EcoDepth: Effective Conditioning of Diffusion Models for Monocular Depth Estimation

Suraj Patni* Aradhye Agarwal* Chetan Arora
Indian Institute of Technology Delhi
https://ecodepth-iitd.github.io

1. Ablation Study

1.1. Effect of `ViT` Architecture

Table 1 investigates the impact of varying `ViT` sizes on the generation of embeddings from RGB images. Our results for the `NYU Depth v2` [4] dataset suggest that ViT-base yields optimal performance. Additionally, our observations in the `KITTI` dataset align with a similar trend.

Table 1. **Ablation Study on ViT Sizes:** Performance comparison of different ViT variants in terms of parameters and depth error metrics on the NYUv2 [4] dataset. The results guide the selection of ViT-base in our final architecture. Best results are in **bold**.

Classifier	#Parameters	RMSE↓	Abs Rel↓	δ_1 ↑
ViT-base	86.6 M	**0.218**	**0.059**	**0.978**
deit-base	86.6 M	0.218	0.059	0.978
ViT-large	303.3 M	0.218	0.060	0.978
ViT-huge	630.8 M	0.219	0.060	0.978

Table 2. **Ablation Study on dimension of Learnable Scene Embeddings (N):** The table shows the impact of varying the dimension of learnable scene embeddings on the depth error metrics. We observe a decrease in error with increasing N until saturation occurs at N=100, prompting us to limit the model parameters to N=100. Best results are highlighted in **bold**.

N	RMSE↓	Abs Rel↓	$\log_{10}$ ↓	δ_1 ↑
10	0.219	0.061	0.027	0.978
50	0.219	0.060	0.026	0.978
100	**0.218**	**0.059**	**0.026**	**0.978**
200	0.218	0.060	0.026	0.978

1.2. Additional Qualitative Ablation

In Fig. 2, we present supplementary qualitative ablation results that highlight the correlation between value of ViT logits and the improvement in the predicted depth. The visualization demonstrates that elevated value of ViT logits for specific objects contribute to our model's ability to focus on those objects, enhancing the accuracy of predicted depth in corresponding regions.

*Equal contribution.

2. Architectural Details

2.1. Image Encoder

Similar to Latent Diffusion [6], we employed the VQVAE's encoder to transition from image space to latent space.

2.2. Upsampling Decoder

After obtaining the hierarchical feature map from denoising UNet, the concatenated feature map undergoes upsampling, transitioning from a resolution of 64×64 back to $H \times W$. Refer to Fig. 1 for a detailed view of the upsampling decoder architecture.

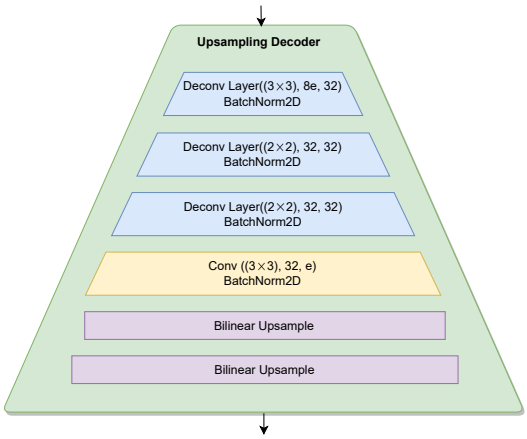

Figure 1. Detailed architecture of the upsampling decoder, responsible for upsampling the concatenated feature map to obtain the final feature map at a resolution of $H \times W$, $e = 192$

