High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, Jaegul Choo

ECCV 2022

Presenter - Gyojung Gu

Image-based Virtual Try-on

Target Clothes

Reference Images

Source: “VITON: An Image-based Virtual Try-on Network”

Dataset

VITON dataset

VITON-HD dataset

256 x 192

512 x 384

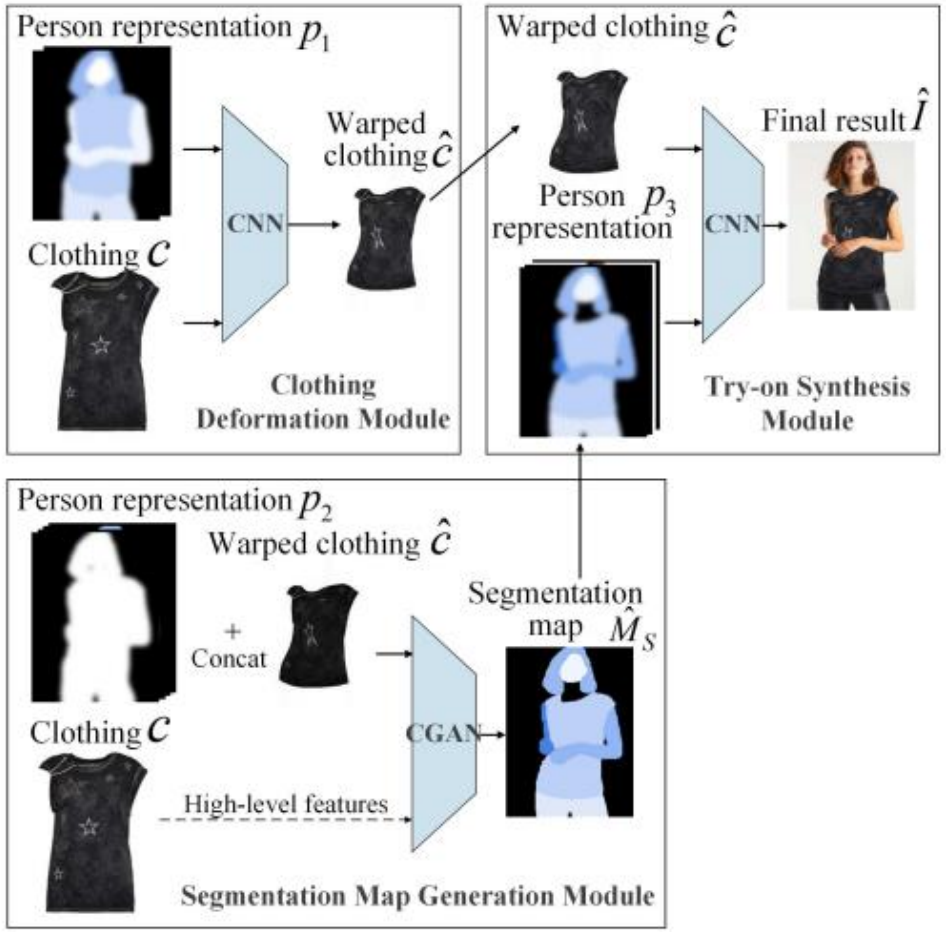
1024 x 768

Virtual Try-on architectures

VTNFP

Warping clothes ->

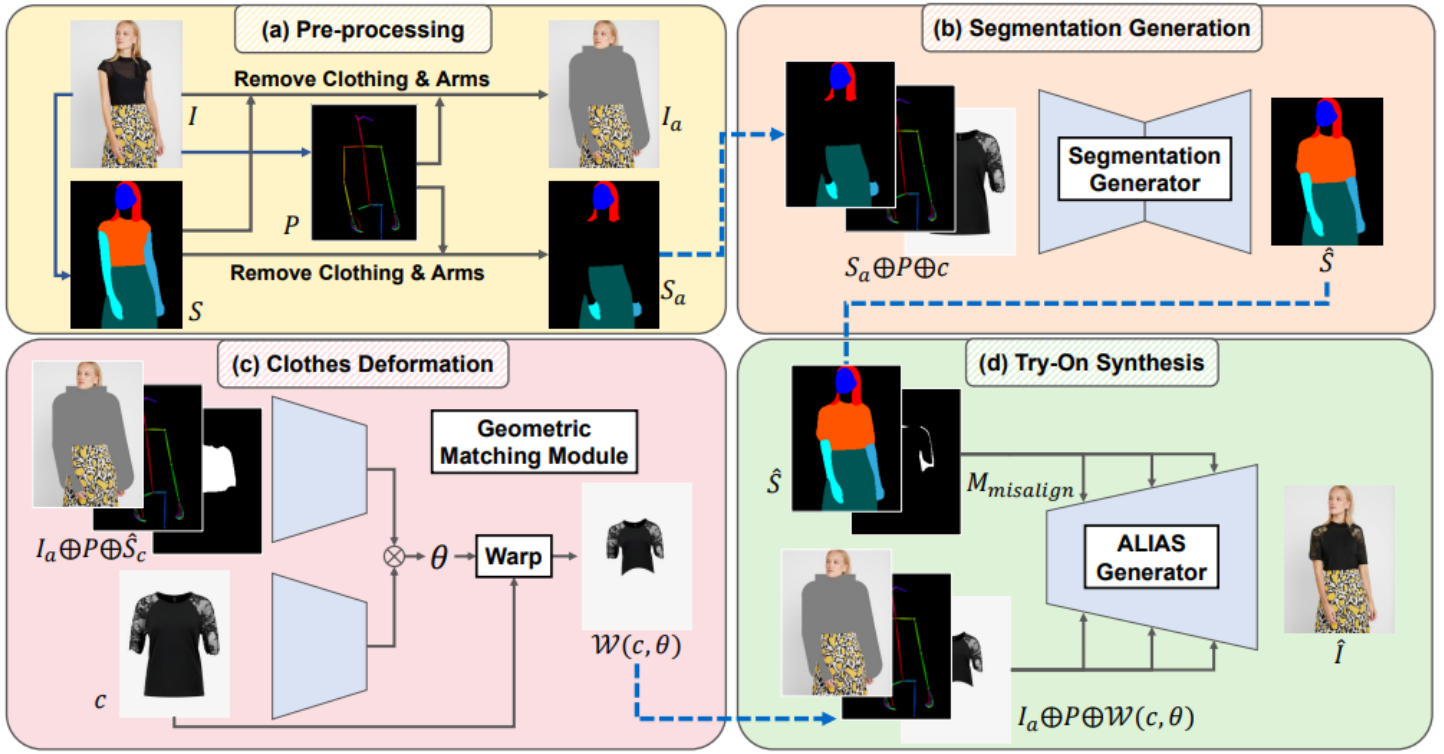
Segmentation generation ->



Yu et al. “VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation”, ICCV 2019

