Visual Prompting for Depth Estimation

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

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

Abstract

We study the effect of *visual prompting* on the downstream vision task of depth estimation. In this paper, we establish a proof of concept of the efficacy of adding an optimized visual prompt in input space to the accuracy of generated depth maps by a pre-trained depth estimation model. We perform ablation studies on prompt size and learning rate to determine the most optimal hyperparameters for training a prompt in pixel space. We further perform experiments on out-of-distribution datasets and analyze their zero shot performance versus performance after visual prompting. Our studies produce promising results, opening up prospects of new ways to adapt different downstream tasks to large pre-trained vision models.
Github: github.com/taqiyaehsan/depth-estimation-visual-prompting

1. Introduction

Depth Estimation is the task of measuring the distance of each pixel relative to the camera. Depth is extracted from either monocular (single) or stereo (multiple views of a scene) images. Traditional methods use multi-view geometry to find the relationship between the images. Newer methods can directly estimate depth by minimizing the regression loss, or by learning to generate a novel view from a sequence [6]. The motivation behind this study was exploring ways to improve upon already existing depth estimation benchmarks by using state-of-the-art pre-trained models for the task.

The current standard method for full fine-tuning of a pre-trained vision model is to make adjustments to the weights and parameters. This method is expensive in terms of GPU capacity and procedural complexity. We may also not have access to the model at all. This makes fine-tuning already existing vision models or adapting them to distribution shifts challenging and costly.

In proposing our solution, we try to answer the question: Can we improve the performance of a pre-existing depth estimation model and make it robust to distribution shifts without access to the model? We propose *visual prompting* to train a prompt in pixel space of the input data and optimize it through backpropagation for the specific downstream task of depth estimation. The optimized visual prompt improves zero shot performance of a pre-trained depth estimation model keeping the model weights and parameters frozen. Evidently, our solution does not require access to the inner workings of the model; rather the foundation of our technique is taking a pre-trained model and keeping it intact throughout.

Prompting is a well-known concept in Natural Language Processing for adapting large language models to new downstream tasks. Recently, there have also been works trying to implement prompting in computer vision [1, 3]. So far there have been studies for adapting new image classification

tasks to pre-trained vision models. Our work studies depth estimation models that perform favorably to state-of-the-art and trains a visual prompt in pixel-space on top of a pre-trained monocular fully-convolutional depth estimation network called **MiDaS** [4] as well as its vision transformer backbone [5]. This study examines the effect of visual prompting compared to zero-shot performance on both pre-training [7] and out-of-distribution datasets [6].

