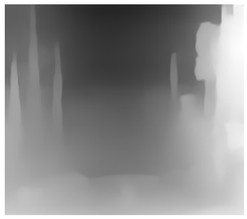
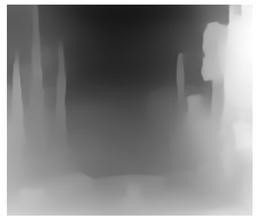
Visual Prompting for Depth Estimation

Taqiya Ehsan[1], Hyojin Bahng[2] and Phillip Isola[2]

[1]Dept. of Electrical & Computer Engineering, Rutgers University

[2]Dept. of Electrical Engineering & Computer Science, Massachusetts Institute of Technology

Introduction

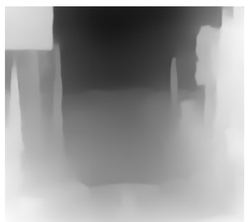


Figure: Depth estimations from occluding different prompt sizes; Original image (top left); Depth map - without prompting (top middle), from a **1x1** prompt (top right), from a **10x10** prompt (bottom left), from a **50x50** prompt (bottom right).

This work studies –

- Large depth estimation models that perform favorably to state-of-the-art
- Training a visual prompt in pixel-space on top of a pretrained monocular depth estimation model called **MiDaS**
- **Analysis of the effect of attaching a prompt to stereo images and the relative depth maps from those**

Methods

We trained and validated --

- **6** different prompt sizes: 1, 10, 50, 75, 100, 150, 200, 256
- at learning rate: **0.1**
- for **100** epochs

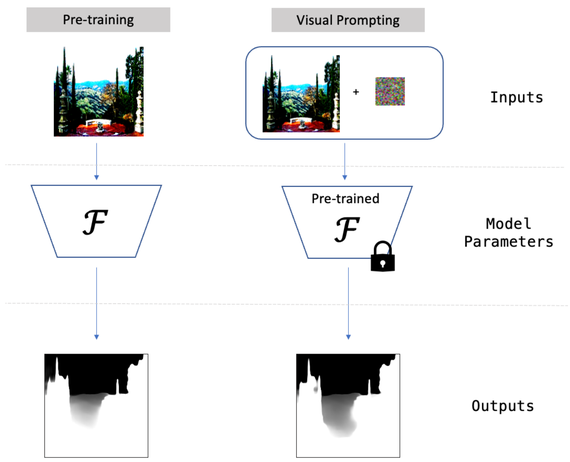


Figure: Method for training a visual prompt for depth estimation in pixel space.

Evaluation & Results

Evaluation Metrics

1. Root Mean Square Error (rms)
2. Absolute Relative Error (rel)
3. Average $\log_{10}$ Error (log10)
4. Accuracy (δ) with Threshold 1.25
5. Structural Similarity Index Measure (ssim)

Empirical Representation of Evaluation Metrics

	Accuracy			Error			ssim
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	rel	rms	log10	
Unprompted	0.101	0.116	0.135	4.2E+15	416.93	0.859	0.17636
Prompted	**0.179**	**0.283**	**0.368**	1.3E+16	**237.37**	1.285	**0.23268**

Table: Comparison of error, accuracy, and structural similarity of prompted and unprompted images with the ground truth of the dataset. **Prompting outperforms in 5 out of 7 metrics**.

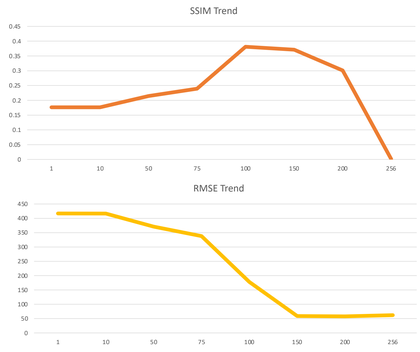


Figure: Evaluation metrics for prompt size ablation study

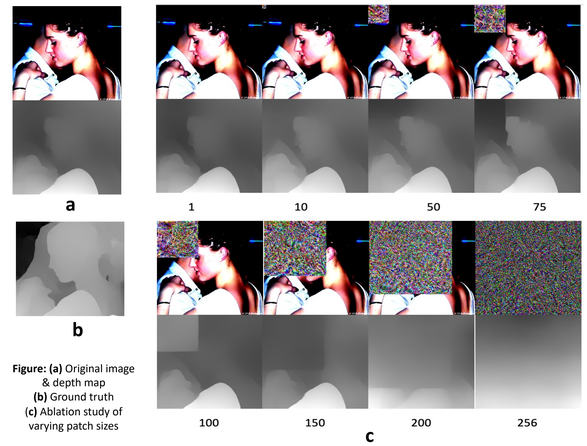


Figure: (a) Original image & depth map **(b)** Ground truth **(c)** Ablation study of varying patch sizes

Analysis & Inference

Key Observations:

- larger prompt size **improves** accuracy by reducing error, but **obscures** the image
- structural similarity **improves** progressively up to prompt size 100 and then **plummets**

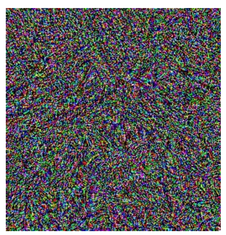


Figure: Visual representation of an optimized prompt

Final Deductions:

- Increasing patch sizes gives the prompt more scope to manipulate the original depth map and emulate relative pixel density of ground truth
- Smaller patch sizes can improve upon the original depth prediction, without obscuring, making it more structurally like the ground truth

Future Work

Robustness experiment:

- Location
  - Randomize location during training
  - Test time: place the prompt in different locations and visualize
- Randomize location, rotation, size of the prompt

Ablation on model scale:

- Run large Dense Prediction Transformer (DPT) model

References

- Bahng, H., Jahanian, A., Sankaranarayanan, S., & Isola, P. (2022). *Visual Prompting: Modifying Pixel Space to Adapt Pre-trained Models. arXiv preprint arXiv:2203.17274*. https://arxiv.org/pdf/2203.17274.pdf
- Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). *Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence*. https://arxiv.org/pdf/1907.01341v3.pdf
- Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., & Luo, Z. (2018). *Monocular relative depth perception with web stereo data supervision.* In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 311-320). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8578138&tag=1

Acknowledgements

I would like to thank everyone at Phillip Isola's group for their support, feedback, and insightful discussions throughout the summer. I would also like to extend my heartfelt gratitude to Noelle and Maria for fostering a supportive and welcoming community for the whole MSRP cohort. This study is part of an ongoing project led by Hyojin Bahng, partially supported by funding from MIT STL and an MIT RSC award from the NEC fund.