Virtual Try-on architectures

VITON-HD



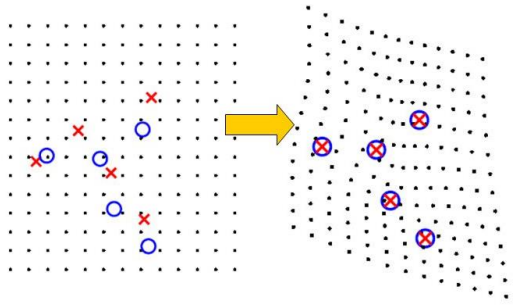
Seunghwan Choi, Sunghyun Park, Minsoo Lee, Jaegul Choo. "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization", CVPR 2021

Motivation

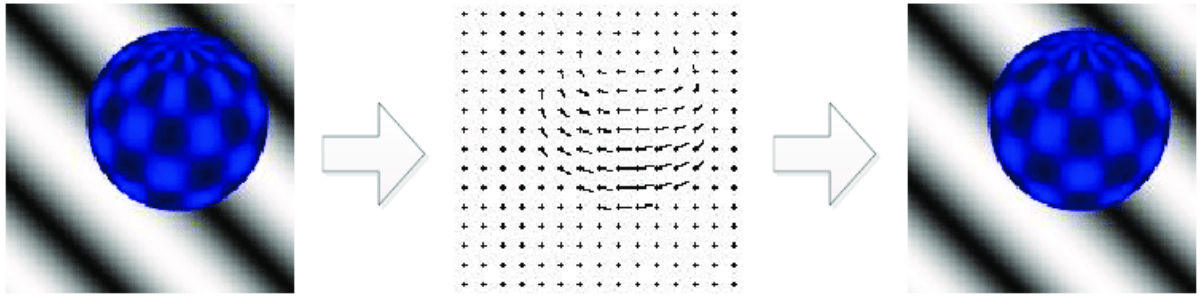
Misalignment

Motivation

TPS transformation

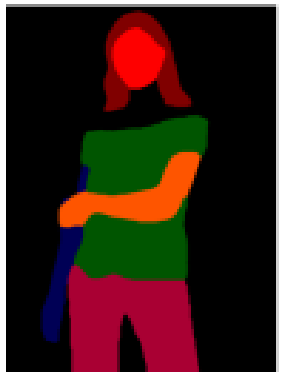


Appearance flow

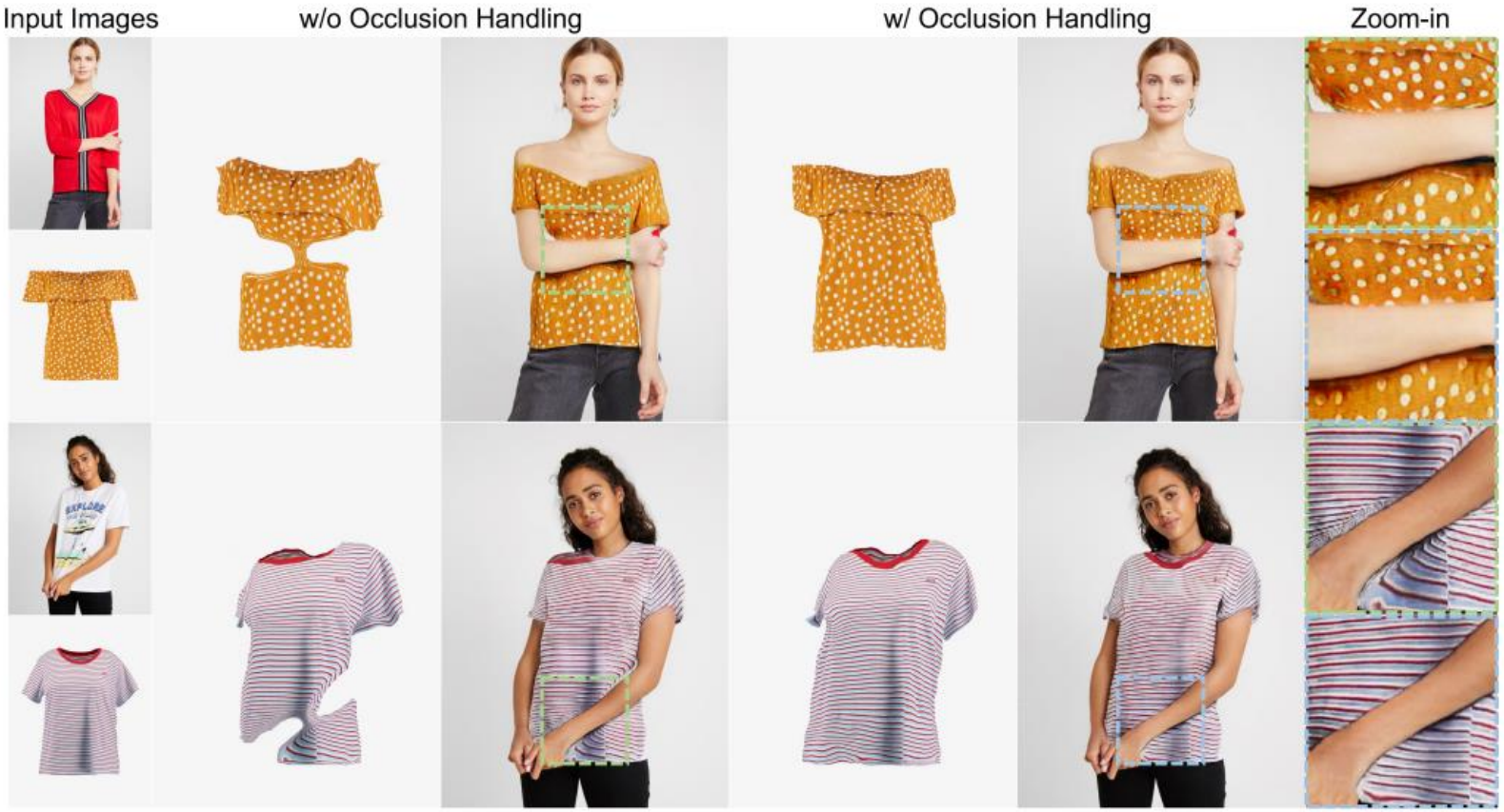


Han et al. “VITON: An Image-based Virtual Try-on Network”, CVPR 2018

Motivation



Body part occlusion



Architecture

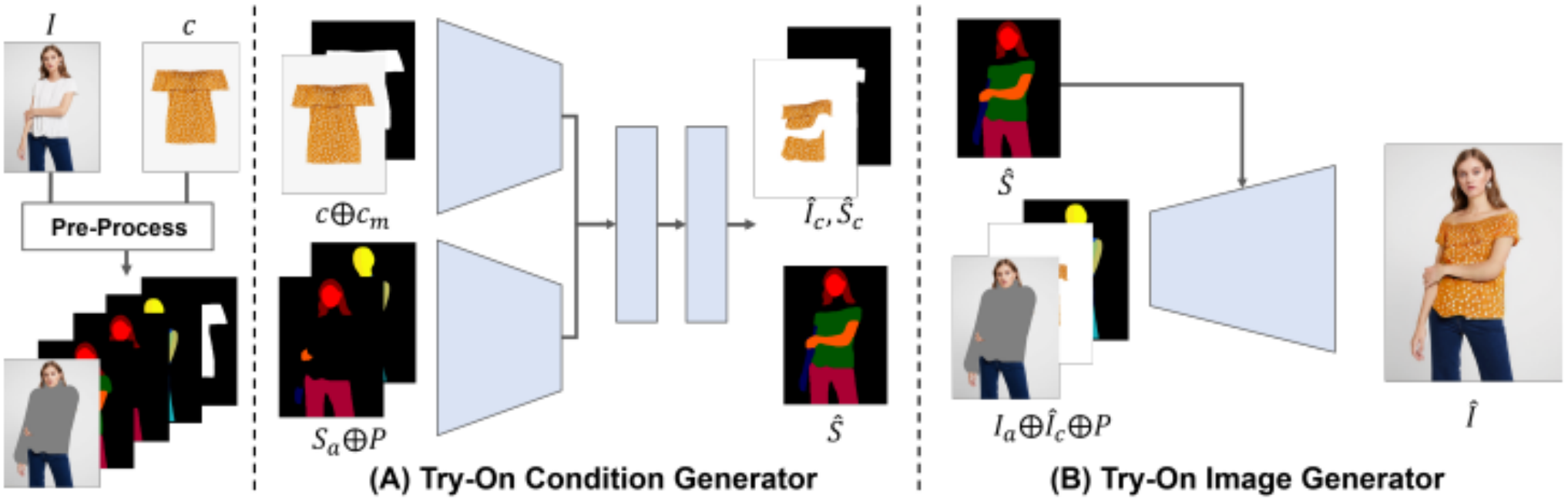


Fig. 2: Overview of the proposed framework.