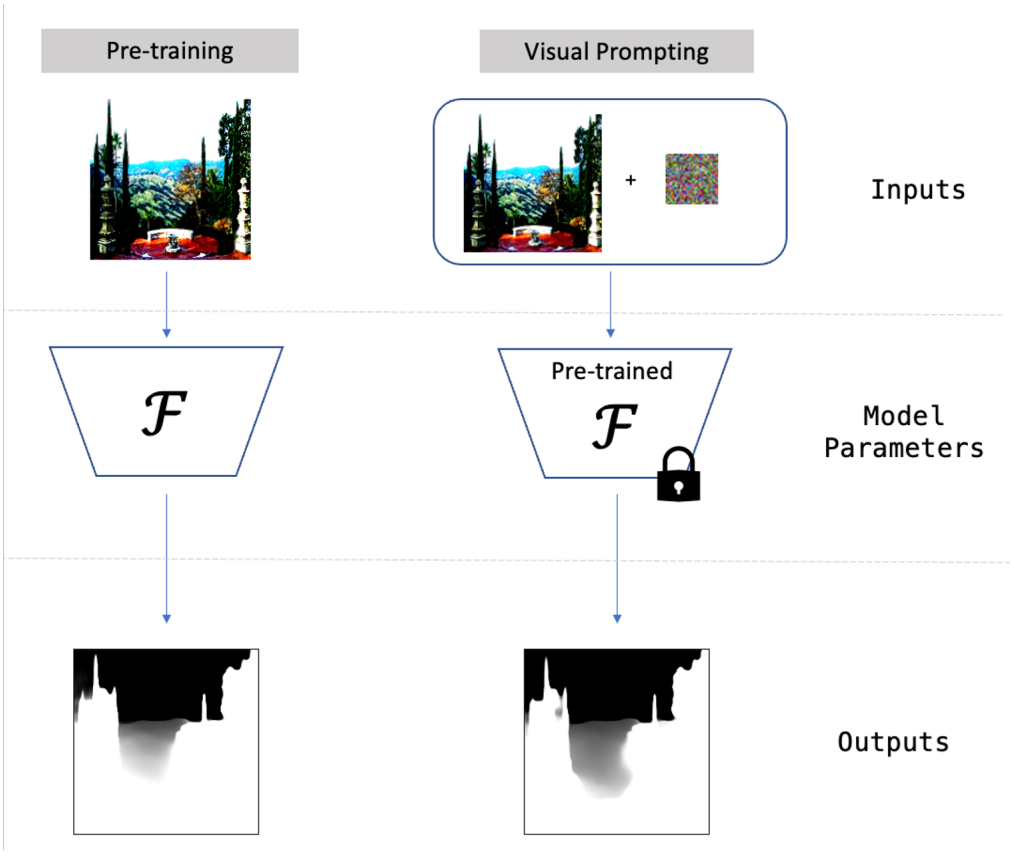

Fig. 1: Method for training a visual prompt in pixel space for depth estimation

2. Methods

The foundation of our method is centered around selecting a robust pre-trained model and optimizing a unique prompt design for the downstream task of depth estimation. In this section, we define different design choices for our work.

2.1. Pre-trained Model

Prompts in pixel space can be attached to basically any visual representation. We, therefore, start with a monocular depth estimation model called *MiDaS* [4]. MiDaS has been trained on five diverse depth datasets with a robust training objective that is invariant to changes in depth range and scale, advocating the use of principled multi-objective learning to combine data from different sources, and highlighting the importance of pre-training encoders on auxiliary tasks. This approach outperforms

competing methods across diverse datasets, setting a new state-of-the-art for monocular depth estimation. MiDaS provides both a convolutional neural network and a vision transformer [5] backbone. For monocular depth estimation, an improvement of up to 28% in relative performance was observed in the vision transformer when compared to a state-of-the-art fully-convolutional network.

We start by implementing the simplest fully-convolutional network from MiDaS and then move on to adapting our pipeline to the vision transformer backbone.

2.2. Prompt Engineering

For the visual prompt, we closely follow the protocol from Bahng et al. [1]. Their work proposes three different ways of prompting – square patch at fixed location, square patch at random location, and padded prompt. We start with the simplest square patch added to the top left corner of the input image.

Fig. 2: Prompting in Computer Vision

The prompt is initialized as a tensor with a set of randomized pixels constrained between (-1, 1). The model then optimizes this prompt through backpropagation using the Adam optimizer. During this optimization process, the model only updates the prompt; the pre-trained model and its parameters remain untouched. During testing, the optimized prompt is added to all test images and processed through the frozen model.

2.3. Implementation Details

During initial experiments with the CNN model, we apply the simplest mean squared error loss and Adam optimizer. We perform ablation over varying prompt sizes (1, 10, 50, 75, 100, 150, 200, 256) and learning rates (0.001, 0.01, 0.1, 1, 10). All images are resized to 256x256 and transformed so as to match the input parameters of the fully-convoluted backbone [4].

For experiments using the vision transformer backbone, we try to recreate the training conditions proposed in [5]. We implement Scale and Shift Invariant loss with Adam optimizer at learning rate 0.1 and prompt size 75 obtained from our initial studies. All images are resized to 384x384 and transformed like before to match the input parameters of the dense prediction transformer.

