VITON: An Image-based Virtual Try-on Network
Supplemental Material

Detailed Network Structures

We illustrate the detailed network structure of our encoder-decoder generator in Figure 1, and that of our refinement network in Figure 2.

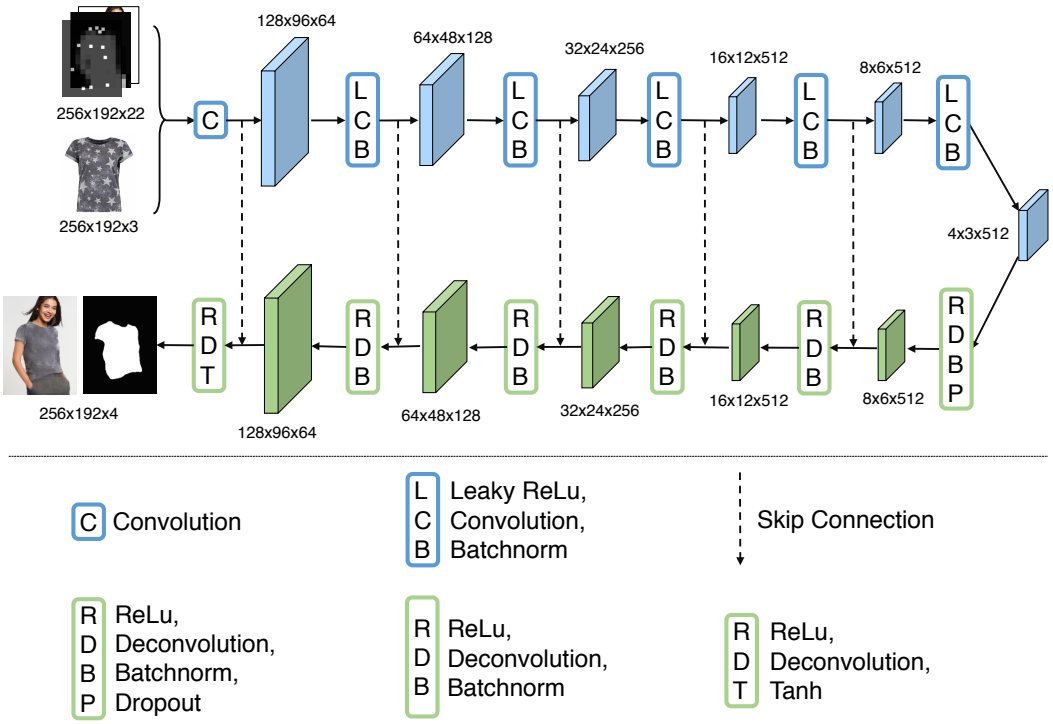

Figure 1: Network structure of our encoder-decoder generator. Blue rectangles indicate the encoding layers and green ones are the decoding layers. *Convolution* denotes 4×4 convolution with a stride of 2. The negative slope of *Leaky ReLu* is 0.2. *Deconvolution* denotes 4×4 convolution with a stride of 1/2. The *dropout* probability is set to 0.5.

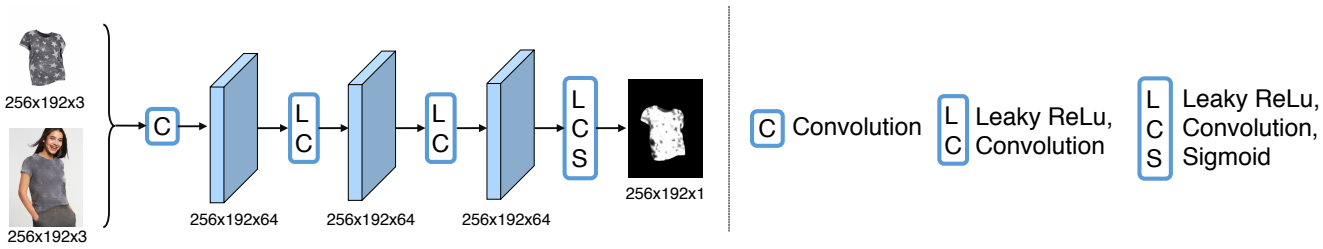

Figure 2: Network structure of our refinement network. *Convolution* denotes 3×3 convolution with a stride of 1. The negative slope of *Leaky ReLu* is 0.2.

