Training a UNET Network to Segment Drapery and Object Contact Surfaces

Jibak Sarkar

This report is a short description of the steps performed to complete the final part of the Computational Visual Perceptron project where I trained a UNET [2] network to properly segment the parts of the image where the cloth touches a given 3D object.

1 UNET: Network Description

In this project, I trained the UNET network using a dataset generated in phase II of the project which consists of 512 by 512 grayscale images. The target images are distances between the rendered cloth and the objects. The whiter shades in target images denote less distance between the cloth and the object. The objective was to generate output images of the same dimensions, which did not involve segmentation but rather regressed pixel values (distance measure but inverse transformed). To achieve this, a sigmoid activation function was applied to the last convolution layer, treating the task as a regression problem. The mean squared error (MSE) loss function was used instead of the categorical cross-entropy loss commonly employed in segmentation tasks. The network was trained with a batch size of 16 and a learning rate of 1e-04, over a span of 50 epochs. These settings were carefully chosen to optimize the network's performance and achieve accurate regression results. In addition to the original architecture specifications, I have included batch normalization after each Convolution and Convolution Transpose operation. I have included both the source code of the implementation and trained network weights ('`model.pth.tar`' format) in the folder named **uNet** as part of the project uploads.
Unfortunately, I forgot to keep track of the training and validation losses of the network for them to use in a plot. As retraining the network is quite computationally expensive, I have not included the loss curve plot in this report. However, I manually chose the two models with the least training and validation loss and used them for further evaluation.

2 Method & Results

After the training, I tried to interpret what the models learned by plotting their predictions on unseen test images. Fig 1 shows the result for one example. I went ahead with the model with minimum validation loss. Next, I re-scaled the model outputs for all the test images by a factor of 255. Now there is another minor hyperparameter that I can tune to get the best results in the downstream fabric/contact segmentation task. This means that if the predicted pixel values of the rescaled images exceed the threshold the model classifies that pixel as 'contact', and 'fabric' otherwise. I tried threshold values between 60 and 180. 73 yielded the best average IoU score of 0.354 over the entire test images (see Fig 2) It is important to note that the ground truth annotation for these segmentation masks is also done by hand. Hence, it is subject to perception bias.

3 Dicussion and Future Work

The training and validation data for this experiment was synthetically generated using 3D computer graphics rendering software Blender and then capturing 2D images from random viewpoints in the upper hemisphere in the 3D space around the object. Whereas, the test data used are also 3D objects but in this case the cloth is already draped around the object and no cloth simulation was needed. This led to a significant difference in various cloth properties such as hardness. translucency etc. Also, a major difference between the cloth simulation and the test data is that in the simulation cloth fell freely on the object whereas the test image objects are perfectly draped using the cloth. This might be a major challenge to get a more generalized model. Hence, the model is yielding comparatively better results in the validation compared to the testing results. So we can say that the model works good when considering very basic visual stimuli but it fails to

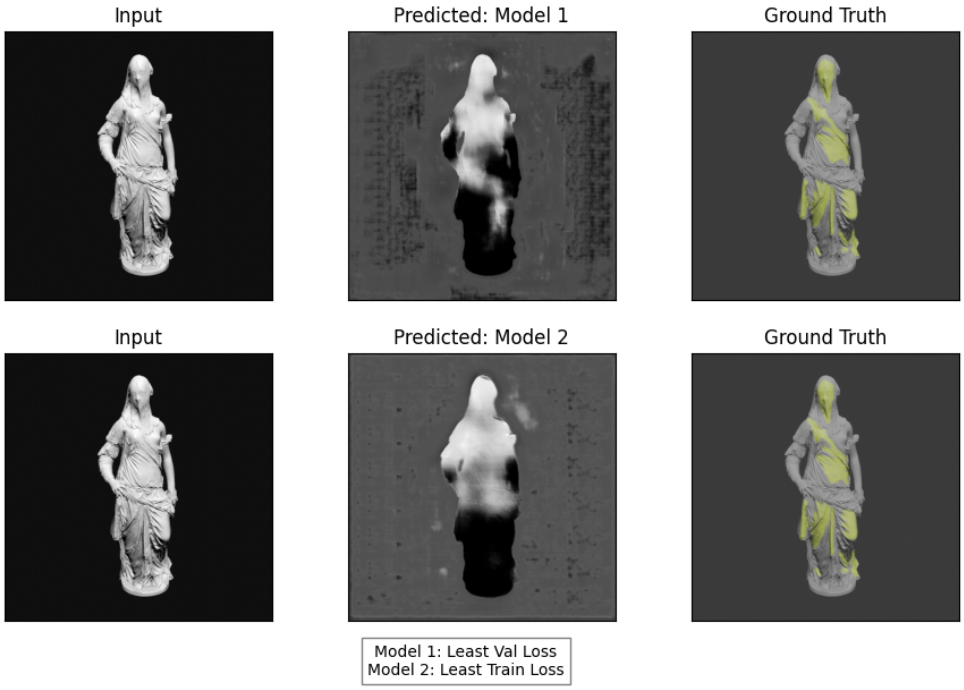

Figure 1: Input, Model Outputs & Manually Annotated Ground Truth

Figure 2: Threshold v/s IoU

capture the more complex ones.
I think the challenge of predicting fabric vs. contact from 2D training data is quite challenging in nature, but it should be solvable by using enough data augmentation on the input data. These augmentations could potentially be geometric transformations such as crop, zoom, etc. or some kernel filters could also be used to change the sharpness.
The first paper [3] discusses human perception which is done by shading the most obvious intelligible part of the image, which is similar to the model predictions after observed after the training. Another observation that is worth mentioning is that the non-draped part of the test object is predicted as not draped by the model but not the other way around.
According to the second paper [1], although unfamiliar stimuli appear opaque, both transparency and the Veiled Virgin effect involve breaking down sensory signals into multiple layers. In transparency, a single visual element is divided into two subjective layers: a background seen through a transparent layer. However, with the limited data at hand, I think, we were not able to replicate such perceptual behaviour in the model.

References

[1] Flip Phillips and Roland W. Fleming. The veiled virgin illustrates visual segmentation of shape by cause. *Proceedings of the National Academy of Sciences*, 117(21):11735–11743, 2020.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 9351:234–241, 2015.

[3] Ilker Yildirim, Max Siegel, Amir Soltani, Shraman Chaudhari, and Joshua Tenenbaum. 3d shape perception integrates intuitive physics and analysis-by-synthesis, 01 2023.