WaldoGen

Nathan Chanak, Ruikang Liu, Nick Nachtmann

ECE/CS 766 Computer Vision Final Project (Spring 26)

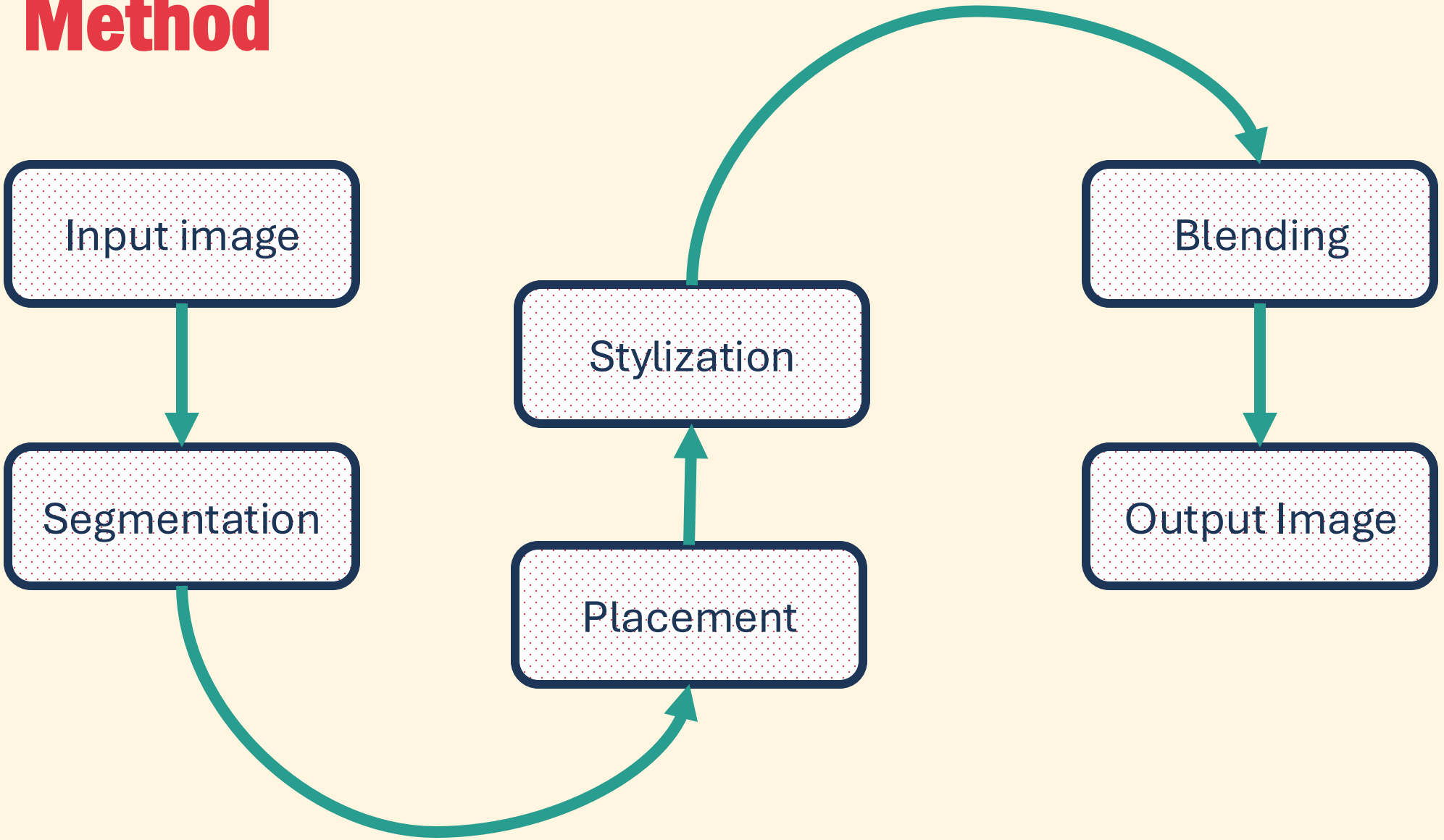
Motivation/Problem

Source: https://salonspanetwork.org/wheres-waldo/

