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 v3.1[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 v3.1, dpt_swin2_large_384:

Description of the MDE model release version MiDaS 3.1 is provided by the reference of footnote 1. The DPT-SWIN2L version (dpt_swin2_large_384) of the model, which trades quality for speed-performance, was chosen to process images for model familiarity and performance comparison with Depth Anything. This DPT-SWIN2 version of the model was trained by the authors of the accompanying research paper[3] by mixing datasets in a method termed "zero-shot cross-dataset transfer" to introduce diversity in backgrounds, lighting, environments, indoor vs outdoor, static vs dynamic, and density of depth annotations to the training images while avoiding overtraining due to distinct characteristics and biases inherent in a single dataset collected in the same manner.
In describing the MDE model dpt_swin2_large_384, "DPT" stands for "Dense Prediction Transformer", which is the architecture enabling pixel-wise depth predictions in monocular depth estimation tasks. This architecture is integrated with the "Swin Transformer V2" encoder as its backbone where the Swin transformer yields relatively high depth estimation quality. The model processes images at a resolution of 384 x 384 pixels.
The following list is a detailed description of dpt_swin2_large_384:

- DPT: Dense Prediction Transformer architecture used as the base framework
- Swin2: Refers to the Swin Transformer V2, which is a more advanced and efficient transformer model specifically designed for vision tasks. It introduces hierarchical feature processing and is used in this model as the backbone for feature extraction.
- Large: Indicates the model size or capacity (number of parameters), which is optimized for higher performance at the cost of increased computational requirements.
- 384: Refers to the input resolution of the model, typically 384 x 384 pixels for the input images.

[1] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2307.14460
[2] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2401.10891
[3] extension://efaidnbmnnnibpcajpcglclefindmkaj/http://vladlen.info/papers/midas.pdf

B. Depth Anything:

Depth Anything is an advanced Monocular Depth Estimation (MDE) model designed to predict depth information from a single image. The research described in the paper of footnote #2 highlights the latest contributions of the authors in up-scaling large unlabeled datasets to train the MDE model with datasets having more diversity in environments, scenery, and failure modes (e.g., reflections) of previous MDE models. So while the study of MiDaS outlined in footnote #2 aimed at having a diverse collection of labeled datasets, Depth Anything was most recently trained on a diverse collection of labeled and unlabeled datasets. Unlabeled datasets were fed through a baseline MDE model to produce the depth maps. Depth Anything outperforms MiDaS as described in the Depth Anything research paper, and also as evidenced by the output images of section 7 Outputs.

Here is a general description of the Depth Anything MDE model, as sourced through ChatGPT[4]:

Key Features:

- Extensive Training Data: Trained on a vast dataset comprising approximately 1.5 million labeled images and over 62 million unlabeled images, enhancing its generalization capabilities across different scenes and objects.
- Zero-Shot Relative Depth Estimation: Capable of estimating relative depth without prior fine-tuning on specific datasets, outperforming models like MiDaS v3.1 (BEiT L-512).
- Zero-Shot Metric Depth Estimation: Provides accurate metric depth estimations, surpassing models such as ZoeDepth in performance.
- Fine-Tuning Capabilities: When fine-tuned with metric depth information from datasets like NYUv2 and KITTI, it achieves state-of-the-art results in both in-domain and zero-shot metric depth estimation.
- Enhanced ControlNet Integration: Improves depth-conditioned ControlNet models, offering more precise synthesis compared to previous versions based on MiDaS.

Model Variants:

- Depth Anything is available in multiple scales to cater to different computational needs and application scenarios:
- Depth-Anything-V2-Small: A lightweight model suitable for applications requiring efficiency. HUGGING FACE
- Depth-Anything-V2-Large: A larger model offering enhanced accuracy for tasks where performance is critical.

[4] ChatGPT (2025). Explanation of Depth Anything MDE model and zero-shot depth estimation capability. OpenAI. Accessed January 5, 2025.

3. REPRODUCTION PROCESS:

For reproducing the results of the MiDaS[5] and Depth Anything[6] 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.

A series of eight custom input images were created to highlight some deficiencies of Depth Anything. Since Depth Anything outperforms MiDaS, I did not include the MiDaS outputs for these custom images. These images contain flat shapes that may or may not overlap, and a cartoon image. The flat shapes that do not overlap with each other or an environment highlight the difficulty (if not impossible) to determine the relative depths of objects if the camera angle is perpendicular to the objects in the image. Also, the colors of objects in the images impact the depth maps of both Depth Anything. The cartoon images[7] were to determine if there is a significant performance degradation between natural and unnatural lighting. Figure 48-49 illustrate potential aspect of improvement for Depth Anything. They contain either reflections in a mirror or a painting, both of which should have constant depth.

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

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

[7] Fotor. *GoArt AI Photo Effects & Filters*. Fotor, https://goart.fotor.com/. Accessed 6 Jan. 2025.

Fig 1. MiDaS execution:

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

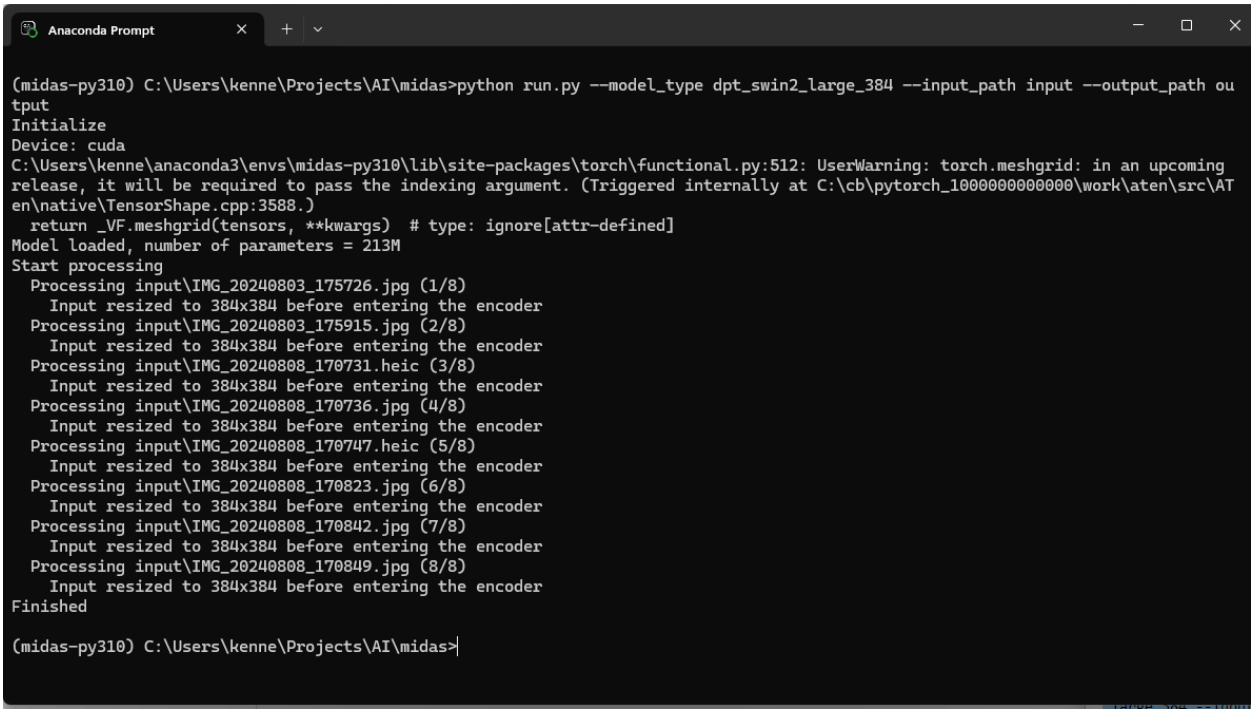
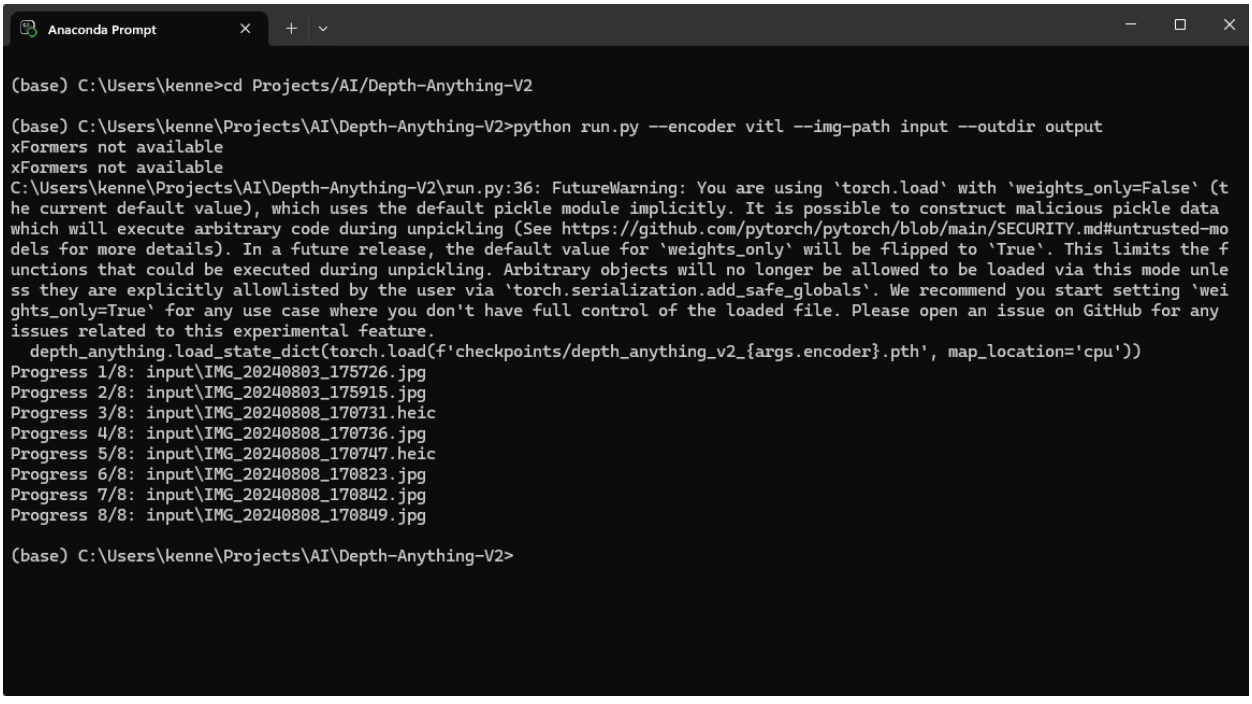


Fig 2. Depth Anything execution:

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



4. ANALYSIS:

The following analysis is focused on the output images of section 7 Outputs.

