Using GANs for Outpainting of Objects in a Scene

Sukrit Arora, Jordan Grelling, Dre Mahaarachchi, Scott McCrae

The Problem

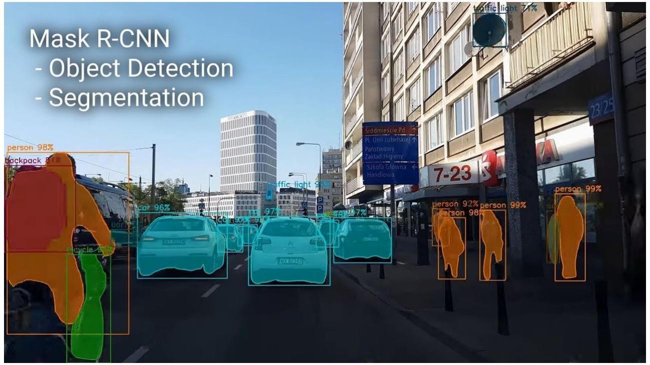
- Scenes can contain objects that may lower the quality of the image or obstruct the purpose of the image
- Examples:
  - Nature photography containing manmade objects (e.g. vehicles, power lines, people, etc.)
  - Removing people/objects from an image to declutter the focus of the scene
  - Photos of people who wish to not be in the photo due to privacy/security reason
- This is a 2-stage problem:
  - Identify the target object(s) accurately and precisely
  - Remove and fill in the missing space without introducing noticeable distortion

The Framework

- The two stages, Detection/Removal and Reconstruction, can be accomplished independently
    - Can employ a separate solution for each subproblem
    - Need to pipeline the result of Detection/Removal ⟶ Reconstruction
- Object Detection and Removal ⟶ Mask R-CNN
    - Accurately detects and identifies common objects in a scene (given a confidence threshold)
    - Somewhat accurately performs classification on these objects
- Reconstruction ⟶ Generative Image Inpainting
    - Uses the relatively new Generative Adversarial Network to fill in missing chunks of photos reasonably well
- Both systems are open-source and have pretrained models

Recap of Mask R-CNN

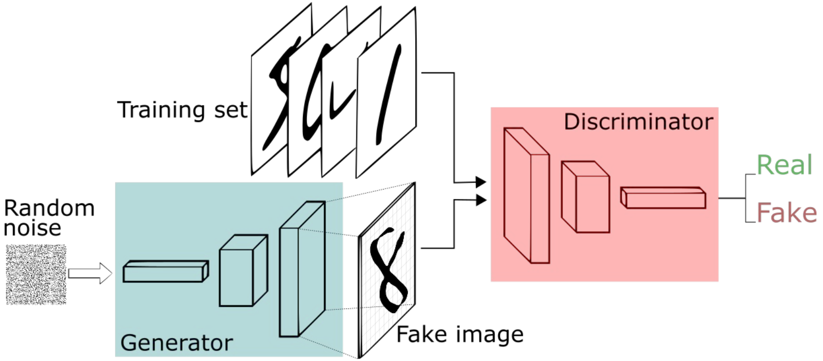
- Region-based CNN for object detection using region proposals
- Generates high-quality, accurate segmentation masks for objects in images
- Usage:
    - Input: any RGB image
    - Output: A segmented mask, bounding box, and categorical confidence for each detected/recognized object in the image

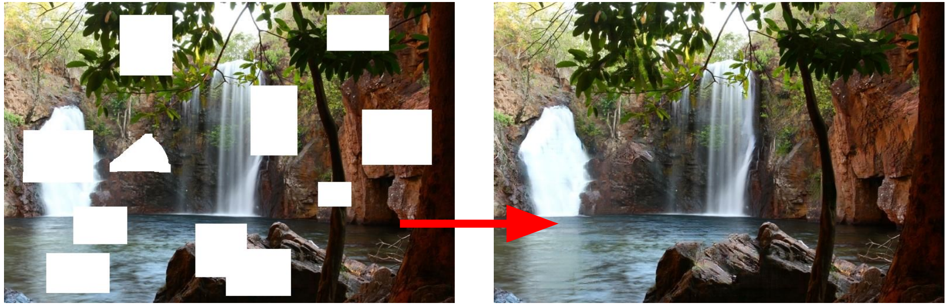


More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (T

Overview of GANs and Generative Inpainting

- Generative Adversarial Network
  - an unsupervised framework
  - two networks, generator and a discriminant, engaged in a zero-sum game
  - generator attempts to 'fool' discriminant by synthesizing authentic-looking outputs
  - Training continues
- Generative Inpainting
  - Input: an image with a missing/blank section
  - Output: a new image with black section filled with generator's best guess of the missing section
  - Paper: Generative Image Inpainting with Contextual Attention, J. Yu et. al., 2018