Person Representation Analysis

To investigate the effectiveness of pose and body shape in the person representation, we remove them individually from our person representation, and train the corresponding encoder-decoder generators to compare with the generator learned by using our full representation (as in Figure 7 in the main paper). Here, we present more qualitative results and analysis.

Person Representation without Pose

In Figure 3, we show more examples where ignoring the pose representation of the person leads to unsatisfactory virtual try-on results. Although body shape representation and face/hair information are preserved, the model without capturing the person's pose fails to determine where the arms and hands of a person should occur. For example, in the first, third, fifth example in Figure 3, the hands are almost invisible without modeling pose. In the second and forth example, only given the silhouette (body shape) of a person, the model without pose produces ambiguous results when generating the forearms in the virtual try-on results, because it is possible for the forearm to be either in front/back of body or on one side of the body. Note that during these ablation results, we only show the coarse results of each model for simplicity.

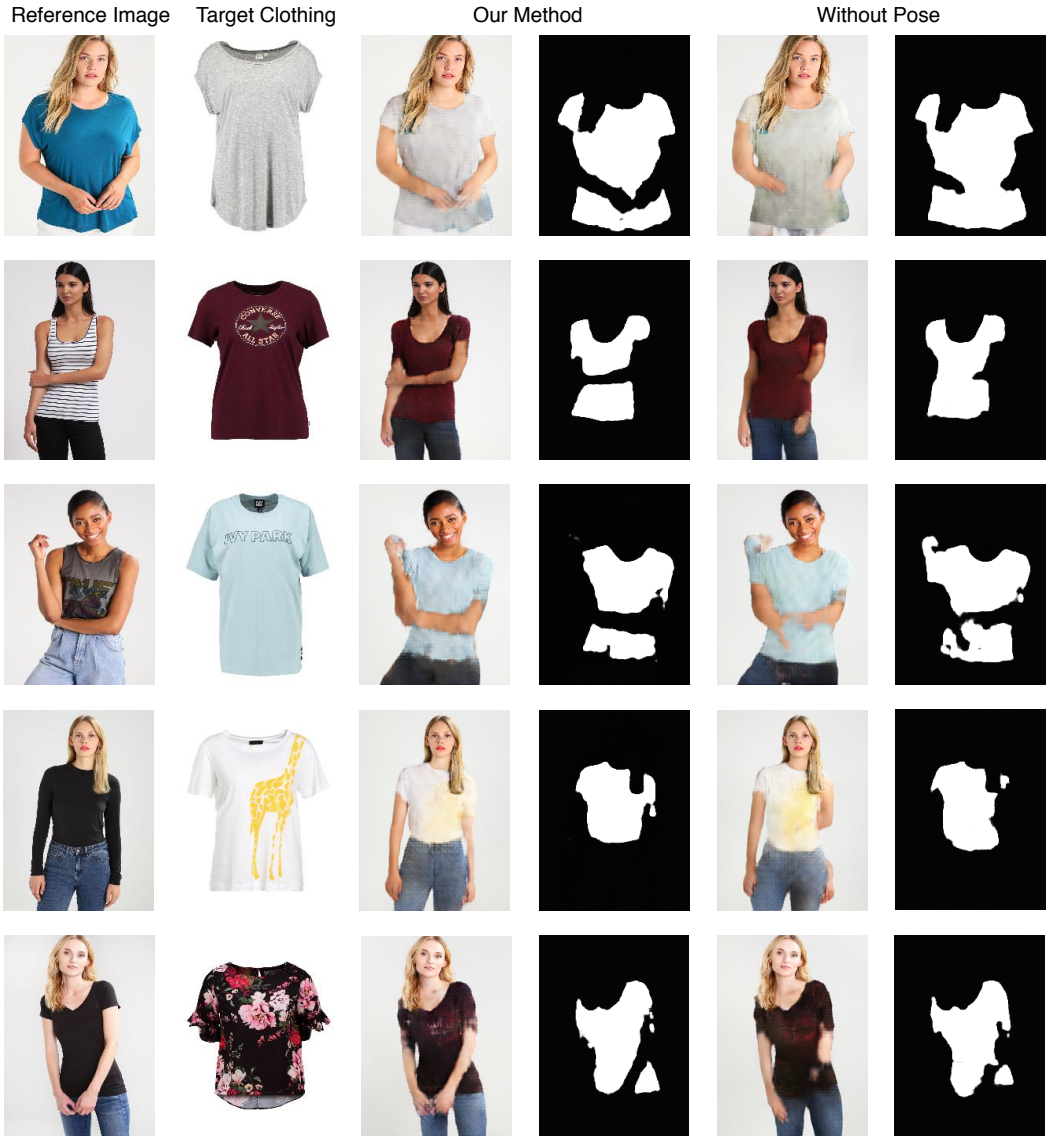

Figure 3: Comparison between the outputs (coarse result + clothing mask) of the encoder-decoder generator trained using our person presentation with the generator trained using the representation without pose.

Person Representation without Body Shape

Figure 4 demonstrates the cases where body shape is essential to generate a good synthesized result, *e.g.*, virtually trying on plus-size clothing. Interestingly, we also find that body shape can help preserving the poses of the person (second and fifth examples), which suggests that body shape and poses can benefit from each other.

Figure 4: Comparisons between the outputs (coarse result + clothing mask) of the encoder-decoder generator trained using our person presentation with the generator trained using the representation body shape.

Quantitative evaluation

Furthermore, we conducted the same user study to compare our person representation with the representation without body shape and without pose, respectively. Our method is preferred in 67.6% trials to the representation without body shape, and 77.4% trials to the representation without pose.

More Qualitative Results

In the following, we present more qualitative comparisons with alternative methods in Figure 5, and more results of our full method in Figure 6. Same conclusions can be drawn as in our main paper.