A. Input I1: In this scene, the girl on the right is holding a wind wheel with a black stem which is clearly at the same depth of the girl. However, MiDaS is not sensitive enough to detect the black stem which overlaps the girl and is an indication of its depth (with the wind wheel) relative to the girl. A subsequent modfied image of the same photograph with a thickened stem enabled MiDaS to correctly map the depth of the wind wheel. Depth Anything did not have such an issue.

B. Input I7: 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
In this scene, MiDaS showed the boundaries indicating slightly different depths of the individual tiles on the floor. Since this is a depth map, the tiles should have resulted in layers of depth colors with only horizontal degradations. Depth Anything did not have this issue and demonstrates an expected result.

C. Input I2: This scene shows the susceptibility of MiDaS to objects with high transparencies, such as the wind wheel. The wind wheel held by the grandma is very faint when it should have a similar depth color as the grandma.

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. Input I5: 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. Input I5: Neither model would distinguish the depth difference between clouds and the sky.

H. As highlighted in the last section of the MiDaS paper of footnote 3, failures also include images in a mirror reflection or painting are erroneously distinguished as separate objects with different depths as the substrate. Moreover, Images have a natural bias where the lower parts of the image are closer to the camera than the higher image regions, strong edges can lead to hallucinated depth discontinuities, thin structures can be missed and relative depth arrangement between disconnected objects might fail in some situations, and results tend to get blurred in background regions (which might be explained by the limited resolution of the input images and imperfect ground truth in the far range).

I. Figure 48: There is a faint outline of the lady in the mirror. Since the mirror is flat, it should have resulted in a constant depth gradient instead of an outline of the lady. The image of the chess piece in the mirror fared better.

J. Figure 49: This scene contains a painting of farm animals, as evidenced by the male and female figures behind the painting (hence it cannot be a window). The painting should have resulted in a constant depth assignment, and outlines of the animals have not have resulted. Also, it is impossible to

determine the relative depths of the soccer balls – it may be preferable to assign them a color to mark them as impossible to determine..

5. PLAN FORWARD (in-work)

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

B. Test model with generated images of non-overlapping objects in parallel planes separated by at least several feet, and line-of-sight is perfectly perpendicular to these planes

C. Test model with existing or generated images with small distinct objects in the foreground and large objects in the background, with objects appearing as the same size, and not have any boundaries be overlapped by the environment

D. Test model with images of clouds and skies

E. In the same vein as Depth Anything, find an unlabeled monocular dataset that may exhibit specific characteristics to expose the deficiencies of the trained models of Depth Anything to improve the trained model

E. Test model with images with irregularly shaped objects that have an obstruction in front separating the object, where one half of the object is larger than the other half

6. Inputs:

Fig 3. Input I1:

Fig 4. Input I2:

Fig 5. Input I3:

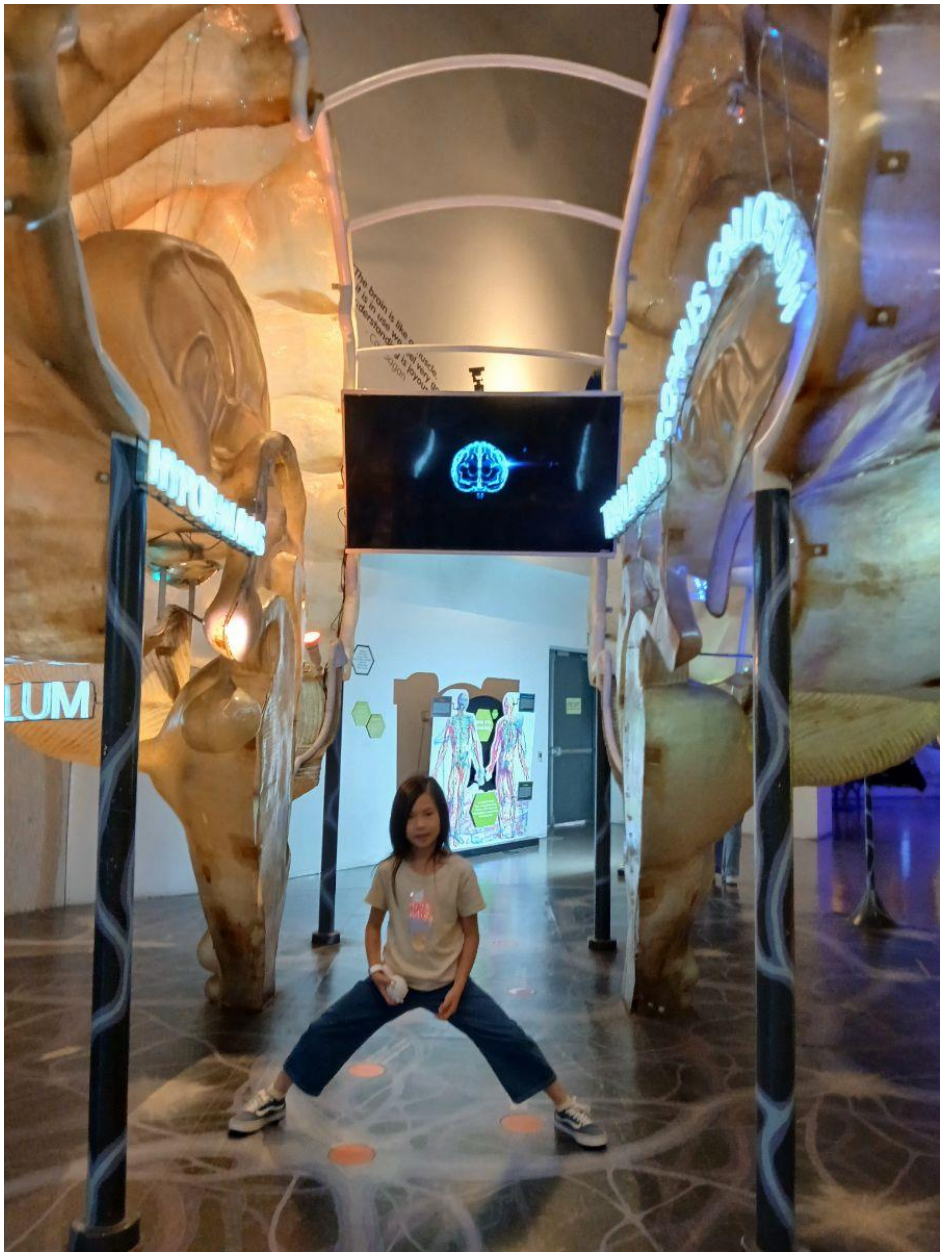


Fig 6. Input I4:

Fig 7. Input I5:

Fig 8. Input I6:

Fig 9. Input I7:

Fig 10. Input I8:

Fig 11. Input I9:

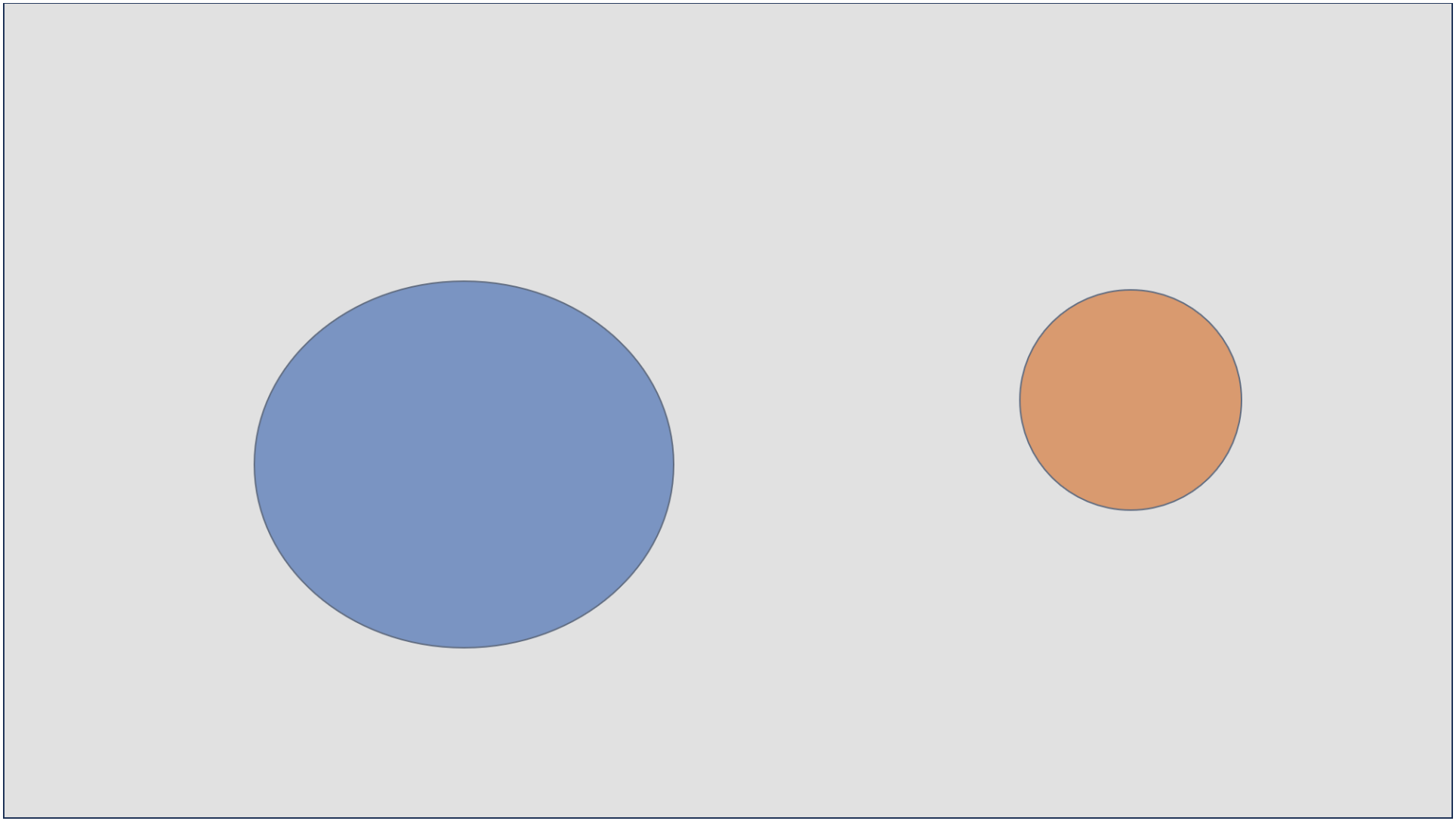


Fig 12. Input I10:

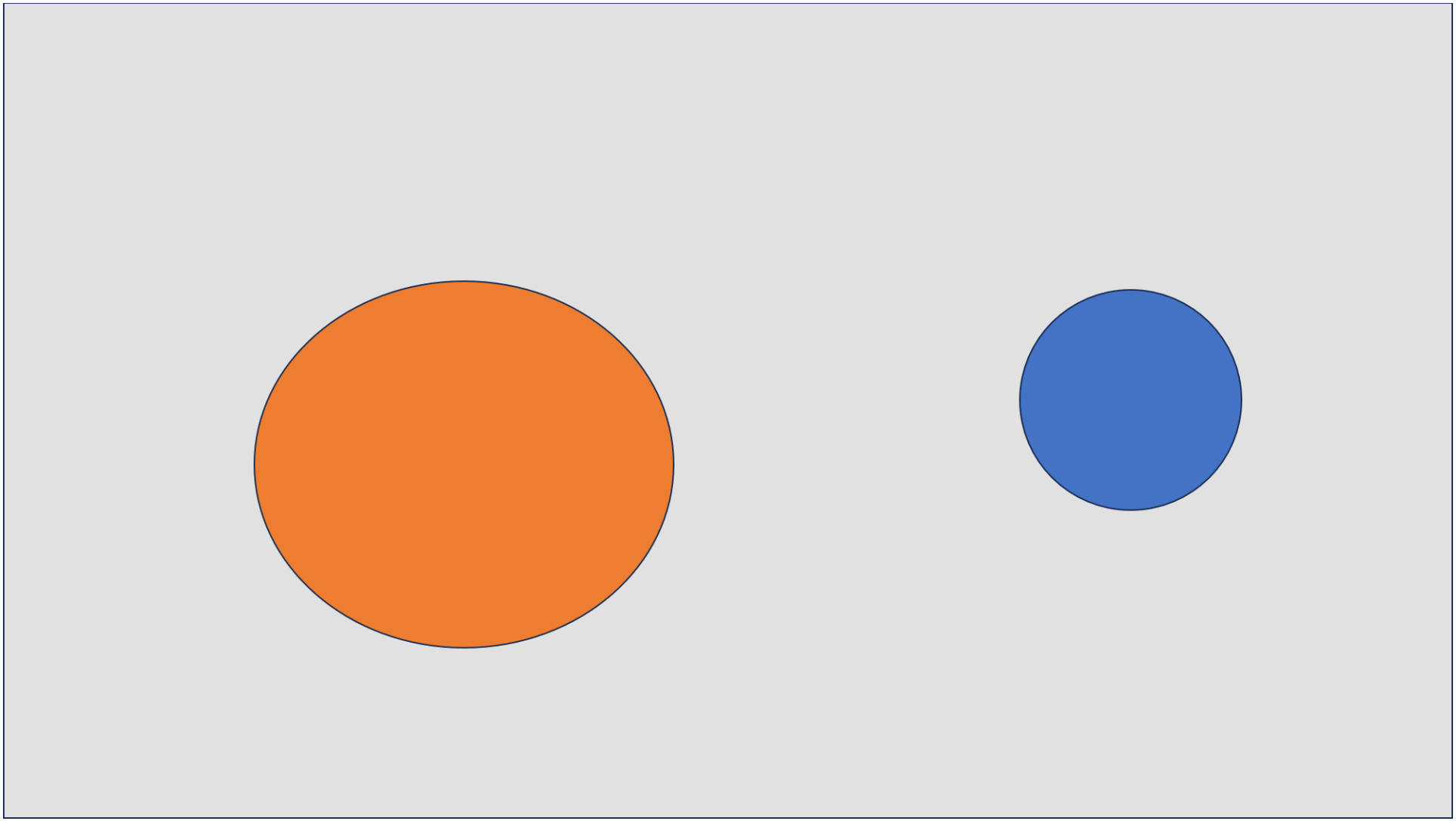


Fig 13. Input I11:

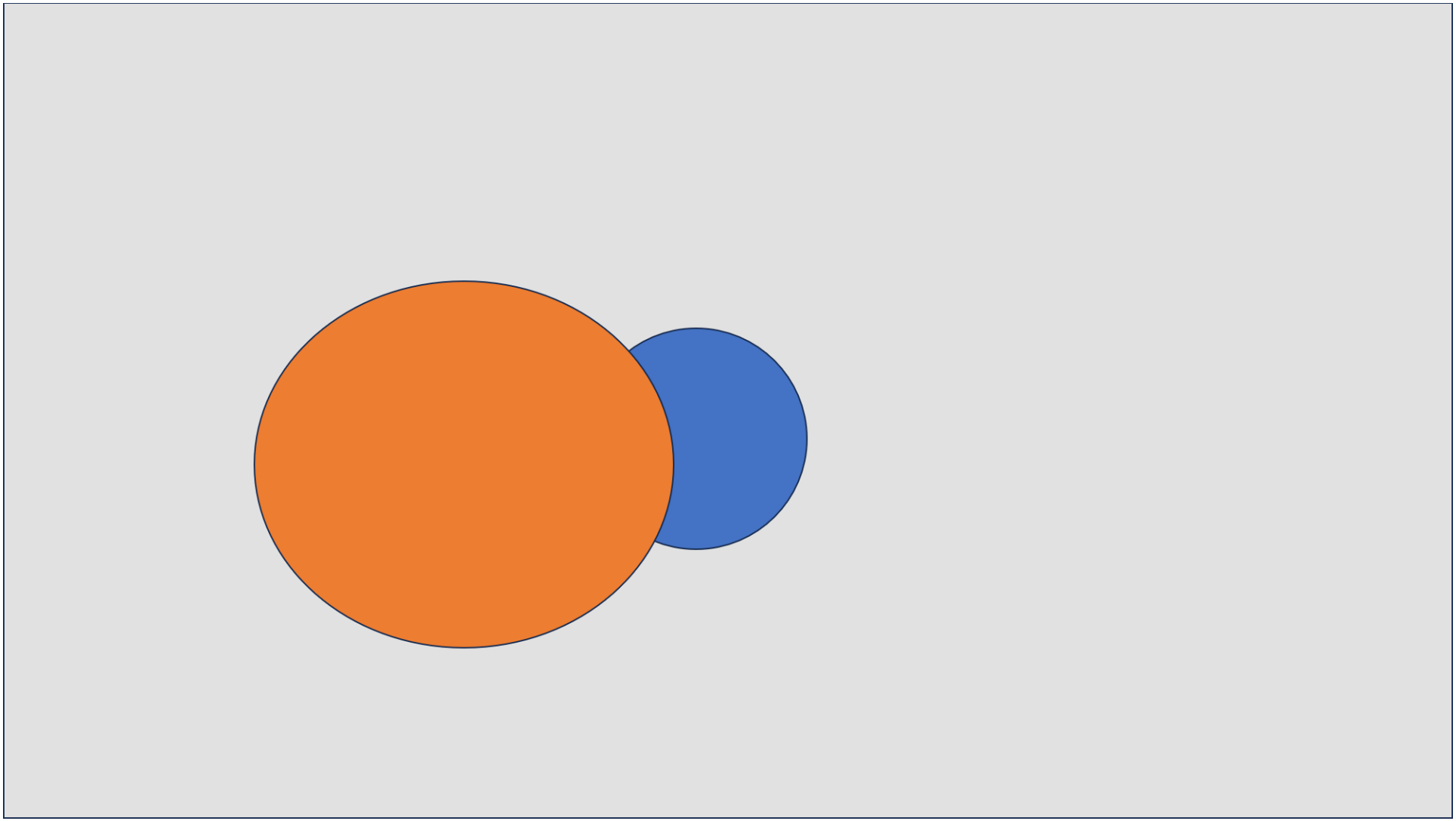


Fig 14. Input I12:

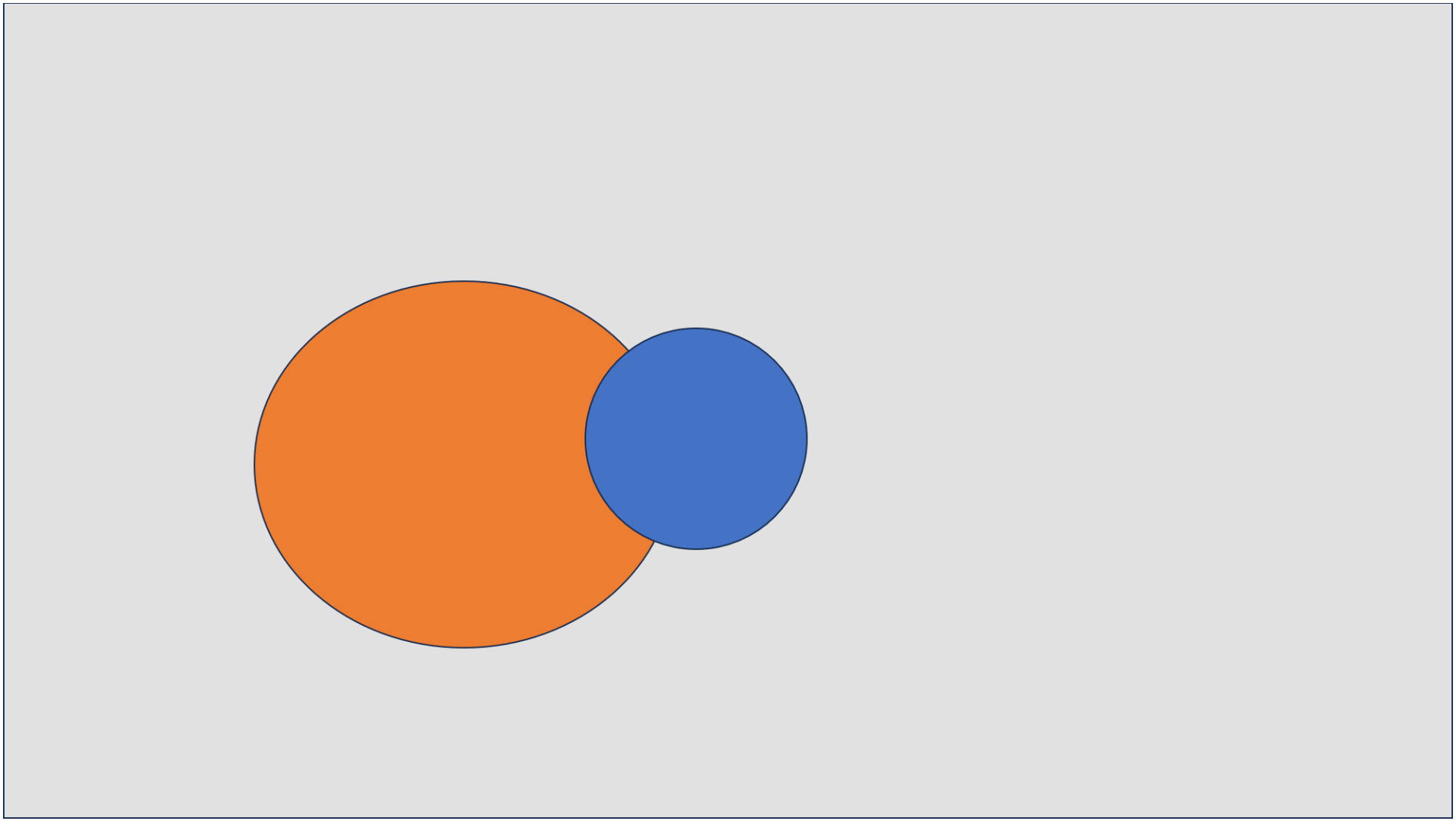


Fig 15. Input I13:

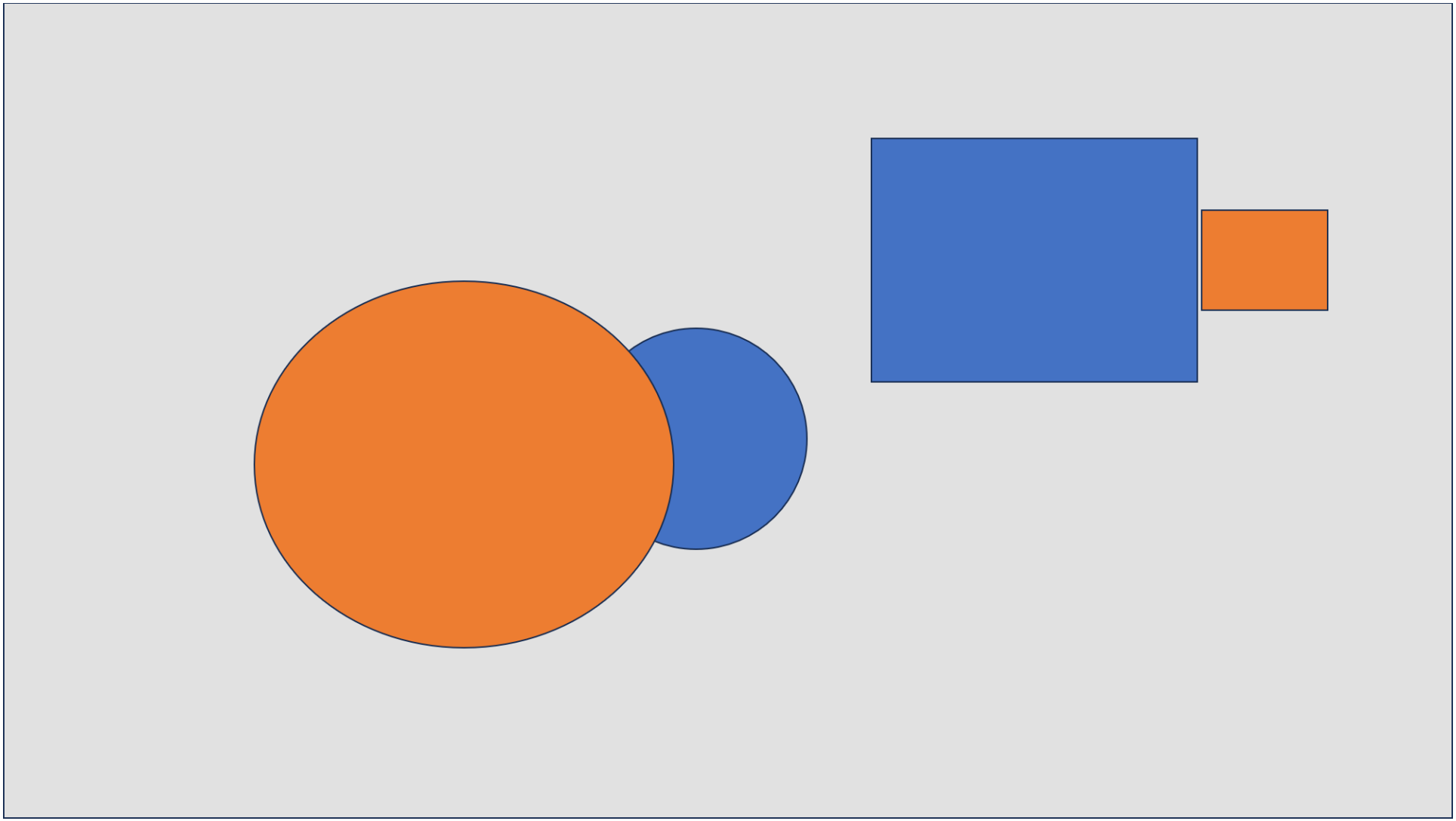


Fig 16. Input I14:

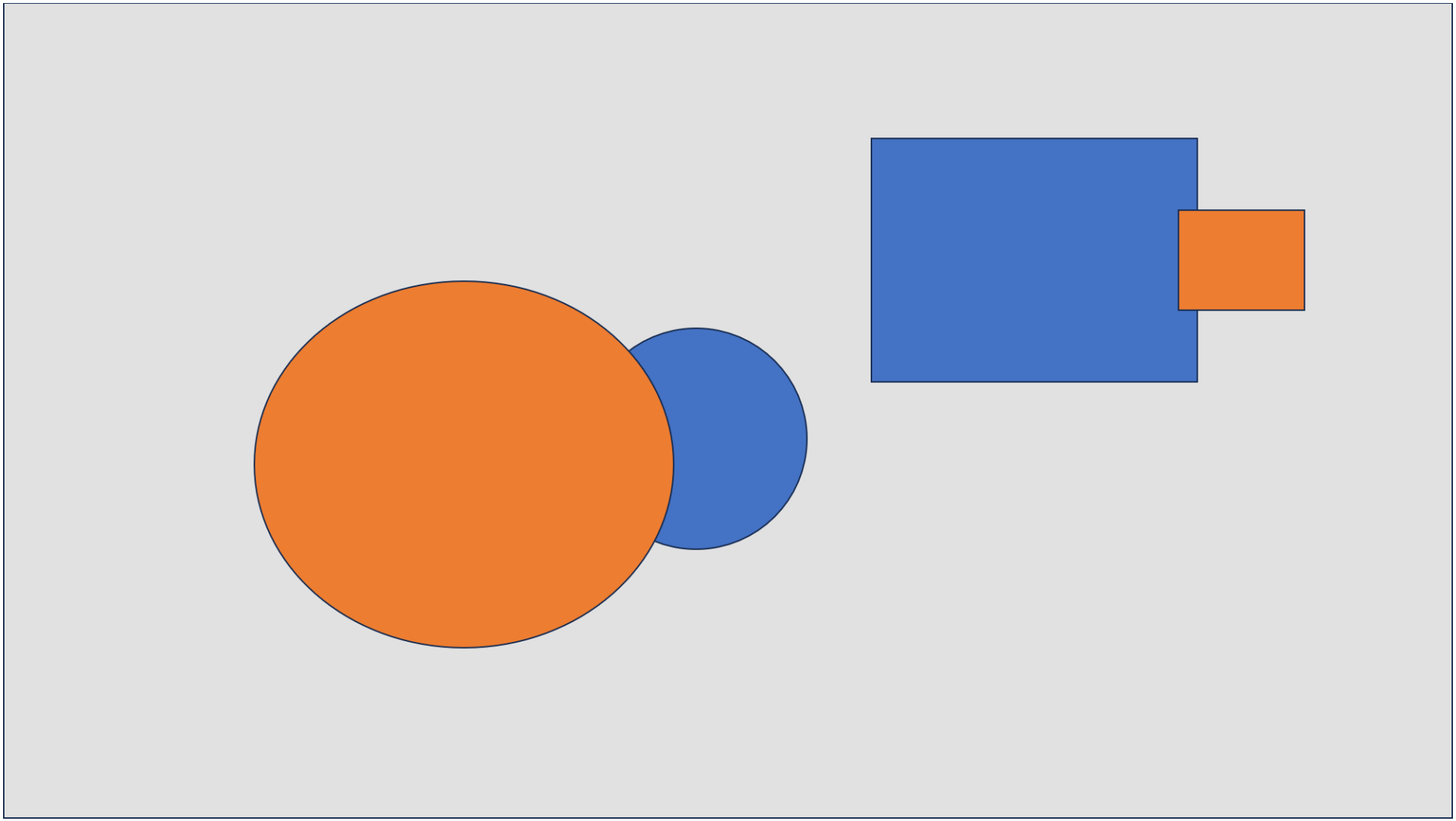


Fig 17. Input I15:

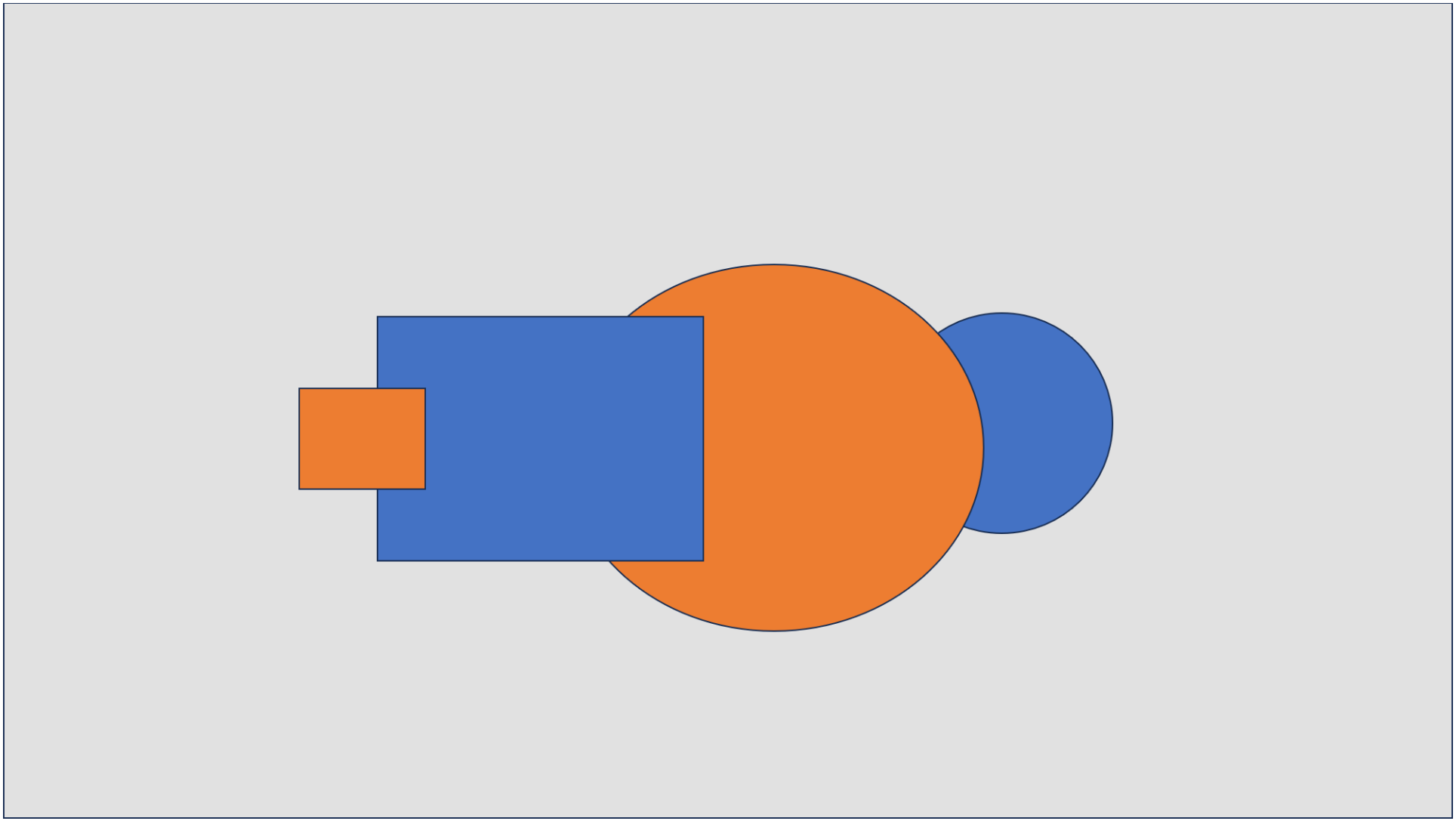


Fig 18. Input I16:

7. Outputs

Fig 19. MiDaS output for I1:

Fig 20. Depth Anything output for I1:

Fig 21. MiDaS output for I2:

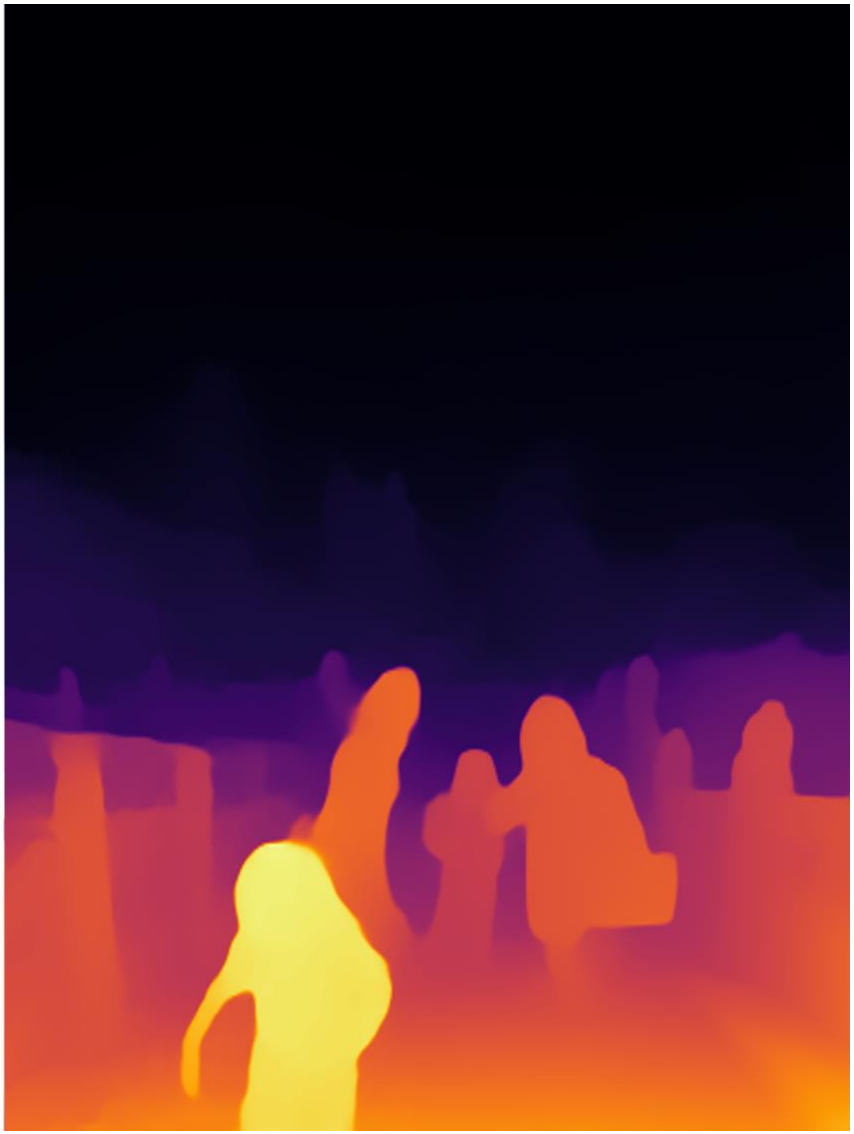
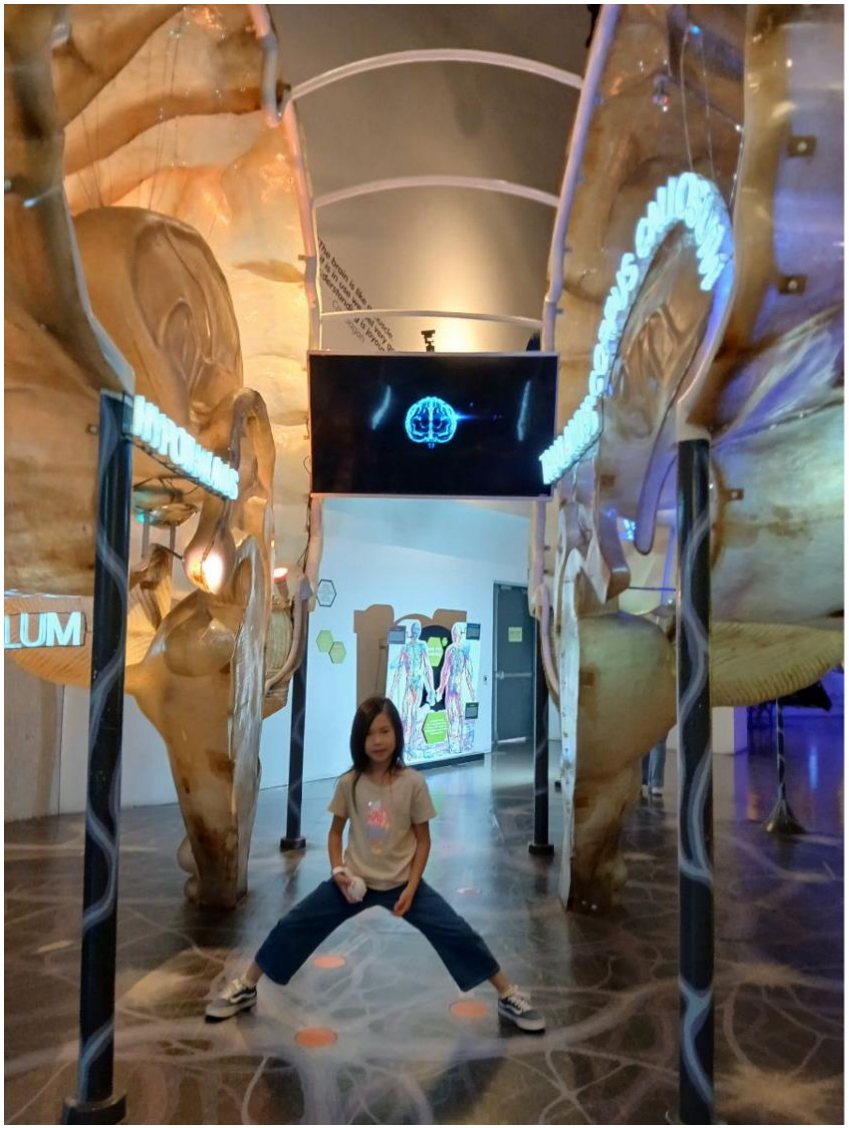


Fig 22. Depth Anything output for I2:

Fig 23. MiDaS output for I3:



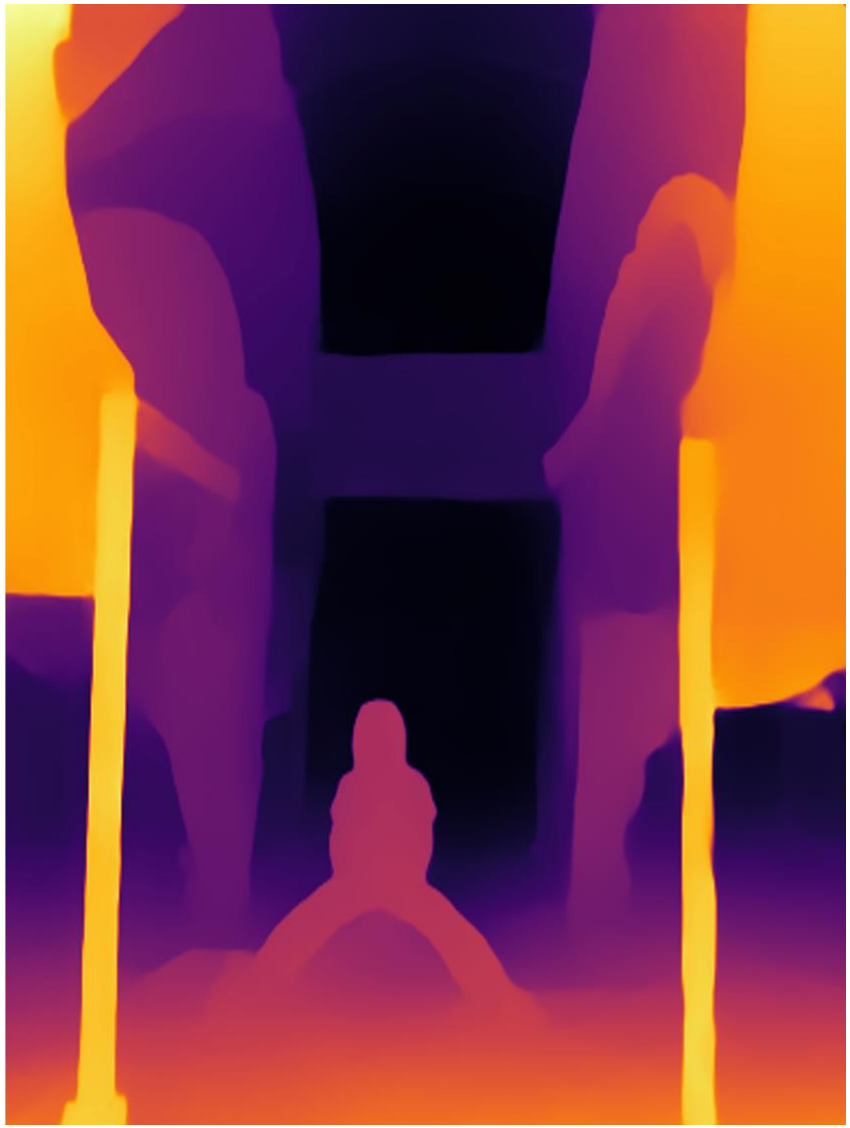
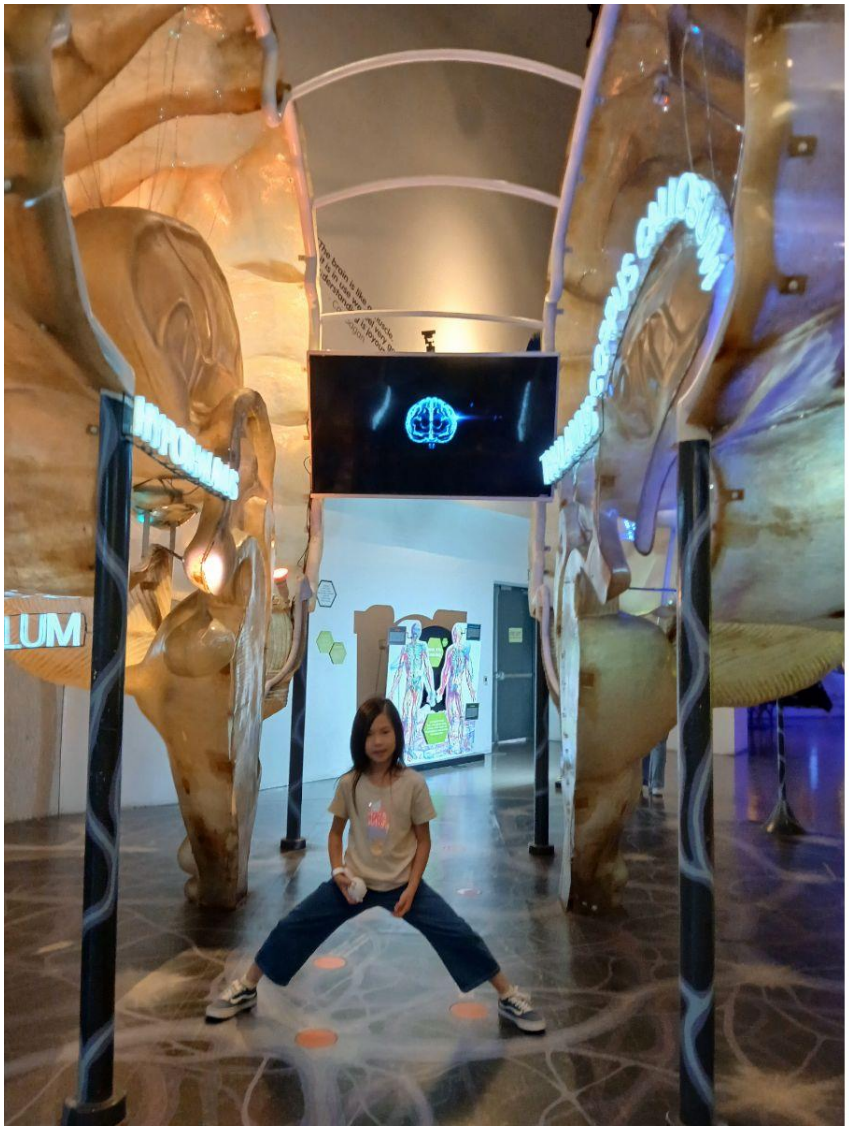


Fig 24. Depth Anything output for I3:



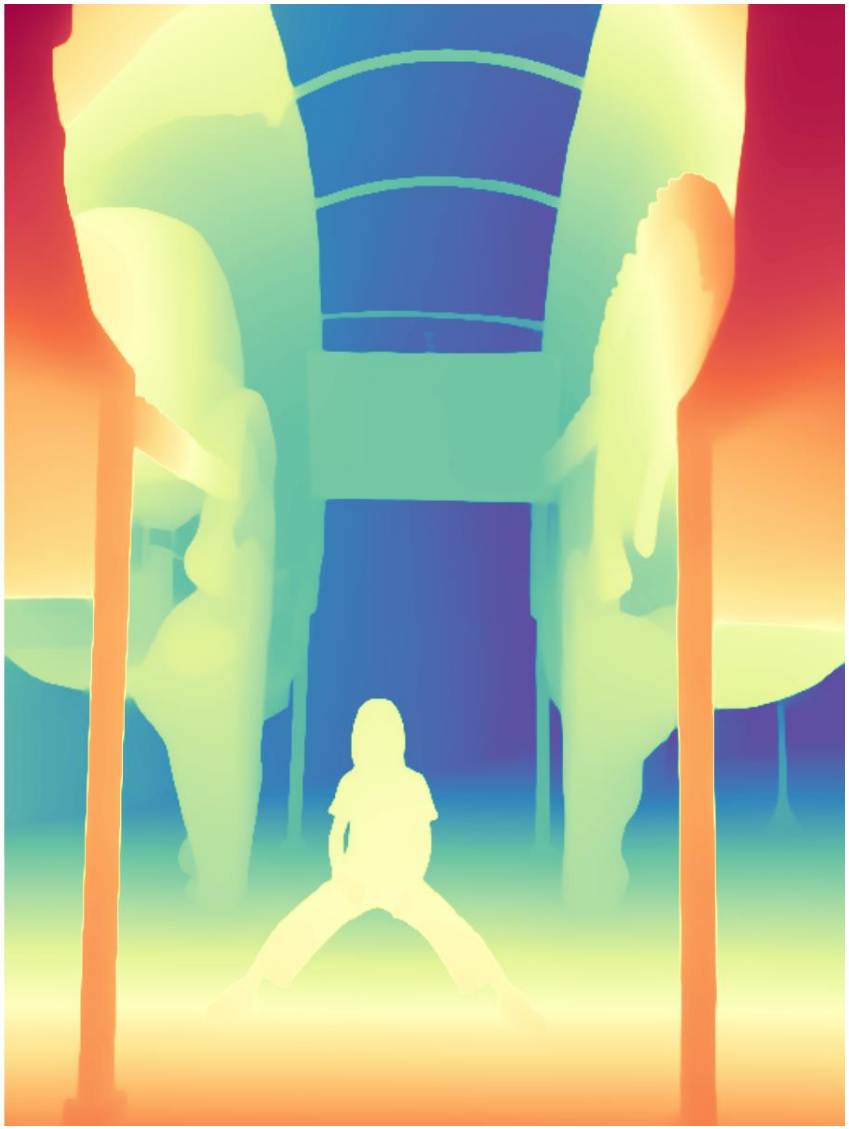


Fig 25. MiDaS output for I4:

Fig 26. Depth Anything output for I4:

Fig 27. MiDaS output for I5:

Fig 28. Depth Anything output for I5:

Fig 29. MiDaS output for I6:

Fig 30. Depth Anything output for I6:

Fig 31. MiDaS output for I7:

Fig 32. Depth Anything output for I7:

Fig 33. MiDaS output for I8:

Fig 34. Depth Anything output for I8:

Fig 35. Depth Anything output for I9-I10:

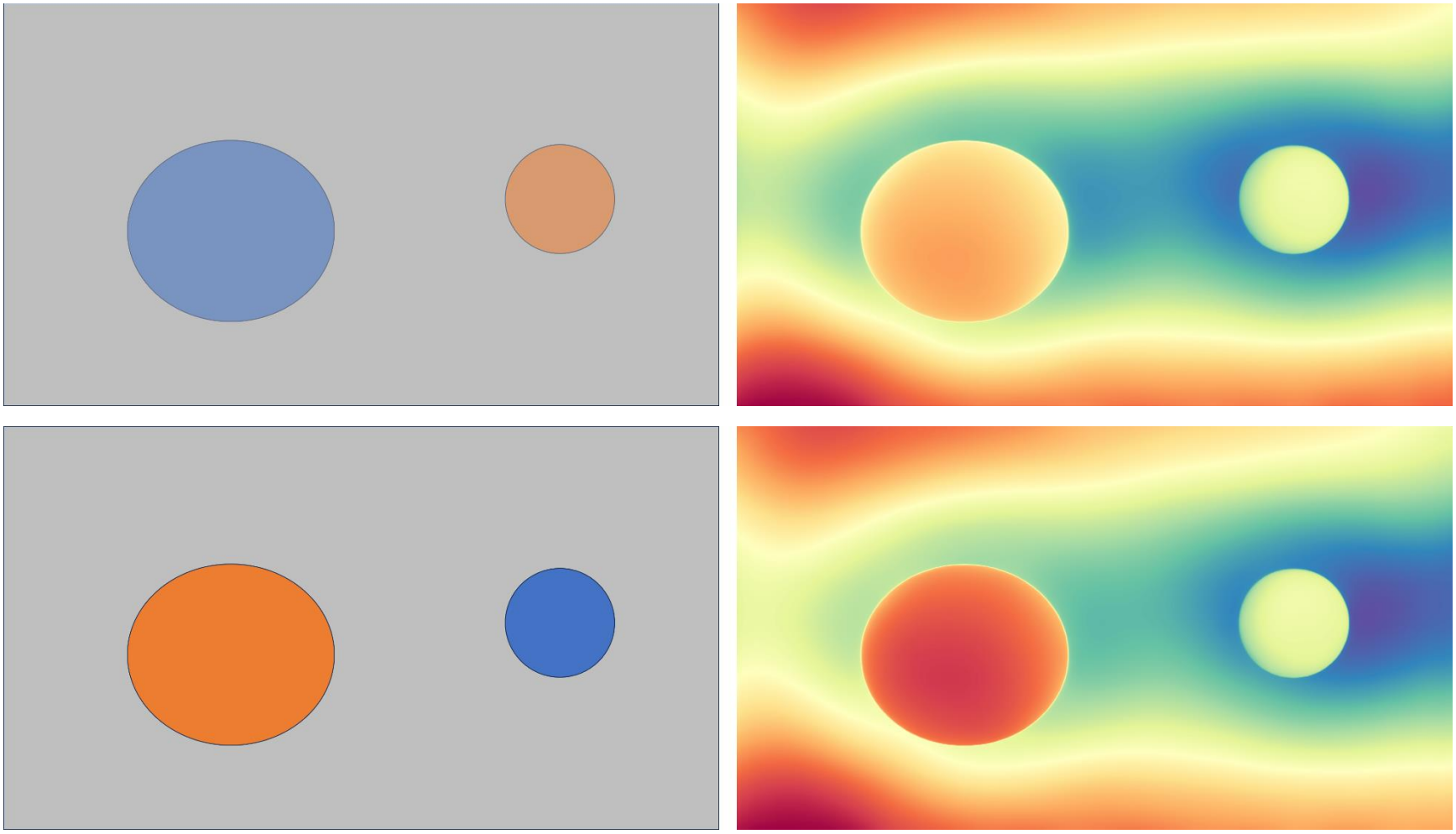


Fig 36. Depth Anything output for I11-I12:

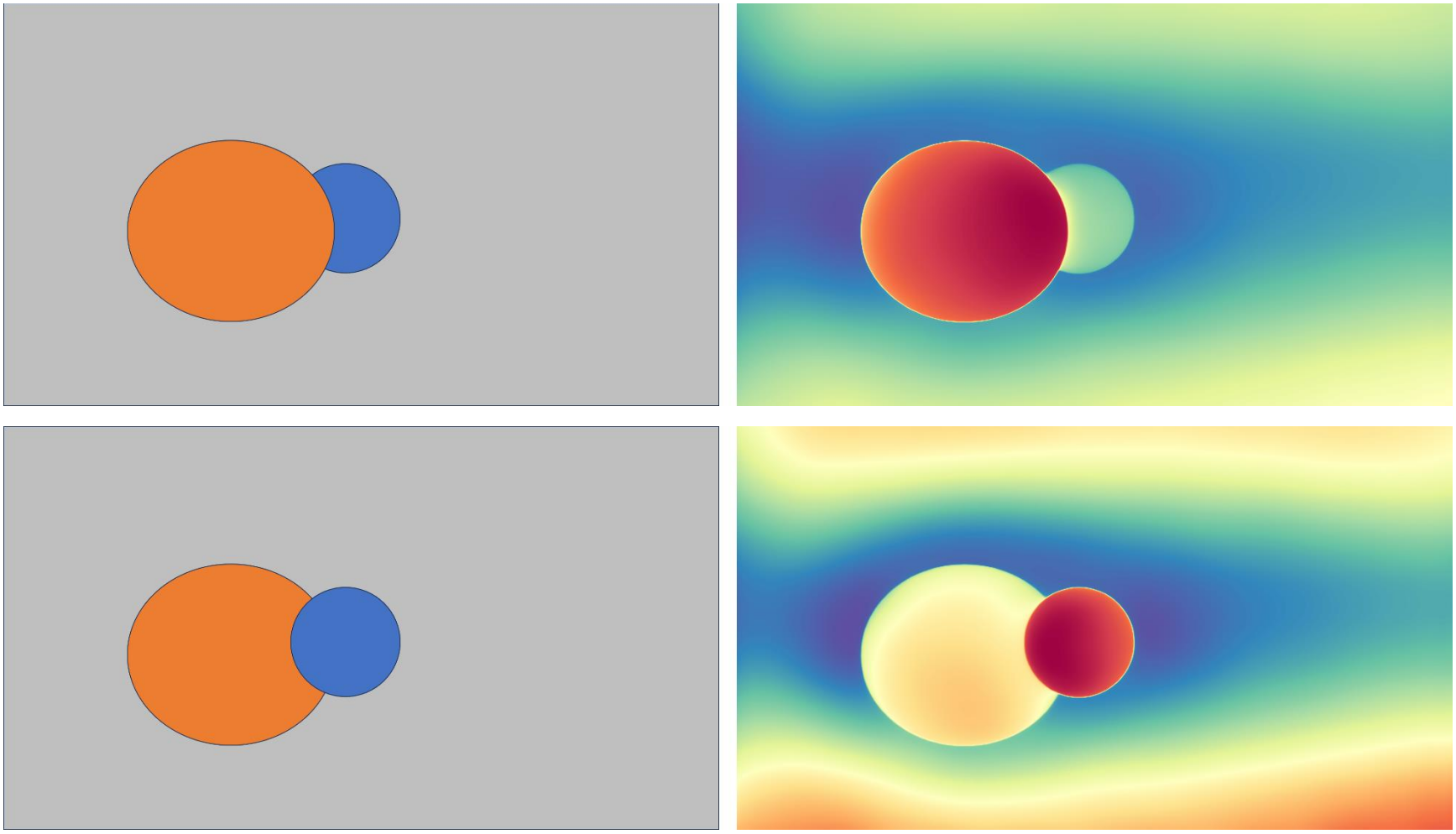


Fig 37. Depth Anything output for I13-I14:

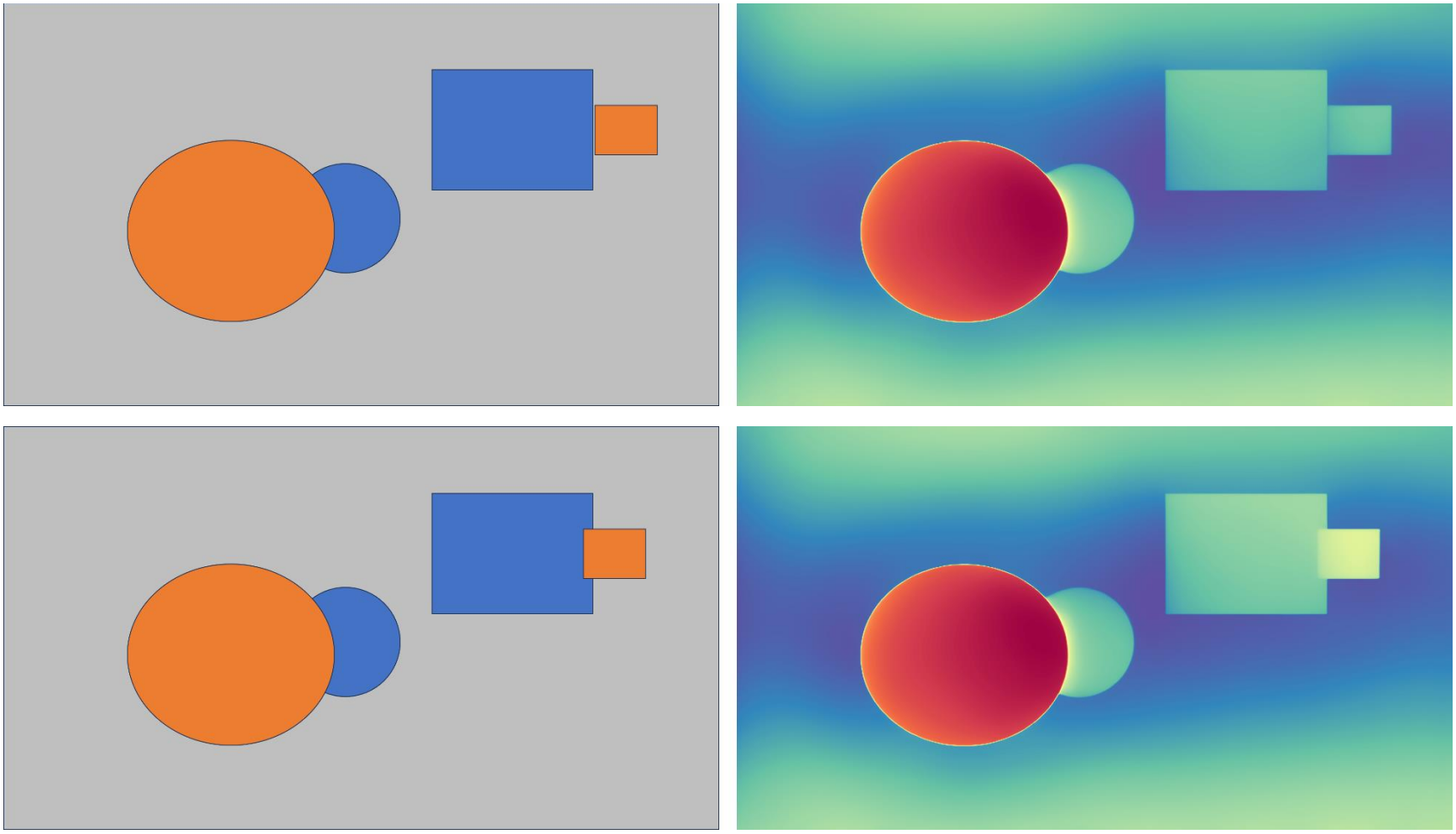


Fig 38. Depth Anything output for I15:

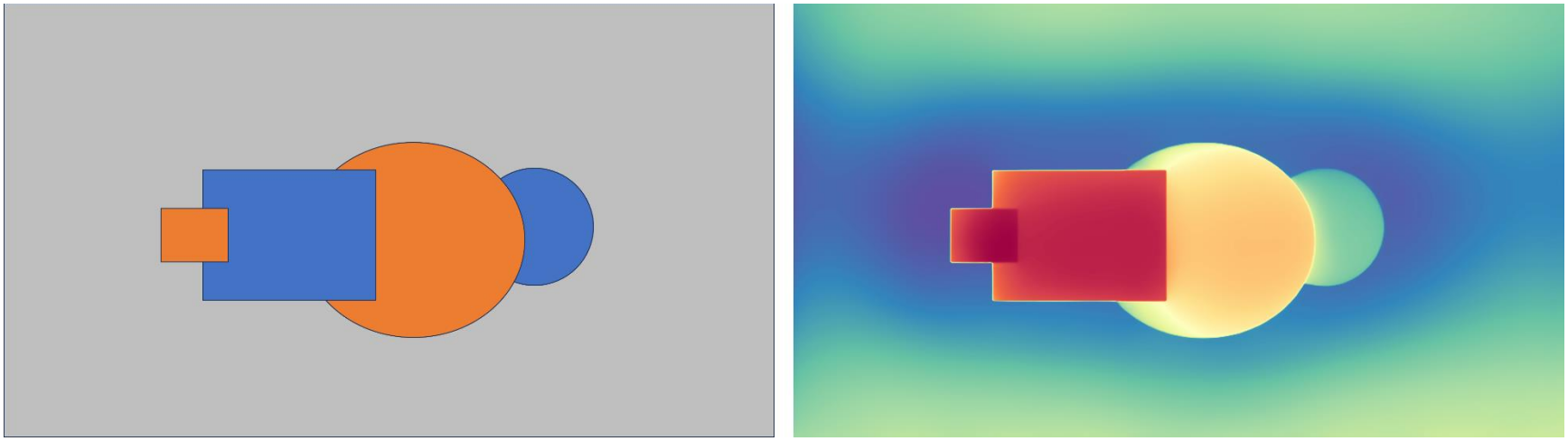


Fig 39. Depth Anything output for I16:

Fig 40. Depth Anything output for I1:

Fig 41. Depth Anything output for I2:

Fig 42. Depth Anything output for I3:

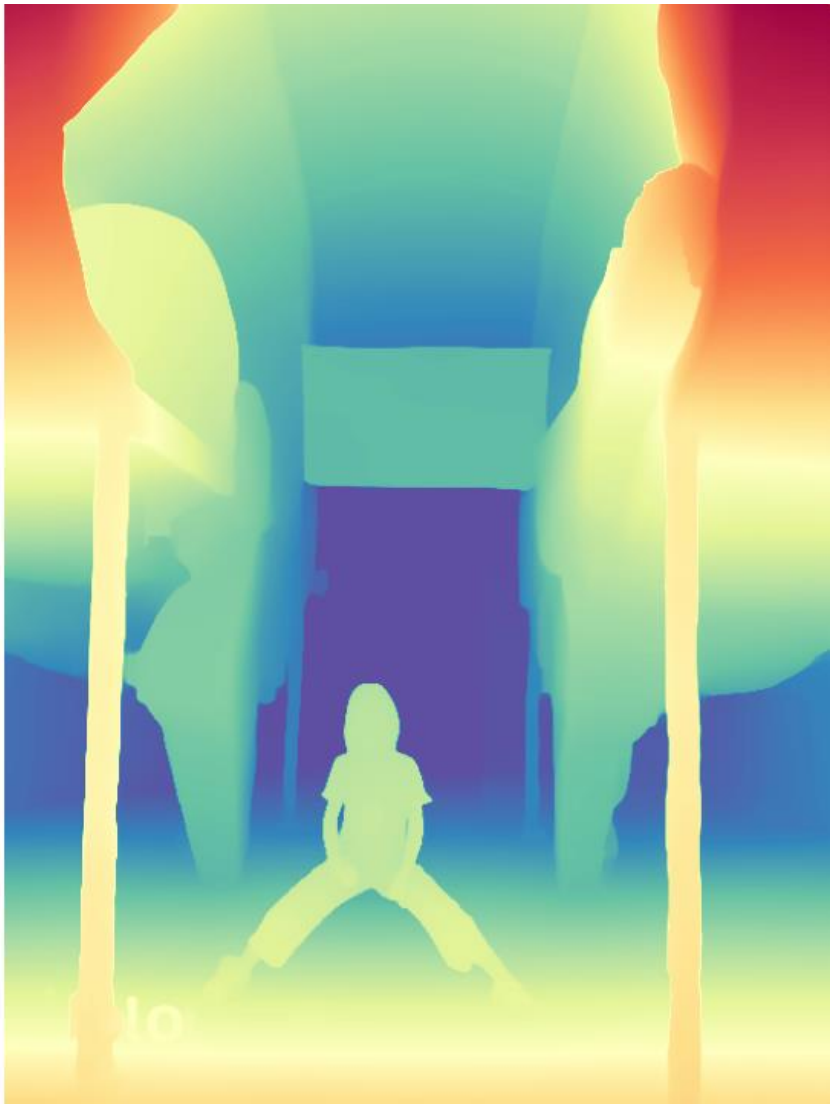


Fig 43. Depth Anything output for I4:

Fig 44. Depth Anything output for I5:

Fig 45. Depth Anything output for I6:

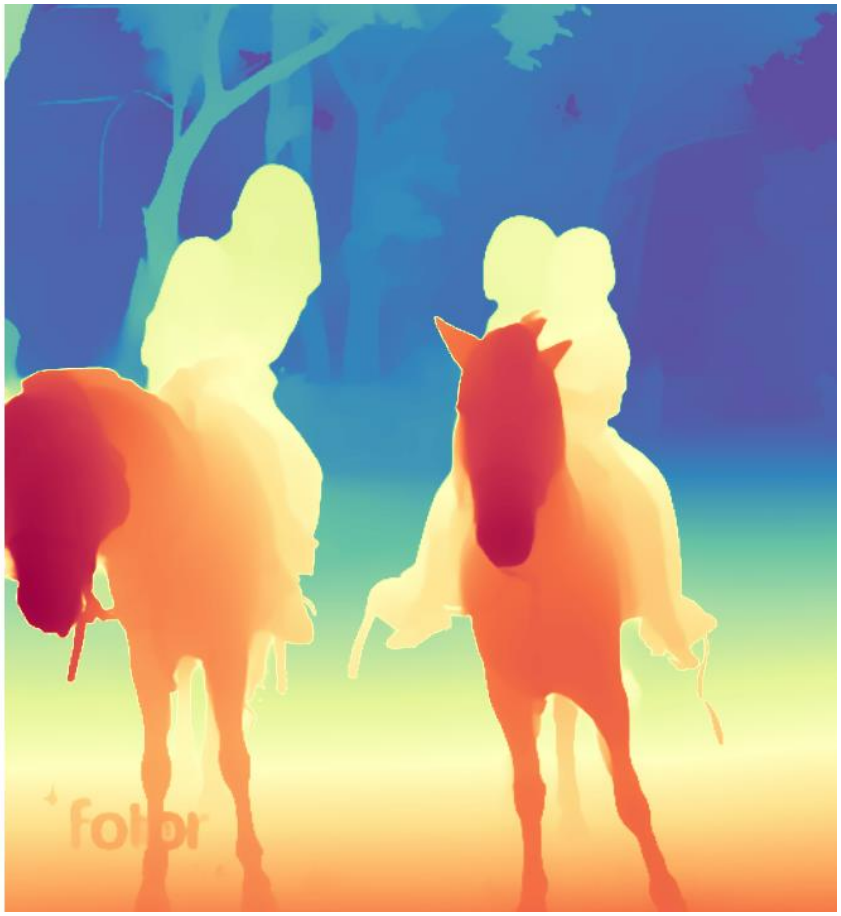


Fig 46. Depth Anything output for I7:

Fig 47. Depth Anything output for I8:

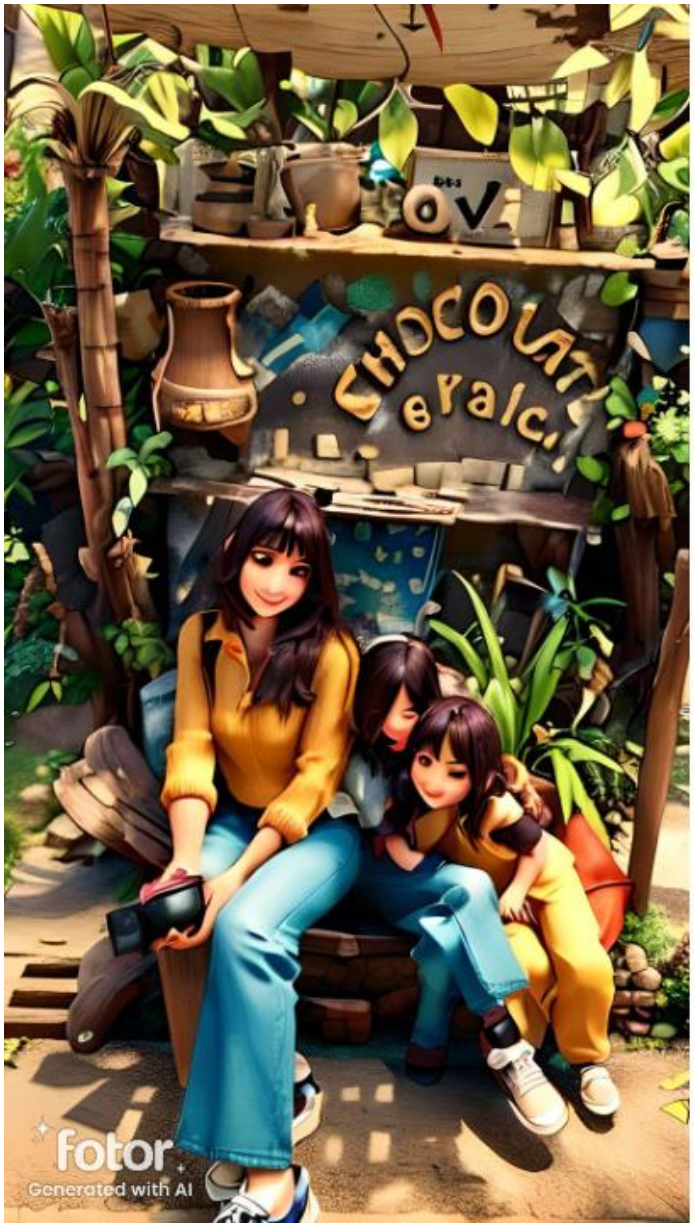


Fig 48. Depth Anything output – area for improvement 1:

Fig 49. Depth Anything output – area for improvement 2: