Machine Learning Engineering Bootcamp
Capstone, Step 4: Survey Existing Research and Reproduce Available Solutions
Student: Kenneth Fung
Date: 01 January 2025

1. INTRODUCTION:

In this step 4 of my Capstone project, I researched existing models for Monocular Depth Estimation (MDE) which I had proposed as my Capstone project in Step 3. The two models I researched are MiDaS[1] and Depth Anything[2]. The two models have comparable performances, with Depth Anything being subtly more advanced. Both models were trained on low-resolution and high-resolution labeled and unlabeled images. Both models have evolved through multiple releases. In section 4 - Analysis, I discuss some deficiencies in the models based on the output color mappings of the depth estimations.

2. DESCRIPTIONS OF EXISTING RESEARCH

A. MiDaS (quoted):

(source: extension://efaidnbmnnnibpcajpcglclefindmkaj/http://vladlen.info/papers/midas.pdf)

"The success of monocular depth estimation relies on large and diverse training sets. Due to the challenges associated with acquiring dense ground-truth depth across different environments at scale, a number of datasets with distinct characteristics and biases have emerged. We develop tools that enable mixing multiple datasets during training, even if their annotations are incompatible.In particular, we propose a robust training objective that is invariant to changes in depth range and scale, advocate the use of principled multi-objective learning to combine data from different sources, and highlight the importance of pretraining encoders on auxiliary tasks. Armed with these tools, we experiment with five diverse training datasets, including a new, massive data source: 3D films. To demonstrate the generalization power of our approach we use zero-shot cross-dataset transfer, i.e. we evaluate on datasets that were not seen during training. The experiments confirm that mixing data from complementary sources greatly improves monocular depth estimation. Our approach clearly outperforms competing methods across diverse datasets, setting a new state of the art for monocular depth estimation.
Our extensive experiments, which cover approximately six GPU months of computation, show that a model trained on a rich and diverse set of images from different sources, with an appropriate training procedure, delivers state-of-the-art results across a variety of environments. To demonstrate this, we use the experimental protocol of zero-shot cross-dataset transfer. That is, we train a model on certain datasets and then test its performance on other datasets that were never seen during training. The intuition is that zero-shot cross-dataset performance is a more faithful proxy of "real world" performance than training and testing on subsets of a single data collection that largely exhibit the same biases [13]. In an evaluation across six different datasets, we outperform prior art both quantitatively and qualitatively, and set a new state of the art for monocular depth estimation. Example results are shown in Figure 1."

---

[1] extension://efaidnbmnnnibpcajpcglclefindmkaj/http://vladlen.info/papers/midas.pdf
[2] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2401.10891

B. Depth Anything (quoted):
(source: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2401.10891)

"The field of computer vision and natural language processing is currently experiencing a revolution with the emergence of "foundation models" [6] that demonstrate strong zero-/fewshot performance in various downstream scenarios [45, 59]. These successes primarily rely on large-scale training data that can effectively cover the data distribution. Monocular Depth Estimation (MDE), which is a fundamental problem with broad applications in robotics [66], autonomous driving [64, 80], virtual reality [48], etc., also requires a foundation model to estimate depth information from a single image. However, this has been underexplored due to the difficulty of building datasets with tens of millions of depth labels. MiDaS [46] made a pioneering study along this direction by training an MDE model on a collection of mixed labeled datasets. Despite demonstrating a certain level of zero-shot ability, MiDaS is limited by its data coverage, thus suffering disastrous performance in some scenarios. In this work, our goal is to build a foundation model for MDE capable of producing high-quality depth information for any images under any circumstances. We approach this target from the perspective of dataset scaling-up. Traditionally, depth datasets are created mainly by acquiring depth data from sensors [18, 55], stereo matching [15], or SfM [33], which is costly, time-consuming, or even intractable in particular situations. We instead, for the first time, pay attention to large-scale unlabeled data. Compared with stereo images or labeled images from depth sensors, our used monocular unlabeled images exhibit three advantages: (i) (simple and cheap to acquire) Monocular images exist almost everywhere, thus they are easy to collect, without requiring specialized devices. (ii) (diverse) Monocular images can cover a broader range of scenes, which are critical to the model generalization ability and scalability. (iii) (easy to annotate) We can simply use a pre-trained MDE model to assign depth labels for unlabeled images, which only takes a feedforward step. More than efficient, this also produces denser depth maps than LiDAR [18] and omits the computationally intensive stereo matching process."

