking the potential of a pretrained diffusion model is to keep its latent space intact. We find this can be done efficiently by modifying and fine-tuning only the denoising U-Net. Turning Stable Diffusion into Marigold requires only synthetic RGB-D data (in our case, the Hypersim [10] and Virtual KITTI [2] datasets) and a few GPU days on a single consumer graphics card. Empowered by the underlying diffusion prior of natural images, Marigold exhibits excellent zero-shot generalization: Without ever having seen real depth maps, it attains state-of- the-art performance on several real datasets. To summarize, our contributions are:

2 Re-purposing Image Diffusion Models for Metric Depth

We present a method that leverages text-to-image diffusion models as an initialization for training a diffusion model to predict metric depth from an image. Specifically, we reformulate monocular metric depth estimation as a generative task, translating RGB images to depth maps using denoising diffusion. Additionally, we demonstrate qualitative results and describe the training process for repurposing components from the text-to-image diffusion model.

2.1 Diffusion Models

Diffusion models are probabilistic models that assume a forward noising process, progressively transforming a target distribution into a noise distribution. A neural denoiser is then trained to reverse this process iteratively, converting noise samples back into samples from the target distribution. These models have demonstrated exceptional effectiveness with images, videos, and relative depth. They are well-suited for this task because they achieve strong performance on regression problems without the need for specialized architectures, loss functions, or task-specific training procedures.

2.2 Using Photo-realistic Synthetic Datasets

Building on the pioneering work of MiDaS [9] in zero-shot relative depth estimation, recent approaches have focused on creating large-scale training datasets to improve estimation accuracy and generalization. Notably, Depth Anything V1, Metric3D V2 [15], and ZeroDepth [4] have compiled datasets ranging from 1M to 16M images sourced from various sensors. However, the data collected from different sensor suites inherently includes sensor-based noise, depth-related inaccuracies, illumination-induced uncertainty, and depth sensor pattern biases. Although preprocessing and cleaning are applied, these biases persist, acting as hidden contaminants that hinder learning a true depth distribution. Depth Anything V2 [14] addresses this issue by exclusively training on photo-realistic synthetic datasets, demonstrating sharper and qualitatively superior depth predictions. In our work, we train our model using the Hypersim and Virtual Kitti 2 datasets.

2.3 Joint Depth scaling for Indoor-Outdoor

Training a unified model for both indoor and outdoor environments is challenging due to significant variations in depth distribution, color, and scale. Indoor datasets typically contain depths under 10 meters, while outdoor datasets include ground truth depths up to 80 meters. Furthermore, learning a latent space for the diffusion model's metric depth is crucial. To address these issues, we propose using log depth as input to the VAE. Log depth directly addresses the indoor-outdoor depth distribution by applying a scaling factor early.

Log Depth Most neural networks usually model data distribution in [-1, 1]. One might naively convert metric depth to this range with linear scaling, given we know d_{min} and d_{max}, maximum and minimum range of depth, i.e.,

$$normalize = clip(2 * d - 1, -1, 1) \tag{1}$$

$$d_{\text{linear}} = normalize((d_r - d_{\text{min}})/(d_{\text{max}} - d_{\text{min}})) \tag{2}$$

We find this does not adequately account for the indoor-outdoor depth distribution. To address this, we model the depth using a logarithmic scale, shifting the uncertainty so that lower depth values receive more attention, while larger distances are relatively less important. By using log-scaled depth (d_{log}) as the target for inference, we allocate more representation capacity to indoor scenes.

$$d_{\text{log}} = normalize(log(d_r/d_{\text{min}})/log(d_{\text{max}}/d_{\text{min}})) \tag{3}$$

This adjustment not only improves the representation of depth across both indoor and outdoor distributions, but also leads to enhanced performance of the Variational Autoencoder (VAE). However, we identify that the reconstruction of depth maps from latent embeddings remains the primary performance bottleneck in the pipeline.

2.4 Model

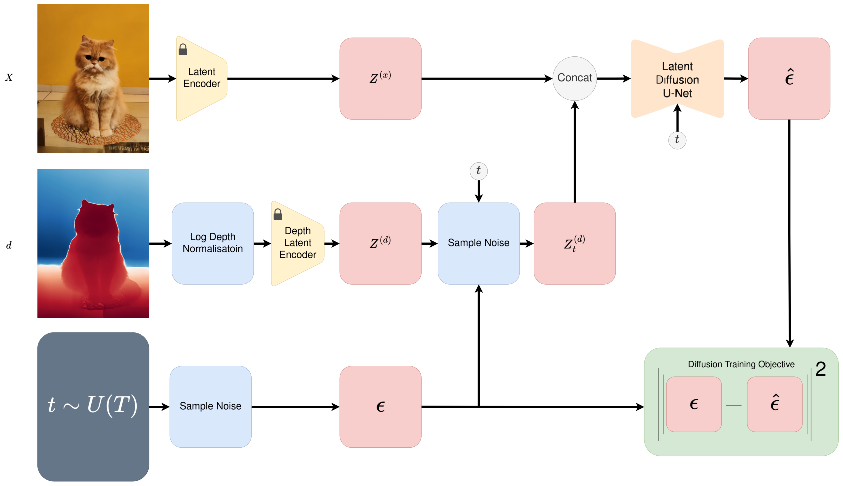

Figure 2: Your caption here

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023.

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

[3] 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 *ICCV*, pages 10786–10796, 2021.

[4] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023.

[5] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation, 2024.

[6] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. VA-DepthNet: A variational approach to single image depth prediction. In *ICLR*, 2023.

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[8] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

[9] 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 TPAMI*, 2020.

[10] 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 *ICCV*, 2021.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35, 2022.

[13] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam, 2020.

[14] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024.

[15] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023.

[16] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. NeWCRFs: Neural window fully-connected CRFs for monocular depth estimation. In *CVPR*, 2022.