Algorithm 1 DCTON

Input: Training dataset VITON [15] or VITON-HD;
Output: Highly-realistic Try-on Results;

Randomly initialize models $\psi_{D1}(\cdot)$ and $\psi_{D2}(\cdot)$;
Pre-process raw data for data augmentation;
for $T = 1, 2, 3, ..., max\_epoh$ **do**
  for *each mini-batch* **do**
    //CNN1
    Predict the clothes and skin region by $M_1^{\text{clothes}}, M_1^{\text{skin}} = \psi_{m1}(D, C_2)$, where $\psi_{m1}(\cdot)$ is the function of MPN;
    Feed STN $\psi_{s1}(\cdot)$ with the clothes C_2 and mask, and predicted clothes region M_1^{skin} to obtain C_2^{warp};
    Segment the skin region via $S_1 = M_1^{\text{skin}} \odot I_1$, where M_1^{skin} is the skin mask of I_1;
    Capture the pyramid features of C_2^{warp}, S_1 and I_1, and concatenate them at each feature level;
    Generate the try-on results I_2 via the CNN1;

    //CNN2
    Predict the clothes and skin region by $M_2^{\text{clothes}}, M_2^{\text{skin}} = \psi_{m2}(D, C_1)$, where $\psi_{m2}(\cdot)$ is the function of MPN;
    Feed STN $\psi_{s2}(\cdot)$ with the clothes C_1 and mask, and predicted clothes regionM_2^{skin} to obtain C_1^{warp};
    Segment the skin region via $S_2 = M_2^{\text{skin}} \odot I_2$, where M_2^{skin} is the skin mask of I_2;
    Capture the pyramid features of C_1^{warp}, S_2 and I_2, and concatenate them at each feature level;
    Generate the try-on results $\overleftarrow{I}_1$ via CNN2;

    //Update the Training Network
    Calculate the cycle consistency loss via $\mathcal{L}_{cyc} = ||\overleftarrow{I}_1 - I_1||_1 + ||\overleftarrow{S}_1 - S_1||_1$;
    Calculate the other losses, i.e., $\mathcal{L}_{adv}, \mathcal{L}_{pre}$ and $\mathcal{L}_{vgg}$, via Eq.4, Eq.6 and Eq.7, respectively;
    Update the DCTON networks $\psi_{D1}(\cdot)$ and $\psi_{D2}(\cdot)$ by minimizing the overall loss $\mathcal{L}_{all}$ in Eq.8;
  end for
end for

6. Pesudo Code

In Algorithms 1, we first present the training procedure of the overall DCTON. The detailed processes of disentanglement and cycle consistency training will also be explained. Besides, since our STN module is pre-trained independently to help clothes warping, we here give an additional explanation on the training process of STN as well as the process of incorporating the proposed regularization term in Algorithms 1.

Algorithm 2 STN with Regularization Term

Input: Training dataset VITON [15] or VITON-HD;
Output: Pre-trained Spatial Transformer Network $\psi_s(\cdot)$;

Randomly Initialize the CNN model $\psi_s(\cdot)$;
Obtain the input triplet data pairs (i.e., the in-shop clothes C, its corresponding mask M, the warped mask M^{clothes});
Pre-process raw data for data augmentation;
for $T = 1, 2, 3, ..., max\_epoh$ **do**
  for *each mini-batch* **do**
    Select n (batch size) training triplet pairs;
    //Transformation Matrix
    Obtain the transformation matrix T for each training pair by $\psi_p(C, M, M^{\text{clothes}})$;
    Obtain the warped clothes C^{raw} via TPS;
    Obtain the processed clothes C^{fake} and the alpha blending mask M^{α} via $\psi_r(\cdot)$;
    Perform the alpha blending, $C^{\text{warp}} = M^{\alpha} \odot C^{\text{fake}} + (1 - M^{\alpha}) \odot C^{\text{raw}}$;
    Calculate the ℓ_1 distance $\mathcal{L}_a$ on warped clothes via Eq.1, and the regularization term R_b via Eq.2;
    Update the STN network $\psi_s(\cdot)$ by minimizing the loss $\mathcal{L}_{STN} = \mathcal{L}_a + R_b$;
  end for
end for

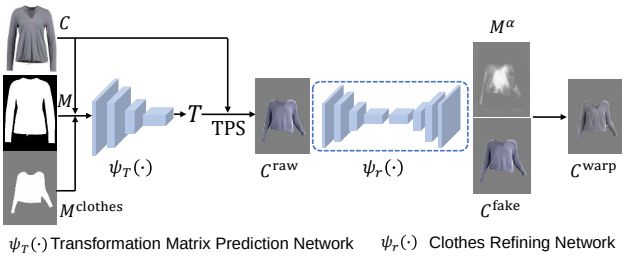

Figure 1. The pipeline of our pre-trained STN $\psi_s(\cdot)$. The overall STN consists of a transformation matrix prediction network $\psi_T(\cdot)$ and a clothes refining network $\psi_r(\cdot)$. It is capable of outputting the warped clothes with the input triplet data pairs (i.e., the in-shop clothes C, its corresponding mask M, the warped mask M^{clothes}).

7. Visual Results

To further validate the effectiveness and robustness of the proposed try-on model, we here present extra try-on results on VITON [15] and VITON-HD dataset. The extensive visual comparisons in Fig. 2 further show our advantages over other prior arts in generating highly-realistic try-on results. Fig. 3 shows the try-on results of the same person in different clothes, while Fig. 4 gives the results of the same clothes worn by different people. Results in Fig. 3 and Fig. 4 validate that DCTON is robust in various try-on scenarios. We also conduct the similar experiments on the datasets with higher resolution (i.e., VITON-HD), which is shown in Fig. 5, Fig. 6 and Fig 7.

Figure 2. Visual comparisons on VITON test set. Compared with the previous works (e.g., [18,40]), DCTON is capable of generating the highly-realistic try-on results, while preserving clothing textures, retaining clothing characteristics and adapitvely generating the human skin at the same time.

Figure 3. Extensive try-on results of DCTON on VITON dataset [15]. Each group of the experiments shows the results of the same person in different clothes. We choose the reference persons and the target clothes randomly. DCTON performs robustly with various kinds of clothes. The clothes characteristics are retained to the greatest extent.

Figure 4. Extensive try-on results of DCTON on VITON dataset [15]. Each group of the experiments shows the results of the same clothes worn by various persons, which validates that DCTON performs robustly with different person poses. DCTON is able to warp the clothes properly according to the specific pose.

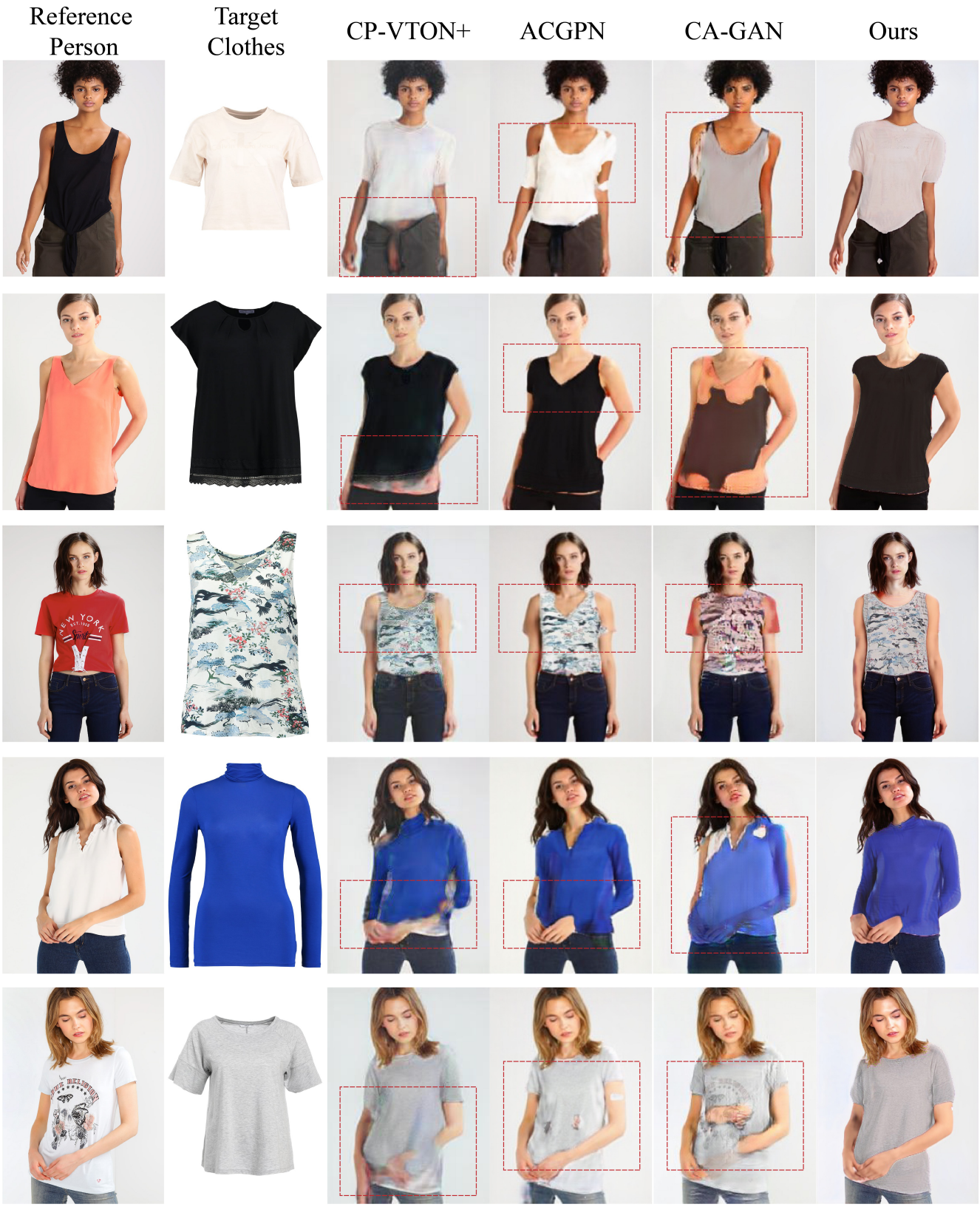

Figure 5. Visual comparisons on VITON-HD test set to show the further advantages of our DCTON. The results validate that DCTON holds the abilities to generate photo-realistic try-on images with the higher resolution of 512x384.

Figure 6. Extensive try-on results of DCTON on VITON-HD dataset. Although there are huge differences in the target clothes, DCTON is capable of fitting different types of clothes to the same reference person properly.

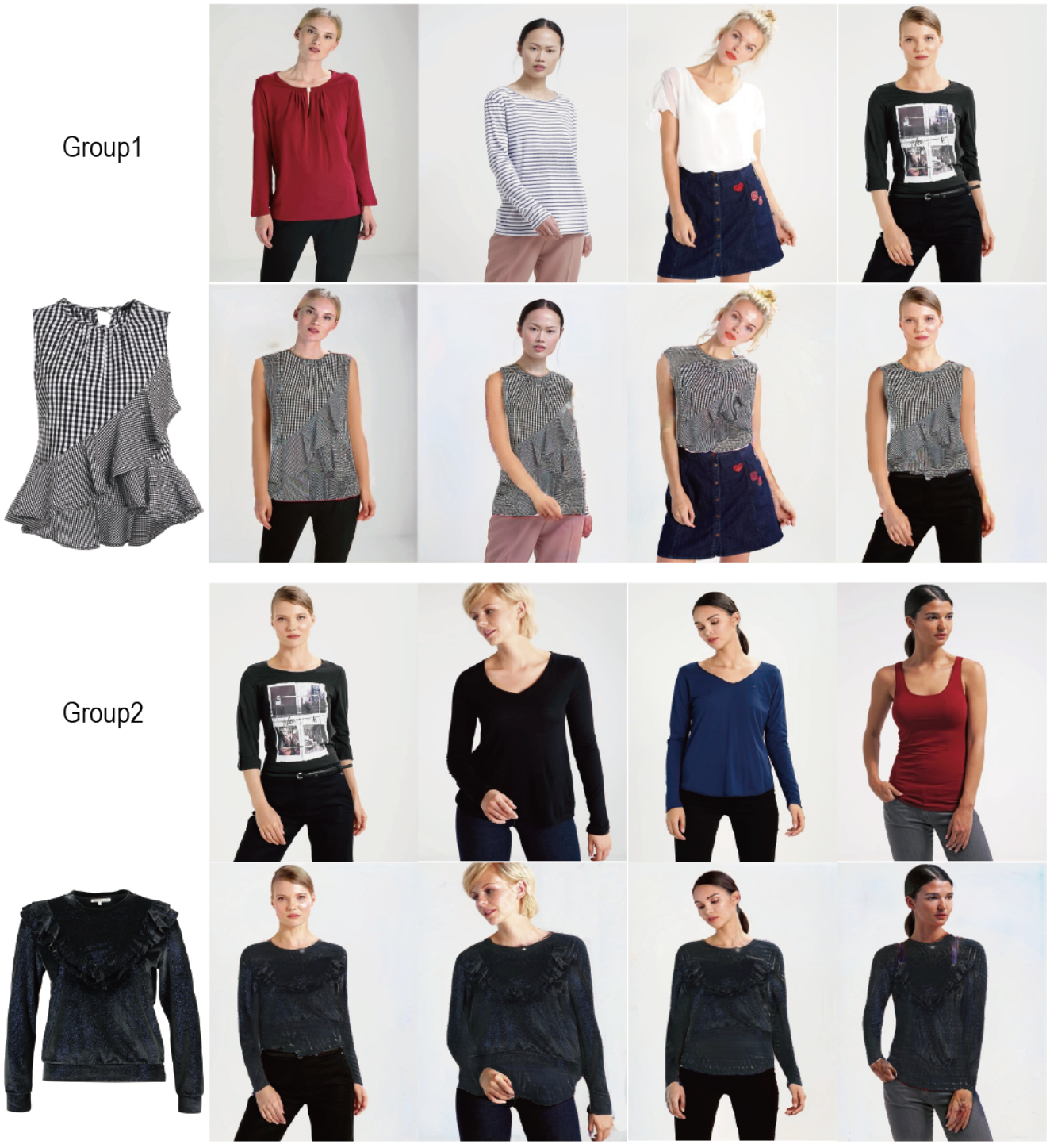

Figure 7. Extensive try-on results of DCTON on VITON-HD dataset. The results shows that our proposed DCTON holds the abilities to robustly handle with different poses. Artifacts on the generated images are reduced to the minimum.