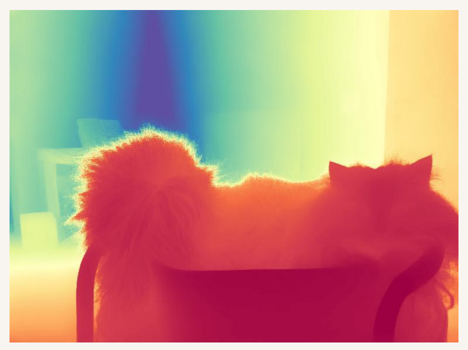
Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

Ridin Datta

EE798R Course Project

INTRODUCTION



ABSTRACT

Monocular depth estimation, transforming a 2D image to a depth map, remains a challenging task. Marigold, a diffusion model built on Stable Diffusion, leverages visual knowledge for zero-shot depth estimation on unseen datasets. Fine-tuned with synthetic data, Marigold demonstrates substantial improvements in monocular depth estimation across a wide variety of datasets. This method outperforms other methods on both indoor and outdoor scenes in most cases, without ever seeing a real depth sample.

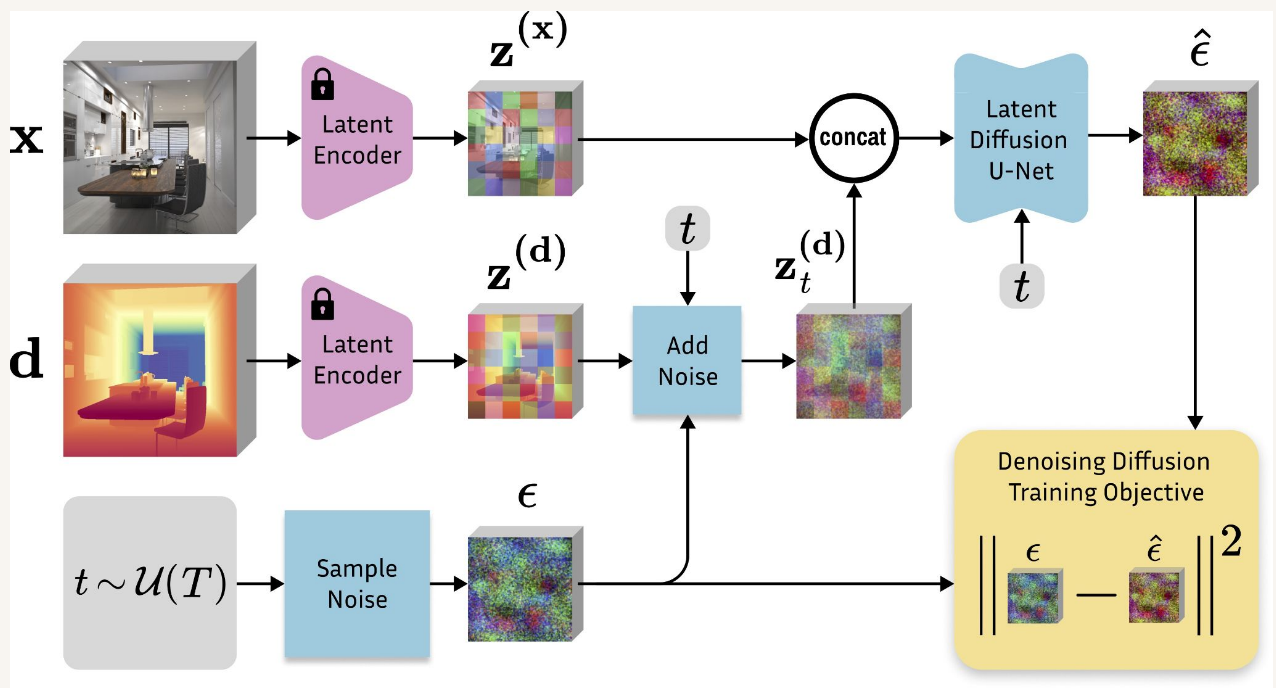
APPLICATIONS

- **Autonomous Driving**: Enabling accurate depth estimation in real-time without expensive LiDAR systems.
- **Augmented Reality (AR)**: Providing depth perception to enable realistic object placement and interaction in AR applications.
- **Robotics**: Assisting robots in depth-aware navigation and manipulation in unstructured environments