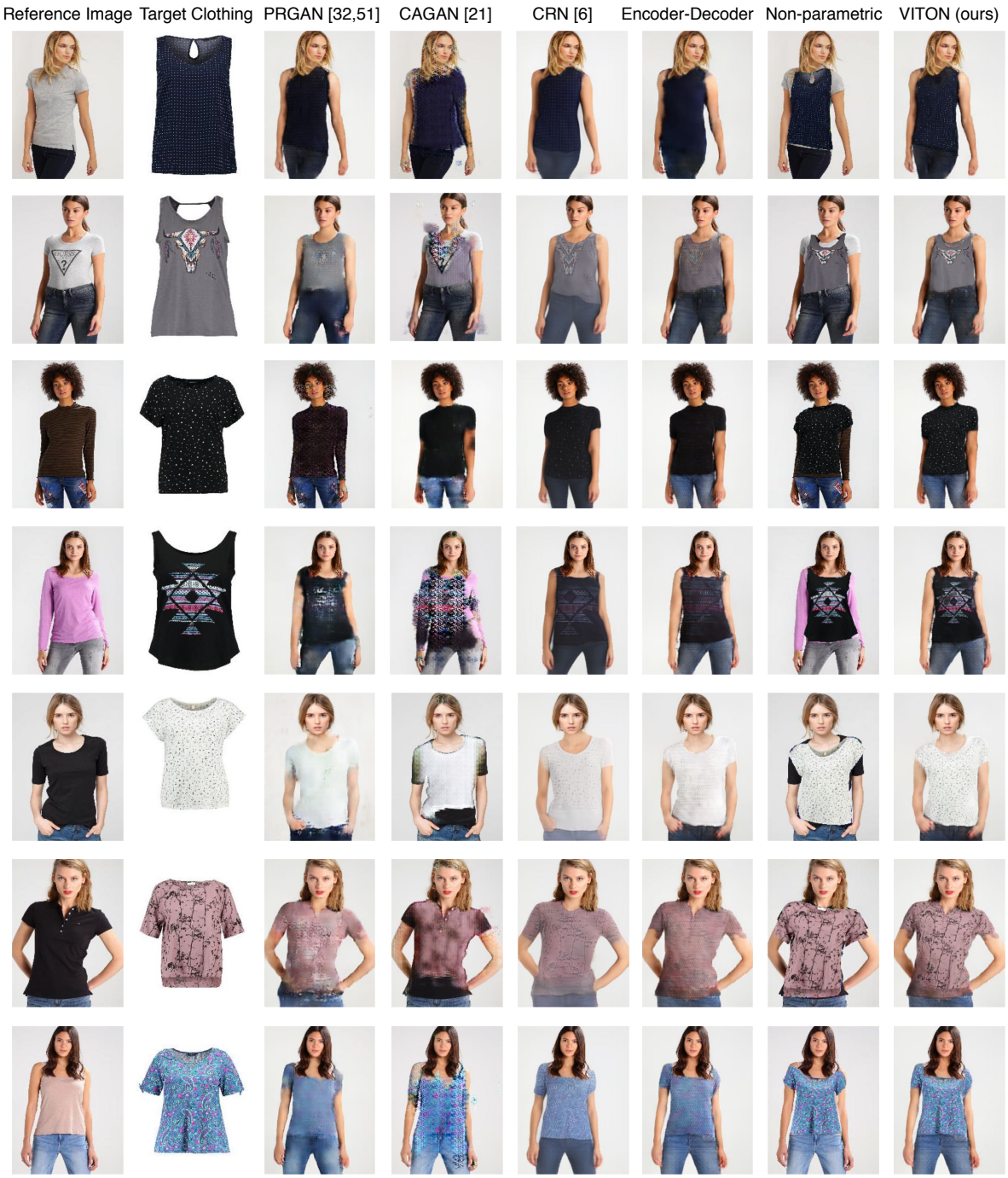

Figure 5: Comparisons of VITON with other methods. Reference numbers correspond to the ones in the main paper.

Figure 6: Virtual try-on results of VITON. Each row corresponds to the same person virtually trying on different target clothing items.

Artifacts near Neck

As mentioned in the main paper, some artifacts may be observed near the neck regions in the synthesized images. This stems from the fact that during warping (See Figure 4 in the main paper) our model warps the whole clothing foreground region as inputs to the refinement network, however regions like neck tag and inner collars should be ignored in the refined result as a result of occlusion.

One way to eliminate these artifacts is to train a segmentation model [2] and remove these regions before warping. We annotated these regions for 2,000 images in our training dataset to train a simple FCN model [2] for this purpose. At test time, this segmentation model is applied to target clothing images to remove the neck regions that should not be visible in the refined results. In this way, we can get rid of the aforementioned artifacts. Some examples are shown in Figure 7. Note that in the main paper, the results are obtained without this segmentation pre-processing step to achieve fair comparisons with alternative methods.

Figure 7: We train an FCN model to segment and remove the regions that cause artifacts during testing. All original results shown in these examples are also included in Figure 6. The artifacts near neck regions (highlighted using blue boxes) are eliminated by removing the unwanted neck areas.

One may notice that there are some discrepancies between the collar style and shape in the predicted image and the target item image. The main reason behind these artifacts is the human parser [1] does not have annotations for necks and treats neck/collar regions as background, and hence the model keeps the original collar style in the reference image. To avoid this, we can augment our human parser with [3] to correctly segment neck regions and retrain our model. Fig 8 illustrates examples of our updated model, which can handle the change in the collars.

Keep the original pants regions

A straightforward way to keep the original pants regions would be simply keeping the original pixels outside the clothing mask. However, this will cause two problems: (a) if the target item is shorter than the original one, the kept region will be too small, producing a gap between the leg and the target (see the non-parametric results in the 1st and 4th row in Figure 6 of our main paper). Therefore, we re-generate the legs for a seamless overlay. As for the face and hair regions, they are isolated since they are rarely occluded by the clothing; (b) comparisons with baselines like PRGAN and CRN might not be fair, since

Figure 8: We combine the human parser in [1] and [3] to correct segment neck regions. As a result, the inconsistency between the collar style in the target clothing and the result is properly addressed. Only the coase results are shown for simplicity.

they generate a whole image without using a mask.

For completeness, we segment the leg region and use it as an input to VITON, the original pants are preserved with some artifacts near the waist regions as shown in Figure 9.

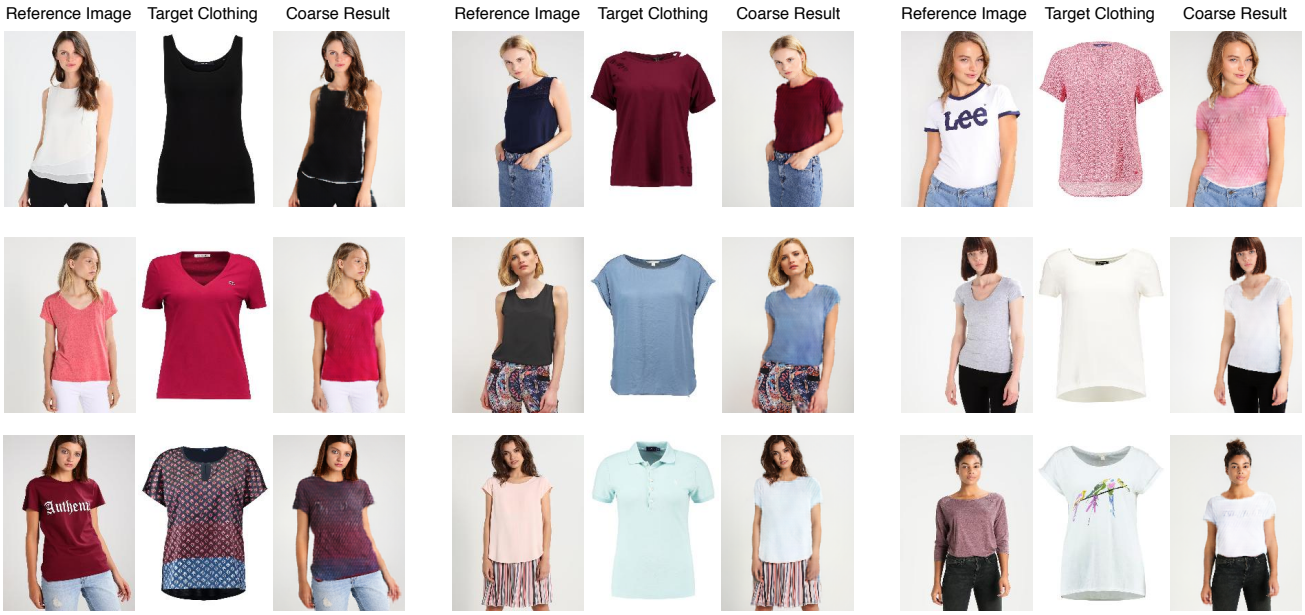

Figure 9: By treating the pants regions in a similar fashion as face/hair, VITON can keep the original pants regions. Only the coase results are shown for simplicity.

More results

In Figure 10, we show more final (refined) results of VITON with the aforementioned updates - artifacts and inconsistency near neck regions are removed, and original leg regions are kept.

Figure 10: More results of VITON with artifacts near neck region removed and original pants regions preserved. Top two rows are from Zalando dataset and the last row contains in-the-wild results from COCO.

References

[1] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 6, 7

[2] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 6

[3] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 6, 7