Importance

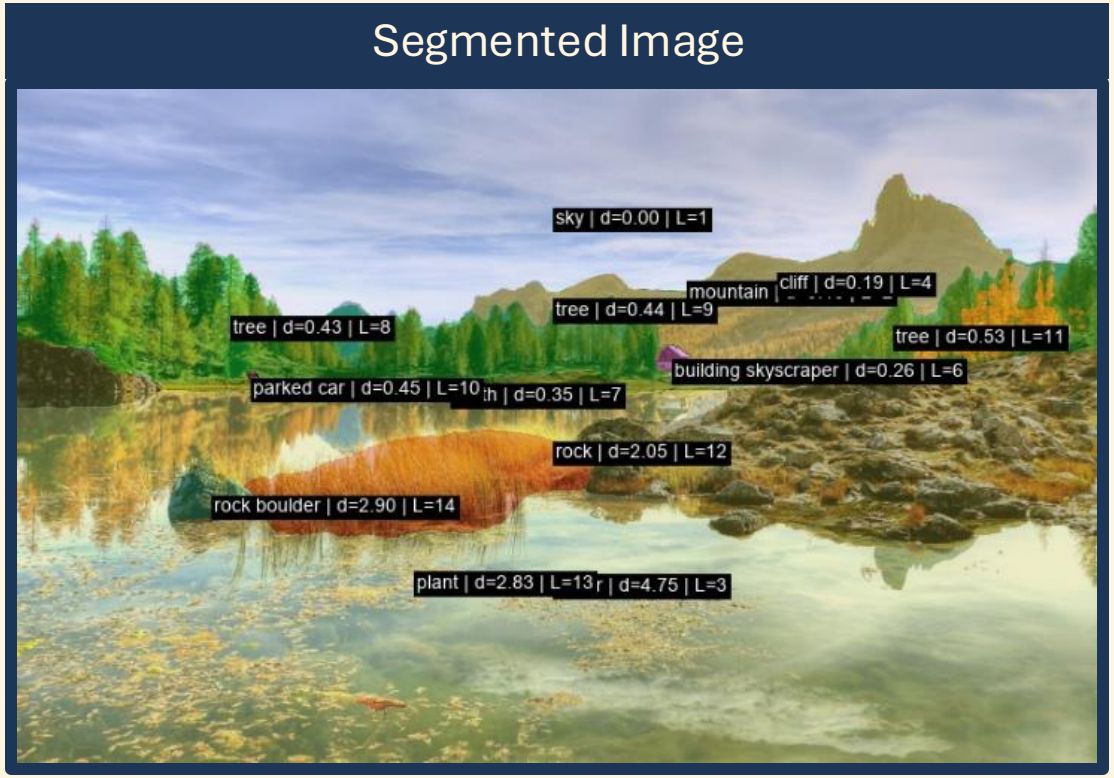
- Good chance to combine several Computer Vision techniques
- Challenging task with several steps: - Segmentation
  - Placement
  - Blending
  - Stylization
- Also, its fun!!!