3. REPRODUCTION PROCESS:

For reproducing the results of the MiDaS[3] and Depth Anything[4] models, I followed the instructions from their respective GitHub repositories, which included cloning their repositories. I also had to install various platforms and packages including PyTorch, Anaconda, and Cuda. Environment setups and executions were performed within an Anaconda environment via an Anaconda prompt. The input images are labeled I1-I8 with eight images in total. Each image in an input directory had specific labels with date-timestamp of when the images were originally captured. The images are sequentially attached (Fig 3-Fig 10) in section 6 – Inputs. Fig 1 shows a snippet of the Anaconda Prompt command window via which commands for the MiDaS execution were performed. Fig 2 shows a snippet of the Anaconda Prompt command window via which commands for the Depth Anything execution were performed. Section 7 – Outputs has figures Fig 11-Fig 26 for the output color depth maps of MiDaS and Depth Anything, alternating between each model. MiDaS outputs individual color depth maps of the input images, while Depth Anything outputs a product with the original image on the left and the color depth map on the right. The MiDaS output images were similarly pasted on the right next to their original images on the left to facilitate comparisons and analysis.

[3] https://github.com/isl-org/MiDaS

[4] https://github.com/DepthAnything/Depth-Anything-V2

Fig 1. MiDaS execution:

GitHub: https://github.com/isl-org/MiDaS

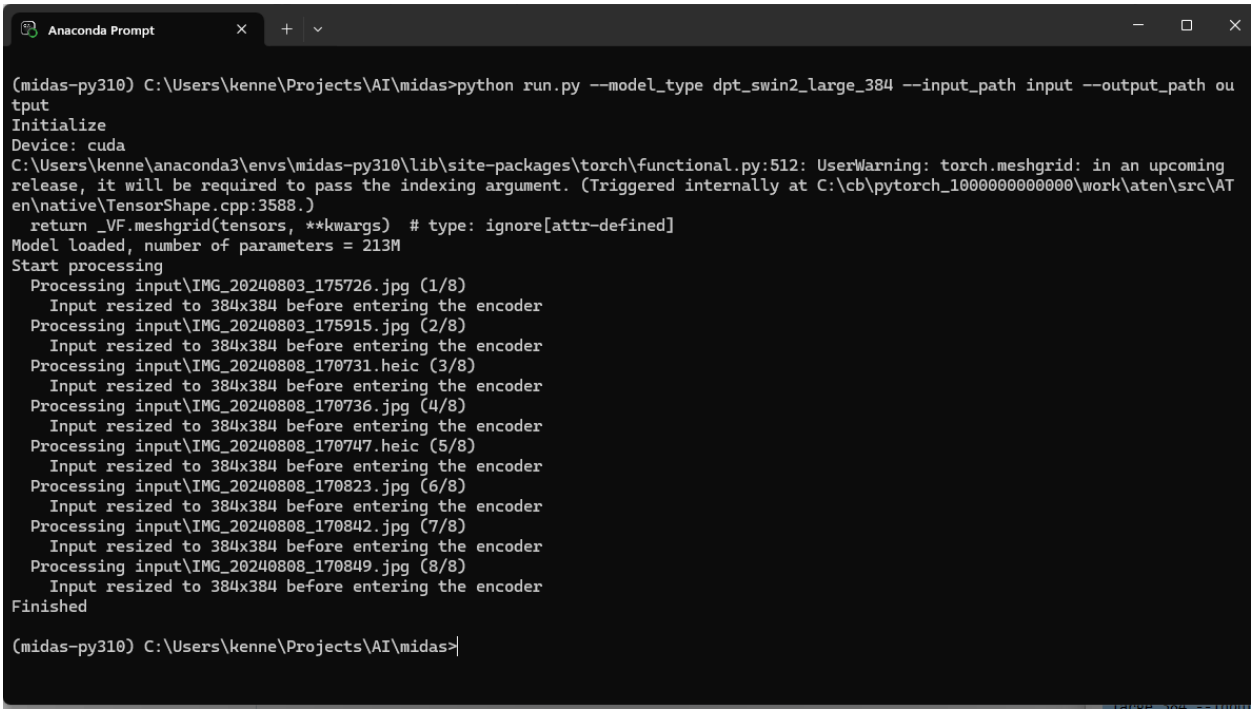
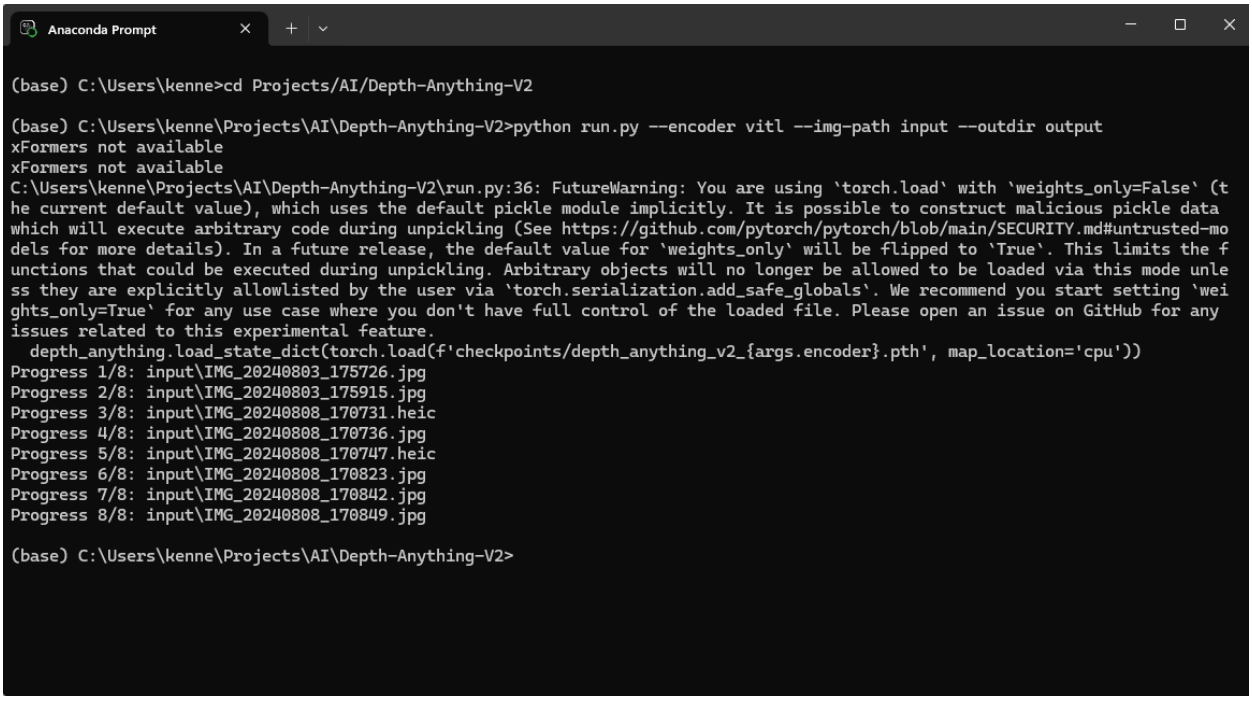


Fig 2. Depth Anything execution:

GitHub: https://github.com/DepthAnything/Depth-Anything-V2



4. ANALYSIS:

A. color wind wheel that Aurora is holding in the restaurant scene

B. the tiles in the museum scene with the glass cases behind – individual tiles had distinct patterns in the MiDaS outputs despite being at the same depth, thus suggesting color of the original image negatively impacted the depth estimation by MiDaS

C. windwheels such as the one held by grandma in the stairs to heaven scene was very faint in MiDaS

D. in general, MiDaS was insensitive to slight changes in shades of colors in the background, in particular, and the depth of objects seemingly at different depths were merged

E. Depth Anything is generally more sensitive and produce crisper depth maps

F. neither model is good with images that are very deep in the background – it is not a matter of size, because small objects in the foreground are picked up

G. neither model would distinguish the depth difference between clouds and the sky

5. PLAN FORWARD

A. Process non-photographs (cartoon images, etc.)

6. Inputs:

Fig 3. Input I1:

Fig 4. Input I2:

Fig 5. Input I3:

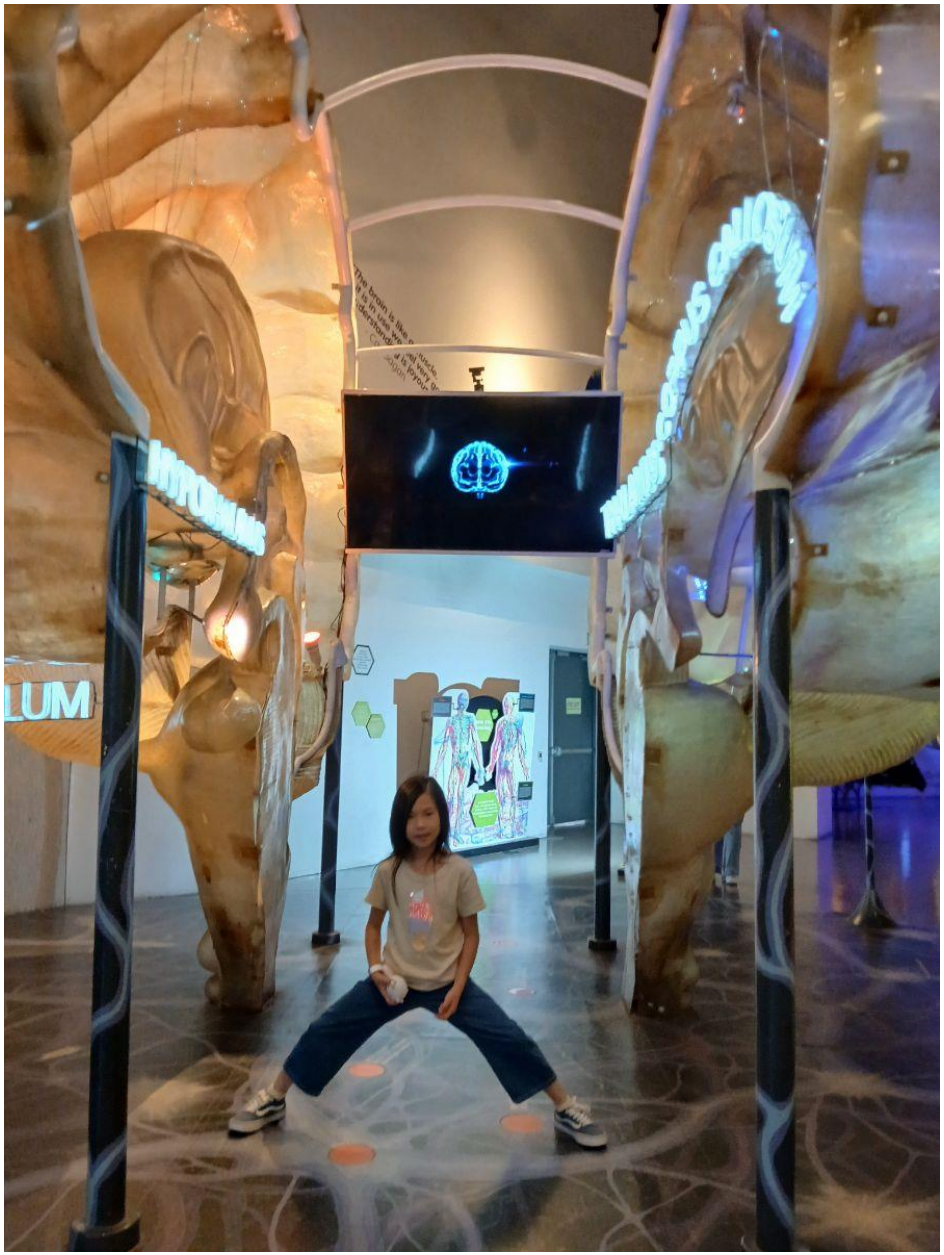


Fig 6. Input I4:

Fig 7. Input I5:

Fig 8. Input I6:

Fig 9. Input I7:

Fig 10. Input I8:

