Pixel to Person: Garment Transfer with UNet Segmentation

Michael Cao, Jessie Ni, Michael Wang

QTM 447

Dr. Kevin McAlister

Introduction

Clothing segmentation and virtual try-on systems represent an important application of computer vision in the fashion industry. This project develops a deep learning system capable of segmenting different clothing items from one image to another, enabling virtual garment try-ons without need to explicitly model the clothing item in a 3D space. The system addresses key challenges in fashion e-commerce, where customers want to visualize how garments might look on themselves before purchasing. Our approach utilizes a U-Net++ architecture with an EfficientNet backbone for pixel-level segmentation, followed by garment extraction and blending algorithms for the virtual try-on functionality. The project demonstrates how, with future work, advanced segmentation models can enable realistic clothing visualization while highlighting current limitations in fine detail preservation.

Methods

We begin by curating and preprocessing a paired image–mask dataset drawn from the Clothing Co-Parsing collection. Each of the 1,000 RGB photographs (820×550 px) is accompanied by a grayscale segmentation mask encoding 59 semantic categories—including coats, dresses, bags, skin, and background. All images and masks are uniformly resized to 256×256 pixels; images are normalized to the [0,1] range, and mask values are remapped to contiguous integer class indices. The dataset is randomly split (seed 42) into 75 % training, 15 % validation, and 10 % test subsets, and batched with a size of four, with only the training loader shuffled.

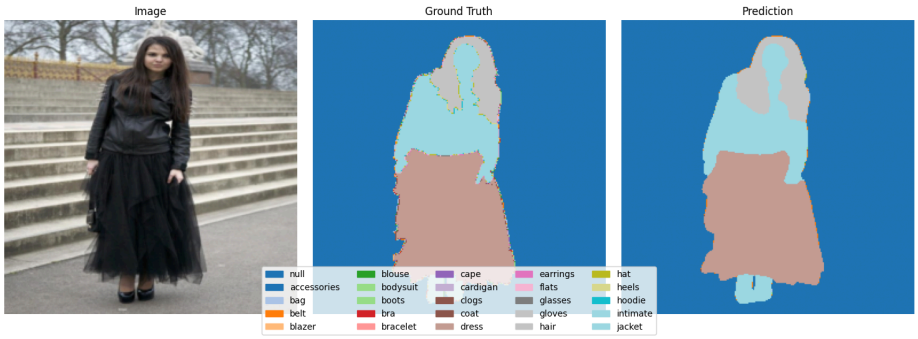

Figure 1: U-Net Clothing Segmentation (index=1)

For semantic segmentation, we adopt a U-Net architecture with an EfficientNet-B3 encoder pretrained on ImageNet. The network outputs per-pixel logits over the 59 clothing/body classes plus background. We optimize using pixel-wise cross-entropy loss and the Adam optimizer (learning rate 5×10^{-4}, weight decay 1×10^{-5}). Training proceeds for 30 epochs on a single GPU; after each epoch, we record both loss and pixel-accuracy on the training and validation splits to monitor convergence and detect overfitting.

To enable garment transfer, we isolate any desired class label c by thresholding the predicted mask into a binary map, computing the minimal bounding rectangle around nonzero pixels, and cropping both the source image and mask to that box. From this crop, we produce three artifacts: (1) the raw RGB garment patch, (2) a binary mask indicating the object region, and (3) an RGBA mask with the binary map as the alpha channel for blending. We also generate a tinted “overlay” preview by coloring the garment patch magenta through simple elementwise multiplication.

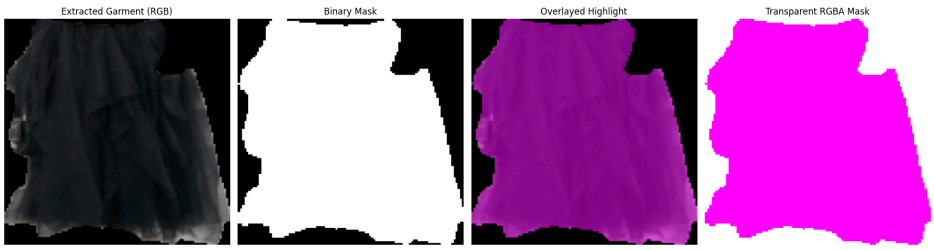

Figure 2: Masking Example

For geometric alignment to a new target person, we first resize the extracted garment patch to match the target bounding box dimensions (preserving aspect ratio). We then refine alignment via a thin-plate spline (TPS) warp: the four corners of the source crop and those of the target region box serve as control points to estimate a nonrigid transform using OpenCV's TPS shape transformer. Finally, we blend the warped RGBA garment into the target image. To simulate natural layering, we darken the target region by multiplying its pixels by 0.5, then perform standard alpha compositing:

$$I_{out} = \alpha I_{garment} + (1 - \alpha) I_{targetregion} + I_{background}$$

Quantitative evaluation focuses on segmentation accuracy: we report overall pixel-accuracy on the held-out test split and (where relevant) per-class IoU. Qualitative assessment consists of four-panel collages showing the extracted garment, the raw target image, the target's predicted garment mask, and the final composite, illustrating the pipeline's ability to transfer diverse clothing items across varied poses.