Architecture – condition generator

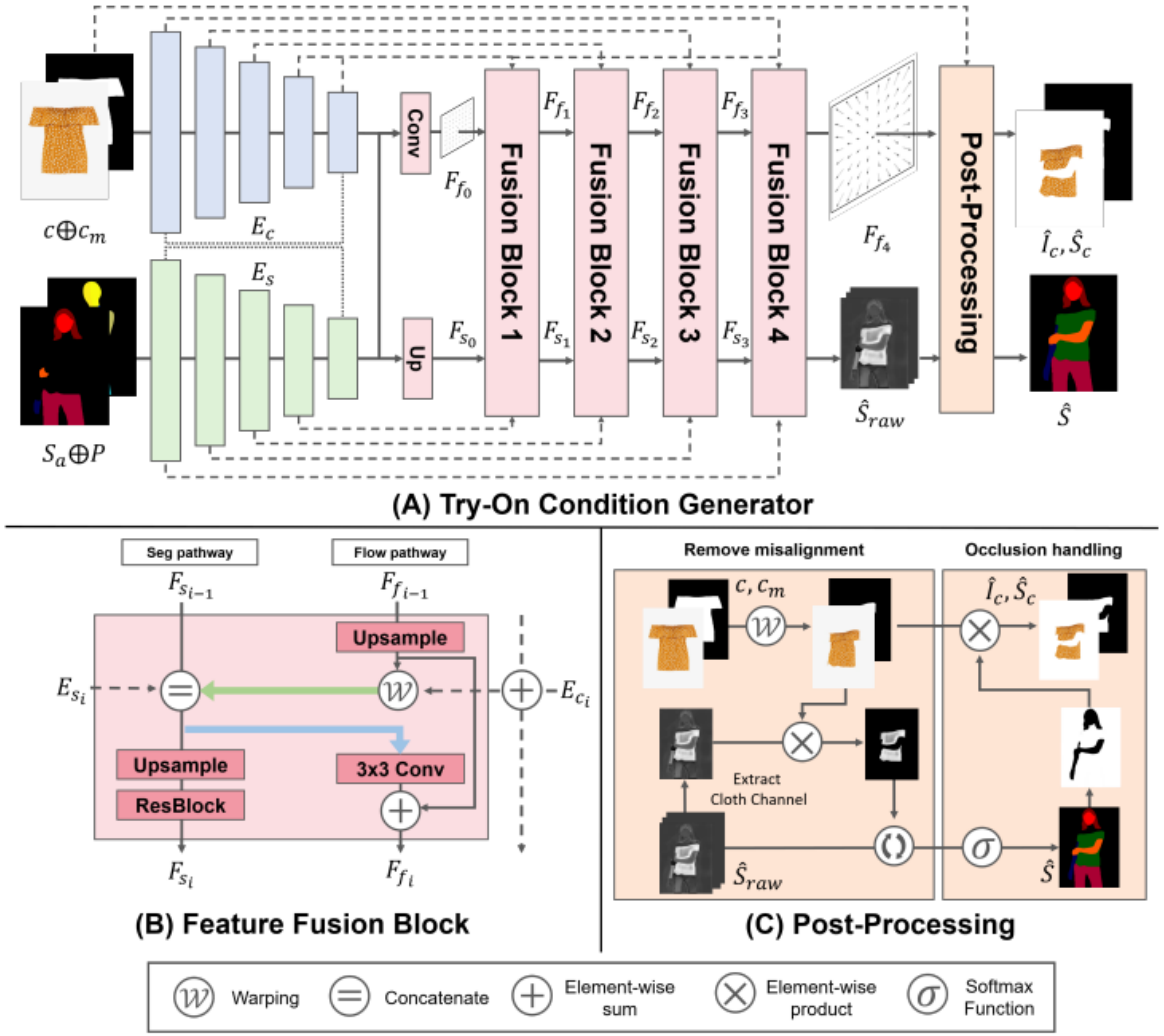
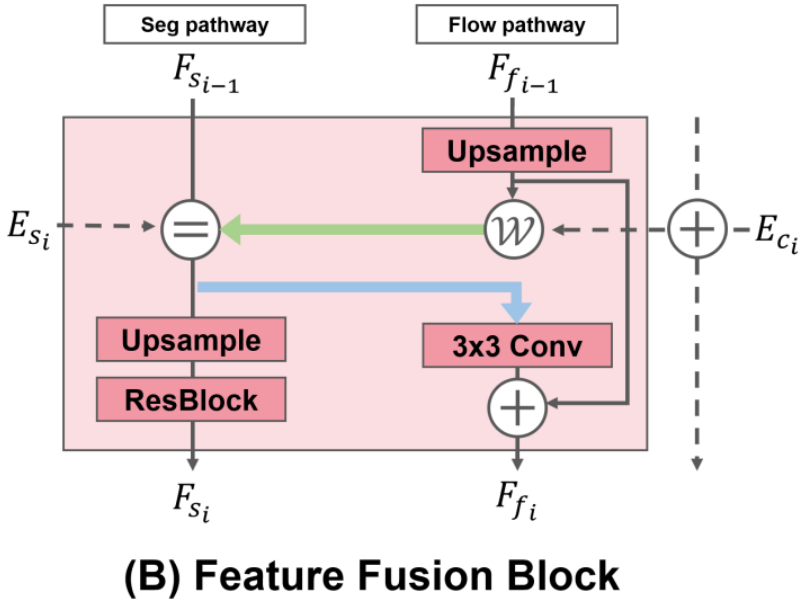
