ECE766 Project Midterm Report

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

Depth Aware Segmentation

One major component of this project is segmenting the image so that we can place Waldo partially in front/behind objects in a manner that makes some sense. Doing this with regular segmentation won't be effective, as we not only need labels, so Waldo isn't placed in the sky, but depth to know what objects are occluding other objects.

We combined 4 models into one pipeline to create a depth aware semantic segmentation. Grounding DINO detects smaller objects like people, signs, and poles. We then use SAM to create segmentation masks of those objects. From there we used SegFormer to segment groups of objects and large features like the sky, groups of buildings, and tree lines. Finally, we ran the image through Depth Anything and compared our segments against the depth at those segments' locations to give each segment a depth score. In the images on the first page, you see inputs on the left, and on the right segments are colored with object class, depth, and label.

Stylization

The primary objective of the stylization stage is to convert realistic photographic input into a dense, hand-drawn aesthetic while preserving the global spatial layout. This structural consistency is vital, as it provides a coherent canvas for the subsequent stages to accurately place Waldo and other distractor characters within the scene.

Our technical approach utilizes a diffusion-based img2img framework, specifically leveraging Stable Diffusion 1.5 coupled with ControlNet (Canny). By extracting edge maps from the source photograph, ControlNet ensures that the generated illustration maintains structural integrity. To achieve the “Waldo” look, we employ a fine-tuned model that encodes the specific palette of bright primary colors and the characteristic pen-and-ink line art bias.

We have successfully established a functional end-to-end stylization pipeline, moving from a single input file to a stylized output. Preliminary testing on representative datasets—such as dense outdoor plazas and theme-park crowds—indicates that the model effectively preserves the overall composition and shifts the appearance toward a flat-color, outline-heavy illustration with high-contrast, primary-heavy colors. A side-by-side comparison is provided below, showcasing the original photograph on the left and the stylized output from our pipeline on the right.

Despite plausible global style and layout, several granular challenges remain to be addressed in the second half of the project. The stylized results read well at a glance, but they still fall short on fine detail. In crowded areas, faces and limbs dissolve into soft, blob-like shapes rather than the crisp, readable figures in Martin Handford’s originals. The familiar red-and-white striped motif and other signature visual cues appear only inconsistently, often reduced to flat reds or generic color blocks. Where many people overlap, outlines turn thick and muddy, so individuals are harder to separate. We also work at a moderate resolution to keep the model stable, which leaves very small figures with few pixels to work with and caps how much detail can ever appear. Our upcoming work will focus on enhancing the visual clarity of output. We aim to refine our control mechanisms to better preserve fine-grained details and investigate post-processing techniques to improve overall resolution.

Blending

This part of the project focuses on blending and integrating Waldo into the target scene after the segmentation stage is done, and the location and scale are determined.

Since the segmentation stage is still under development and to get some initial testing done, manual scaling and placement are being used.

To make sure that Waldo does not stick out a pipeline of methods are being used to adapt waldos appearance by adjusting color, brightness and edges. The current pipeline is designed to handle a variety of scene conditions, as different insertion locations and images can vary significantly in lighting and texture. Local color and brightness matching adjusts Waldos' appearance to the statistics of the surrounding image region. In addition, sharpness and noise adjustment are applied to match the image quality of the background to avoid Waldo from appearing overly sharp compared to the rest of the scene.

For the final integration, different blending strategies are being explored to integrate waldo seamlessly into the scene. Specifically, so far, we experimented with alpha blending with a feathered mask to create a smooth transition at the boundaries without losing Waldos characteristic features as well as Poisson-based blending that enforces gradient consistency with the background. They both show different scenarios in which they perform better or worse than one another. Beyond these “classical” approaches we are also looking into some learning-based methods such as DoveNet, which can refine the composite by adjusting the color inconsistencies and illumination as well as diffusion-based approaches that can further increase the realism.

The whole pipeline is modular, which allows the segmentation outputs to later be plugged in and modify Waldos' visible region, and enabling the integration of the aforementioned more advanced methods as an additional refinement step.

Revisions

1. Current plan for the website is to have a suite of generated Where’s Waldos that the user can play.
2. We are holding off on letting the user generate their own Where’s Waldos through the website due to technical difficulty, but you can always make your own by running the code through GitHub.
3. Due to images being such a large domain, some Where’s Waldo puzzles would be easier than others. We decided it’s best to focus on urban area images, but technically all images would be supported.
4. Making the CNN will be held off to focus on getting a full working pipeline.
5. The metric on loss of trained CNN vs human time to find Waldo may be amended as it doesn’t have much utility.