3. Experiment Setting

3.1. Dataset

For our preliminary proof of concept, we started with a dataset that was used in the training set of MiDaS – ReDWeb [7]. This is a monocular relative depth perception dataset compiled from stereo images from the web. There are 3600 images with a varied set of scenarios – indoors and outdoors, with a moderate accuracy of ground truth depth maps. We used ReDWeb to test our hypothesis of whether visual prompting improves depth predictions or not on the fully-convoluted network.

For the second phase of our study, we used two out-of-distribution datasets – DIODE [6] and CityScapes [2] with the vision transformer backbone. DIODE is a public dataset with diverse high-resolution RGB images of indoor and outdoor scenes with accurate, dense, far-range depth measurements obtained with one sensor suite. CityScapes is a new large-scale dataset that contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities, with high quality pixel-level annotations of 5000 frames in addition to a larger set of 20,000 weakly annotated frames. Both datasets consist of approximately 3500 color images along with their depth map labels.

3.2. Baseline Methods

To evaluate how our proposed technique performs on depth predictions made by MiDaS, we compare zero-shot performance with performance after visual prompting. MiDaS' base models outperform the baselines of its test datasets by a comfortable margin in terms of zero-shot performance [4]. Further improvement on top of this by visual prompting would validate our proposed hypothesis.

4. Results

4.1. Effectiveness on Prompting

ReDWeb was used in initial training of the fully-convolutional network backbone of MiDaS. Despite that, we were able to observe improvements in depth maps produced from the images after applying our visual prompt. The prompted images outperformed the original in 5 out of 7 metrics, thus validating that visual prompting can improve upon state-of-the-art.

The evaluation metrics used for our analyses are:

- Root Mean Square Error (rms)
- Absolute Relative Error (rel)
- Average log10 Error (log10)
- Accuracy with Threshold 1.25
- Structural Similarity Index Measure (ssim)

We also performed an ablation study of different patch sizes and their corresponding depth estimations. From this, we observed that the depth predictions got better (closer to ground truth) as prompt size increased from 1 to 75, but then the larger prompts obscured parts of the original image, thus obscuring the generated depth map.

Table 1: Comparison of evaluation metrics for prompting on ReDWeb

Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	rel	rms	log10	ssim
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**

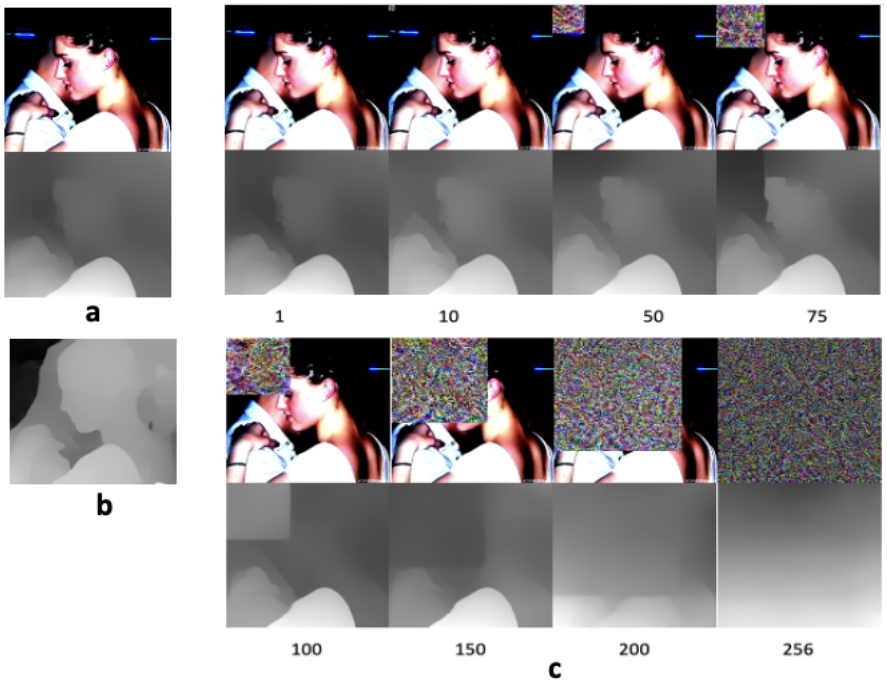

Fig. 3: (a) Original image & depth map (b) Ground truth (c) Ablation study of varying patch sizes