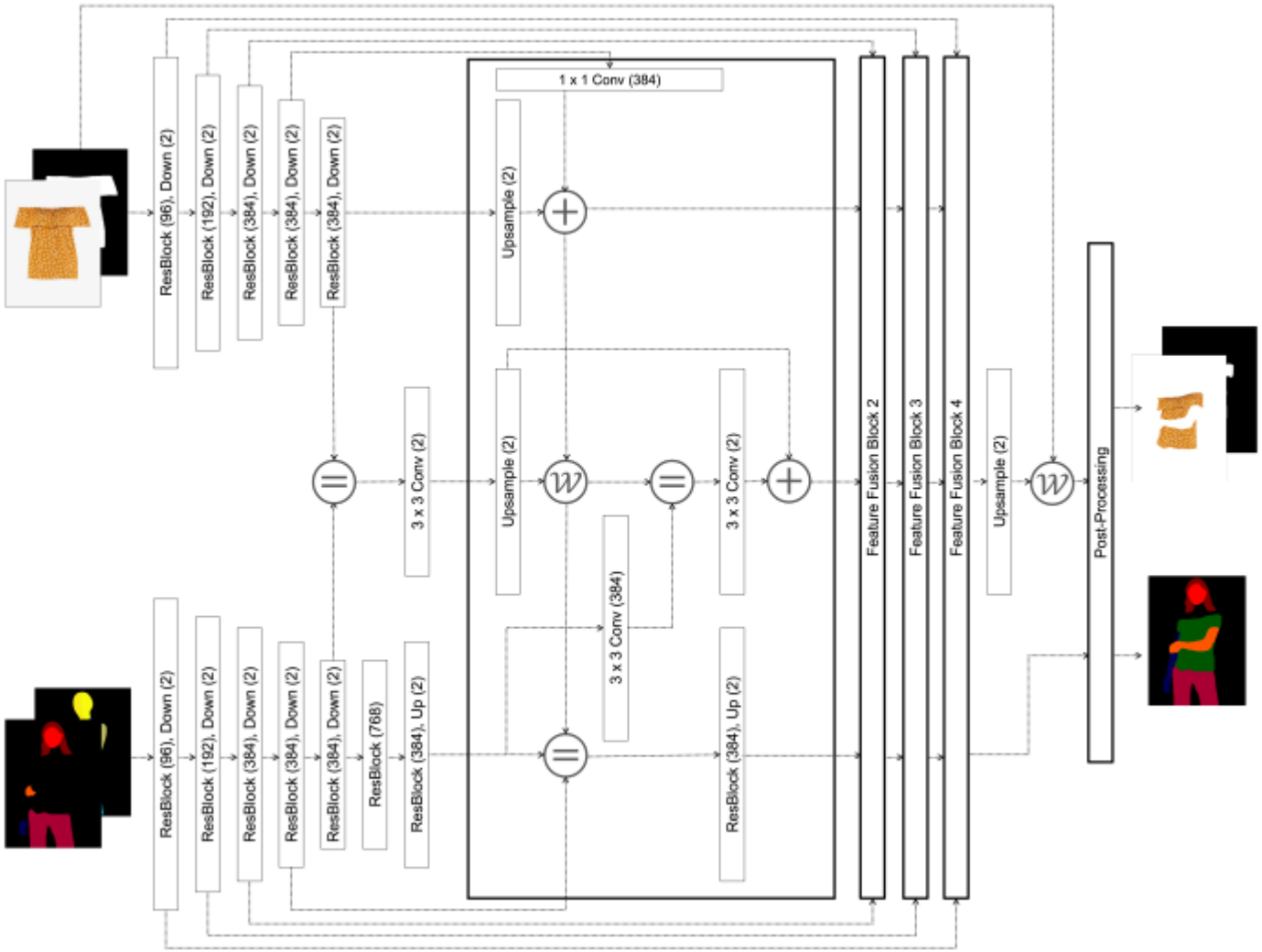


Fig. 3: Architecture of try-on condition generator.

