CS230

Clothing Segmentation for Virtual Try-On

Sanjana Sarda*
Department of Electrical Engineering
Stanford University
ssarda@stanford.edu

Tony Sun*
Department of Computer Science
Stanford University
suntony@stanford.edu

Abstract

Machine learning models for virtual try-on attempt to warp and overlay clothing items on top of a target person. While state-of-the-art models such as VITON-HD can produce impressive looking results on test data, we find that they face difficulty adapting to real-world images. In this work, we look at improving the segmentation masking for such models to improve model robustness and adaptability.

1 Introduction

Online shopping for clothing apparel is a quickly growing global market. The COVID-19 pandemic has led consumers to realize a latent demand for virtual try-on as a more accessible alternative to the fitting room. Prior work in this field looks at the problem of overlaying a garment on top of a 2D image of a person. We find that while state-of-the-art models can generate impressive results on test images, these same models are brittle to change and face difficulty generalizing to everyday examples from the real world. In particular, we observe that popular models such as VITON-HD[4] and ACGPN [16] can hallucinate artifacts, distort body shape, and blur out clothing and body parts.

Figure 1: Virtual try-on models face difficulty adapting to images from the real world. *Left:* ACGPN [16] overlays a t-shirt over images of a student (top) and fashion model (bottom), but incorrectly distorts both the body shape and clothing item. *Right:* VITON-HD [4] overlays a Uniqlo sweater over images of a student (top) and fashion model (bottom), but grossly misconfigures the result.

*Equal Contribution

CS230: Deep Learning, Fall 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

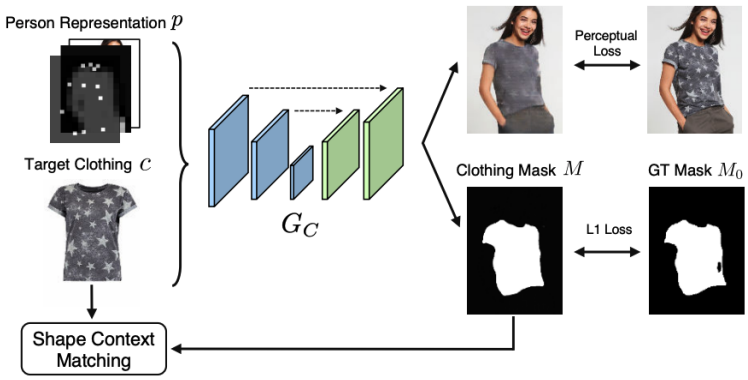

Figure 2: VITON overlays a warped clothing item on top of a target person using a clothing mask generated from a pose representation. The pose representation is assumed to already be present, either manually annotated or from a separate third-party model.

The disparity in performance between selected examples and real-world examples inspire us to make virtual try-on more robust to shifts in data distribution. While VITON-HD [4] is capable of producing accurate and faithful images, the model is dependent on a pre-existing segmentation mask of the target person. A segmentation mask that is inaccurate due to different lighting conditions or poses can lead to sub par results when overlaying the target clothing item on top of the user (Figure A.1).

In this paper, we improve improve on the pre-processing step of creating a segmentation mask of the target person. Specifically, we use the open-source mmdetection toolbox [3] built on PyTorch to label clothing and human attributes of an input image of the given user.

We find that we are able to generate accurate, high-resolution segmentation masks using a Mask R-CNN [5] as the backbone of our model.

2 Related Work

Atrous ResNet-50. Segmentation Task for Fashion and Apparel [2] aims to develop different CNN architectures for the purpose of generating segmentation masks. The paper find Atrous ResNet-50 with an SGD optimizer to have the best accuracy and IoU metrics compared to other architectures such as SegNet. The final accuracy was reported to be 93% on the iMaterialist dataset.

Feature Pyramid Networks. Semantic Segmentation of Fashion Images using Feature Pyramid Networks [12] uses fully convolutional neural networks based on FPNs with ResNeXt backbones. The paper reports a final accuracy of 93.26% on the Refined Fashionista dataset without a conditional random field (CRF) and a final accuracy of 93.62% on the same dataset with a CRF.

Evaluation Frameworks. Jouanneau et al. aims to change the evaluation approach for fashion segmentation architectures [8] by using mAP for instance segmentation evaluation. The paper evaluates its results for Mask R-CNN and Yolact after training for 5 epochs and reported AP50 results of 0.567 and 0.687 correspondingly.

VITON-HD [4]. Virtual Try-On Network HD (VITON-HD) aims to synthesize a new image in high definition from a reference image *I* of a clothed person and a target clothing item *c*. Han et al. combine a pose representation of the person along with a clothing mask to generate a new image with the same person now in a warped version of the target clothing item *c*. The pose is determined using an early version of OpenPose [1], and the clothing mask is created with a multi-task encoder-decoder generator. The result is then adjusted with a refinement network to produce the final image.