7. Outputs

Fig 11. MiDaS output for I1:

Fig 12. Depth Anything output for I1:

Fig 13. MiDaS output for I2:

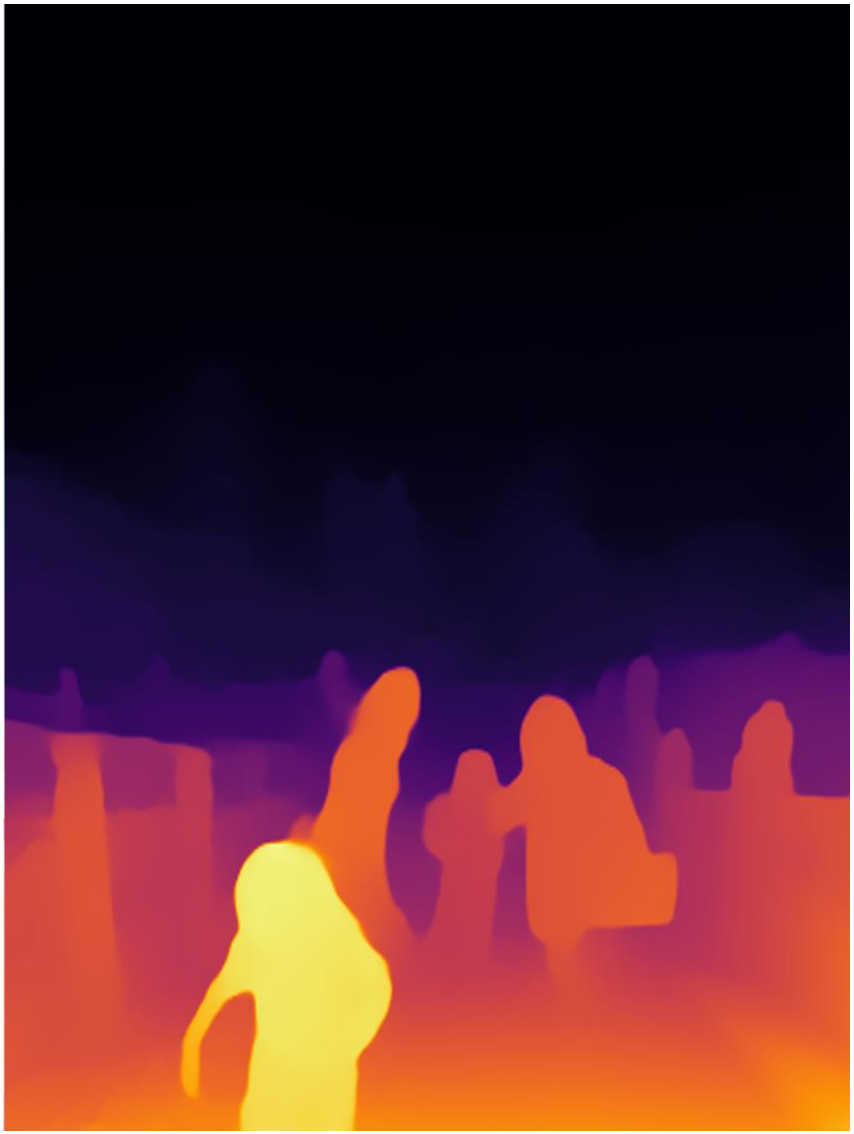
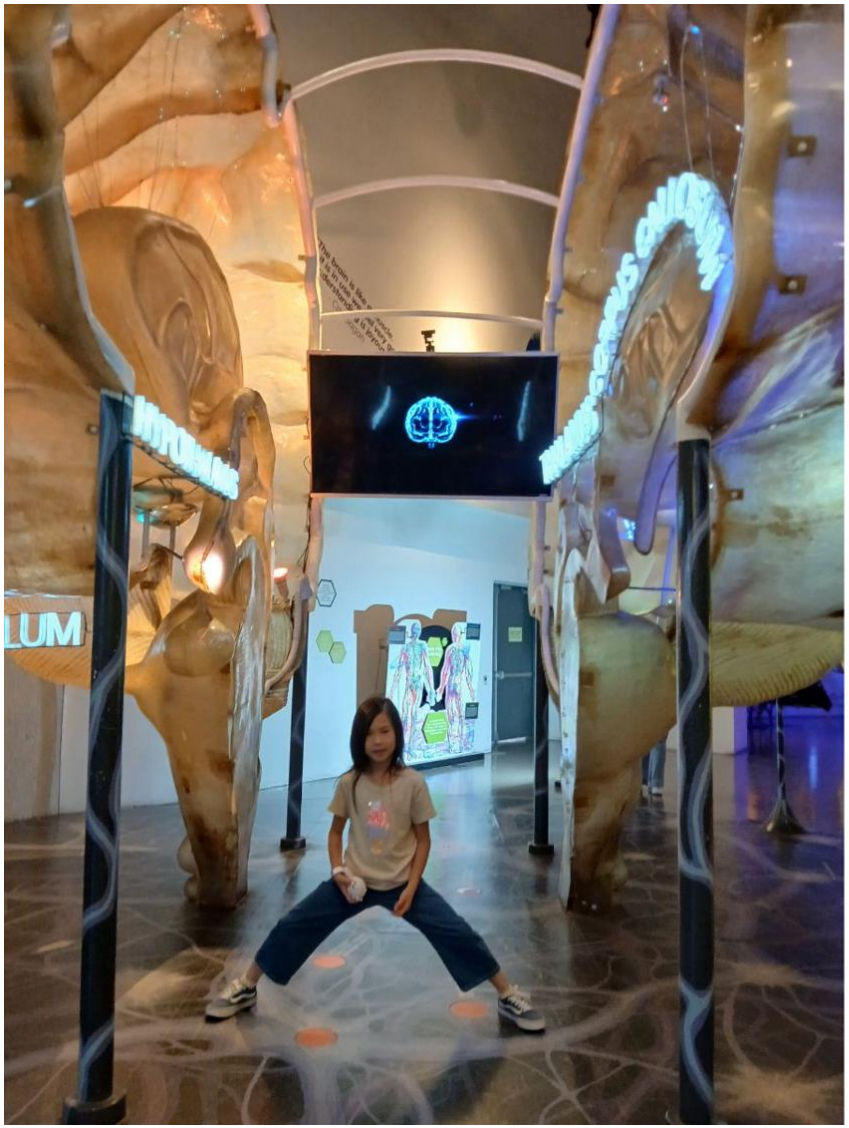