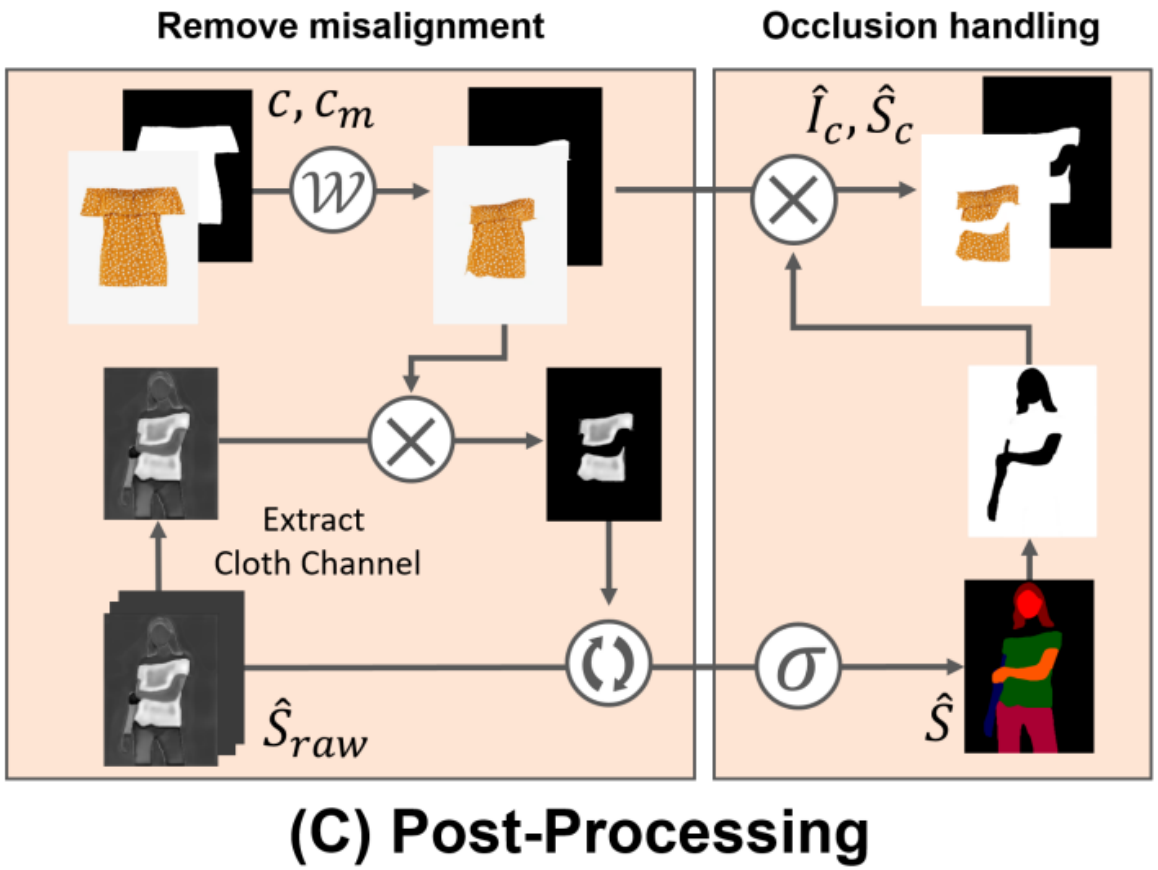
Architecture – condition generator



(B) Feature Fusion Block



Architecture – condition generator



(C) Post-Processing

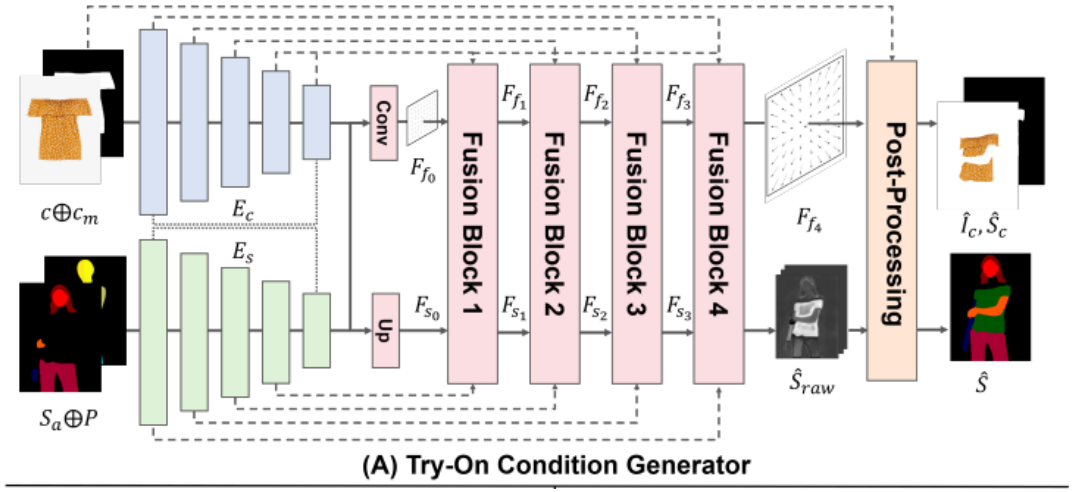
$$\hat{S}^{k,i,j}_{logit} = \begin{cases} \hat{S}^{k,i,j}_{raw} & \text{if } k \neq C \\ \hat{S}^{k,i,j}_{raw} \cdot W(c_m, F_{f_4}) & \text{if } k = C \end{cases}$$

Training – condition generator

$$\mathcal{L}_{L1} = \sum_{i=0}^{3} w_i \cdot ||W(c_m, F_{f_i}) - S_c||_1 + ||\hat{S}_c - S_c||_1,$$

$$\mathcal{L}_{VGG} = \sum_{i=0}^{3} w_i \cdot \phi(W(c, F_{f_i}), I_c) + \phi(\hat{I}_c, I_c),$$

$$\mathcal{L}_{TV} = ||\nabla F_{f_4}||_1$$



(A) Try-On Condition Generator

$$\mathcal{L}_{TOCG} = \lambda_{CE}\mathcal{L}_{CE} + \mathcal{L}_{cGAN} + \lambda_{L1}\mathcal{L}_{L1} + \mathcal{L}_{VGG} + \lambda_{TV}\mathcal{L}_{TV},$$

Training – condition generator

$$\mathcal{L}_{L1} = \sum_{i=0}^{3} w_i \cdot ||W(c_m, F_{f_i}) - S_c||_1 + ||\hat{S}_c - S_c||_1,$$

$$\mathcal{L}_{VGG} = \sum_{i=0}^{3} w_i \cdot \phi(W(c, F_{f_i}), I_c) + \phi(\hat{I}_c, I_c),$$

$$\mathcal{L}_{TV} = ||\nabla F_{f_4}||_1$$

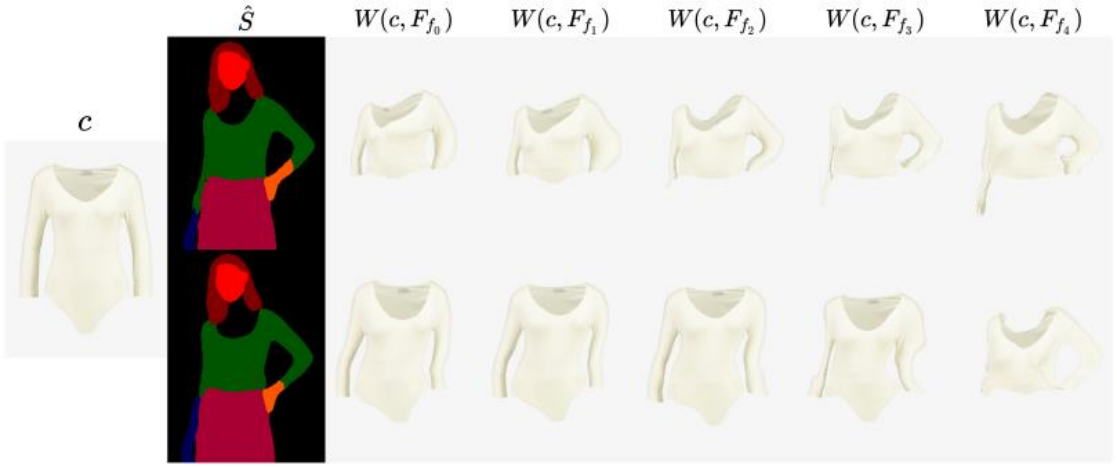
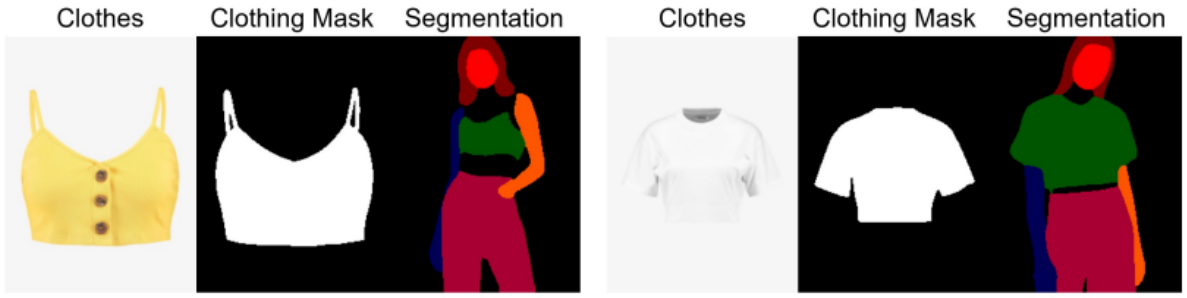


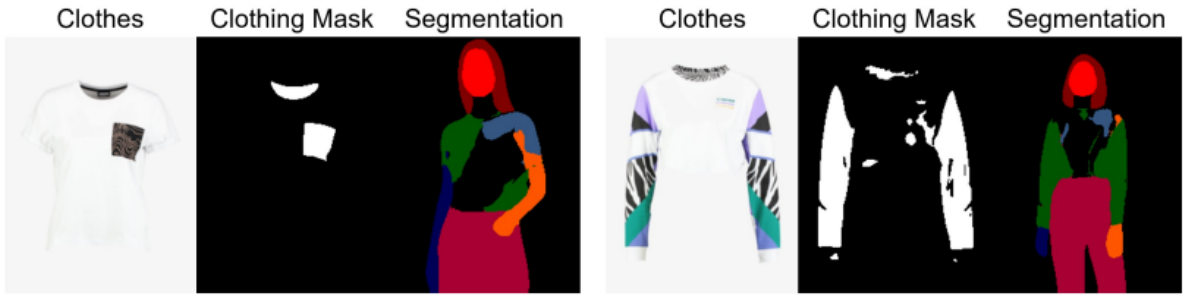
Fig. 15: Effects of the multi-scale $L1$/VGG losses. $1st$ row: w/ multi-scale losses. $2nd$ row: w/o multi-scale losses.

$$\mathcal{L}_{TOCG} = \lambda_{CE}\mathcal{L}_{CE} + \mathcal{L}_{cGAN} + \lambda_{L1}\mathcal{L}_{L1} + \mathcal{L}_{VGG} + \lambda_{TV}\mathcal{L}_{TV},$$

Discriminator rejection



(A) Accepted Samples



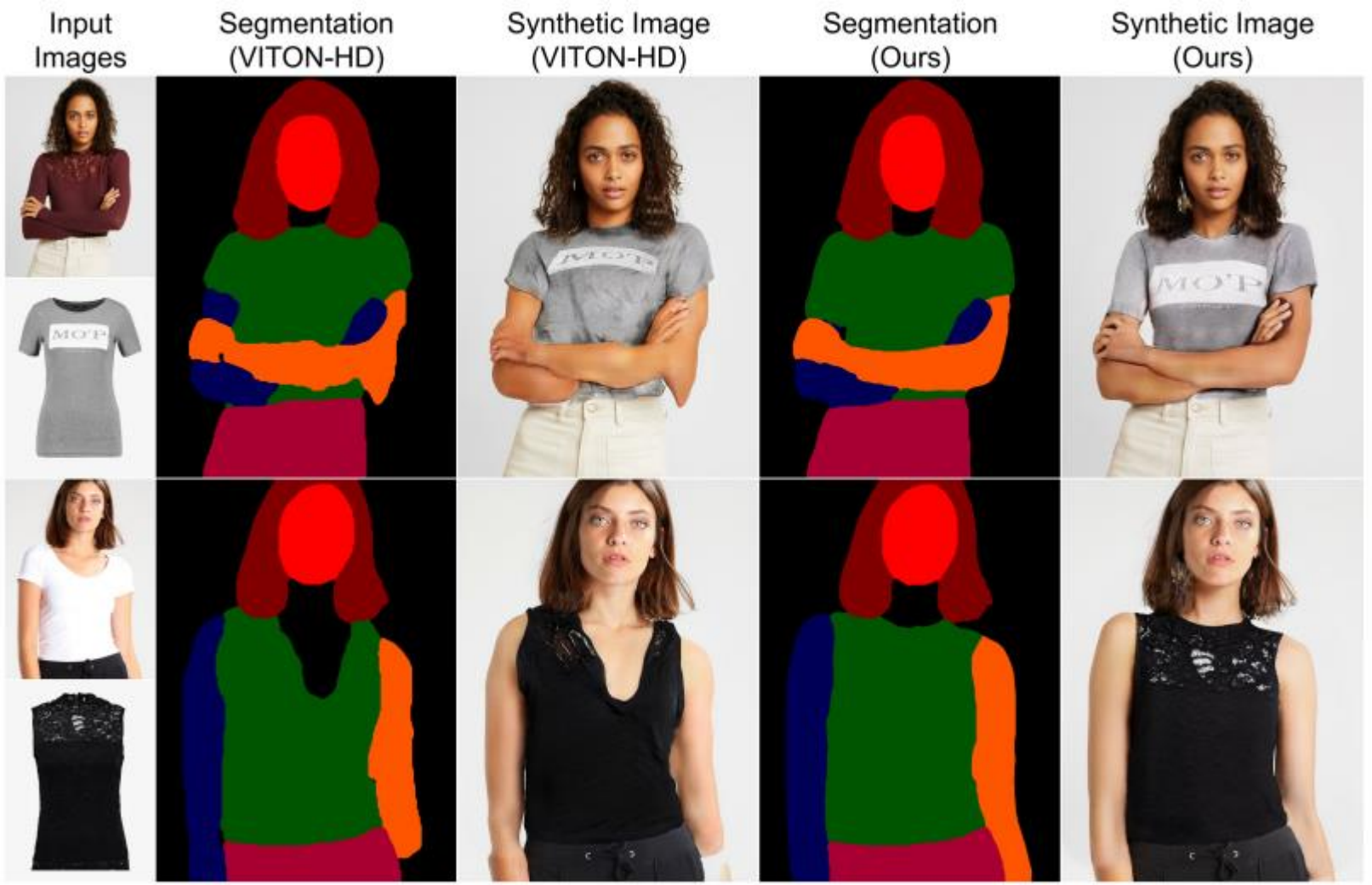
(B) Rejected Samples

Experiments

	256×192				512×384				1024×768			
	LPIPS$_\downarrow$	SSIM$_\uparrow$	FID$_\downarrow$	KID$_\downarrow$	LPIPS$_\downarrow$	SSIM$_\uparrow$	FID$_\downarrow$	KID$_\downarrow$	LPIPS$_\downarrow$	SSIM$_\uparrow$	FID$_\downarrow$	KID$_\downarrow$
CP-VTON	0.159	0.739	30.11	2.034	0.141	0.791	30.25	4.012	0.158	0.786	43.28	3.762
ACGPN	0.074	0.833	11.33	0.344	0.076	0.858	14.43	0.587	0.112	0.850	43.29	3.730
VITON-HD	0.084	0.811	16.36	0.871	0.076	0.843	11.64	0.300	0.077	0.873	11.59	0.247
Ours	**0.062**	**0.864**	**9.38**	**0.153**	**0.061**	**0.878**	**9.90**	**0.188**	**0.065**	**0.892**	**10.91**	**0.179**

Table 1: Quantitative comparison with baselines. We describes the KID as a value multiplied by 100.

