Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content

Appendix

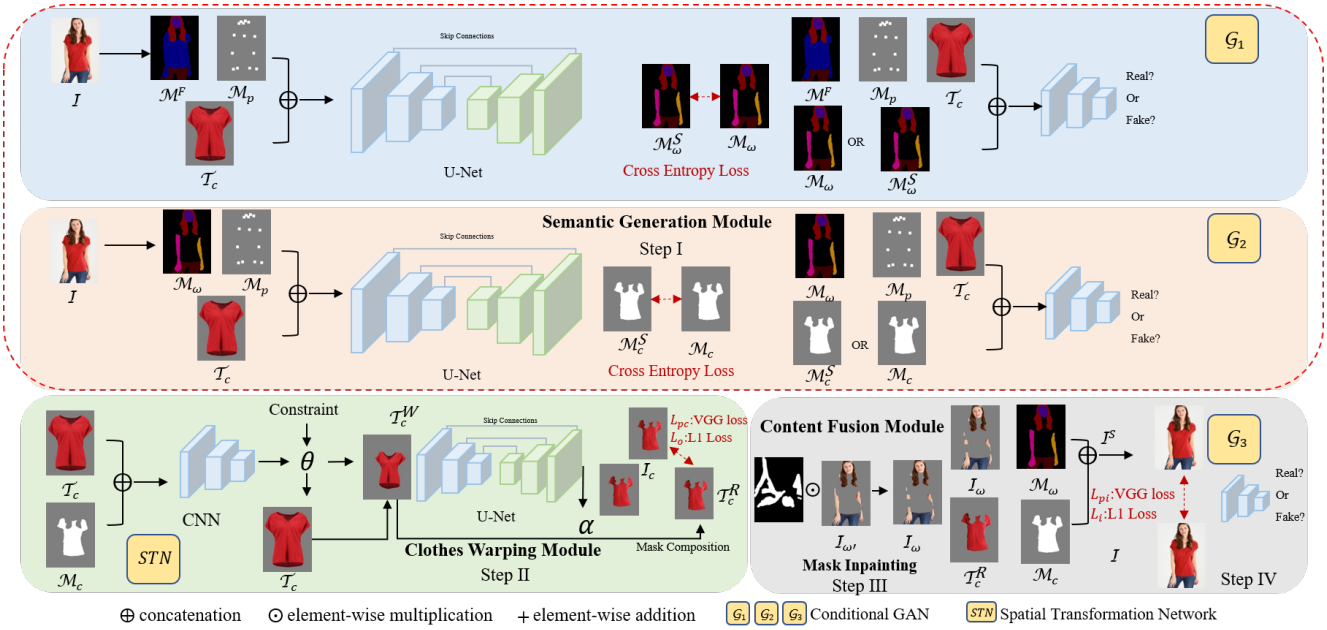

Figure 1. The overall training pipeline of ACGPN. Two major differences from the inference are that, first, the target clothes $\mathcal{T}_c$ is the in-shop version of the clothes on reference image $\mathcal{I}$, second, the Step III masks the limbs of $\mathcal{I}_{\omega'}$ (Reference person $\mathcal{I}$ removing clothes) to form $\mathcal{I}_\omega$ which is later fed into Step IV as input. The reconstruction loss for semantic masks and RGB images (*i.e.* clothing images, clothed person images) are cross entropy loss [1] and perceptual loss [2] (*i.e.* VGG loss) combined with L1 loss (pixel-wise $\mathcal{L}_1$ distance).

1. The Reconstruction loss

We can see the whole training pipeline in Fig. 1 which includes all the losses, the interactions between generator and discriminator and the compositions of inputs as well as outputs. Here we introduce the reconstruction losses for training in Step II and Step IV, which are widely used in most of the image-to-image translation tasks.

In step II shown in Fig. 1, in order to refine the $\mathcal{T}_c^W$, a U-Net is used to refine the warped clothes to fit the mask $\mathcal{M}_c$. $\mathcal{T}_c^W$ and $\mathcal{M}_c$ are fed into the refinement network and a coarse result as well as a composition mask α will be produced to perform the composition,

$$\mathcal{T}_c^R = (1-\alpha) \odot \mathcal{T}_c^W + \alpha \odot \mathcal{T}_c^R, \tag{1}$$

where $\odot$ indicates element-wise multiplication.

The loss for the refinement operation is the combination of L_1 loss and perceptual loss [2] (*i.e.* VGG loss which computes the distance of the features extracted by VGG19 [3]).

$$\mathcal{L}_o = \|\mathcal{T}_c^R - \mathcal{I}_c\|_1, \tag{2}$$

where $\mathcal{L}_o$ indicates the L_1 loss in refinement of $\mathcal{T}_c^W$, and $\mathcal{I}_c$ is the ground-truth. And the full reconstruction loss to refine $\mathcal{T}_c^W$ is

$$\mathcal{L}_{rc} = \lambda_o \mathcal{L}_o + \lambda_{pc} \mathcal{L}_{pc}, \tag{3}$$

where $\mathcal{L}_{rc}$ indicates the reconstruction loss of $\mathcal{T}_c^R$, and $\mathcal{L}_{pc}$ is the perceptual loss between $\mathcal{T}_c^R$ and $\mathcal{I}_c$. λ_o and λ_{pc} are weights of each loss.

In Step IV shown in Fig. 1, the reconstruction loss for the inpainting based fusion GAN is also the combination of

Figure 2. Extensive try-on results with three difficulty levels. We can see that ACGPN performs equally well for long-sleeve clothes to short-sleeve reference image (fifth row in the middle) and short-sleeve clothes to long-sleeve reference image (fourth row on the left), which demonstrates the generality of our method.

L_1 loss and perceptual loss between the synthesized image and the ground-truth image. The L_1 loss $\mathcal{L}_i$ is formulated as

$$\mathcal{L}_i = \|\mathcal{I}^S - \mathcal{I}\|_1. \tag{4}$$

The full reconstruction loss is

$$\mathcal{L}_{ri} = \lambda_i \mathcal{L}_i + \lambda_{pi} \mathcal{L}_{pi}, \tag{5}$$

where $\mathcal{L}_{pi}$ is the perceptual loss between $\mathcal{I}^S$ and its ground-truth $\mathcal{I}$, and $\mathcal{L}_{ri}$ indicates the full reconstruction loss of $\mathcal{I}^S$. λ_i and λ_{pi} are weights of each loss.

The weights of each loss are given as $\mathcal{L}_o = \mathcal{L}_i = 1$ and $\mathcal{L}_{pc} = \mathcal{L}_{pi} = 10$.

2. More Try-on Results

We here show more try-on results produced by ACGPN in Fig. 2 and Fig. 3. For more results, **an example video is provided in youtube**: `https://www.youtube.com/watch?v=h-QWM92VLA0`.

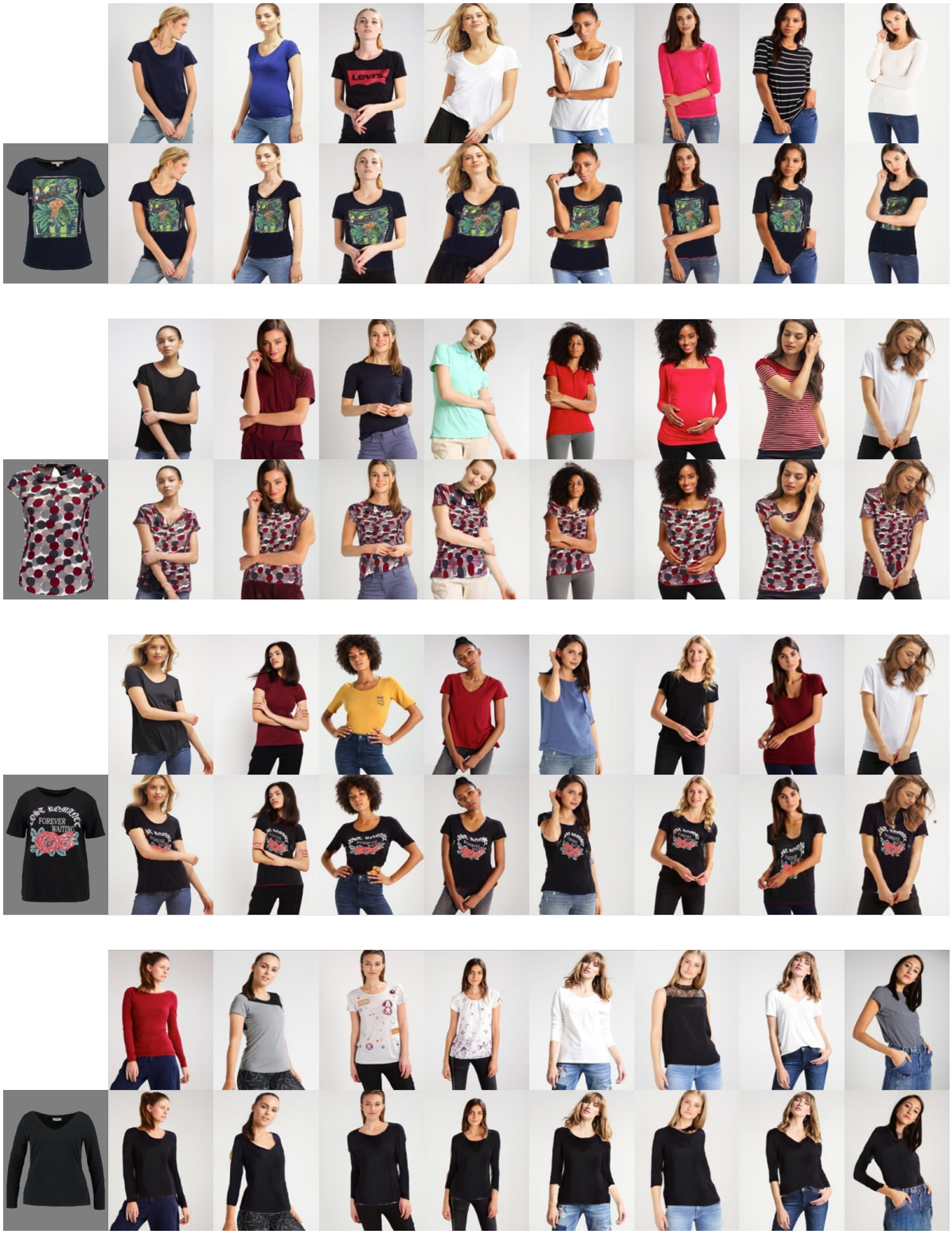

Figure 3. Extensive try-on results with four reference people. ACGPN perform robustly with various poses including occlusions and cross-arms. Artifacts are reduced to the minimum.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[2] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.