METHODOLOGY

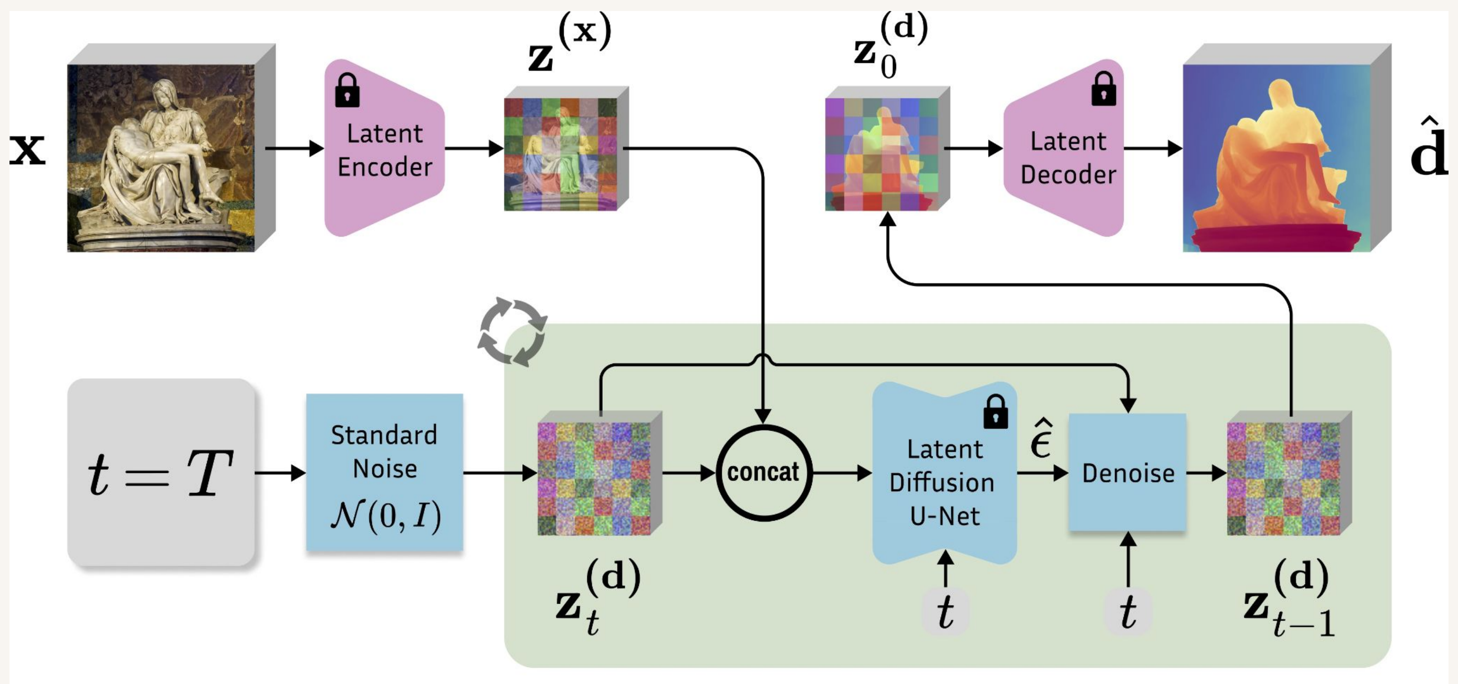
FINE-TUNING

Starting from a pretrained Stable Diffusion, we encode the image x and depth d into the latent space using the original Stable Diffusion VAE. We fine-tune just the U-Net by optimizing the standard diffusion objective relative to the depth latent code. Image conditioning is achieved by concatenating the two latent codes before feeding them into the U-Net. The first layer of the U-Net is modified to accept concatenated latent codes.



Method	# Training samples Real	Synthetic	NYUv2 AbsRel↓	$\delta1$↑	KITTI AbsRel↓	$\delta1$↑	ETH3D AbsRel↓	$\delta1$↑	ScanNet AbsRel↓	$\delta1$↑	DIODE AbsRel↓	$\delta1$↑	Avg. Rank
DiverseDepth [54]	320K	—	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	6.6
MiDaS [33]	2M	—	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	6.3
LeReS [55]	300K	54K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	4.3
Omnidata [11]	11.9M	310K	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	3.8
HDN [58]	300K	—	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	**78.0**	2.4
DPT [34]	1.2M	188K	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	**18.2**	75.8	3.1
Marigold (ours)	—	74K	**5.5**	**96.4**	**9.9**	**91.6**	**6.5**	**96.0**	**6.4**	**95.1**	30.8	77.3	**1.4**

INFERENCE SCHEME



Given an input image x we encode it with the original Stable Diffusion VAE into the latent code z(x) and concatenate with the depth latent zt(d) before giving it to the modified fine-tuned U-Net on every denoising iteration.

After executing the schedule of T steps, the resulting depth latent z0(d) is decoded into an image, whose 3 channels are averaged to get the final estimation d^

OPTIMIZATION

IMAGE COMPLEXITY

$$C = \mathbb{E}\left[\|\nabla x\|^2\right] \quad \text{or}$$

$$C = H(x) = -\sum_{i=1}^{N} p(x_i) \log p(x_i)$$

STEP SELECTION

$$T(C) = T_{\min} + \alpha(C - C_{\min})$$

where α is the scaling factor.

ADAPTIVE DENOISING

$$z_{t-1} = z_t - \alpha_t \epsilon_\theta(z_t, x, t)$$

where αt is a noise schedule parameter.

RESULTS

Adapting T(C) based on image complexity, reduces the number of denoising steps for simpler images, resulting in significant computational savings without compromising depth estimation accuracy.

Original Average Inference Time	Optimized Average Inference Time
172 seconds	145 seconds

This shows a substantial improvement, with average inference time decreased by approximately 15% without compromising performance.