# Robust-MVTON: Learning Cross-Pose Feature Alignment and Fusion for Robust Multi-View Virtual Try-On

## Supplementary Material

In this supplementary material, we provide more implementation details and visual results of our proposed method.

**Latent Diffusion Model** Following the general implementation using the latent diffusion model (LDM) for image generation, we adopt a pre-trained Variational Autoencoder (VAE) [12] to map image inputs $x$ into a lower-dimensional latent code $z$, wherein the encoder of VAE is $\varepsilon$, and $z = \varepsilon(x)$. The UNet-based [46] prediction model$E_{\theta}(o, t)$ is trained and inferred at a reduced computational cost. Following the Denoising Diffusion Probabilistic Model (DDPM) [23], during training, the image latent $z$ is diffused in $t$ timesteps to produce noise latent $z_t$. The optimization process is defined as the following expression.

$$Loss_{diffusion} = \mathbb{E}_{\varepsilon(x),\epsilon,t}[\| \epsilon - \epsilon_{\theta}(z_t, t) \|_2^2]. \qquad (7)$$

**Details of Implementaion** We re-implemented the coarse-to-fine latent diffusion model [34] based on the pre-trained StableDiffusion v1.5 [45, 49] from the inpainting version. We train our method (i.e., Robust-MVTON) using the Adam [28] optimizer with 8 NVIDIA A100-80G GPUs. The batch size is 176 and the base learning rate is $1e{-}4$. Besides, the learning rate undergoes a linear warm-up during the first 1,000 steps and is multiplied by 0.1 at 50 epochs.

**Details of User Study** In our experiments, we enlisted 50 professionals with expertise in image generation to conduct a user study on 30 sets of experimental results. Each set included the outcomes from Robust-MVTON and various comparative experiments, comprising 2-4 different views. Participants were asked to select the results that best preserved the clothing shapes, textures, and model features from among all the options.

**More Visual Results** We provide more visual results of our Robust-VTON, as illustrated from Fig. 10 to Fig. 13.

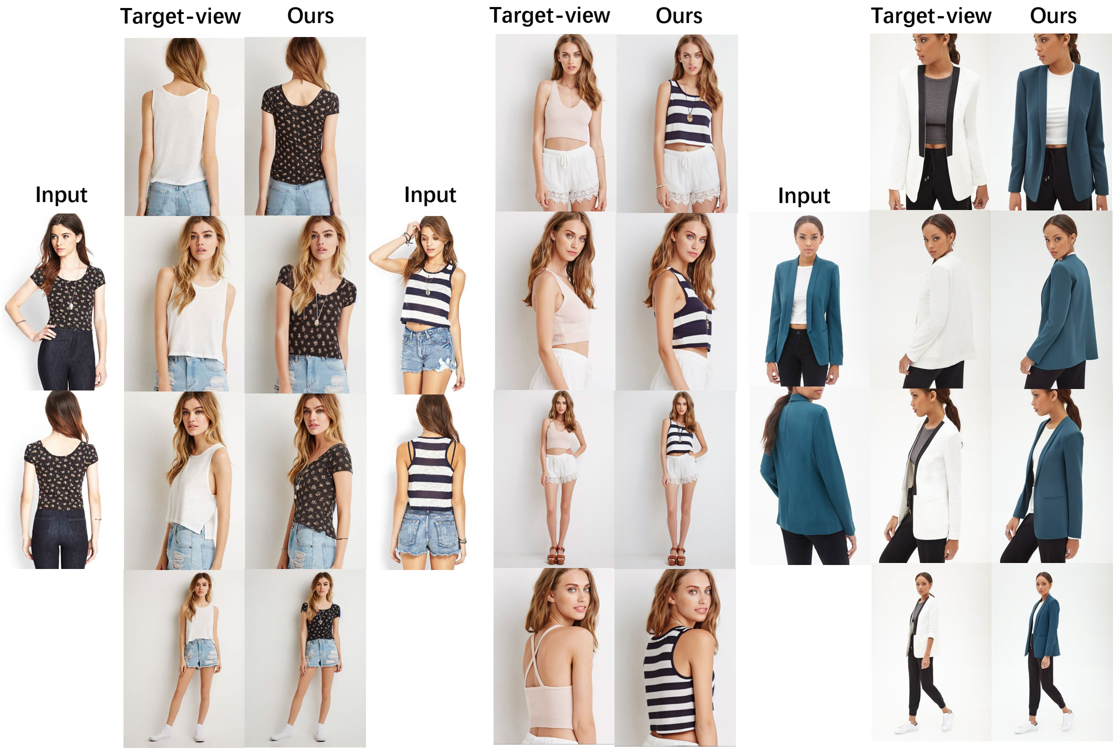

Figure 10. The qualitative results of Robust-MVTON

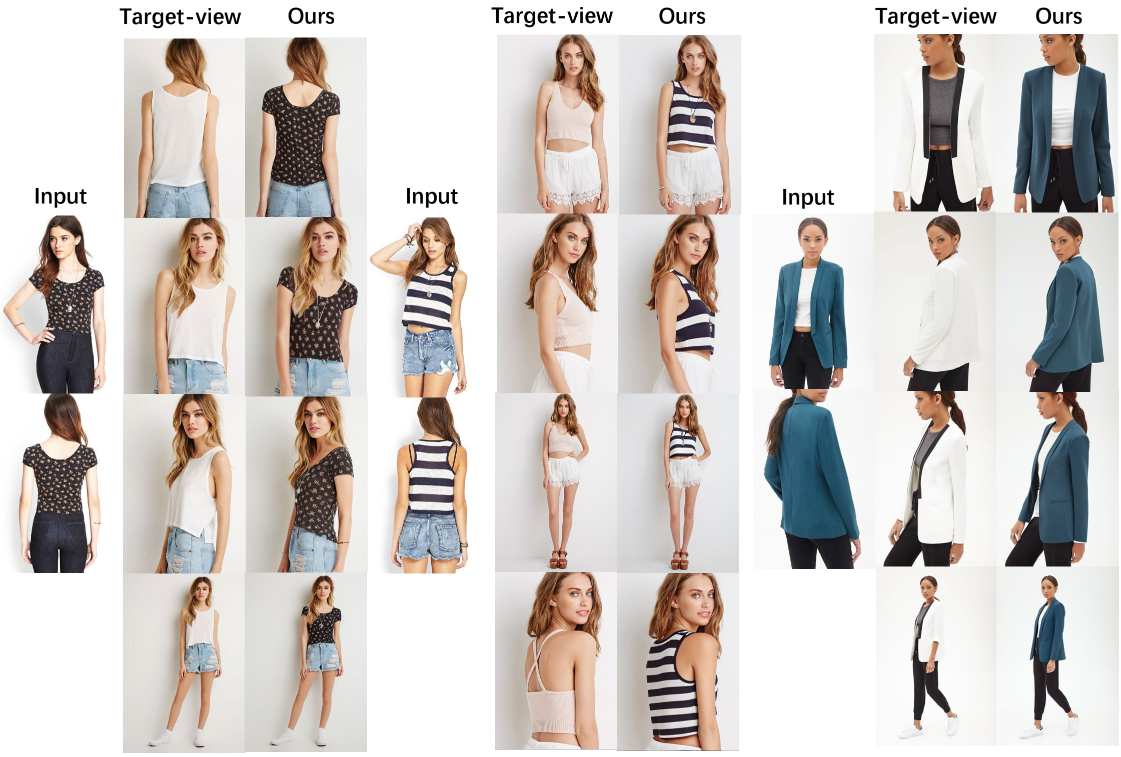

