Semester Project Proposal

Nathan Chanak (nchanak), Nick Nachtmann (nachtmann), Ruikang Liu (rliu356)

Problem

The classic *Where's Waldo?* Books challenges readers to find a small, distinctive character hidden within a cluttered scene. Inspired by this idea, we want to build a system that automatically turns any natural image into a "Where's Waldo" puzzle. Rather than simple object pasting, we aim to generate realistic hidden object puzzle images where the inserted character (Waldo) appears naturally integrated into the scene. The key challenge lies in identifying appropriate hiding locations based on scene structure and visual complexity, and performing context-consistent compositing that avoids obvious synthesis artifacts.

Importance/Interest

This problem is interesting to us because it combines technical computer vision topics with an interactive game as a product. With all the techniques required like image blending, segmentation, and classification, it is an interesting testbed to combine them into a seamless pipeline. There are only about 100 Where's Waldo puzzles out there, and this project could create an endless number. Also, anyone can create their own version of it by uploading an image that they want to have turned into one of the puzzles using our website. We are excited by different ideas of how to implement the game whether it is a daily game like a Wordle or a timed puzzle rush. The combination of technical challenges and game design make it an enjoyable project to work on.

State of the Art

The main difficulty lies in blending the images of Waldo into the scene. Since lighting, brightness and color scheme may be different between the initial image of waldo and the target domain. This is a well-researched issue and there are several state-of-the-art approaches that we found that work on that issue. Zhang et al. [1] suggest a diffusion-based model that can preserve the underlying image details and properties by introducing consistent light transport during training. Chaturvedi et al. [2] suggests using a diffusion-based model that utilizes a physically based rendering engine to simulate lighting-conditioned transformation. Skorokhodov et al. [3] work with an approach based on a diffusion architecture that uses delta denoising score to ensure proper lighting correction.

Additionally, segmentation techniques are required. Kirillov et al. [4] that developed a foundation transformer-based segmentation model that can segment any object without knowing what it is. Another class of segmentation models are instance segmentation models that use some form of CNNs and masks often combined with a detection step like in He et al. [5].

Implementation

Our implementation uses a combination of state of the art approaches to create something novel. Below is a pipeline showing our planned implementation.

1. Input image into diffusion model to stylize to cartoony Waldo style
2. Image segmentation to show whether objects are in front or behind each other to determine Waldo's hiding spots
3. Choose Waldo's hiding spot randomly or by algorithmic choice, ensuring necessary parts of him aren't obscured
4. Add other characters/objects as distractors to the image similarly
5. Blend Waldo and other added objects with diffusion
6. Train a CNN model either from scratch or as an added class to a pretrained model to find Waldo in these images. Data will come either from online or our own generated images.

Bonus ideas:

- Make it a daily game with a picture of the day
- Give the option to hide other characters like Bucky in the pictures instead of Waldo
- Have a timed version that gives you a series of images that allows to compete against other players and get a score on the leaderboard
- Have CNN give out additional score for difficulty of puzzle (Easy, Medium, Hard)

Performance Evaluation and Metrics

Our primary metric for evaluating the Waldo scenes will be done by seeing how long it takes people to find Waldo in the image vs accuracy of trained model on finding Waldo. To put it formally, we have T_{human} which is the average human search time to find Waldo and the trained model's loss. These metrics will be evaluated individually, but we perform $\text{corr}(T_{\text{human}}, \text{loss})$. This will give a general metric that shows if the model as well as humans are aligned on scene difficulty. That way we can tell if the model is poor since humans can easily identify Waldo, or the other way around.

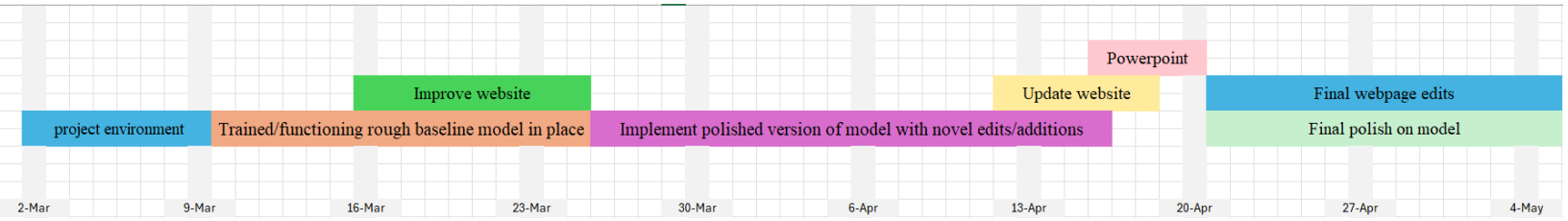
If Waldo is improperly placed and obscured, the previous metric could give a good score. So, a small but important evaluation we need to manually go over is whether Waldo is properly placed such that his face and parts of his clothes aren't completely obscured. This will ensure fairness so

people can find Waldo and the trained model can identify him. We can do this by generating 20 or so samples and seeing if Waldo is properly placed or not through manually classifying.

Finally we need some way of objectively evaluating how well Waldo and other placed objects fit in the scene. We can do this by evaluating the brightness and contrast between the placed objects and the rest of the scene. Since obviously Waldo himself doesn't fit in some scenes well, we will have a generous threshold to evaluate what's considered good.

Timeline

Start Date	End Date	Objective
3/2	3/6	Set up project environment (dataset, repo, website, etc)
3/9	3/25	Rough baseline of project
3/16	3/20	Improve Website
3/25	4/16	Polished Model with novel edits/additions
4/9	4/20	Update website with model changes
4/16	4/20	Create PowerPoint
4/23	5/6	Final website and model polish



References:

[1] SCALING IN-THE-WILD TRAINING FOR DIFFUSIONBASED ILLUMINATION HARMONIZATION AND EDITING BY IMPOSING CONSISTENT LIGHT TRANSPORT- Lvmin Zhang , Anyi Rao , Maneesh Agrawala - https://openreview.net/pdf?id=u1cQYxRI1H

[2] SynthLight: Portrait Relighting with Diffusion Model by Learning to Re-render Synthetic Faces - Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, Zhixin Shu - https://arxiv.org/html/2501.09756v1?utm

[3] D3DR: Lighting-Aware Object Insertion in Gaussian Splatting- Vsevolod Skorokhodov, Nikita Durasov, Pascal Fua - https://arxiv.org/pdf/2503.06740

[4] Segment Anything - Alexander Kirillov et al., https://arxiv.org/pdf/2304.02643

[5] Mask R-CNN- Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshic - https://arxiv.org/pdf/1703.06870