Fig 14. Depth Anything output for I2:

Fig 15. MiDaS output for I3:



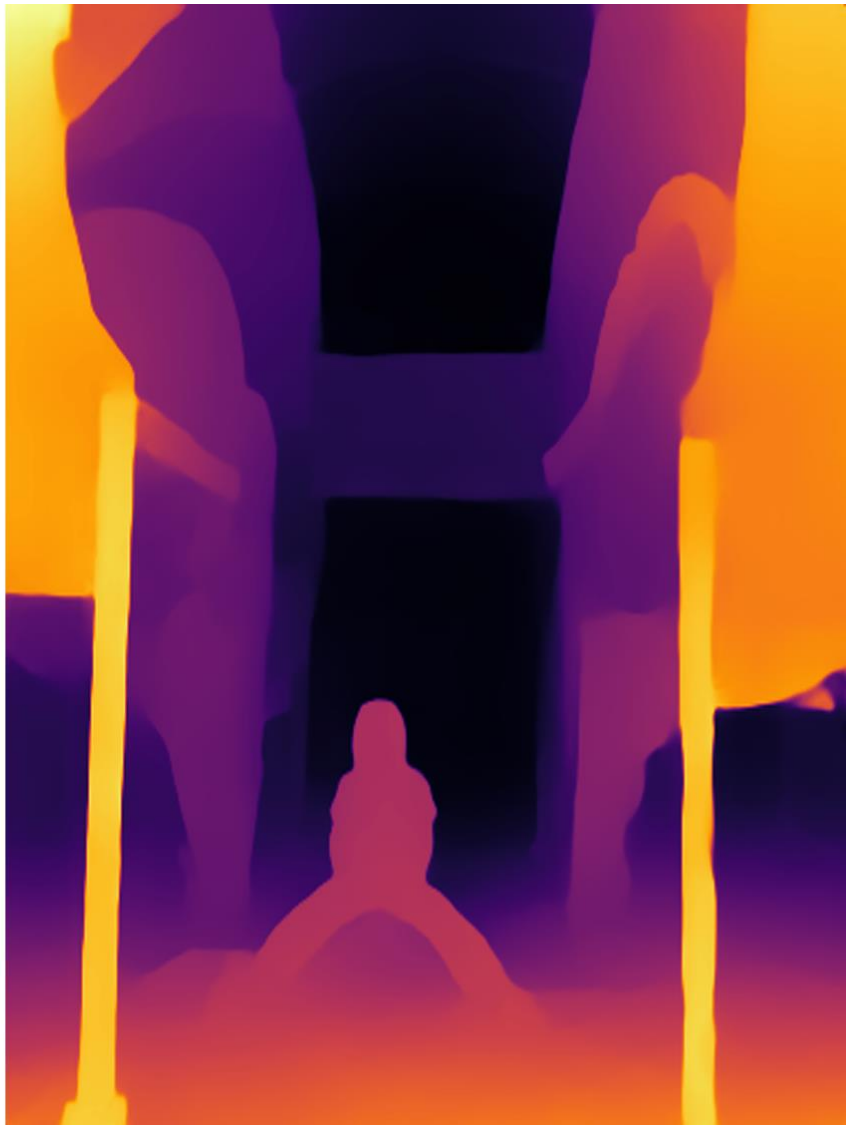
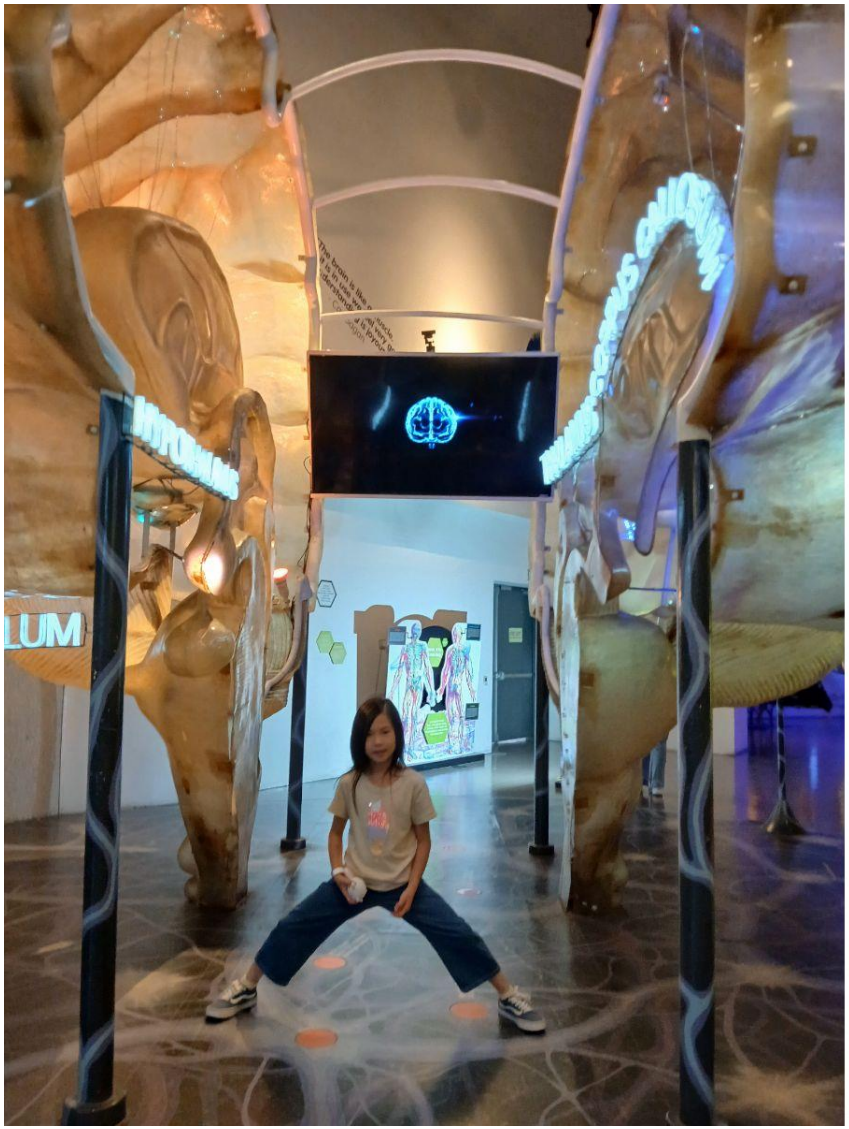


