CS577 Project Proposal

Samuel Golden (A20430084)
Theo Guidroz (A20426895)

Problem statement

With the rise of working from home and attending meetings virtually, people are concerned about privacy concerns. When turning their camera on during an online event, the participants expose their background which might intrude on their personal life. To remediate this solution, most communication platforms adopted a background replacement technology where the users can display a virtual background instead of showing the room in which they are. While the idea is great and certainly solves the privacy issues, most companies did a mediocre job at extracting the background completely. This project focuses on perfecting background matting in real-time. Using two neural networks, the goal is to perform a high-resolution background replacement technique that operates at 30fps in 4K resolution, and 60fps for HD on a modern GPU.

Methodology

The team will take a similar approach as the one discussed in the paper to solve the problem. We will start by implementing the base network which is a fully-convolutional encoder-decoder network inspired by the DeepLabV3 [4] and DeepLabV3+ [5] architectures. The base network consists of three modules:
Backbone: ResNet-50, a convolutional neural network that is 50 layers deep
Atrous Spatial Pyramid Pooling: a semantic segmentation module for resampling a given feature layer at multiple rates prior to convolution.
Decoder: It applies bilinear upsampling at each step, concatenated with the skip connection from the backbone, and followed by a 3×3 convolution, Batch Normalization, and ReLU activation.
Then, a refinement network will be implemented to reduce redundant computation and recover high-resolution matting details. This will be done in a two-stage refinement process. First, the network bilinearly resamples the coarse outputs. Then, it crops out 8×8 patches around the error locations. Finally, it upsamples the coarse alpha matte and foreground residual to the original resolution and swaps in the respective 4×4 patches that have been refined to obtain the final alpha matte and foreground residual.
During training, multiple data augmentation techniques will be implemented to avoid overfitting and help the model generalize to challenging real-world situations.

Data Source

We plan on using the same data sources as referenced in the paper, since most of them are available to the public. These are the Adobe Image Matting dataset, a human-only subset of Distinctions-646, and a novel dataset of web-crawled images from the researchers of the paper.

Description of the paper

The paper attacks the same problem as described in the problem description. Using two datasets comprising matted videos and photos of differing quality, the researchers created neural networks capable of producing real-time high-resolution video matting that separates the person from their background.

The researchers note that the real-time processing required to do this operation in real-time is incredibly expensive, and they could not solve the problem generally for any space in the video. Therefore, they define two neural networks, one to operate at a lower resolution and identify which locations require higher resolution (like hair), and a second to operate at a higher resolution (only at those locations identified as difficult by the first network). They refer to these networks as the base and refinement networks, respectively.

The base network comprises ResNet-50, ASPP (a pooling module with a series of convolution filters dilated at various rates), and an upsampling decoder. This network returns a “coarse” alpha matte, an error prediction map, and a hidden feature map, all of which will be passed along to the refinement network.

The refinement network only operates on values in the image with a particularly high error map value. It resamples the coarse matte, hidden features, and input image, and concatenates them into a tensor. Then run that tensor through a series of convolutional layers for refinement. And finally, concatenate the coarse alpha with the refined alpha to get a final matte.

The paper also compares their results to existing software, such as that of Zoom and Skype.

Individual Responsibilities

The following table contains the specific work to be done, and the person responsible for each:

WORK	PERSON
Collect data from listed datasets	Sam
Collect pre-trained models where necessary (ResNet-50)	Theo
Design and implement base model	Sam
Design and implement refinement model	Theo
Train models on collected data	Sam
Evaluate results and capture new data for examples	Both
Create project report and presentation	Both