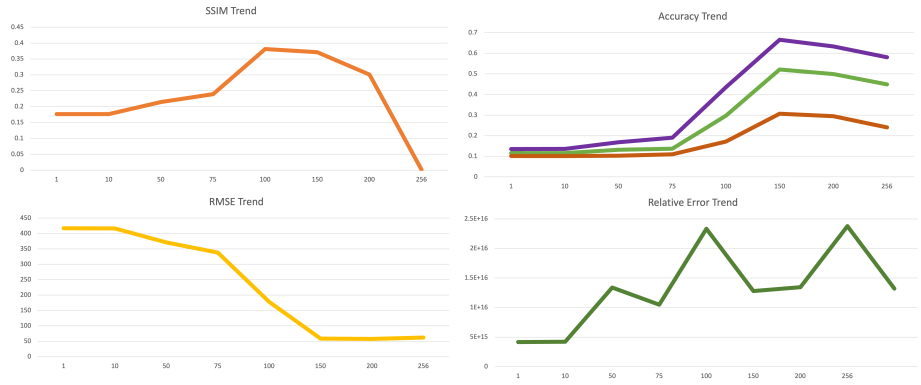

Fig. 4: Evaluation metrics for prompt size ablation study

This trend can be explained using the graphs obtained from analysing the same evaluation metrics

for each different prompt size. Accuracy goes up and error goes down with increasing prompt sizes and the predictions become more similar to ground truth with respect to pixel density. However, structurally, the depth predictions become imperceptible – as a result, the structural similarity index peaks at prompt size 100 and then decreases. More precisely:

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

4.2. Robustness to Distribution Shifts

The technique of visual prompting was based off adapting new tasks on a pre-trained frozen vision model. That is to say, the model parameters remain frozen throughout training of the prompt, thus preventing the prompt from modifying the general knowledge base of the model. Consequently, the possibility of overfitting to wrongful correlations is reduced and robustness to distribution shift is improved. We use training sets from the DIODE [6] and CityScapes [2] datasets to compare the zero shot performance with prompting. Table 2 shows a 0.18% improvement in accuracy of depth predictions and an approximately 18% decrease in errors. More examples of improvement in depth maps are shown in the **Appendix**.

Table 2: Out-of-distribution test error and accuracy

Dataset	Method	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$	rel	rms	log10	ssim
DIODE	Zero-Shot	0.977	0.951	0.93	7.8633E+13	18.96858	0.99571	0.04903
DIODE	Prompted	**0.977**	**0.954**	**0.932**	**6.23084E+13**	**11.82858**	**0.86933**	**0.15148**
CityScapes	Zero-Shot	0.9962	0.99619	0.99617	1.53745E+16	18.19244	4.53995	0.013
CityScapes	Prompted	**0.99812**	**0.99806**	**0.99799**	**1.20803E+16**	**16.2054**	**4.40625**	**0.013**

5. Conclusion

In this paper, we present two findings:

- visual prompting fares well compared to zero shot performance for depth estimation
- visual prompting is robust to distribution shifts for depth prediction

The results from our experiments open up the prospects of a new way to fine-tune pre-trained models for depth estimation, and further validates the method proposed in [1] for adapting new downstream tasks to already existing vision models without accessing model parameters or activations. Future work along this line would look into making the visual prompts more robust to size and location changes, as well as producing more generalizable prompts that can be used to adapt more than one task to pre-trained vision models.

Acknowledgments

I would like to thank everyone at Phillip Isola's group for their support, feedback, and insightful discussions. I would also like to thank the MIT Office of Graduate Education and College of Engineering for sponsoring me to conduct this research.

References

[1] Hyojin Bahng et al. *Exploring Visual Prompts for Adapting Large-Scale Models*. 2022. DOI: 10.48550/ARXIV.2203.17274. URL: https://arxiv.org/abs/2203.17274.

[2] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *n Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). URL: http://arxiv.org/abs/1604.01685.

[3] Menglin Jia et al. *Visual Prompt Tuning*. 2022. DOI: 10.48550/ARXIV.2203.12119. URL: https://arxiv.org/abs/2203.12119.

[4] Katrin Lasinger et al. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer". In: *IEEE Transactions On Pattern Analysis and Machine Intelligence* (2019). URL: http://arxiv.org/abs/1907.01341.

[5] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision Transformers for Dense Prediction". In: *International Conference for Computer Vision (ICCV)* (2021). URL: https://arxiv.org/abs/2103.13413.

[6] Igor Vasiljevic et al. "DIODE: A Dense Indoor and Outdoor DEpth Dataset". In: *Computer Vision and Pattern Recognition (CVPR)* (2020). URL: http://arxiv.org/abs/1908.00463.

[7] Ke Xian et al. "Monocular Relative Depth Perception with Web Stereo Data Supervision". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

Appendix

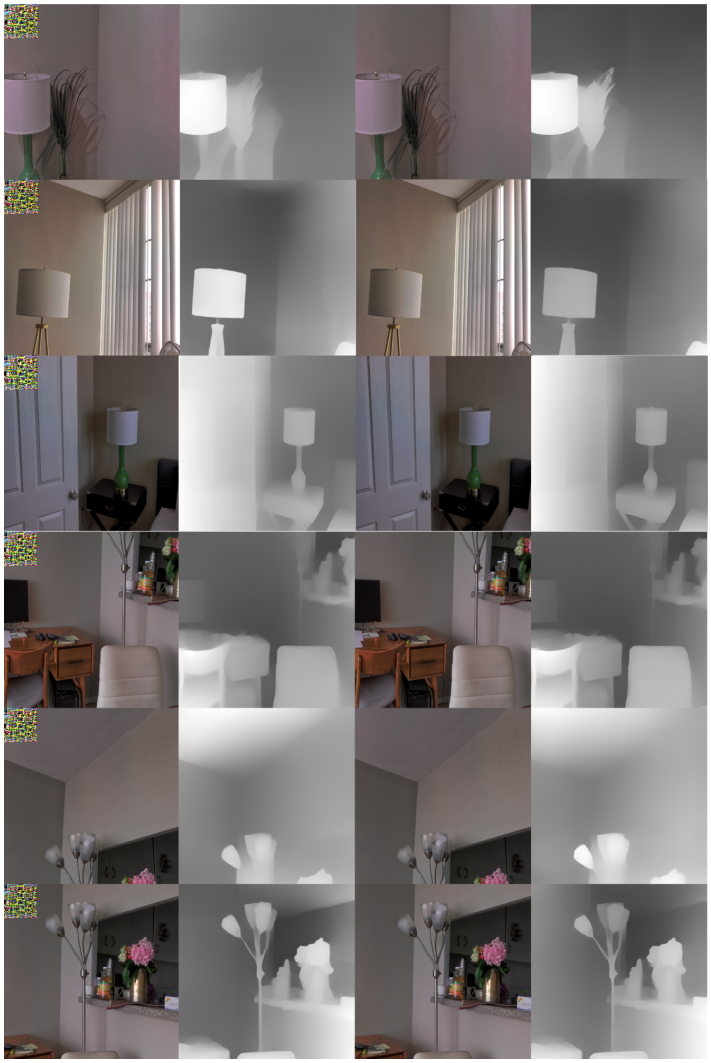

Fig. 5: (L-R) Prompted Image; Depth Map from Prompted Image; Original Image; Depth Map from Original Image