ACGPN [16]. Virtual Try-On by Adaptively Generating and Preserving Image Content (ACGPN) aims to transfer a target clothing image to a reference person by predicting the layout of the reference image that will be changed and then determining whether to generate or preserve consequent body or clothing content.

VOGUE [9]. Try-On by StyleGAN Interpolation Optimization (VOGUE) aims to generate a target person in a given garment using pose-conditioned StyleGAN2 latent space interpolation. This combines area of interest regions from the image of the person such as body shape, hair, and skin color, with area of interest regions from the garment such as folds, material, and shape. By optimizing interpolation coefficients in the latent space per layer, a merged result is generated at a 512 x 512 resolution.

Novel Contributions. This paper adds to the current body of work by building upon the work in [12] by using CARAFE with feature pyramid networks. We also were able to leverage various research in the object detection sphere to surpass the accuracy and mAP scores of previously existing segmentation architectures.

3 Dataset and Features

A number of datasets collected explicitly for the purpose of virtual try-on are no longer publicly available [4, 7]. DeepFashion is a large-scale, publicly available database of clothing images designed with bounding box detection and image annotation in mind.

Figure 3: A sample image from DeepFashion and its corresponding segmented version.

DeepFashion [11]. DeepFashion contains over 800,000 diverse fashion images ranging from well-posed images to unconstrained consumer photos, constituting the largest visual fashion analysis database. Additionally, each image in DeepFashion is labeled with 50 categories, 1,000 descriptive attributes, bounding box and clothing landmarks. For this work, we specifically used the In-shop Clothes Retrieval subset. The subset includes dense pose and parsing mask annotations with bounding boxes for 7,892 clothing items. The dataset provides classification for 14 classes - top, skirt, leggings, dress, outer, pants, bag, neckwear, headwear, eyeglass, belt, footwear, hair, skin, and face. The training set includes 6816 examples, the validation set includes 3111 examples, and the test set includes 3809 examples. Some images are randomly flipped as a part of the pre-processing data augmentation step to prevent overfitting. The images are normalized with a mean of [123.675, 116.28, 103.53] and standard deviation of [58.395, 57.12, 57.375] to match the pre-trained weights for the PyTorch implementation of Mask R-CNN with ResNet-50 as the backbone of our model. All images have a resolution of 750 x 1101.

4 Methodology

The goal of this work is to produce accurate instance segmentation results for users and apparel. To accomplish this, we use a Mask R-CNN framework implemented with the help of mmdetection, an open-source object detection toolbox [3]. Mask R-CNN generates potential regions of interest for object discovery based on the input image and then based on its prediction of the class of the object, refines the bounding box, and generates a pixel-level segmentation mask. This framework is constructed with a convolutional architecture (e.g. ResNet-50) as the backbone for feature extraction over the entire image and the network head for bounding-box recognition (classification and regression) and mask prediction that is applied to each region of interest (RoI) [5]. The backbone can be combined with a Feature Pyramid Network (FPN) structure to extract RoI features from different levels of the pyramid according to their resolution scale [5].

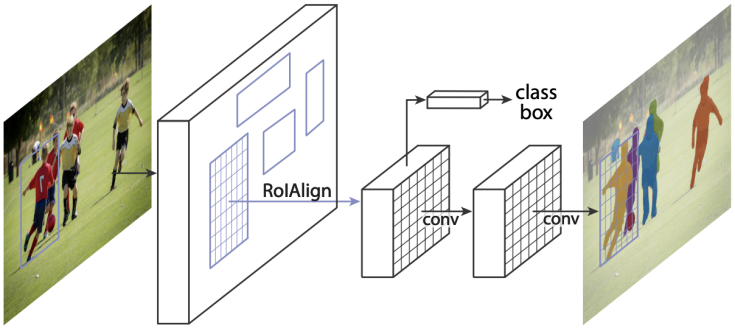

Figure 4: The **Mask R-CNN** framework for image segmentation.

Our implementation uses a ResNet-50 [6] backbone initialized with pre-trained weights on ImageNet due to its high segmentation performance on the COCO dataset. We used an implementation of Content-Aware ReAssemble of FEatures (CARAFE) [15] with FPN on top of the ResNet-50 backbone. CARAFE is an operator that reassembles features as a spatial block inside a predefined region centered at each location via a dynamically generated weighted combination such that feature upsampling can be performed. This method has shown to have better accuracy metrics with Mask R-CNN compared to without and lowers the computation cost when used with FPN.

We use a RoI head and a Region Proposal Network (RPN) head for our implementation of Mask R-CNN. RPN is responsible for generating the proposal for potential objects and consists of a classifier and a regressor [14]. The classifier uses a Cross Entropy Loss with a loss weight of 1.0 while the bounding box detector uses a L1 loss with a loss weight of 1.0 for both heads. The RoI head uses a Cross Entropy Loss with a loss weight of 1.0 for its internal mask head.

5 Results

During training, we tested different optimizers and tuned the learning rate based on mean average precision (mAP) scores for bounding box detection and segmentation on the validation data subset. We find that SGD with momentum of 0.9 performed better than Adam. We also select a learning rate of 0.009 with training warm-up steps for 500 iterations.

Neck	Optimizer	Learning Rate	Epochs	bbox AP50	segm AP50
FPN	Adam	0.0164	2	0.238	0.205
FPN	SGD with momentum	0.02	2	0.595	0.568
FPN	SGD with momentum	0.01	2	0.599	0.574
FPN	SGD with momentum	0.009	2	0.602	0.570
CARAFE FPN	SGD with momentum	0.009	2	0.647	0.603
CARAFE FPN	SGD with momentum	0.009	5	**0.802**	**0.752**

Table 1: We experiment with using different combinations of necks, optimizers, learning rates, and epochs. The average precision (AP-50) scores were calculated for an IoU threshold of 0.5 for objects of all sizes. We find that using CARAFE FPN for the neck leads to better results compared to FPN.

The IoU (Intersection over Union) threshold is given by the ratio of the area of intersection with the area of union of the predicted bounding box and the ground truth bounding box. This is used to determine whether a predicted bounding box is a true positive (IoU > 0.5), false positive (IoU < 0.5), or false negative example. Average Precision calculates the area under the precision-recall curve for a specific IoU threshold, in this case 0.50.

$$AP@0.50 = \int_0^1 p(r)\, dr$$

Compared to the results from prior work, our model performed better for segmentation average precision scores. As expected, both the validation and the test scores decrease with an increased IoU threshold, however, they are still comparable or better than previous results.

bbox AP@0.5:0.95	bbox AP@0.5	bbox AP@0.75
0.504	0.741	0.566
segm AP@0.5:0.95	**segm AP@0.5**	**segm AP@0.75**
0.528	0.695	0.558

Table 2: Final Test Results for model. The final train accuracy for the model was 95.26%, compared to 93.26% of the Feature Pyramid Networks. Refer to Appendix subsection A2 for validation results.

Figure 5: *Left:* Raw input. *Middle Left:* Parsing Mask Annotation. *Middle Right:* Result from baseline model. *Right:* Result from our model.

For qualitative results, we decided to compare our results to a baseline model - Mask R-CNN FPN architecture with ResNet-50 backbone weights pre-trained on the COCO dataset. Visually, it is clear that our model has higher accuracy for most input images (Figure 5). However, our model performs worse on images with noisy backgrounds probably due to the limited availability of such images in our training dataset. Our model also performs better for apparel with solid color (Figure 6).

6 Conclusion and Future Work

In this paper, we propose combining the neck of the original Mask R-CNN FPN architecture with CARAFE and subsequently using a ResNet-50 backbone and mAP score for evaluation purposes. We show that our model achieves better results than previous work for a training period of 5 epochs while postulating that training for longer time periods with improve performance and accuracy.

Virtual try-on is a rapidly growing field of research with a variety of real-life applications. We find that existing try-on models have difficulty with adapting to out-of-domain situations. Segmentation mask generation is a relatively new but fast developing field that has the potential to improve the current state and use case scenarios of virtual try-on.

We envision future work to expand on this to 3D images which could potentially be useful for virtual try in 3D environments such as augmented and virtual reality. Additionally, future work can also expand the classes being identified such as specific types of headwear or accessories.

7 Contributions

All members of the team contributed to each aspect of the project. All team members contributed to the literature review, proposal, and milestone. Sanjana focused on the final report and Tony focused on the final video.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[2] Hassler Castro and Mariana Ramirez. Segmentation task for fashion and apparel. *arXiv preprint arXiv:2006.11375*, 2020.

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[4] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2287–2292, 2017.

[8] Warren Jouanneau, Aurelie Bugeau, Marc Palyart, Nicolas Papadakis, and Laurent Vezard. Where are my clothes? a multi-level approach for evaluating deep instance segmentation architectures on fashion images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3955, 2021.

[9] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Vogue: Try-on by stylegan interpolation optimization. *arXiv preprint arXiv:2101.02285*, 2021.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[11] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[12] John Martinsson and Olof Mogren. Semantic segmentation of fashion images using feature pyramid networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[15] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3007–3016, 2019.

[16] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

A Appendix

A.1 Additional Results

Figure 6: The first two rows use inputs from the dataset while the last row uses a custom input. These input images did not have corresponding parsing mask annotations. *Left:* Raw input. *Middle:* Result from baseline model. *Right:* Result from our model.

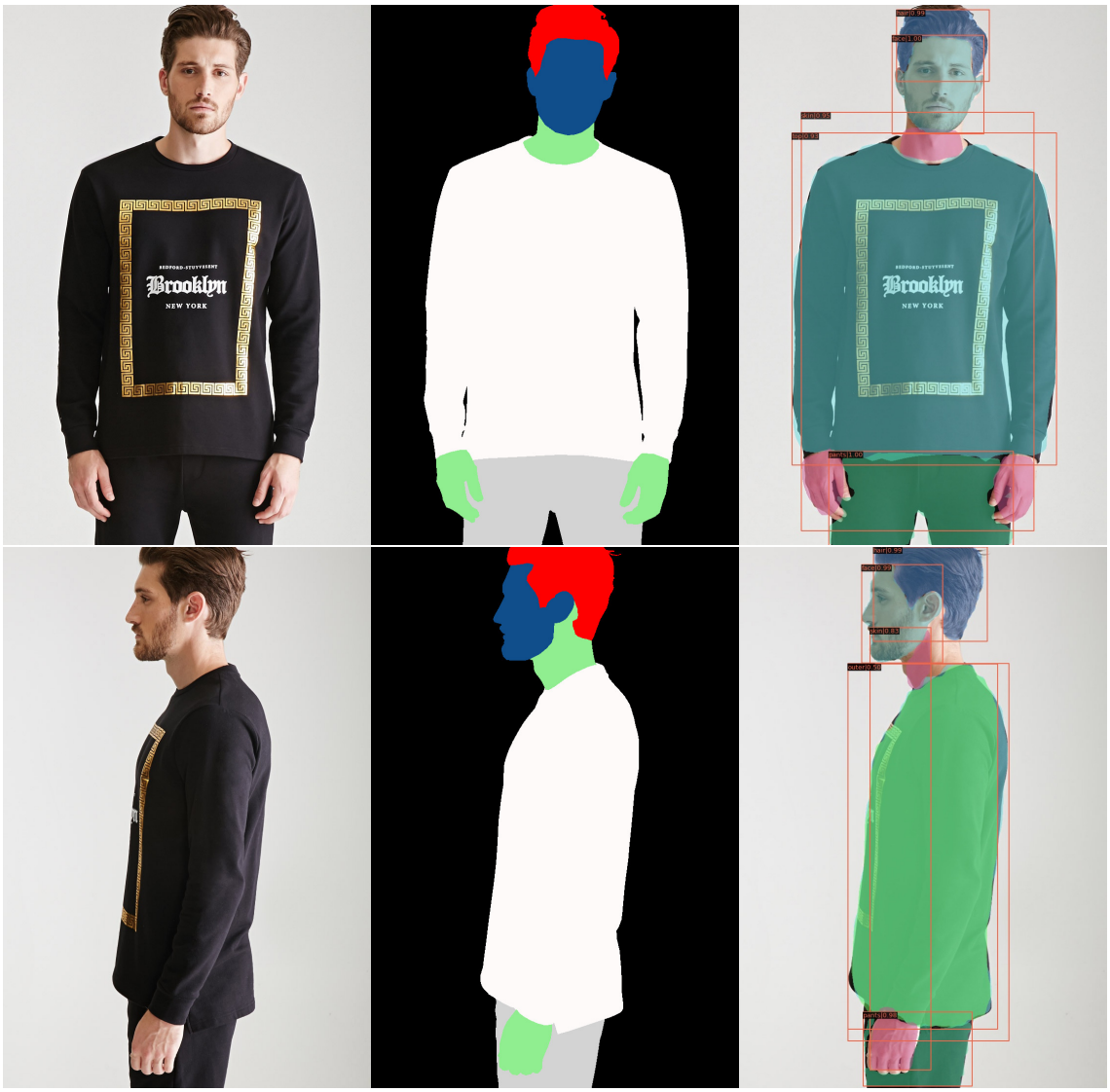

Figure 7: Additional results using inputs from dataset with corresponding parsing mask annotations available. *Left:* Raw input. *Middle:* Parsing Mask Annotation for Comparison. *Right:* Result from our model.

A.2 Validation Results

bbox AP@0.5:0.95	**bbox AP@0.5**	**bbox AP@0.75**
0.566	0.802	0.644
segm AP@0.5:0.95	**segm AP@0.5**	**segm AP@0.75**
0.566	0.752	0.615

Table 3: Final Validation Results for model.