Experiments



Experiments

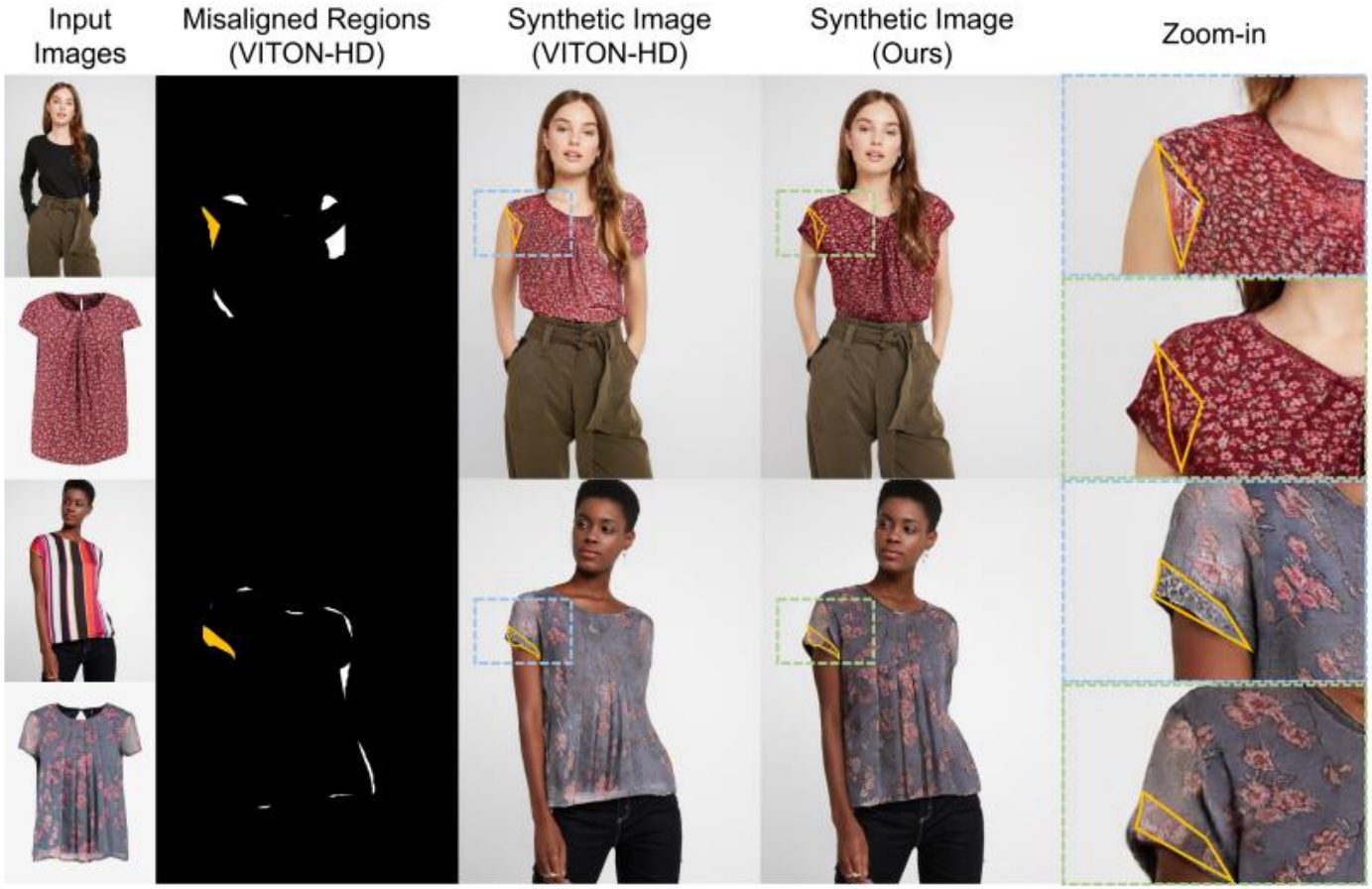


Fig. 6: Synthesis results and corresponding misaligned regions indicated by yellow colored areas. VITON-HD suffers from the artifacts caused by misalignment.

Experiments

Fig. 7: Effects of the body part occlusion handling. The green colored areas indicate the pixel-squeezing artifacts.

Experiments

Method	FID↓	KID↓
PF-AFN	14.01	0.588
Ours	**10.91**	**0.179**
∟ w/o Post-Processing	12.05	0.356
∟ w/o Fusion Block	12.41	0.381
∟ w/o Fusion Block & Post-Processing*	12.73	0.415

Experiments

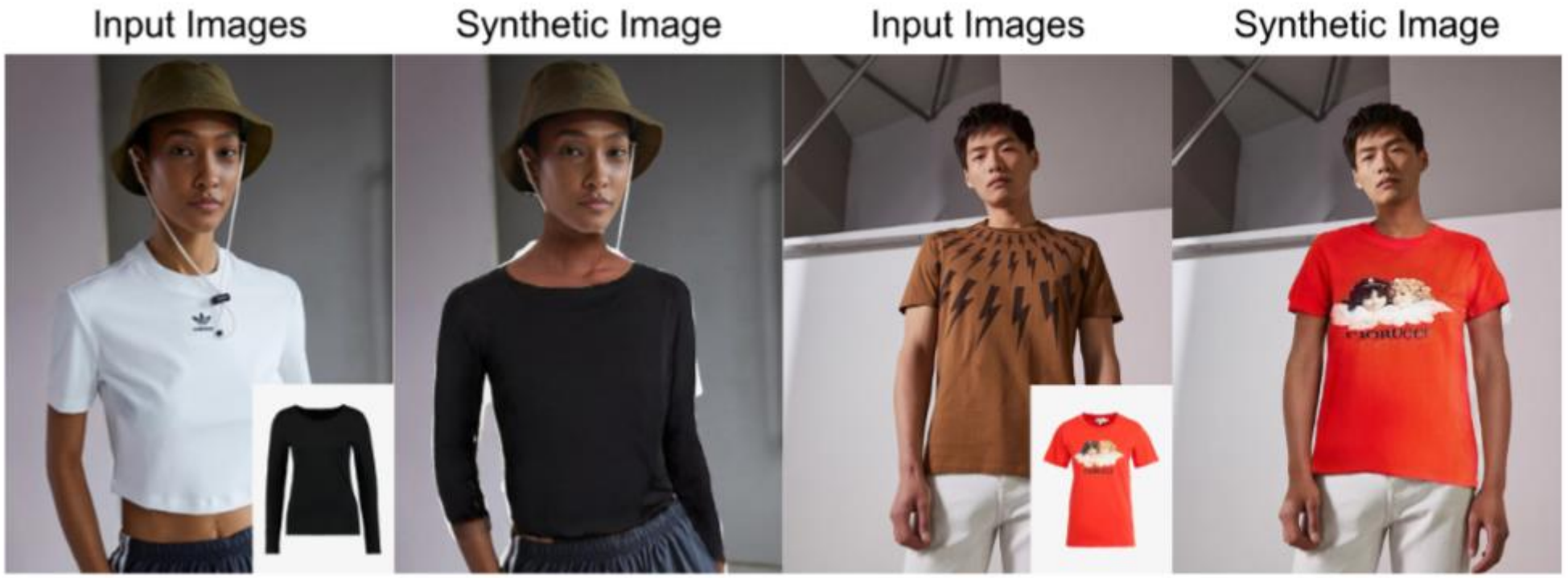


Figure 3. Qualitative results of our model on the images in the wild setting (*i.e.*, complex background).

Thank you