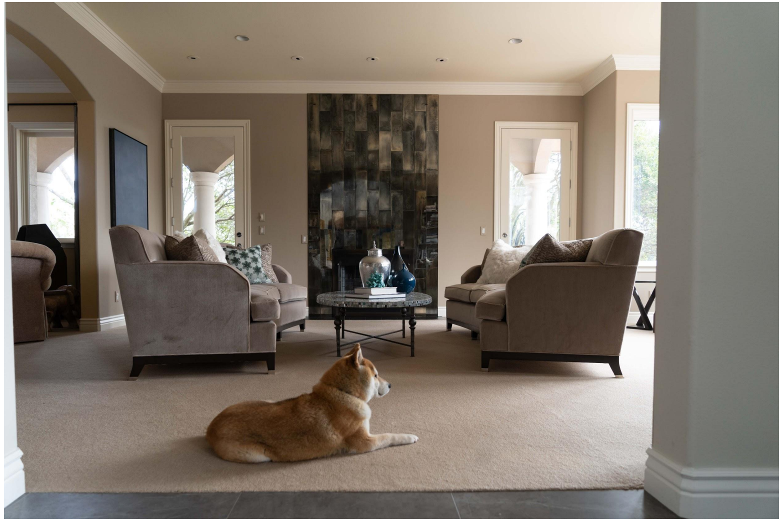


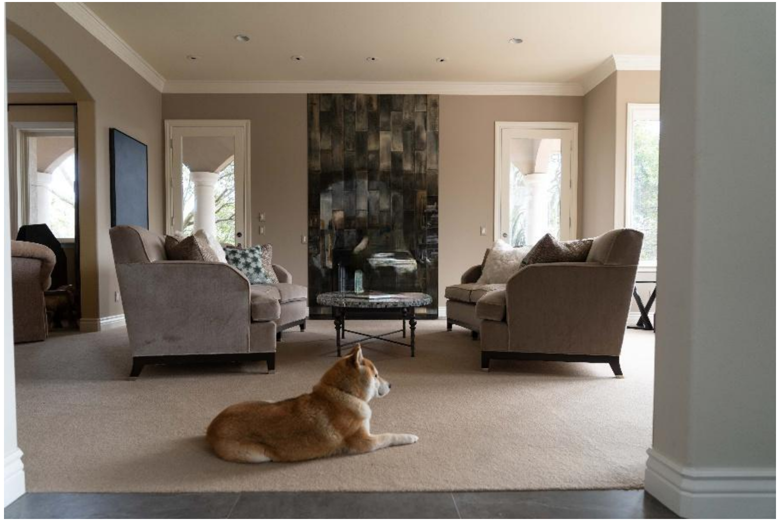


The Network Pipeline

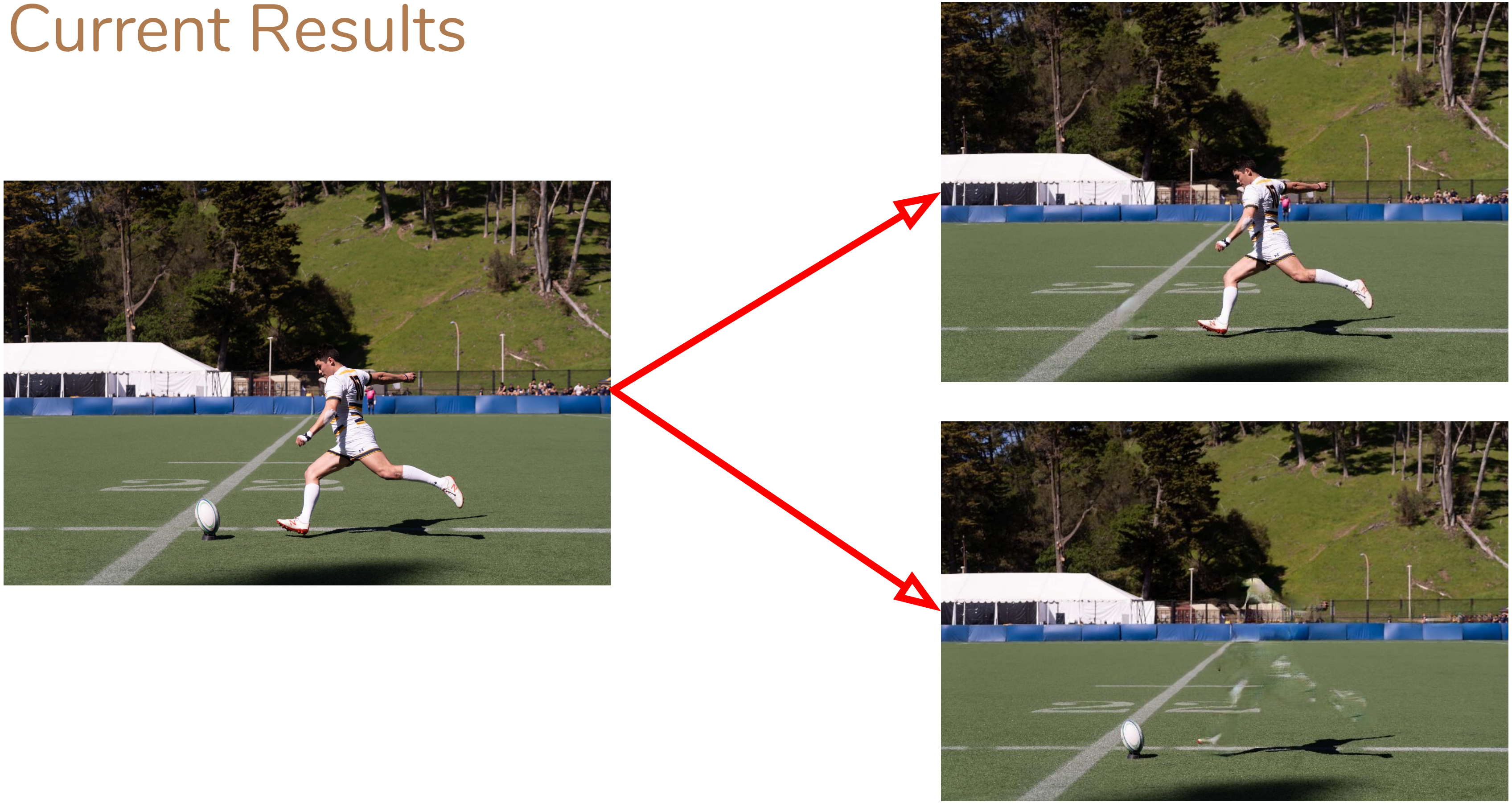
- System Inputs:
    - An image and a set of objects categories that are to be removed
- Stage 1: Mask R-CNN
    - Output downscaled image with highlighted masks and categorical labels of detected objects
    - Create a mapping of the labels to the masks
    - Select only the masks/bounding boxes that have the matching categorical labels defined in the input
    - Run a low-pass filter on the masked pixels within the selected bounding boxes and threshold the output (>0 -> 255) and return this as the new mask
    - Using new mask, set all of the pixels within it to zero ⟶ this is the output
- Stage 2: GAN Inpainting
    - Feed the output of stage one into the GAN Inpainting network

Current Results

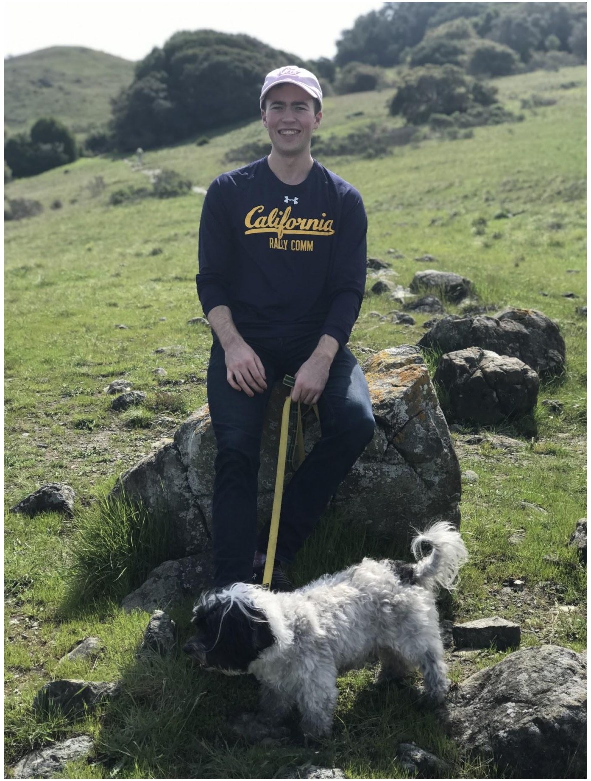


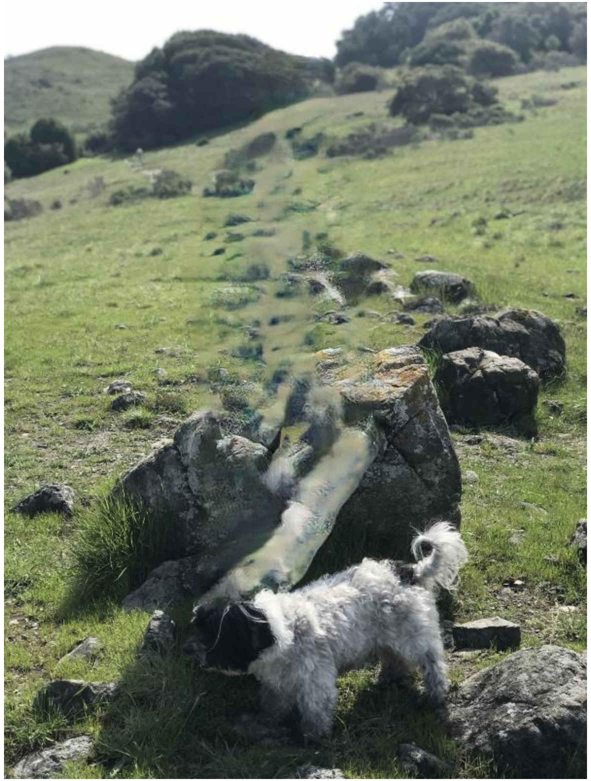


Current Results

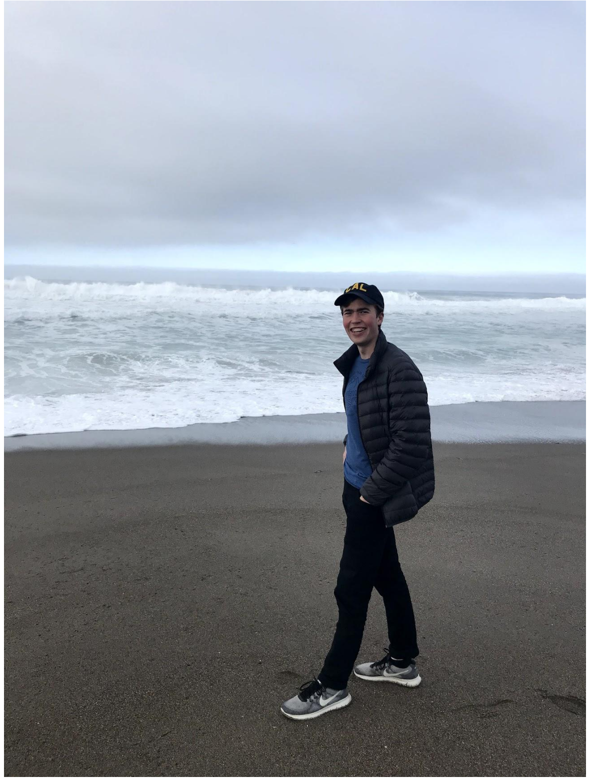


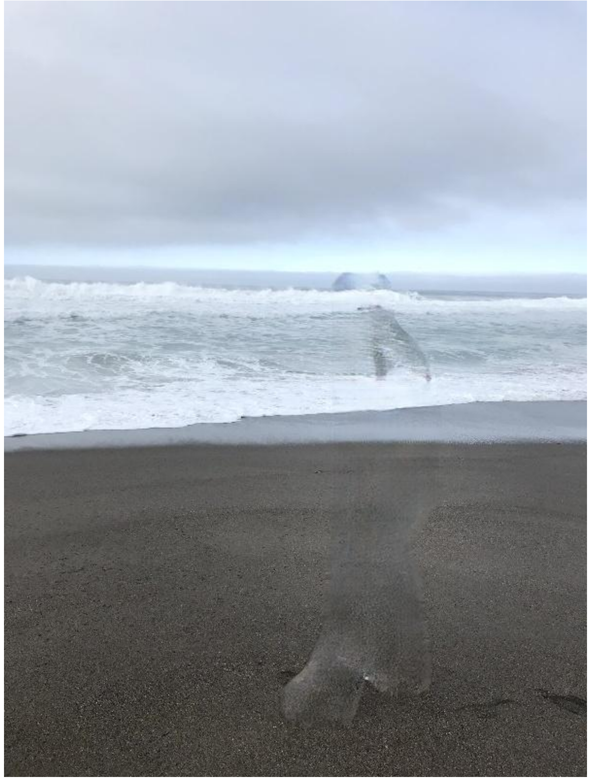
Current Results



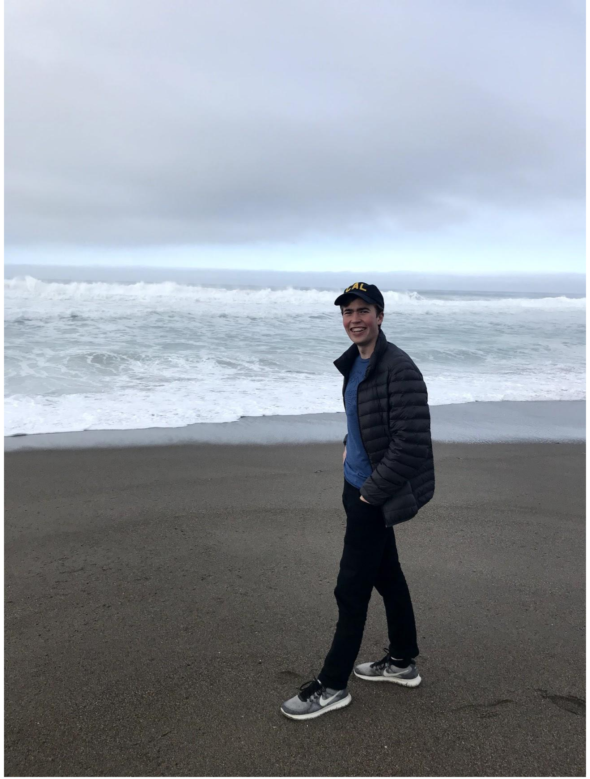


Current Results

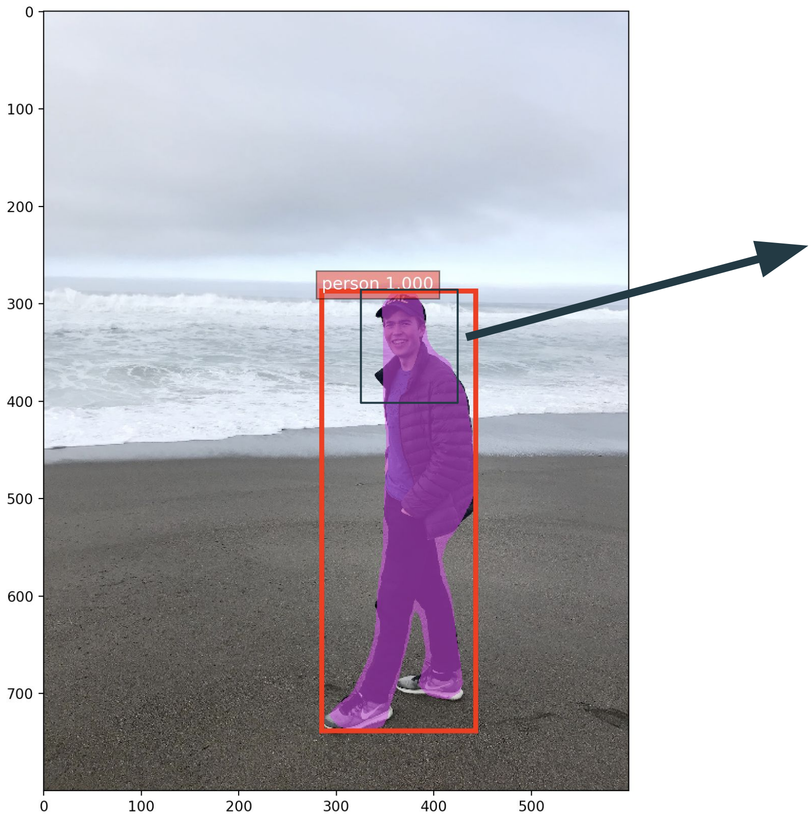


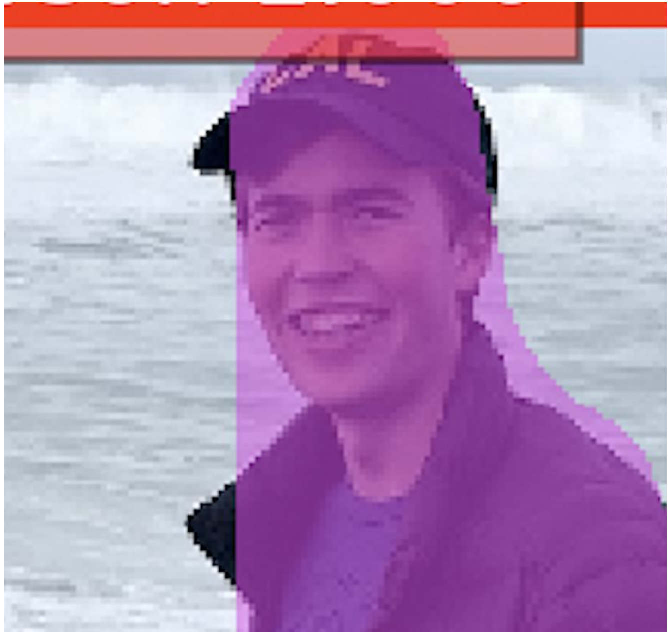


Mask Expansion: Input



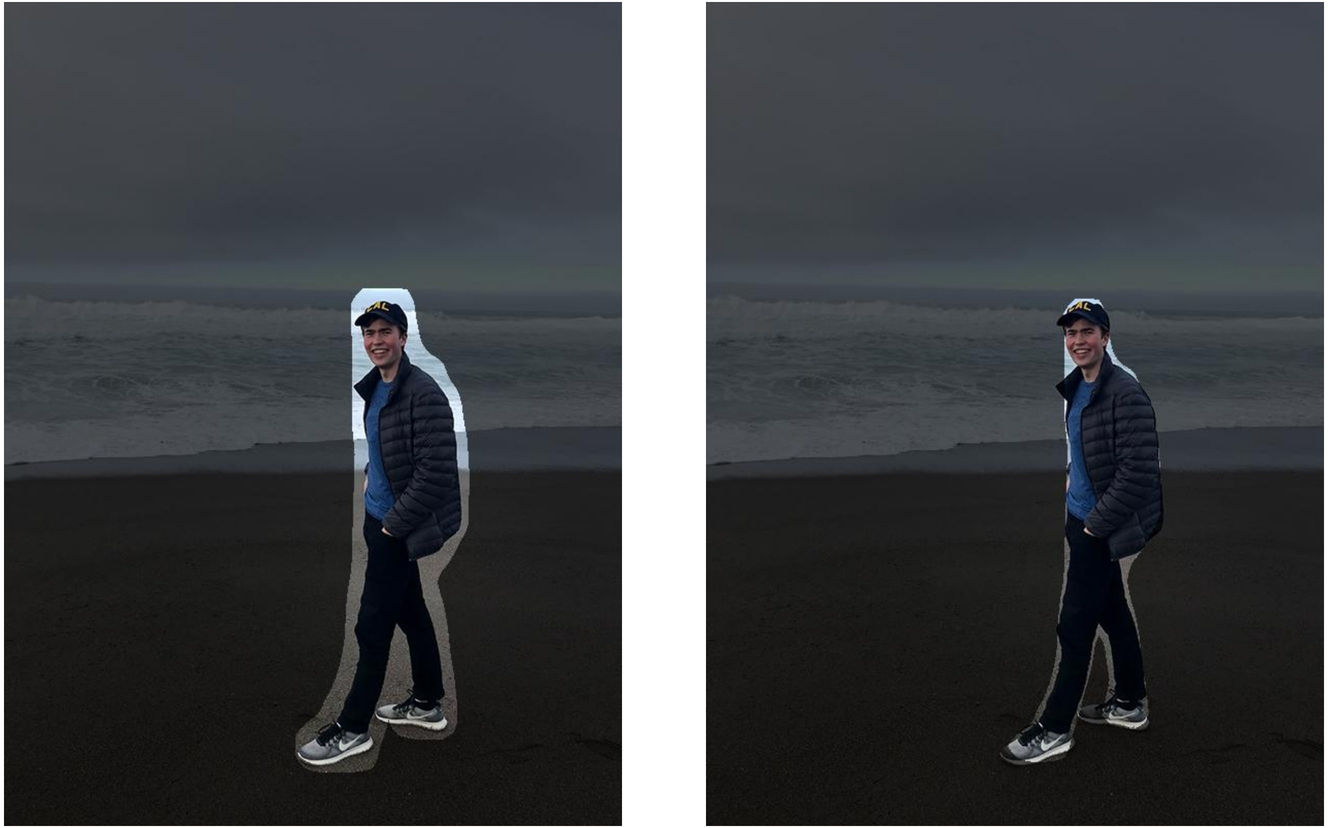
Mask Expansion: First Mask



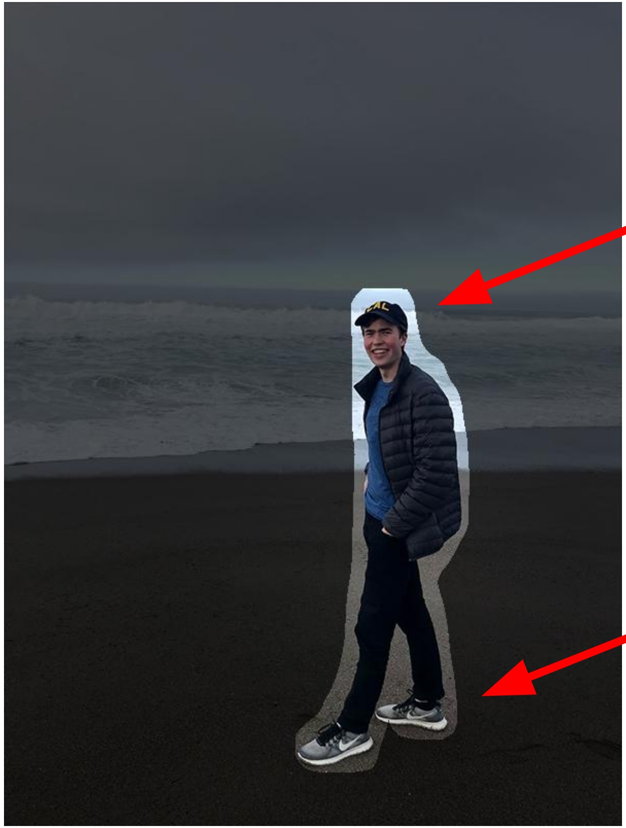


Note that the edges of the hat and jacket are not captured by the mask, as well as the shoes