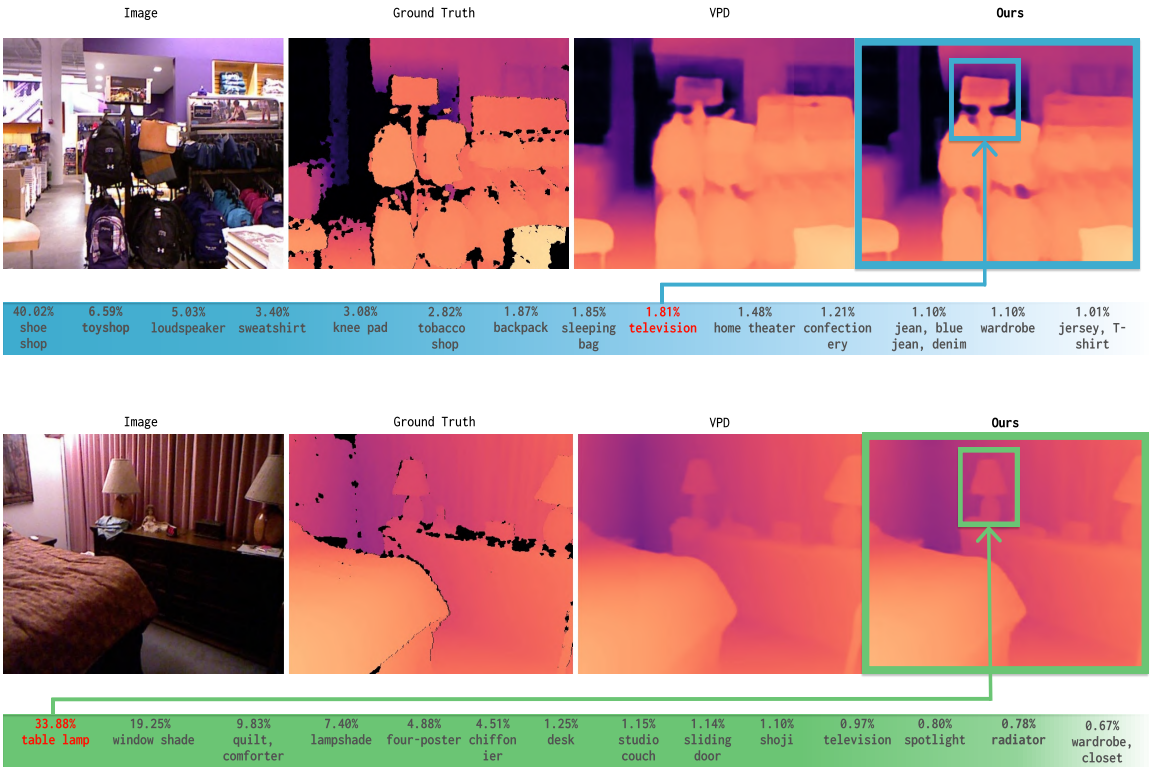

Figure 2. Enhanced visualizations showcasing improvements over VPD [9] in our model, facilitated by ViT embeddings employed as conditional vectors for the denoising procedure. In the presented images, our model demonstrates heightened accuracy in detecting objects, such as the television (blue in first image) and table lamp (green in second image) when these are detected with high probability by ViT.

Table 3. Hyper-parameter settings for our model.

Hyper-parameter	Value
Learning rate schedule	one cycle
Min learning Rate	3×10^{-5}
Max learning Rage	5×10^{-4}
Batch Size	32
Optimizer	AdamW [3]
β_s in optimizer	(0.9, 0.999)
Weight Decay	0.1
Layer Decay Rate	0.9
Embedding Dimension	192
Variance focus in SiLog loss	0.85
ViT Size	ViT-base
Number of learnable emb.	100
epochs	25

3. Additional Experimental Details

3.1. Hyperparameters

For reproducibility of the results presented in the main paper and the supplementary material, we provide a comprehensive list of the hyper parameters employed in our experiments in Table 3.

4. Qualitative Results for Zero-Shot Performance Across Datasets

In the main paper, we presented a quantitative comparison of our method's zero-shot performance. Here, we provide a qualitative assessment of our method's performance in comparison to ZoEDepth [1] across the HyperSim, DIODE, Sun-RGBD and iBims1 datasets in Fig. 3, 4, 5 and 6.

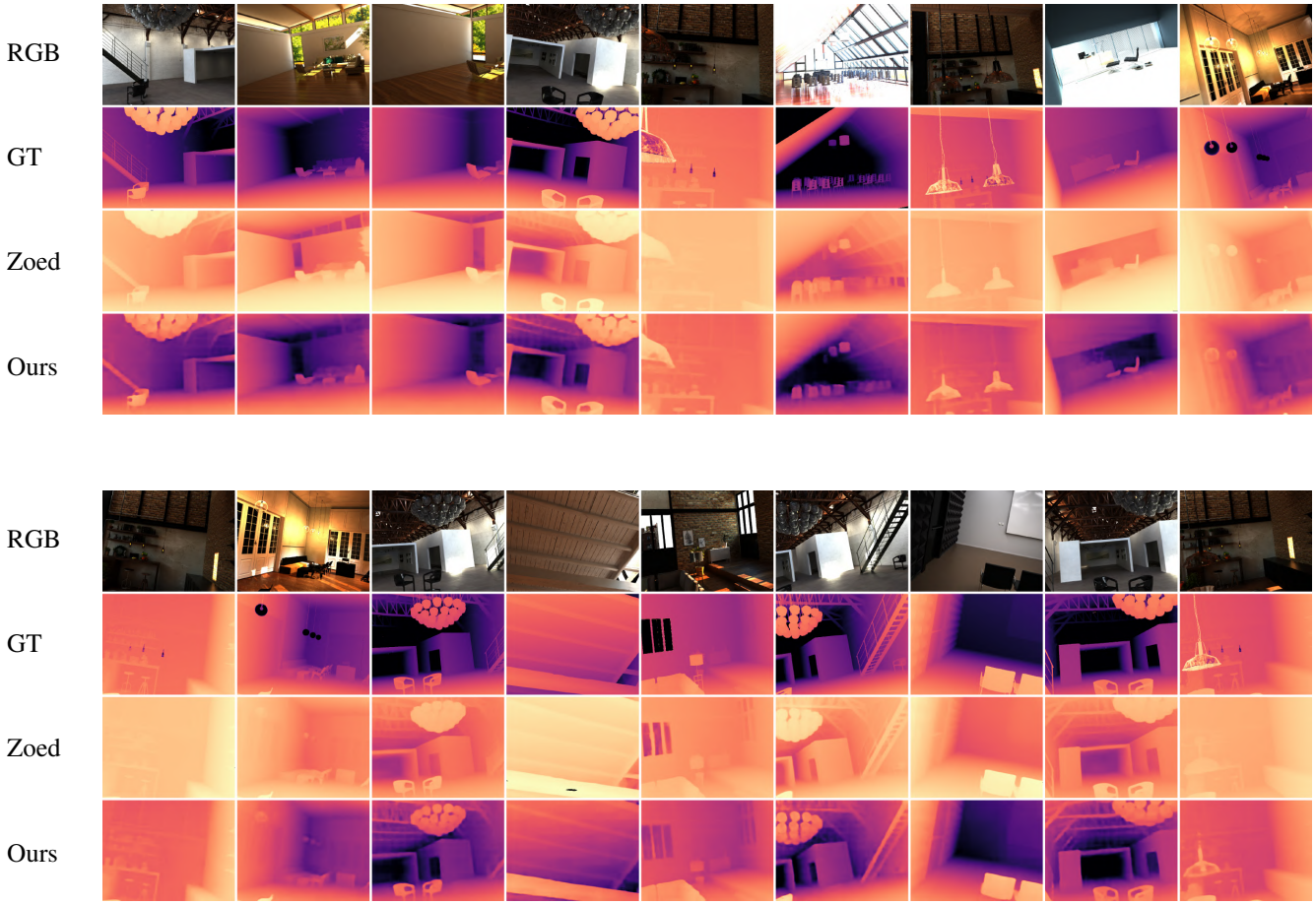

Figure 3. **Qualitative Comparison on the `HyperSim` [5] Dataset.** Our depth predictions are contrasted with those of Zoedepth[1]. The first row displays RGB images, the second row shows groundtruth depth, the third row exhibits Zoedepth[1]'s depth, and the fourth row showcases our depth predictions. To facilitate visual comparison, the colormap scale remains consistent across corresponding depth maps. Our model, trained only on `NYU Depth v2`, is compared with Zoedepth[1], which is trained on 12 datasets and then fine-tuned on `NYU Depth v2`.