Method

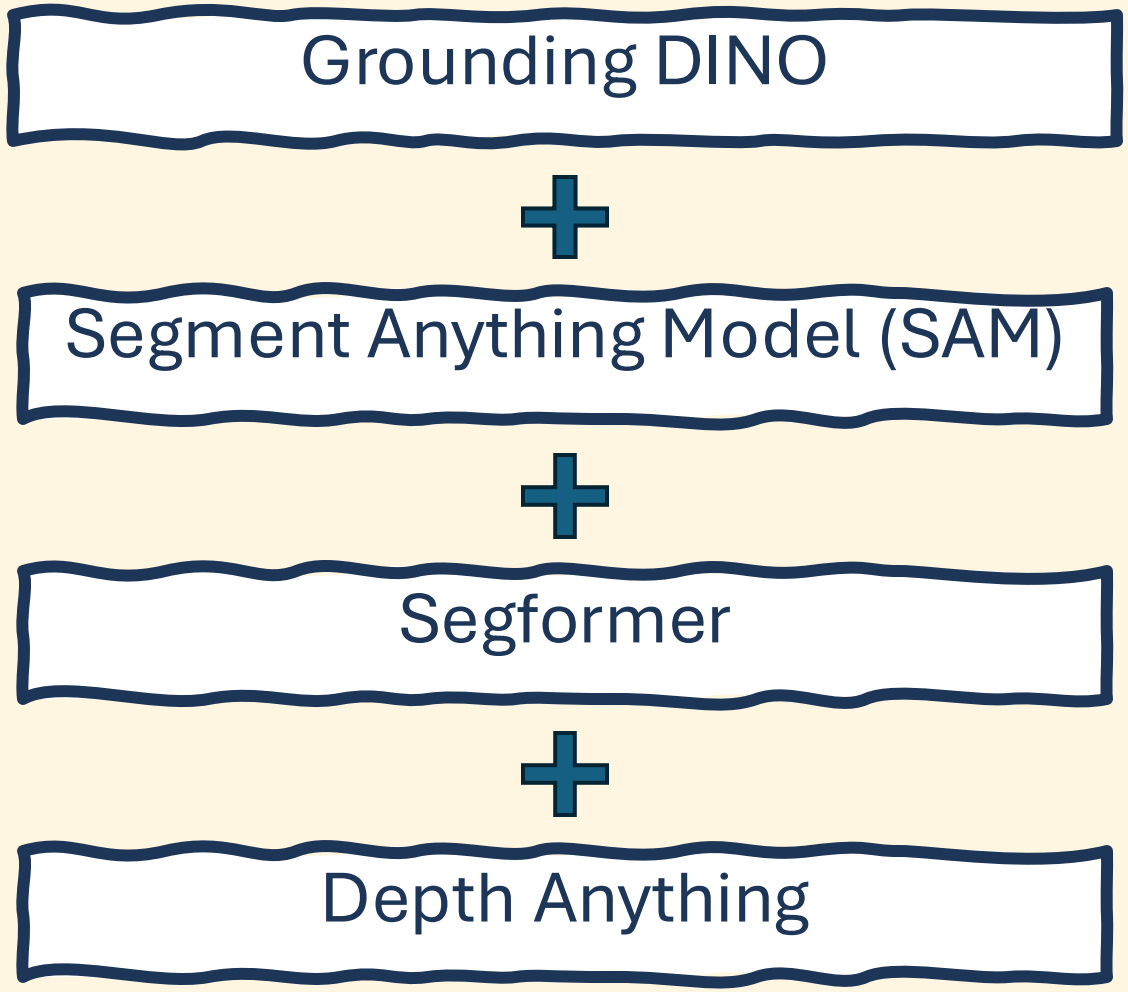

Input image
Segmentation
Placement
Stylization
Blending
Output Image

Segmentation

- Segmentation to enable realistic occlusion
- Avoid unrealistic placements
- Incorporate depth information