Figure 11. The qualitative results of Robust-MVTON

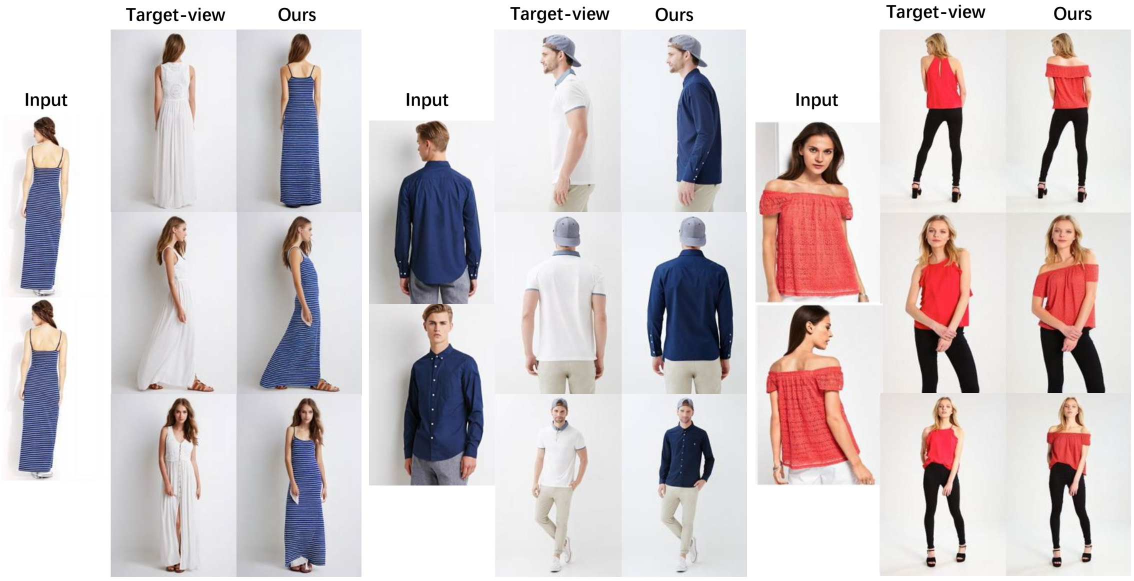

Figure 12. The qualitative results of Robust-MVTON

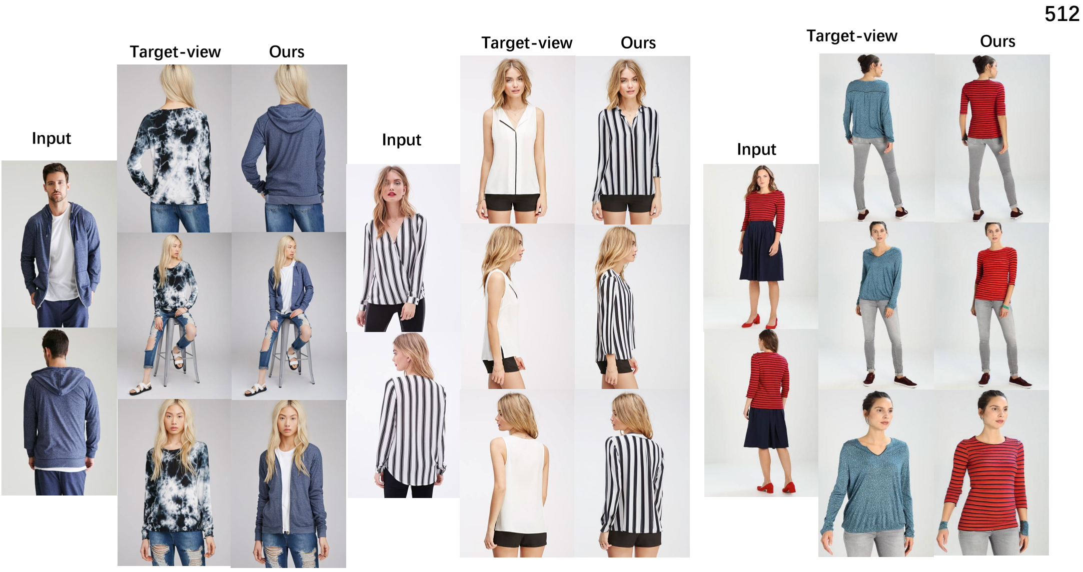

Figure 13. The qualitative results of Robust-MVTON