Algorithm 1 PB-AFN

Input: Training dataset; Human parsing results and pose estimations of person images;

Output: The warping module $\psi_{B_w}(\cdot)$ and the generative module $\psi_{B_g}(\cdot)$ in PB-AFN.;

Randomly initialize the CNN model $\psi_{B_w}(\cdot)$ and $\psi_{B_g}(\cdot)$;
Pre-process raw data;
for $T = 1, 2, 3, ..., max\_epoh$ **do**
  for *each mini-batch* **do**
    Obtain the person representations p^* of the person image I in the input batch with the clothing region masked;
    Predict the appearance flows $u_f = \psi_{B_w}(p^*, I_c)$ between the person representations p^* and the clothes images I_c in the input batch;
    Warp the clothes images I_c to u_w via the appearance flows u_f;
    Synthesize the images $u_I = \psi_{B_g}(p^*, u_w)$ with the person representations p^* and the warped clothes u_w;
    Calculate the L1 loss $\mathcal{L}_{B_l}$ (Eq. 3), the perceptual loss $\mathcal{L}_{B_p}$ (Eq. 4), the second-order smooth constraint loss $\mathcal{L}_{B_{sec}}$ (Eq. 1) and add these losses to $\mathcal{L}_B$ (Eq .2);
    Update $\psi_{B_w}(\cdot)$ and $\psi_{B_g}(\cdot)$ by minimizing $\mathcal{L}_B$;
  end for
end for

6. Pseudocode

Our method contains a parser-based network PB-AFN and a parser-free network PF-AFN. We first introduce the training process of the parser-based network PB-AFN in Alg. 1 following the existing training pipeline [30, 8, 18, 32], which masks the clothing region of the person image and reconstructs the person image with the corresponding in-shop clothes and the person representations. After training PB-AFN, we use the generated fake images as the input of PF-AFN, which is supervised by real images, and further distill the appearance flows to find accurate correspondences between the target clothes and the person image. The complete training process of PF-AFN is presented in Alg. 2. During inference, only a target clothes image and a reference person image will be given to PF-AFN to generate the try-on results.

7. More Try-on Results

Results of VITON-HD Different from VITON [9] dataset with low resolution 256×192, VITON-HD with the image resolution 512×384 poses a severe challenge for the model to retain the details of the target clothes and generate high-resolution try-on images with satisfactory visual quality. The results on VITON-HD is shown in Fig. 8. Since VITON-HD hasn't been tackled before by previous methods [30, 8, 18, 32], we up-sample their low-resolution results to 512×384. Compared with the parser-based methods, our PF-AFN is able to warp the target clothes to the reference person seamlessly where the logo and the embroidery are retained without being distorted, preserve non-target clothes such as trousers and skirt, and keep photo-realistic body details like hands and fingers.

Results of MPV We show more try-on results on MPV [5] dataset in Fig. 9. Compared with WUTON [13] which is also a parser-free method, our PF-AFN generates try-on images with much better visual quality. PF-AFN yields accurate warping for the target clothes, preserves the details of body parts (*i.e.* hands) even in complex postures and retains the characteristics of the target clothes (*i.e.* color, collar and sleeve).

Results of VITON We here show extensive try-on results produced by our PF-AFN on VITON [9] dataset. Fig. 10 shows four reference persons with different target clothes and Fig. 11 shows four target clothes to different persons. Our PF-AFN can adapt to different kinds of clothes with satisfactory performance. It also performs robustly with various poses and generates high-quality results.

Algorithm 2 PF-AFN

Input: Training dataset; Human parsing results and pose estimations of person images; $\psi_{B_w}(\cdot)$ and $\psi_{B_g}(\cdot)$ in PB-AFN;

Output: The warping module $\psi_{F_w}(\cdot)$ and the generative module $\psi_{F_g}(\cdot)$ in PF-AFN;

Randomly initialize the CNN model $\psi_{F_w}(\cdot)$ and $\psi_{F_g}(\cdot)$;
Pre-process raw data;
for $T = 1, 2, 3, ..., max\_epoh$ **do**
  for *each mini-batch* **do**
    `//Generate fake images with PB-AFN`
    Randomly select clothes images $I_{\tilde{c}}$, that are different from the clothes images I_c in the input batch;
    Obtain the person representations p^* of the person image I in the input batch with the clothing region masked;
    Predict the appearance flows $u_{\tilde{f}} = \psi_{B_w}(p^*, I_{\tilde{c}})$ between the person representations p^* and the selected clothes images $I_{\tilde{c}}$;
    Warp the selected clothes images $I_{\tilde{c}}$ to $u_{\tilde{w}}$ via the appearance flows $u_{\tilde{f}}$;
    Synthesize the images $u_{\tilde{I}} = \psi_{B_g}(p^*, u_{\tilde{w}})$ with the person representations p^* and the warped clothes $u_{\tilde{w}}$ ($u_{\tilde{I}}$ are the fake images of person in the input batch changing clothes);
    `//Training PF-AFN with the fake images`
    Predict the appearance flows $s_f = \psi_{F_w}(u_{\tilde{I}}, I_c)$ between the fake images $u_{\tilde{I}}$ and the clothes images I_c;
    Warp the clothes image I_c to s_w via the appearance flows s_f;
    Synthesize the images $s_I = \psi_{F_g}(u_{\tilde{I}}, s_w)$ with the fake images $u_{\tilde{I}}$ and the warped clothes s_w;
    Calculate the L1 loss $\mathcal{L}_{F_l}$ (Eq. 3), the perceptual loss $\mathcal{L}_{F_p}$ (Eq. 4), the second-order smooth constraint loss $\mathcal{L}_{F_{sec}}$ (Eq. 1) and add these losses to $\mathcal{L}_F$ (Eq .2);
    `//Distill the appearance flows`
    Predict the appearance flows $u_f = \psi_{B_w}(p^*, I_c)$ between the person representations p^* and the clothes imageI_c;
    Calculate the adjustable knowledge distillation loss $\mathcal{L}_{F_{kd}}$ (Eq.5, Eq.6, Eq.7 and Eq.8);
    Add $\mathcal{L}_{F_{kd}}$ to $\mathcal{L}_F$ as the final loss $\mathcal{L}_F^{all}$;
    Update $\psi_{F_w}(\cdot)$ and $\psi_{F_g}(\cdot)$ by minimizing $\mathcal{L}_F^{all}$;
  end for
end for

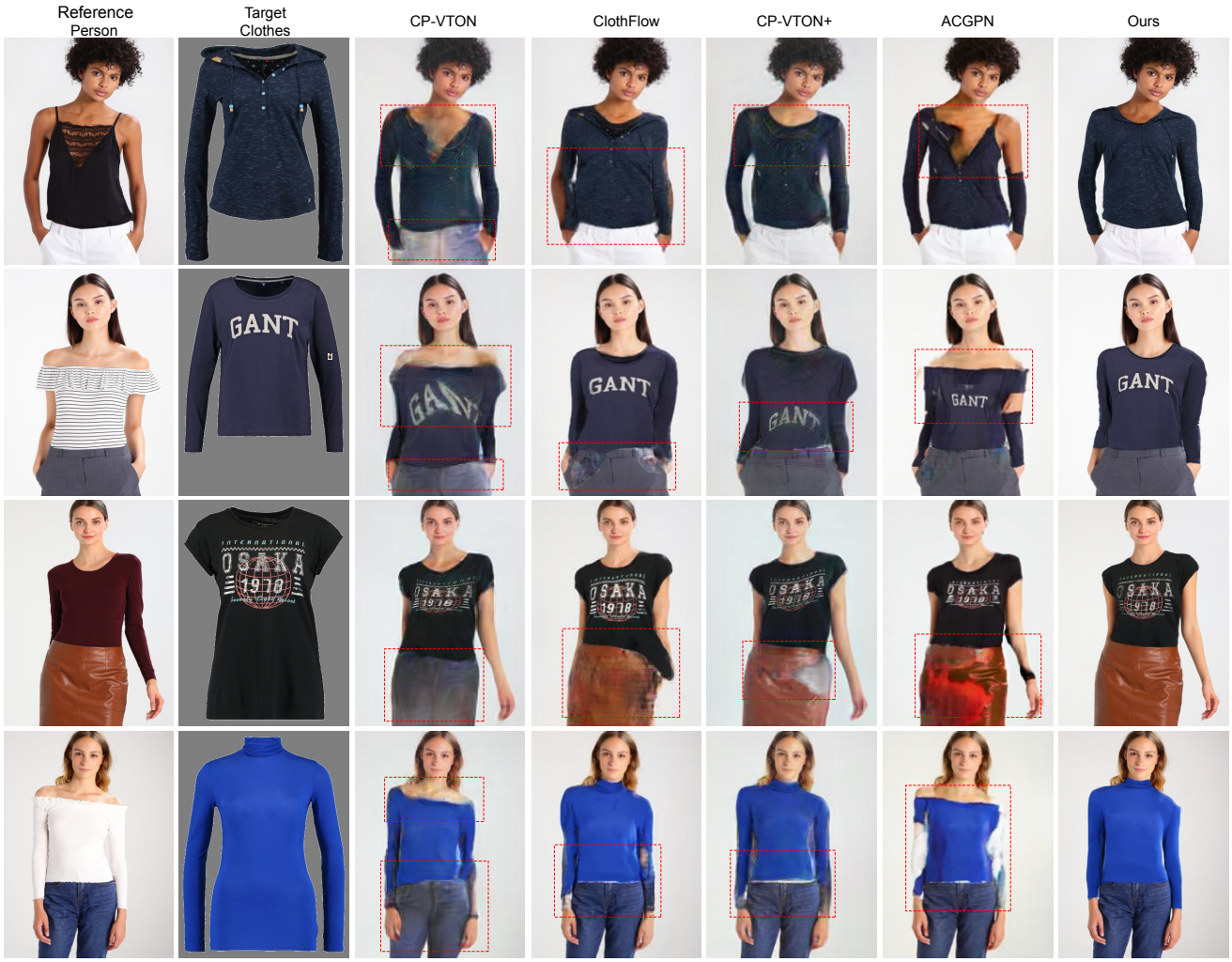

Figure 8. Visual comparison on VITON-HD dataset with the image resolution 512×384. Our model is capable of generating high-resolution try-on images, where details of the target clothes (*i.e.* collar, logo and embroidery), non-target clothes (*i.e.* trousers and skirt) and body parts (*i.e.* hands and fingers) are very well retained, compared with the recent proposed parser-based methods [30, 8, 18, 32].

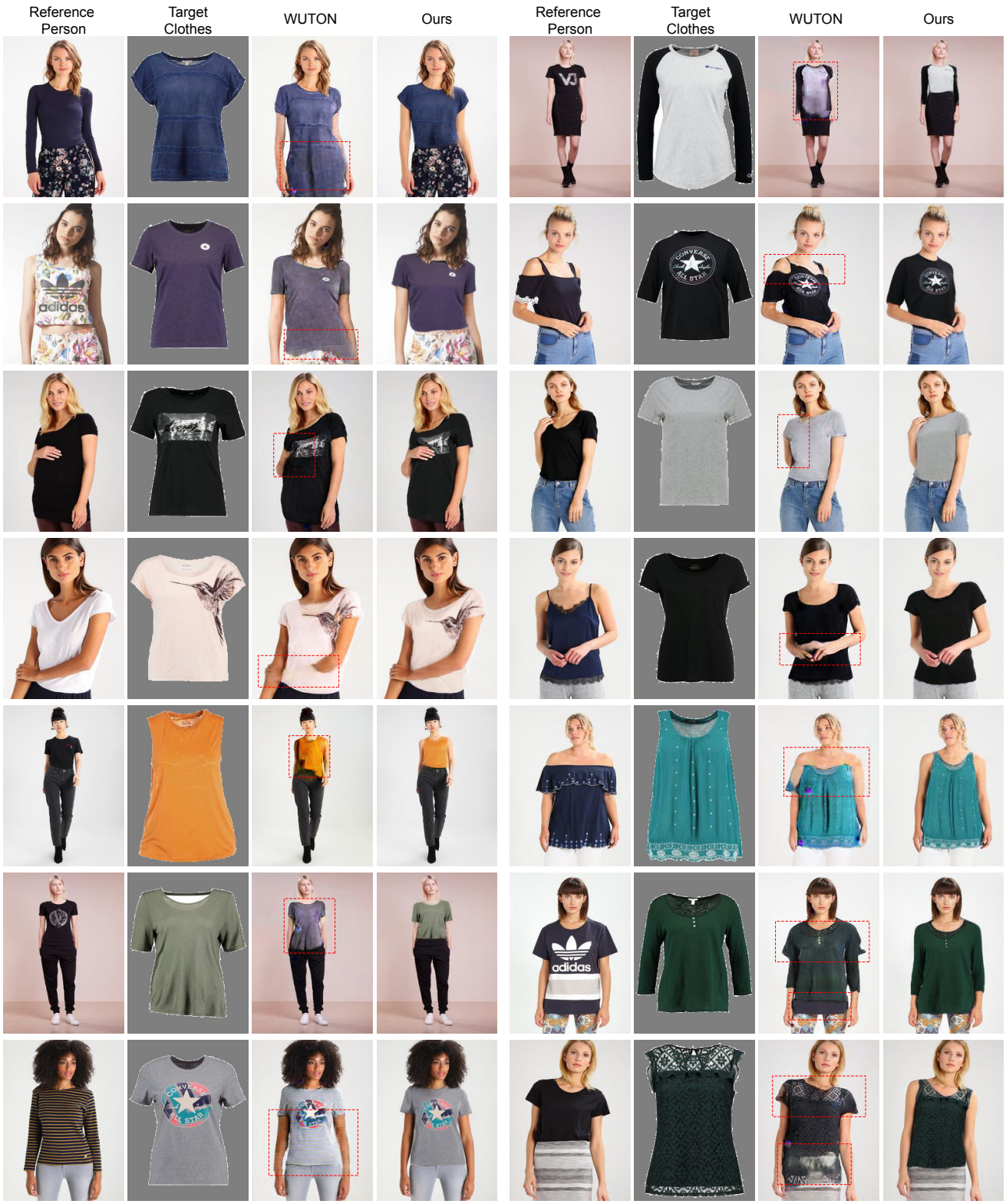

Figure 9. Extensive visual comparison on MPV dataset with parser-free inputs. Compared with WUTON [13], our PF-AFN generates try-on images with much better visual quality, which warps the target clothes to the corresponding region on the person accurately, preserves the details of body parts (*i.e.* hands) even in complicated poses and retains the characteristics of the target clothes (*i.e.* color, collar and sleeve).

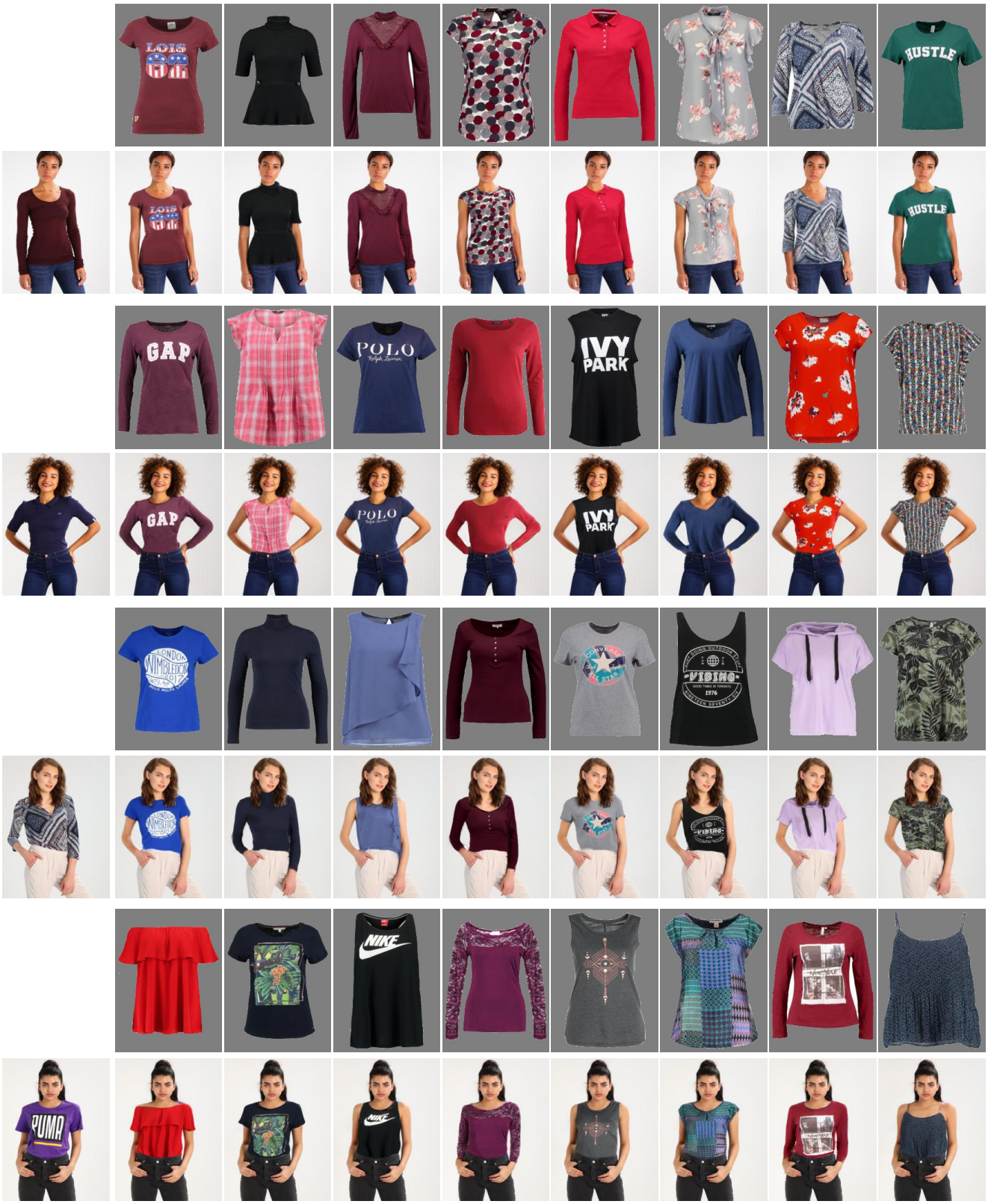

Figure 10. Extensive try-on results of four reference persons with different target clothes on VITON dataset. Our PF-AFN achieves satisfactory performance with various clothes such as short sleeve top, long sleeve top, vest, sling, high-necked top and polo top, where the characteristics of the target clothes are retained without being interfered by the original clothes on the reference person. Also, the texture, logo and embroidery on the target clothes are preserved without being distorted.

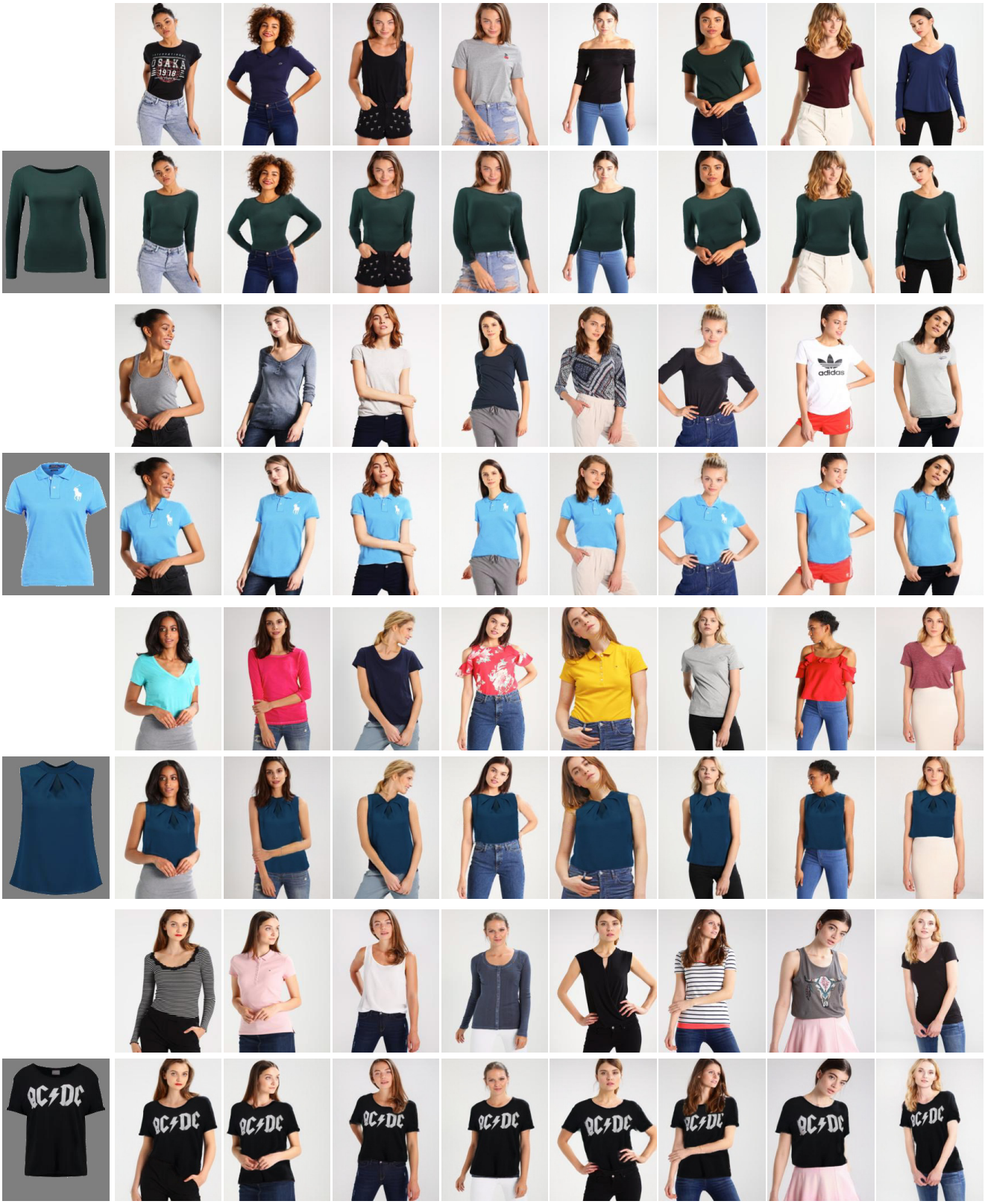

Figure 11. Extensive try-on results of four target clothes to different reference persons on VITON dataset. Our PF-AFN performs robustly with various poses including arms akimbo, two hands blocking in front of the body, cross-arms and one arm behind the back, where large misalignment and deformation occurs between the target clothes and the reference person. Our PF-AFN warps the target clothes to different persons accurately and generates high-quality results.