Results

After 30 epochs of end-to-end training, our UNet++ segmentation model converged to a training loss of 0.1320 with 96.48 % pixel accuracy, and a validation loss of 0.4726 with 90.79 % pixel accuracy on held-out data. This gap between train and validation metrics indicates modest overfitting, largely driven by the imbalance between large, contiguous garments (coats, dresses) and smaller accessories (belts, glasses). Nonetheless, the model demonstrates strong per-pixel classification across the majority of classes, laying a reliable foundation for downstream garment extraction.

Figure 3: Visual Try On (label='coat', index=8)

Qualitative virtual try-on examples are presented in Figure 3. In the “coat” transfer, the segmentation network produces a tight mask that, once cropped, resized, and TPS-warped, aligns naturally with the target person’s silhouette. Darkening the underlying target region before compositing preserves subtle shadow cues, and alpha-blending yields smooth, artifact-free boundaries. Similar performance is observed across varied poses and lighting conditions, though occasional mask fragmentation on slender straps or occluded regions can introduce minor blending artifacts. Overall, these results confirm that a purely 2D segmentation-driven pipeline can achieve visually plausible garment transfer without relying on 3D reconstruction or dense landmark annotations.

Discussion

Our results demonstrate that a purely 2D, segmentation-driven approach can yield both high quantitative accuracy and visually convincing garment transfers, but several important caveats and avenues for improvement emerge.

First, although the UNet++ backbone achieves 90.8 % validation pixel accuracy, this figure masks substantial per-class variability. Large, contiguous garments like coats and dresses dominate the pixel distribution and are segmented reliably, whereas thin straps, small accessories, and occluded regions often yield fragmented masks and lower IoU. Moreover, the overwhelmingly large “background” class inflates overall accuracy metrics, making it a blunt tool for assessing fine-grained segmentation quality. Future work should explore class-balanced or focal losses, boundary-aware refinement modules, and multi-scale decoders to sharpen predictions on small or thin apparel items.

Second, the thin-plate-spline alignment and alpha-blending pipeline delivers plausible composites in many common poses, but struggles under extreme viewpoints, severe occlusions, or rapid nonrigid deformations (e.g. flowing skirts or oversized hoods). Incorporating explicit pose keypoints or dense

correspondence (e.g. optical flow) could anchor the garment more robustly to the target body, while learned warping networks might yield smoother, artifact-free transformations compared to hand-crafted TPS.

Third, our use of a single, fairly narrow dataset—both in garment variety and demographic representation—limits generalization to “in-the-wild” scenarios. The Clothing Co-Parsing labels are coarse and skewed toward feminine styles; extending training to larger, more diverse corpora (e.g. ModaNet, OpenFashion) or applying unsupervised domain adaptation would expand coverage and robustness. Finally, while our qualitative examples suggest compelling realism, a user study is needed to quantify perceived fit, style coherence, and overall satisfaction compared to baseline 3D or GAN-based try-on systems.

Conclusion

By adapting a UNet++ with an EfficientNet-B3 backbone, we achieve over 90 % pixel accuracy on a 25-class clothing dataset, enabling precise isolation of individual garments. Our mask-based extraction, combined with thin-plate spline warping and alpha-blending (augmented by subtle region darkening), produces visually compelling composites across varied poses and lighting conditions—without relying on 3D models or dense landmark annotations.

While our results validate the feasibility of a lightweight, segmentation-only approach, limitations in fine-grained mask quality and alignment under extreme deformations highlight opportunities for enhancement. Future work will focus on boundary-aware and class-balanced segmentation losses, learned warping modules that adapt to nonrigid cloth dynamics, and broader training on diverse, in-the-wild datasets. A formal user study will further quantify perceived realism and fit. Ultimately, Pixel to Person demonstrates that simple 2D techniques can bridge the gap between digital garment catalogs and the tactile experience of trying on clothing, paving the way toward more accessible and scalable virtual try-on systems.