Segmentation



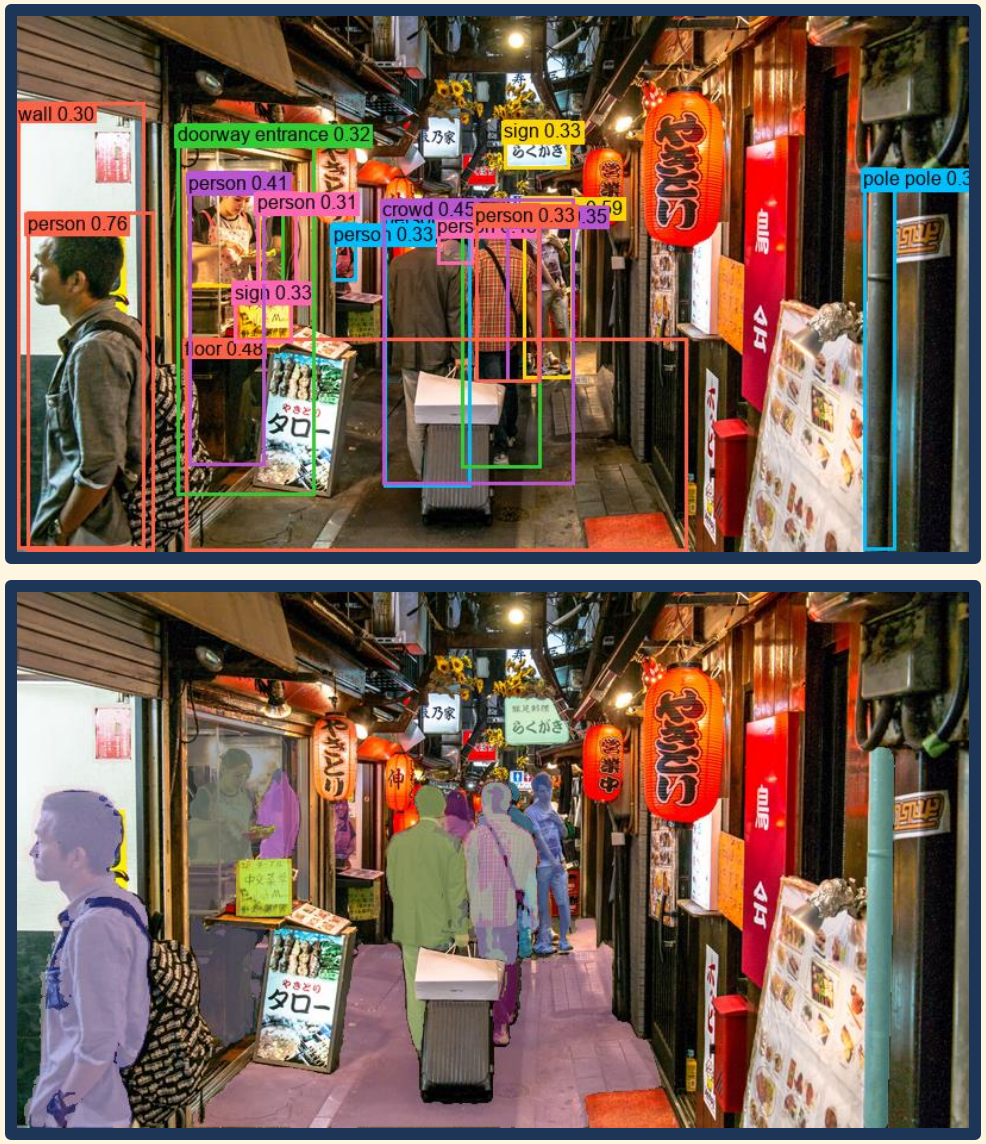
- Grounding DINO detects objects based on text input
- SAM converts detected objects into pixel accurate masks
- Segformer segments out broad regions such as sky or ground
- Depth Anything estimates depth as a pixel value 0-255

Segmentation

Grounding DINO

+

Segment Anything Model (SAM)



Segmentation

Segformer

+

Depth Anything

Segmentation Result

Placement

- Find location for Waldo
- Random but with certain conditions
  - Prefer certain objects and occluders
  - Partial occlusion up to x percent
  - Prefer visually busy areas
- Waldo is composited into image and reoccluded by object masks

Example of Placement Locations

Stylization

- Convert real images into Waldo-like illustrations
- Preserve scene structure and layout
  - Structure map: SoftEdge via PidiNet
  - Generation: SD 1.5 + ControlNet SoftEdge + Img2Img using a Waldo-tuned full SD1.5 checkpoint.

Original

Stylized

Stylization Result

Blending

- Make sure Waldo is inserted realistically
- Adapt: - Color
  - Brightness
  - Texture
- Avoid Waldo popping out

Original

Blended

Results

Full transition from input image to finished Where’s Waldo

Results

- Website with interactive game:

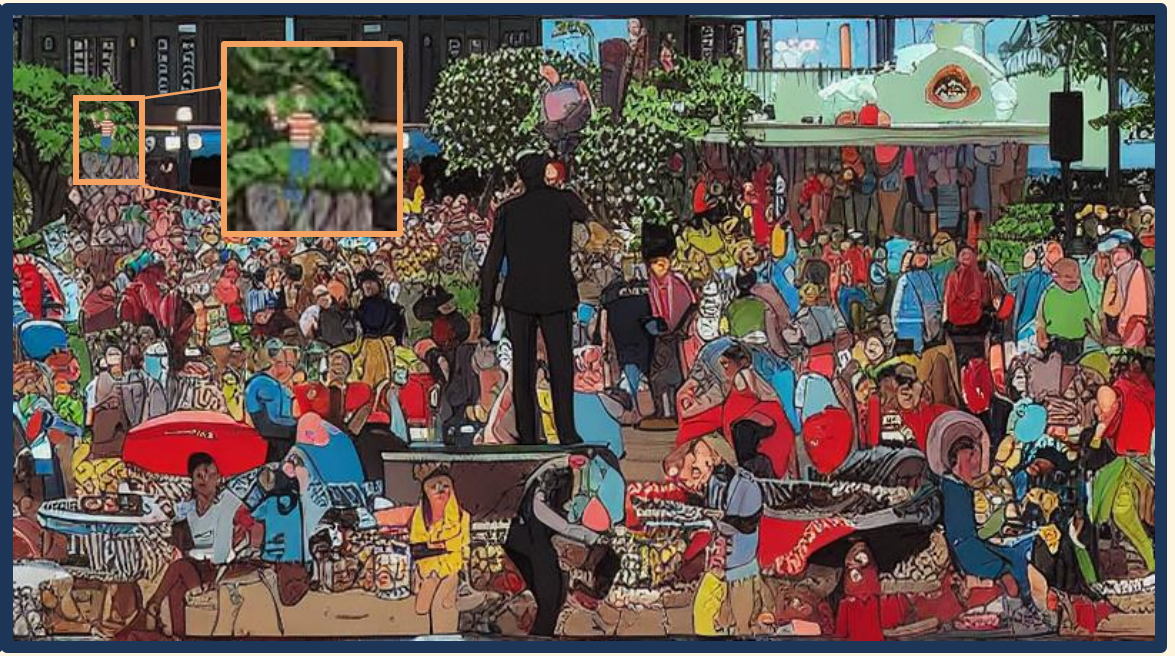
- Image of website or even video/live demo

Results

- Some more examples:

Results

Another example



References

1. S. Liu *et al.*, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” *arXiv.org*, Mar. 09, 2023. https://arxiv.org/abs/2303.05499
2. A. Kirillov *et al.*, “Segment Anything,” *arXiv:2304.02643 [cs]*, Apr. 2023, Available: https://arxiv.org/abs/2304.02643
3. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” *arXiv:2105.15203 [cs]*, Oct. 2021, Available: https://arxiv.org/abs/2105.15203
4. L. Yang *et al.*, “Depth Anything V2,” *arXiv.org*, Jun. 13, 2024. https://arxiv.org/abs/2406.09414
5. L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” arXiv:2302.05543, 2023.
6. Su, Zhuo & Liu, Wenzhe & Yu, Zitong & Hu, Dewen & Liao, Qing & Tian, Qi & Pietikainen, Matti & Liu, Li. (2021). Pixel Difference Networks for Efficient Edge Detection. 5097-5107. 10.1109/ICCV48922.2021.00507.
7. R. Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.