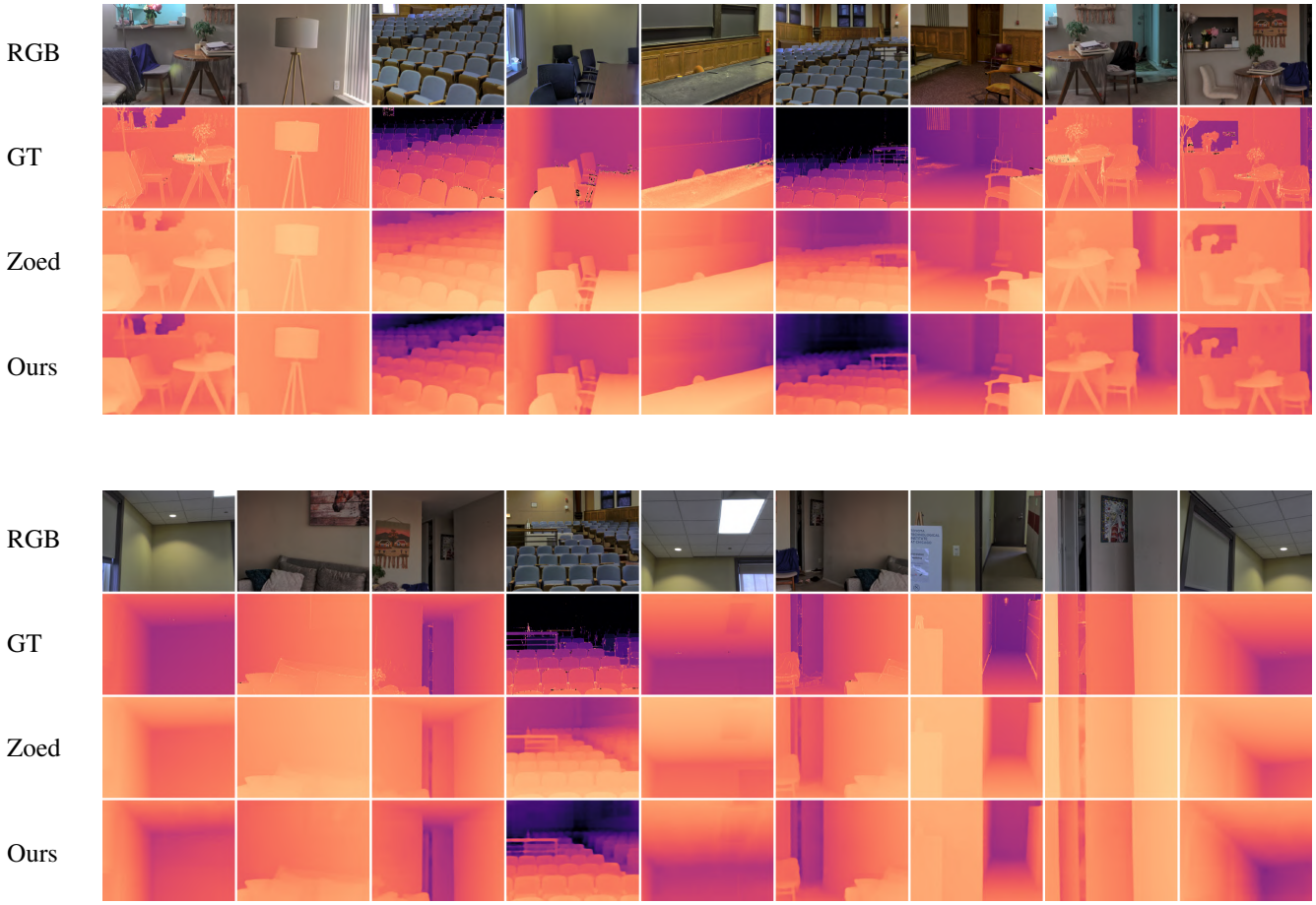

Figure 4. **Qualitative Comparison on the DIODE [8] Dataset.** Our depth predictions are contrasted with those of Zoedepth[1]. The first row displays RGB images, the second row shows groundtruth depth, the third row exhibits Zoedepth[1]'s depth, and the fourth row showcases our depth predictions. To facilitate visual comparison, the colormap scale remains consistent across corresponding depth maps. Our model, trained only on `NYU Depth v2`, is compared with Zoedepth[1], which is trained on 12 datasets and then fine-tuned on `NYU Depth v2`.

Figure 5. **Qualitative Comparison on the `Sun-RGBD` [7] Dataset.** Our depth predictions are contrasted with those of Zoedepth[1]. The first row displays RGB images, the second row shows groundtruth depth, the third row exhibits Zoedepth[1]'s depth, and the fourth row showcases our depth predictions. To facilitate visual comparison, the colormap scale remains consistent across corresponding depth maps. Our model, trained only on `NYU Depth v2`, is compared with Zoedepth[1], which is trained on 12 datasets and then fine-tuned on `NYU Depth v2`.

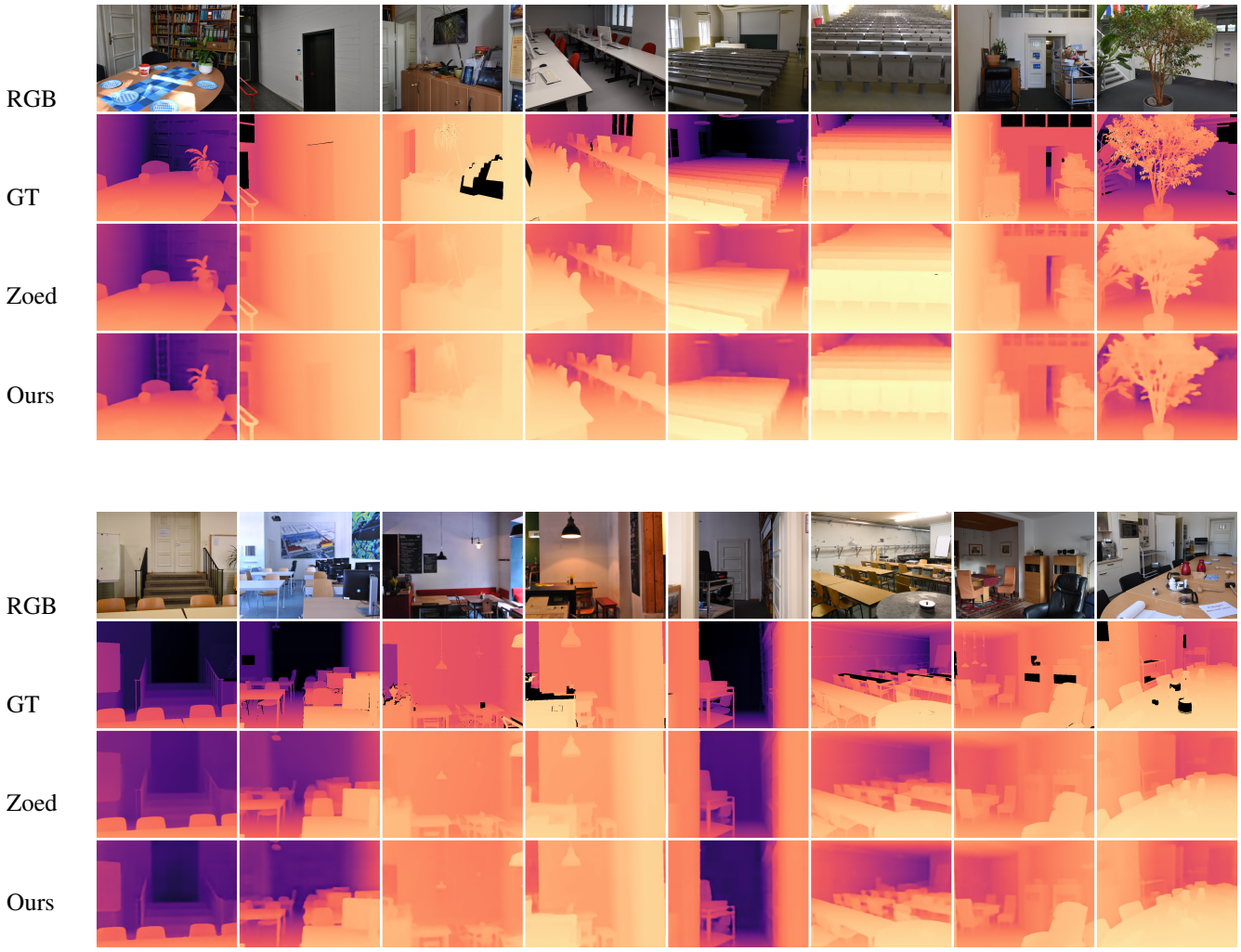

Figure 6. **Qualitative Comparison on the `iBims1` [2] Dataset.** Our depth predictions are contrasted with those of Zoedepth[1]. The first row displays RGB images, the second row shows groundtruth depth, the third row exhibits Zoedepth[1]'s depth, and the fourth row showcases our depth predictions. To facilitate visual comparison, the colormap scale remains consistent across corresponding depth maps. Our model, trained only on `NYU Depth v2`, is compared with Zoedepth[1], which is trained on 12 datasets and then fine-tuned on `NYU Depth v2`.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 2, 3, 4, 5, 6

[2] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6

[3] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICML 2019*. 2

[4] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[5] 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 *International Conference on Computer Vision (ICCV) 2021*, 2021. 3

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[7] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 5

[8] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 4

[9] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 2