Mask Expansion: New vs Old



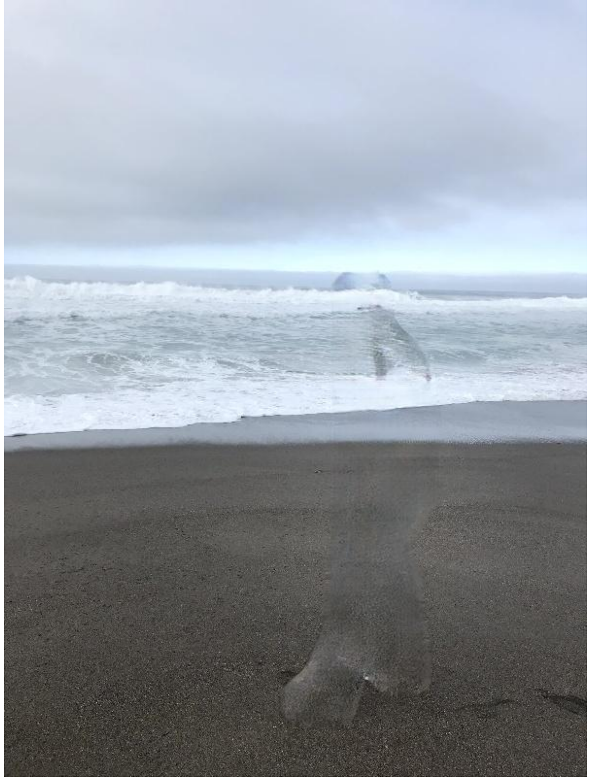
Mask Expansion: Padded Mask

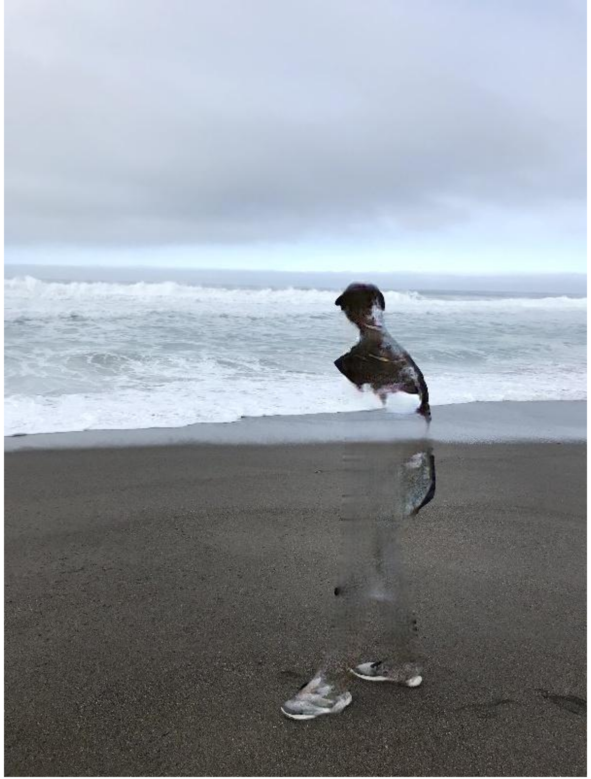


Entire hat, jacket included

Shoes included

Mask Expansion Results: New vs Old





Potential Innovations

- Shadows
  - Target objects may have a shadow that still remains in the photo
    - This is difficult to solve traditionally because it requires an understanding of the lighting quality and angle in the scene
- Distortions on mask edges
  - Caused by:
    - Abrupt discontinuity in the original image and missing section
    - Network not train on the background pattern
  - Potential fix:
    - Low pass filter the image and the unsharp the output
    - Use a network trained on the pattern

Challenges and Solutions

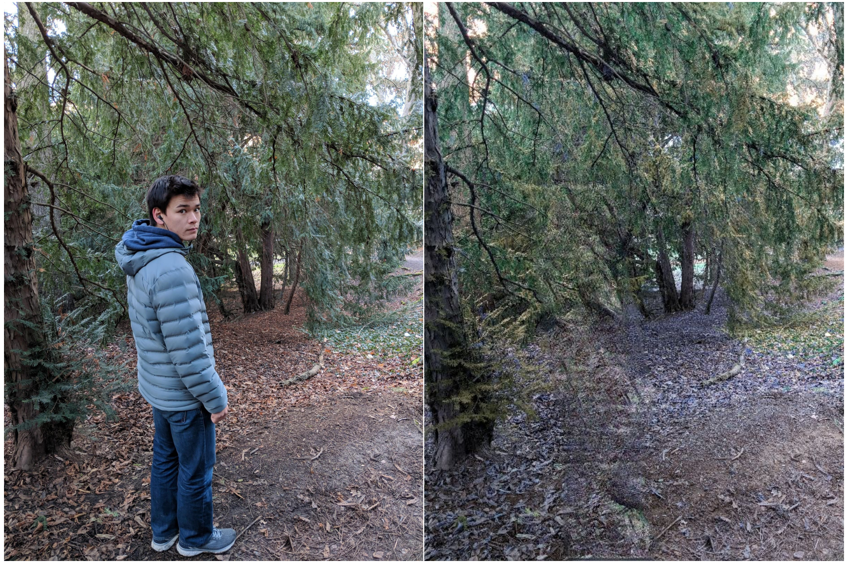
1.	Mask R-CNN tends to generate masks that do not cover the entire object	→	1.	We pad a buffer around all generated masks to ensure we cover the whole object
2.	The outpainting performance is dependent on the dataset the GAN trains on	→	2.	We use the Places2 dataset (pretrained) under the assumption that users have a wide array of use cases
3.	Mask R-CNN and GAN inpainting are slow to run on consumer computers	→	3.	Downsize input images such that the short side is 600 pixels long, this runs in less than a minute

Roadblocks Encountered

- Mask R-CNN timeliness
  - The paper in which He et. al. detail Mask R-CNN states that the algorithm runs at 5 frames per second on an average computer; this is definitely not true. Mask R-CNN is the rate-limiting factor in our network and typically takes minutes to run even on small images
- Places2 Dataset size
  - We wanted to train our inpainting model on the Places2 dataset; however, this set contains over 10 million images and is completely infeasible for us to use. Instead, we will be training on COCO, and at 330,000 images, this is still at the limit of what is possible using the resources we have available

Future Work

- Color balance issues
  - On some inputs, the blue and red color channels seem to be having problems
- Loss of resolution
  - Currently, the network has to downsample large images severely to be able to run both halves of the network, we could work on an algorithm to upsample the output and use the original as context
- Attention Module
  - The GAN inpainting algorithm from Yu et. al. has an attention module that we haven't really explored



Example showing blueshift in output images

Ethical Concerns

Dangers

We also wanted to explore the ethical implications of this work on censorship or IP infringement. There are two main concerns why this might be dangerous:

- Ease of replication
  - Created by college students, not difficult or computationally expensive to recreate
- Modularity
  - There's no reason the RCNN couldn't be modified to detect a certain person's face with no change to the rest of the network

Solutions

Reproducibility means it's impossible to prevent access to, and likely this would have been a problem regardless of whether we demonstrated this algorithm

- Detection of altered images
  - Future work could focus on creating a discriminator that can outperform the generator
- Ethical benefits
  - Could also be used to preserve privacy