

Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

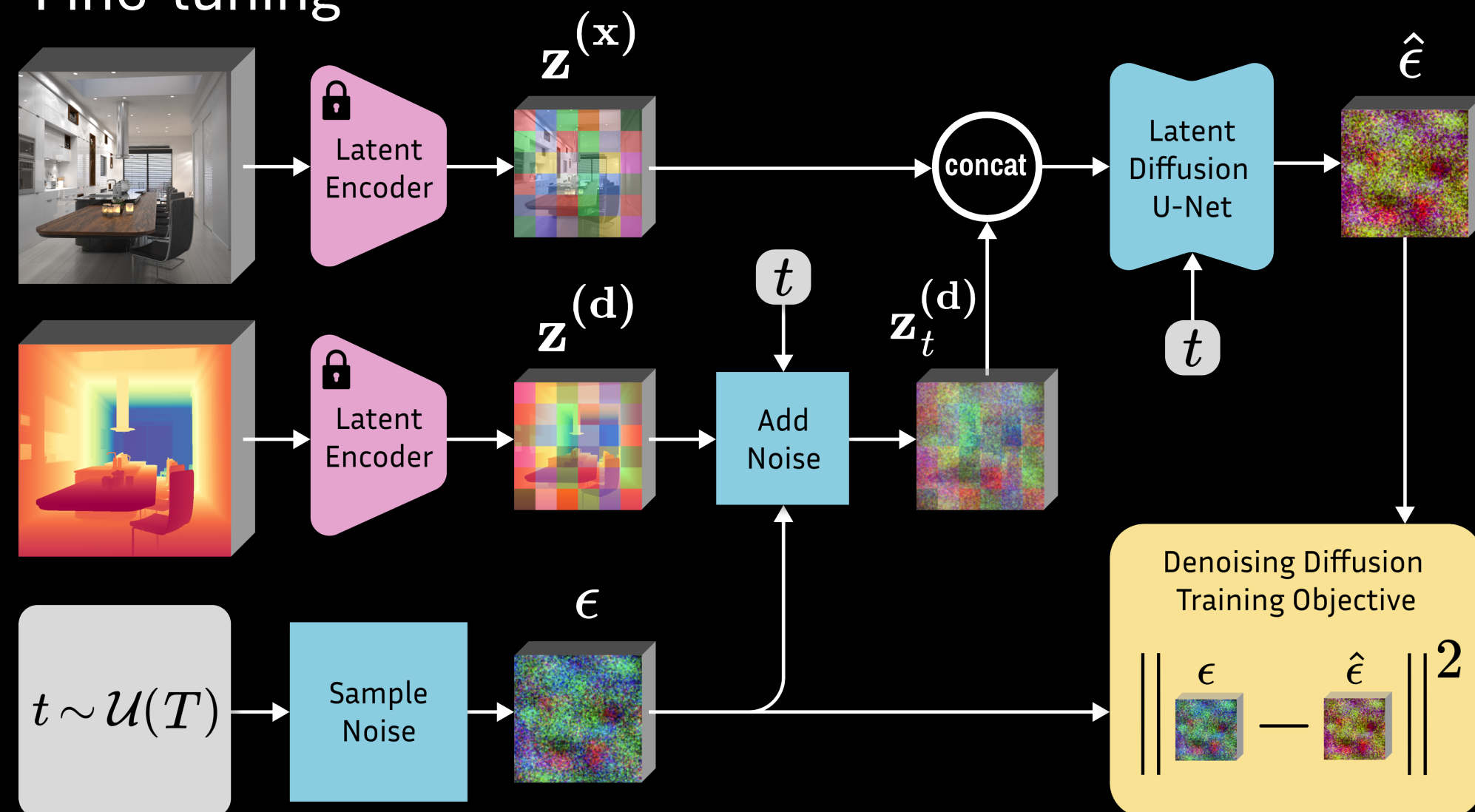
Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, Konrad Schindler

Marigold Recipe:

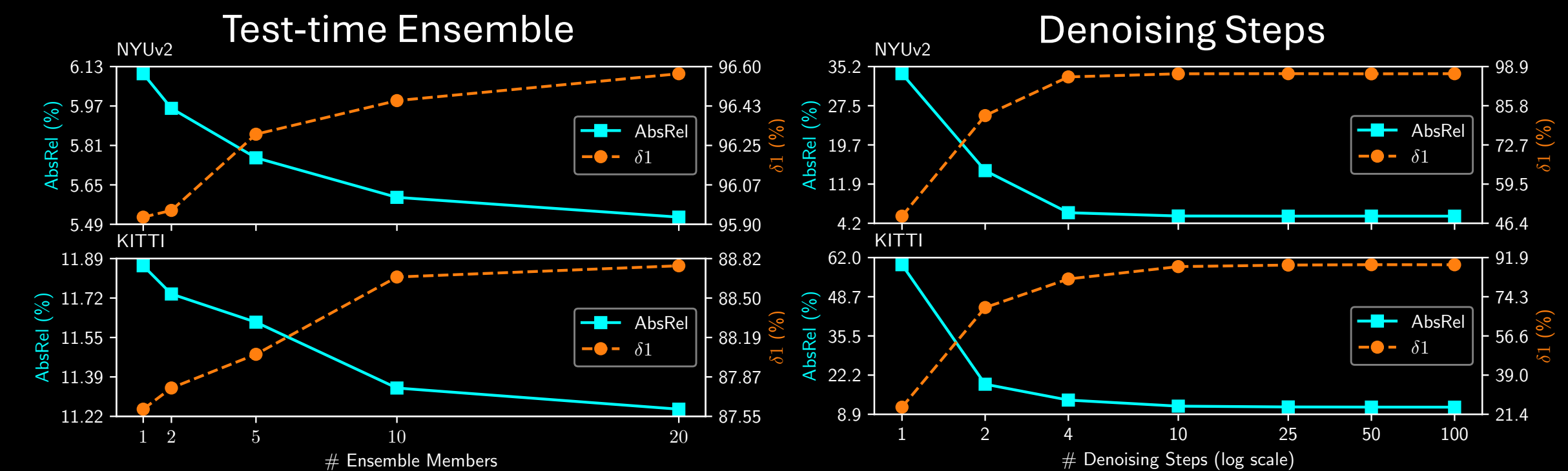
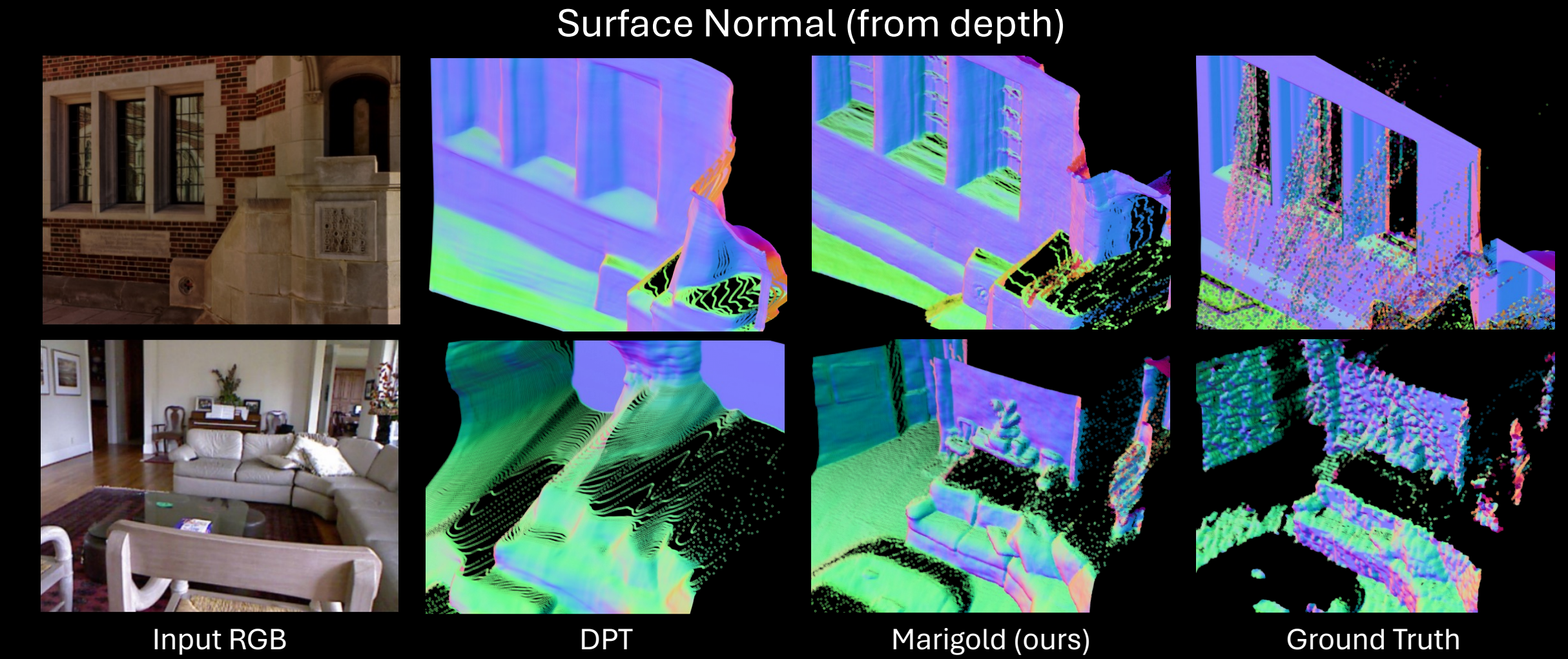
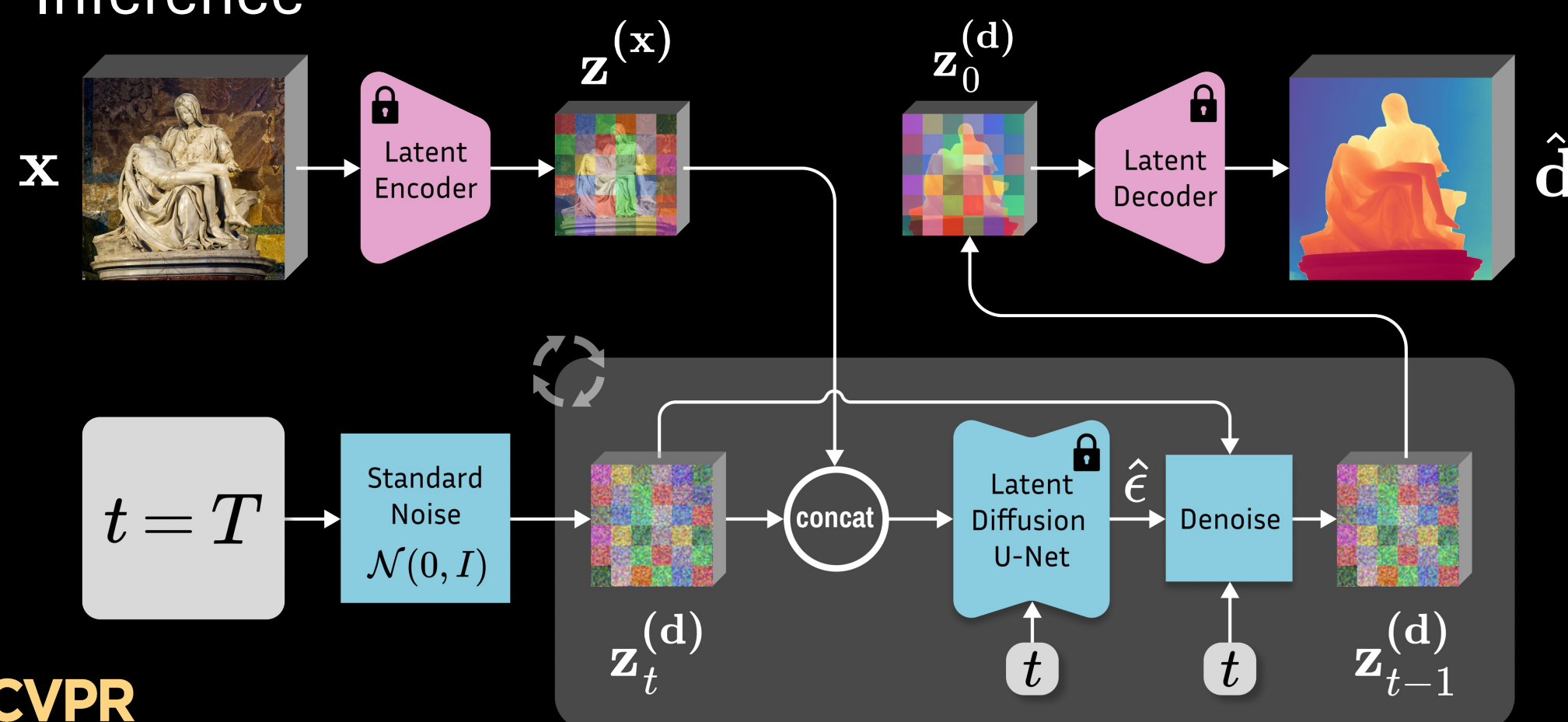
- Rich prior from a pretrained generative image latent diffusion model
- 74K synthetic RGB-D samples
- One consumer-level GPU
- Fine-tune for < 3 days
- Serve within 280ms(1-step LCM) / 85ms(1-step LCM+TAESD)



Fine-tuning



Inference



Method		# Training Samples		NYUv2		KITTI		ETH3D		ScanNet		DIODE		Avg. Rank
		Real	Synthetic	AbsRel↓	δ1↑	AbsRel↓	δ1↑	AbsRel↓	δ1↑	AbsRel↓	δ1↑	AbsRel↓	δ1↑	
DiverseDepth		320K	--	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	7.6
MiDaS		2M	--	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	7.3
Leres		300K	54K	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6	5.2
OmniData		11.9M	310K	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2	4.8
HDN		300K	--	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	78.0	3.2
DPT-Large		1.2M	188K	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	18.2	75.8	3.9
Ours (w/o ensemble)		--	74K	6.0	95.9	10.5	90.4	7.1	95.1	6.9	94.5	31.0	77.2	2.5
Ours (ensemble x10)		--	74K	5.5	96.4	9.9	91.6	6.5	96.0	6.4	95.1	30.8	77.3	1.4

* Image-text data is used in the pretrained model.

Our method outperforms other methods on both indoor and outdoor scenes in most cases, without having seen a real depth sample. Qualitatively, Marigold stands out for its superior reconstruction of detailed structures and flat surfaces while preserving overall layout of the scene.

Extension – Efficient Inference

- Distilled Latent Diffusion Model → 1-step inference
- Further acceleration with smaller autoencoder (e.g. TAESD)