Fig 16. Depth Anything output for I3:



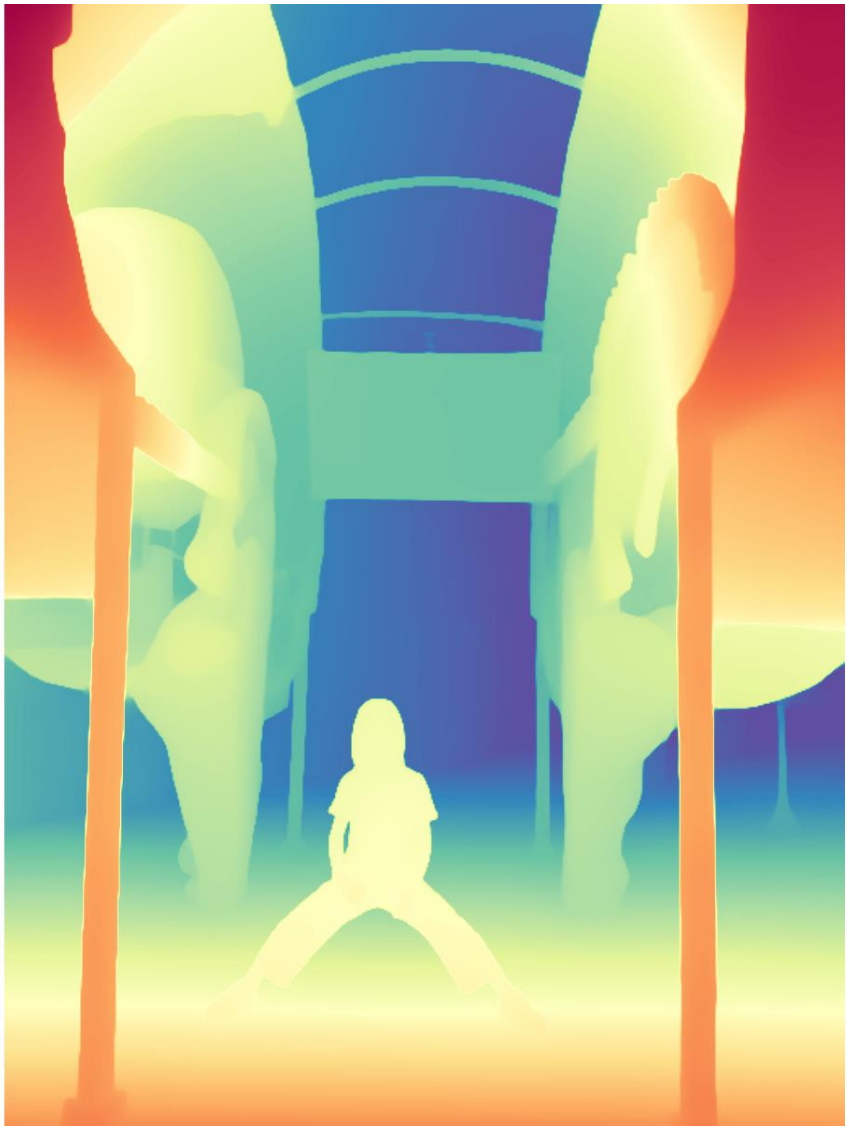


Fig 17. MiDaS output for I4:

Fig 18. Depth Anything output for I4:

Fig 19. MiDaS output for I5:

Fig 20. Depth Anything output for I5:

Fig 21. MiDaS output for I6:

Fig 22. Depth Anything output for I6:

Fig 23. MiDaS output for I7:

Fig 24. Depth Anything output for I7:

Fig 25. MiDaS output for I8:

Fig 26. Depth Anything output for I8: