Zflow: Gated Appearance Flow-based Virtual Try-on with 3D Priors
Supplementary Material

Ayush Chopra*†[2], Rishabh Jain *‡[3], Mayur Hemani[1], and Balaji Krishnamurthy[1]

[1]Media and Data Science Research Lab, Adobe
[2]Media Lab, Massachusetts Institute of Technology
[3]BITS Pilani

1. Additional Qualitative Results

Due to limited space, we focused on presenting diversity of improvement across multiple facets in the main paper. Here, we present extensive additional qualitative results (see pages 4 to 11 here). The results are sampled at random from the test set and encapsulate diversity across body shape, ethnicity, clothing type, pattern, occlusion, pose of the target model and in-shop garment. Results show that *Zflow* consistently and significantly improves over baseline methods. Along with the extensive evidence in the main paper including quantitative & qualitative results, ablation studies and user studies, we believe these results reinforce the superiority of Zflow.

2. Details for Experiments and Results

Implementation Setup This follows from the appendix reference in Section 4. We note the following hyperparameter values for our experiment: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$, $\lambda_1 = \lambda_2 = 2, \lambda_3 = 3, \lambda_4 = 0.5$ and $\alpha_1 = 0.2, \alpha_2 = \alpha_3 = 1$.

Qualitative Results with Edge Loss This follows from the appendix reference in Section 6. The edge loss (L_{edge}), obtained using sobel operators (S_x, S_y), is used when optimizing the dense fusion module and is represented as:

$$L_{edge} = \|\phi(I_{tryon}), \phi(I_m)\|_{smoothL1} \quad (1)$$

$$\phi(i) = |S_x(i)|+|S_y(i)| \quad (2)$$

As mentioned in the paper, results in Figure 1 show that using the edge loss reduced bleeding (row 2) and improved texture and contrast (row 1).

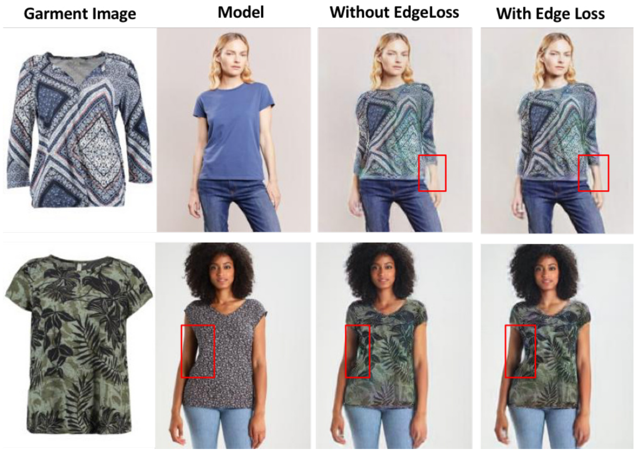


Figure 1. Using L_{edge} during dense fusion reduces bleeding and improves texture.

3. Details for Gated Appearance Flow

This section provides additional details for GAF implementation.

Equation for *GAF* in Zflow Here, we provide supporting details regarding the *GAF* used in Zflow (introduced in Section 3). The aggregation sub-network employs a sequential 4-cell ConvGRU to combine the flow estimates (f_i) from the last $M = 3$ layers to obtain the final appearance flow map (f_{agg}). This computation can be formalised as

*equal contribution
†work done while working at Adobe MDSR Lab
‡work done as part of Adobe MDSR internship

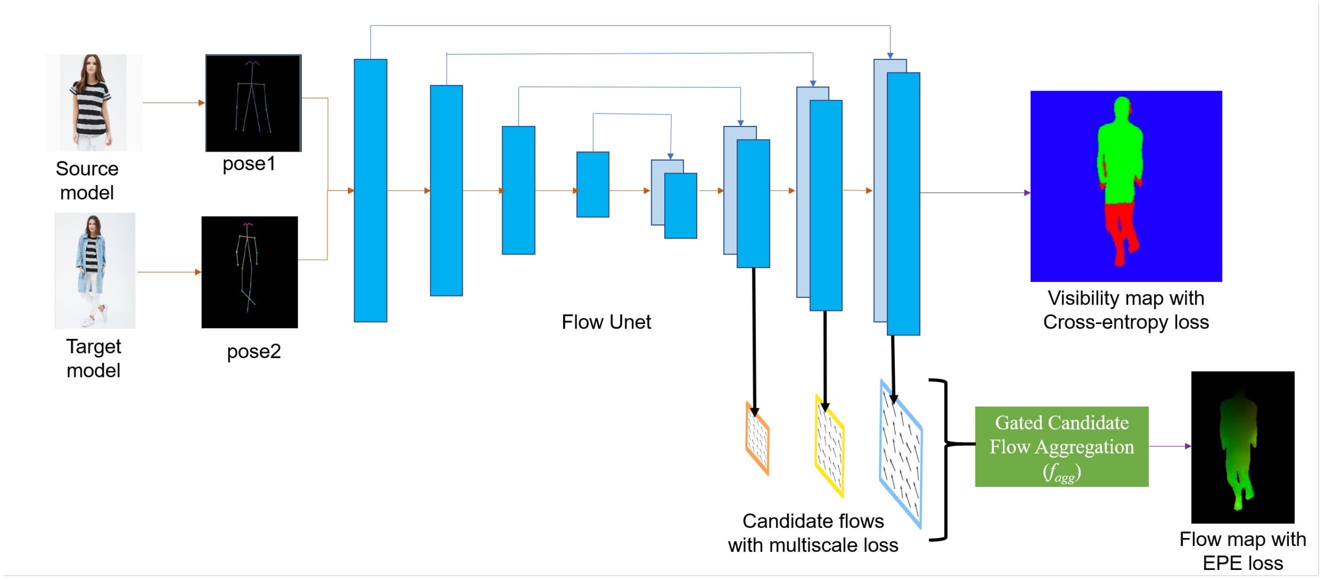

Source model
pose1
Target model
pose2
Flow Unet
Visibility map with Cross-entropy loss
Candidate flows with multiscale loss
Gated Candidate Flow Aggregation (f_{agg})
Flow map with EPE loss

Figure 2. *GAF* based flow regression module in DIF-GAF for the task of human pose transfer. The network is trained to predict a visibility map along with the flow map, as originally mentioned in [1]. *GAF* is specifically introduced to predict the flow map. The network is optimized on a binary cross entropy loss (visibility map) and end point error [1] for the flow map. The ground-truth visibility and flow maps are obtained used the same setup as DIF [1]

($f_{agg} = F_3$):

$$\begin{aligned} F_{i+1} &= CG_{j=3}(f_i, F_i) \text{ where} \\ F_0 &= \vec{0} \text{ and } i \,\epsilon\, [0,2] \end{aligned} \tag{3}$$

$$\begin{aligned} CG_j(f,h) &= \begin{cases} C_{p_{j+2},p_{j+1}}(f, CG_{j-1}(f,h)) & j > 0 \\ C_{p_1,p_0}(f,h) & j = 0 \end{cases} \\ &\text{where } p = [2,64,128,64,2] \text{ and } \; j \,\epsilon\, [0,3] \end{aligned} \tag{4}$$

$$\begin{aligned} C_{n,m}(f,h) &= (1 - U_{n,m}(f,h))h \\ &\quad + U_{n,m}(f,h)\tilde{H}(f,h,R_{n,m}) \end{aligned} \tag{5}$$

$$U_{n,m}(f,h) = sigmoid(Conv2d_{n,m}(f \oplus h)) \tag{6}$$

$$R_{n,m}(f,h) = sigmoid(Conv2d_{n,m}(f \oplus h)) \tag{7}$$

$$\tilde{H}(f,h,r) = \tanh(Conv2d_{n,m}(f \oplus (h \odot r)) \tag{8}$$

As noted in the paper, f_{agg} is used to sample the warped garment image I^p_{wrp} and the warped garment mask M^p_{wrp} from the garment image I_p and M_{cm} respectively.

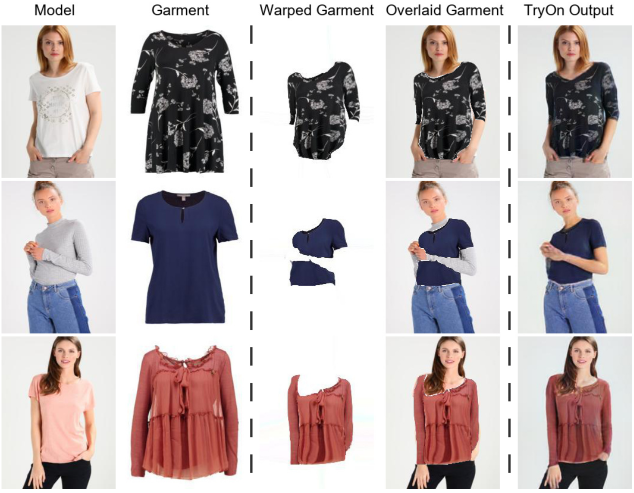


Figure 3. Zflow can handle patterns (row 1), self-occlusion (row 2, 3), and extremepose (row 2) when warping the garment

4. Details for Human Pose Transfer

This follows from the extension of ZFlow to Human pose transfer task. We provide additional details regarding the *GAF* based flow regression with *DIF* [1]. As mentioned in the main paper, [1] is a recent state-of-the-art in pose transfer which generates photo-realistic images in the target pose in two stages i) first regressing 3D appearance flow which map input to target pose and ii) then performing feature warping on the input using the flow estimates. We also showed that swapping in GAF for the flow regression stage (to define *DIF-GAF*) resulted in significant performance improvement. Figure 2 summarizes the flow regression module *DIF-GAF*.

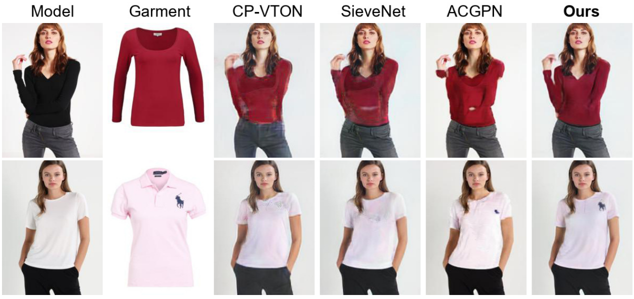


Figure 4. Failure cases due to pre trained networks(row 1) and improper collar generation(row 2)

5. Warping quality

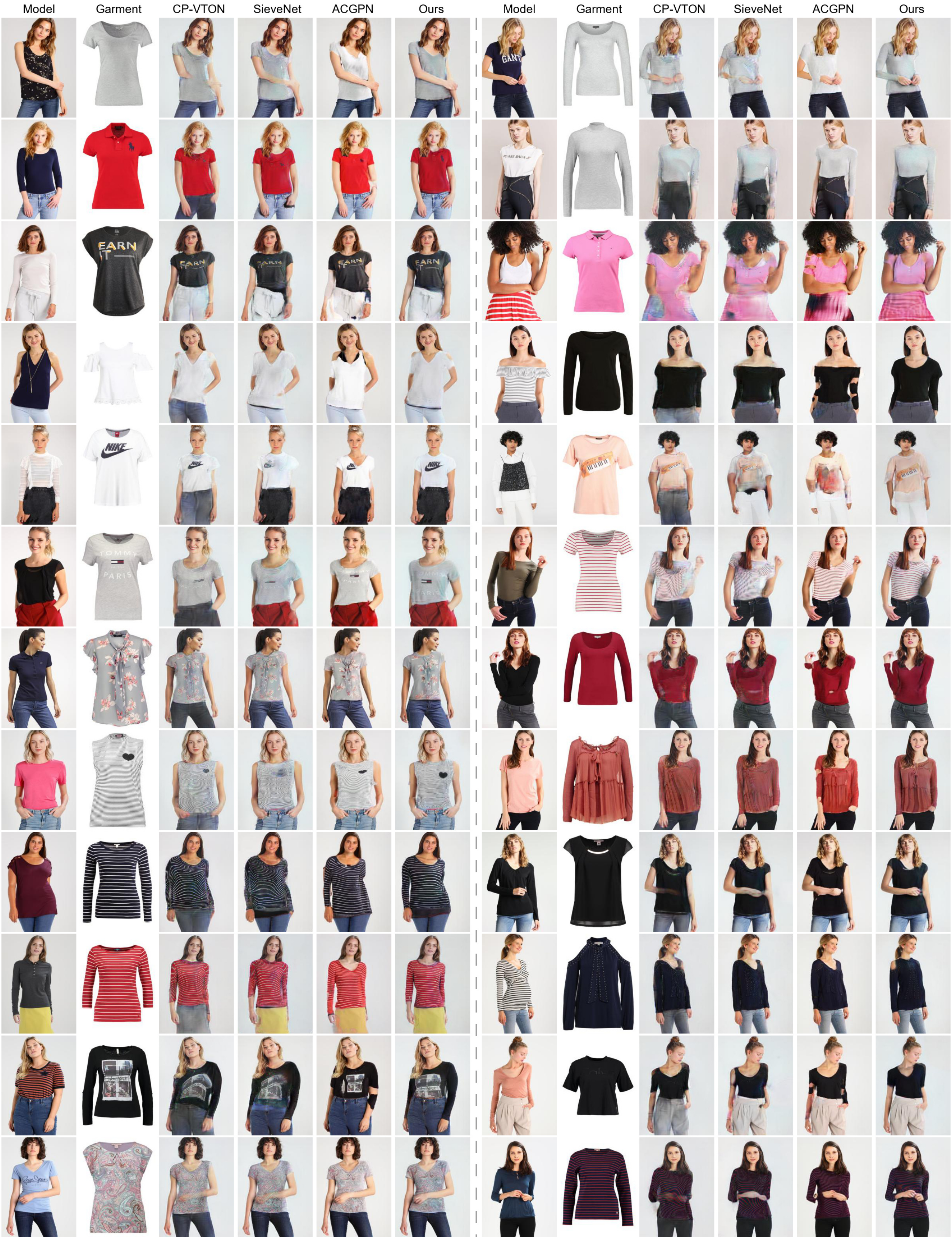
We also showcase our networks ability to generate superior results when warping the garment under difficult conditions. As can be seen from Figure 3, ZFlow can handle complex patterns, self-occlusions and extreme poses with high accuracy.

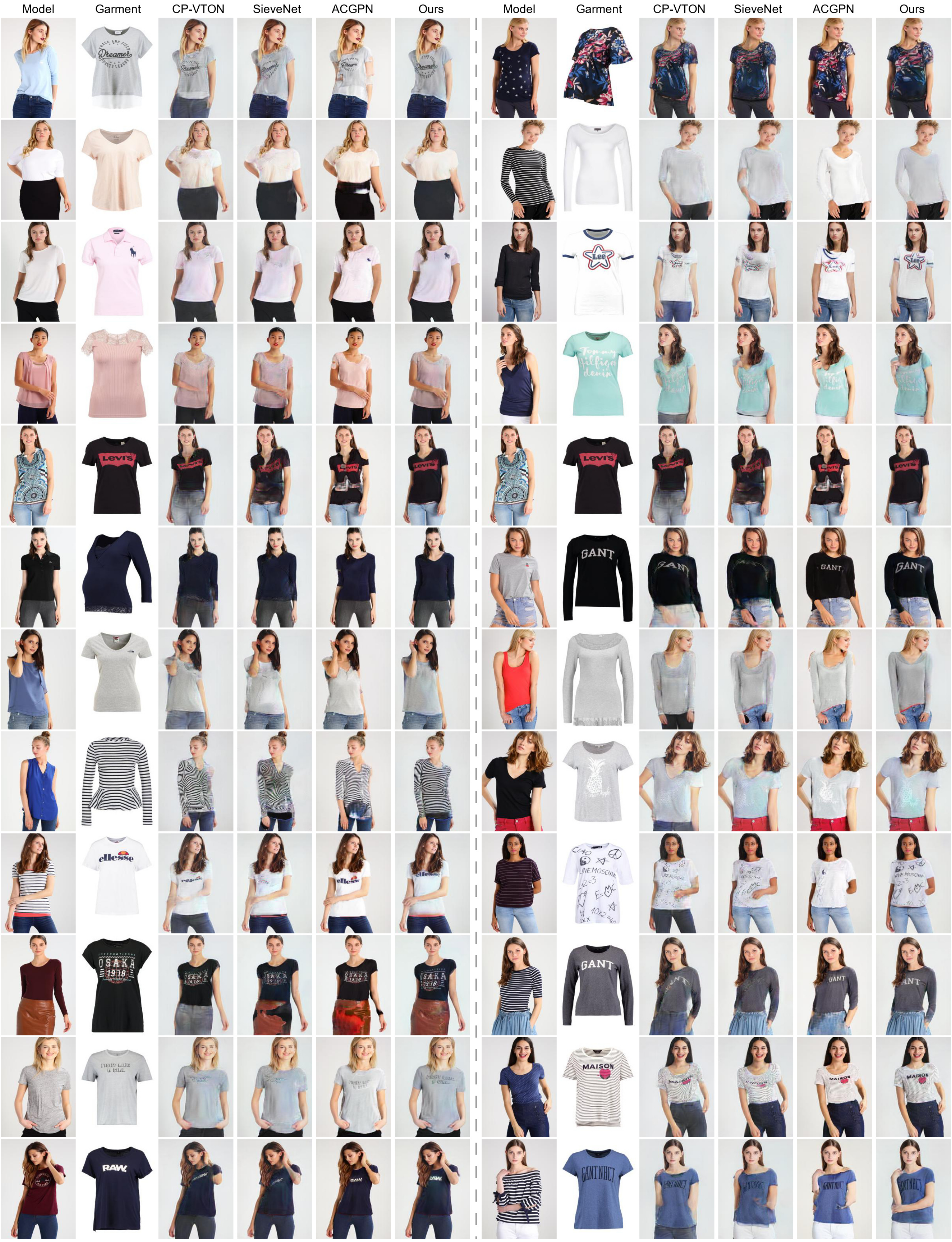
6. Limitations and Failures

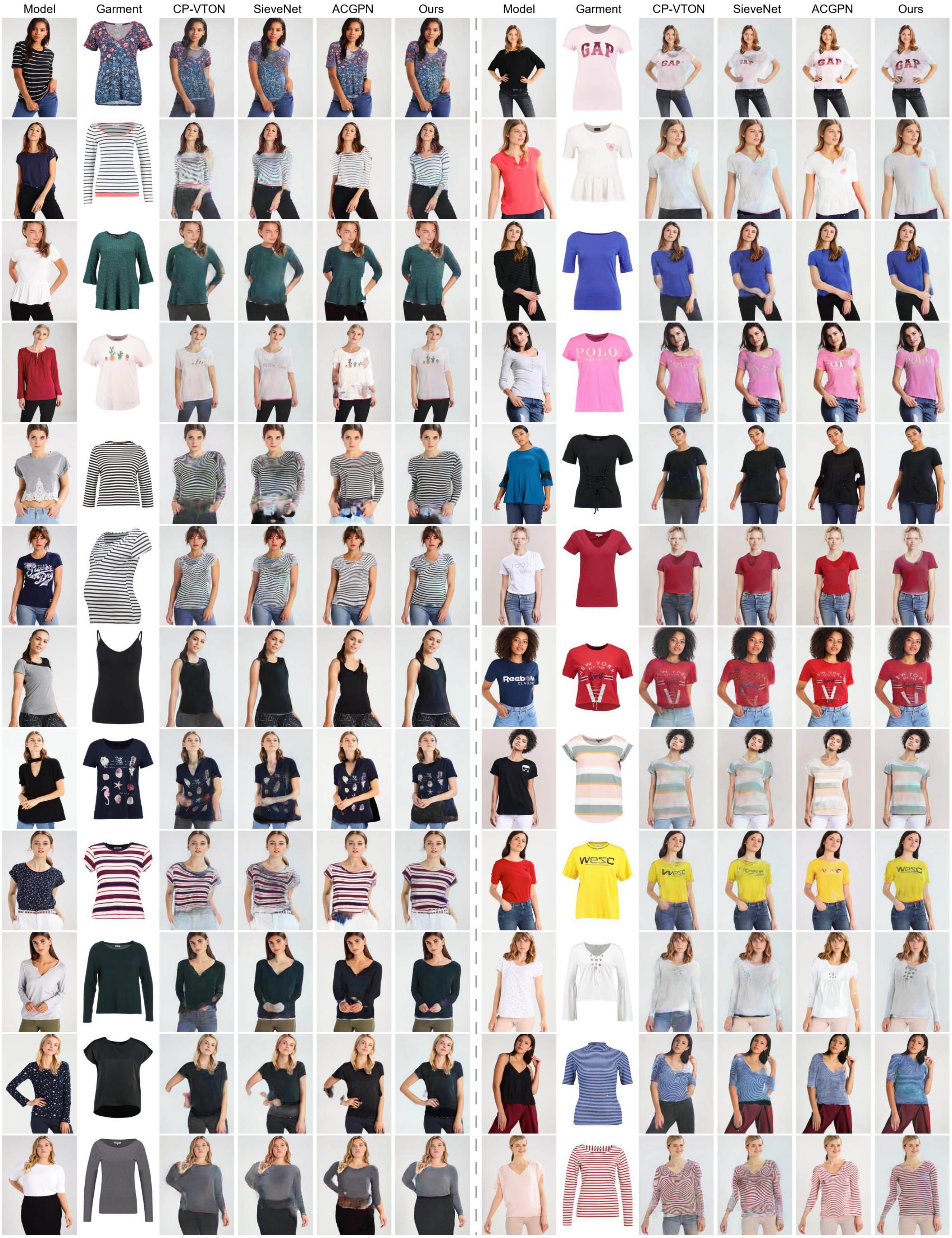
Although Zflow performs significantly better than previous frameworks for virtual try-on, there are still some limitations caused due to the use of pretrained human parsing and densepose networks (Figure 4). Zflow sometimes also misses the collar generation in shirts due to fewer examples (vs. v-/u-neck shirts) in the training set, and the collar being frequently obscured by hair.

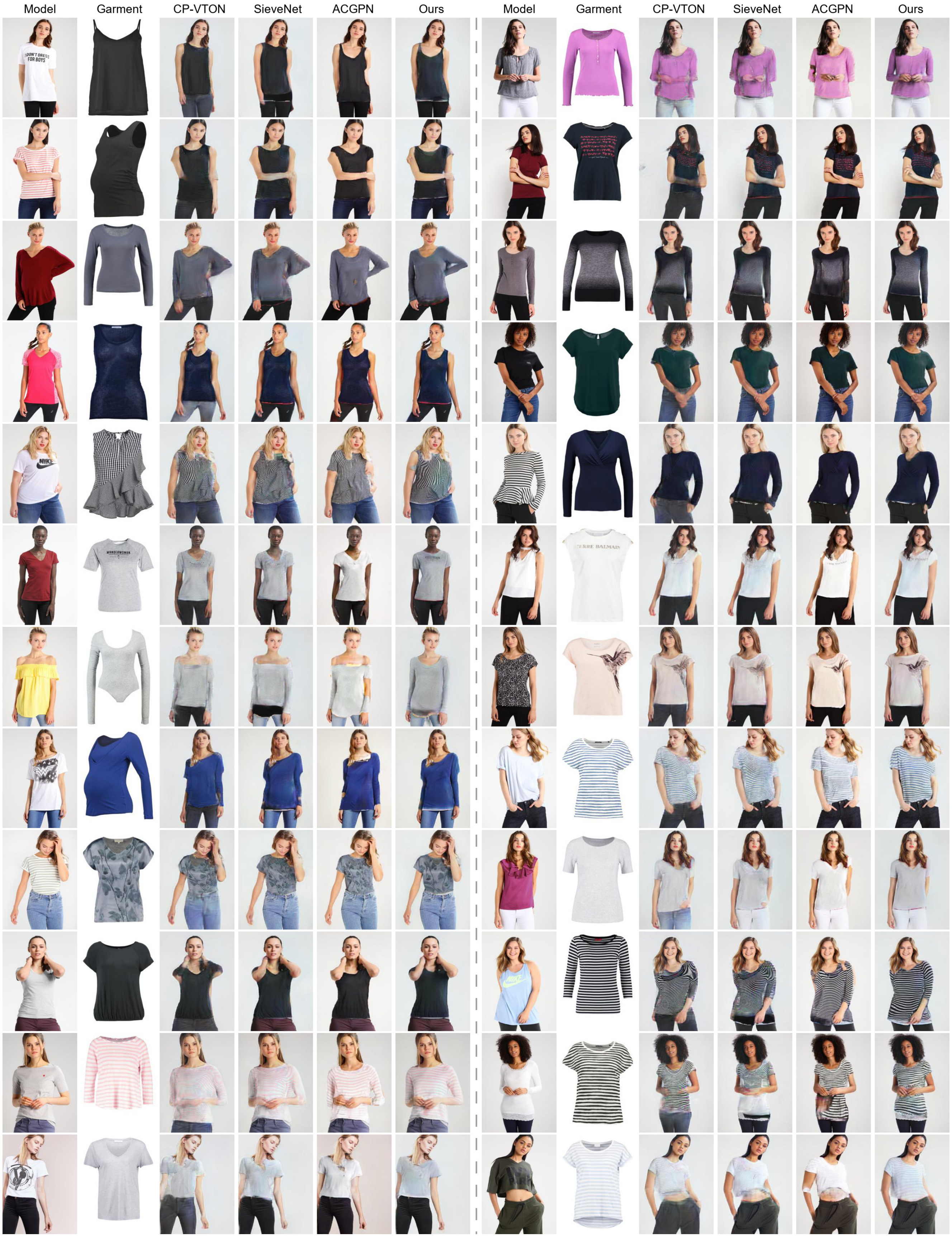
References

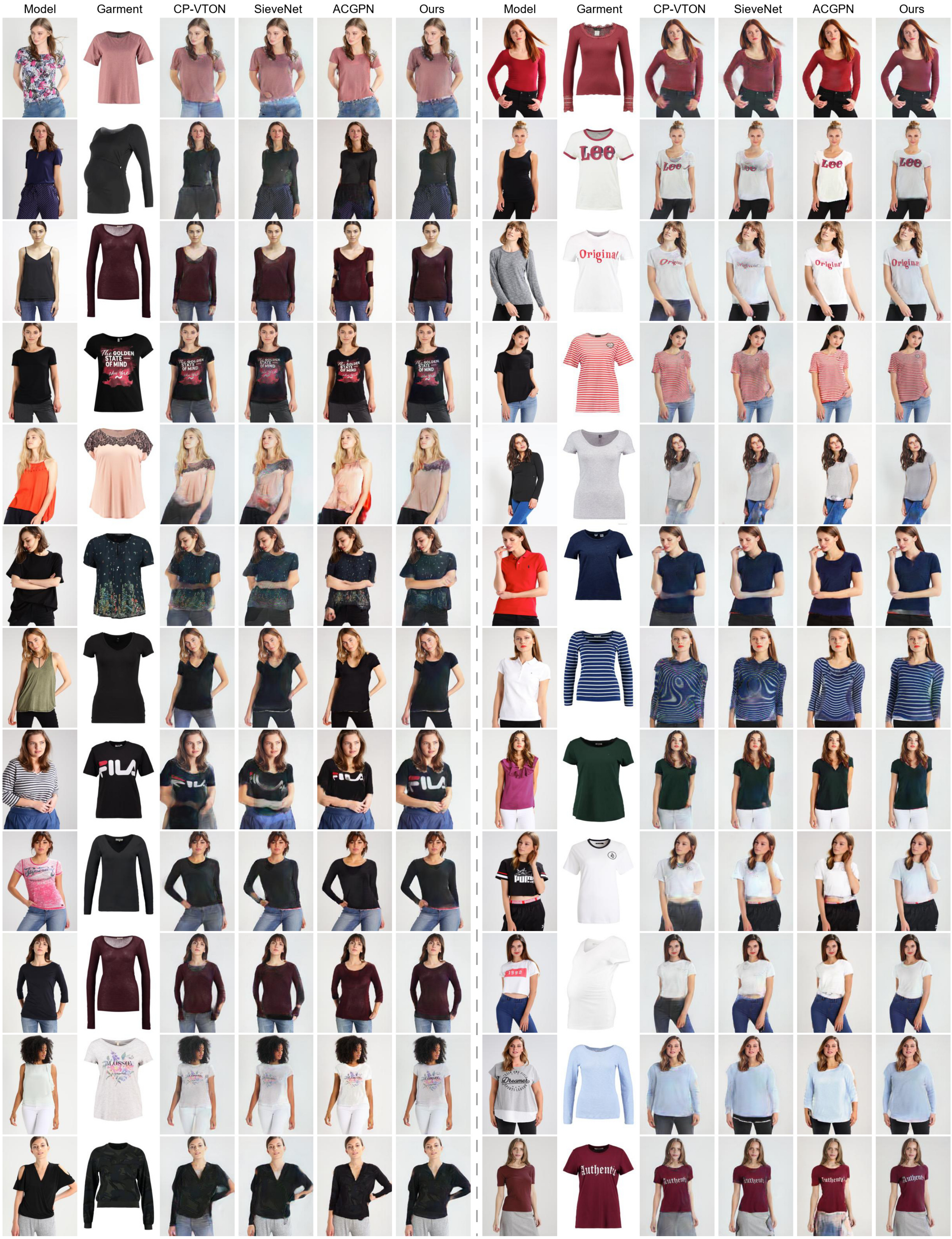
[1] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

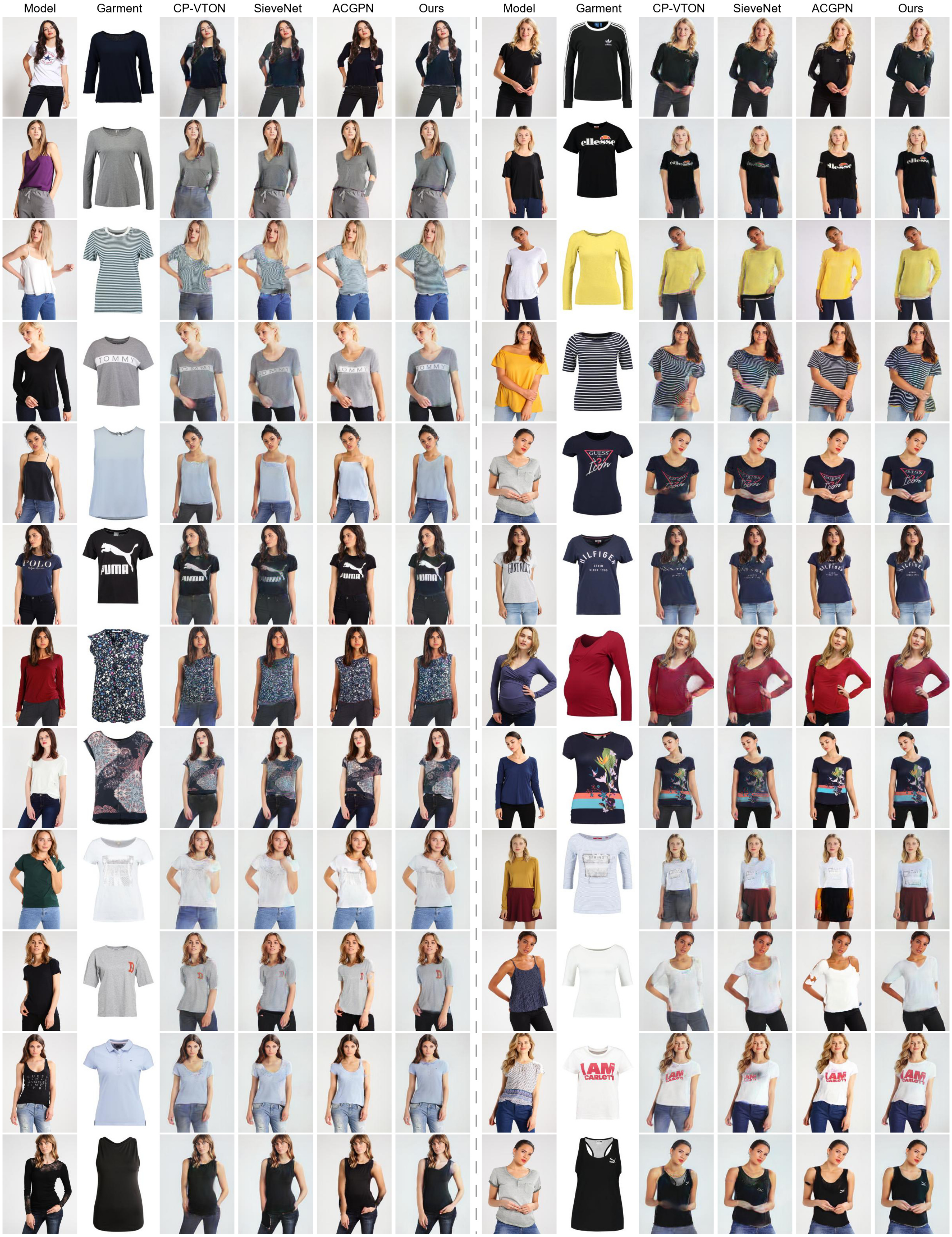

Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours

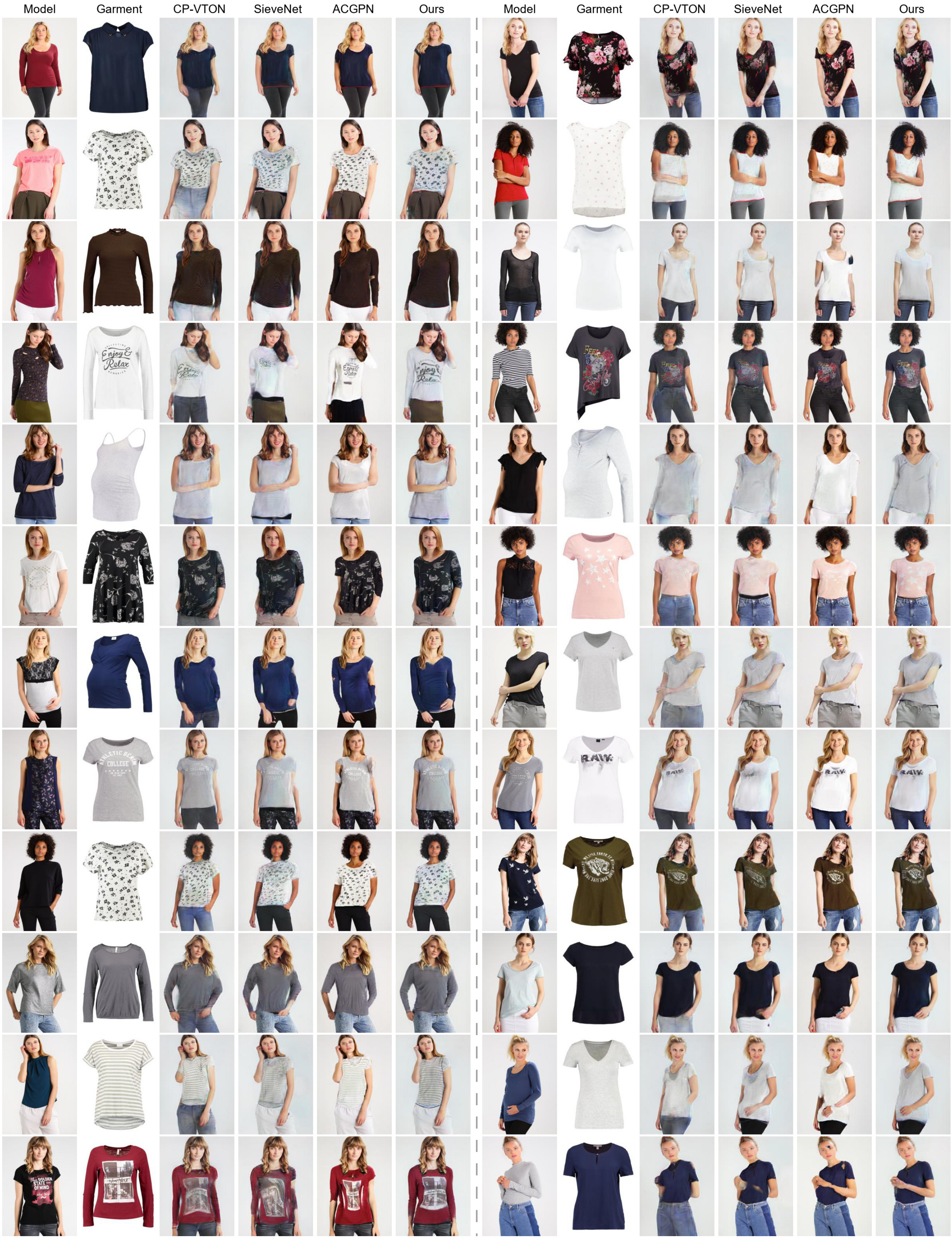

Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours


Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours


Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours


Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours


Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours


Model
Garment
CP-VTON
SieveNet
ACGPN
Ours
Model
Garment
CP-VTON
SieveNet
ACGPN
Ours

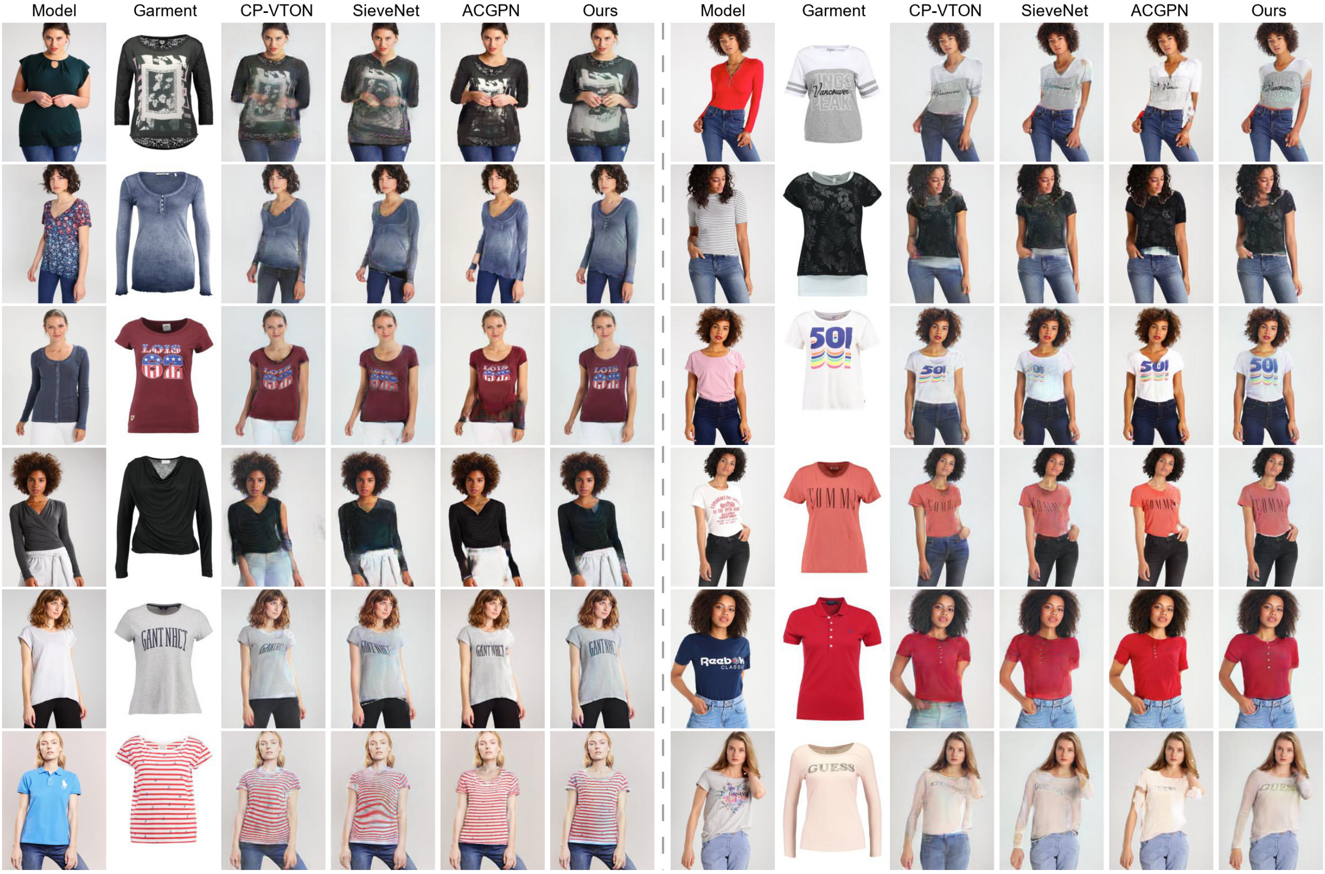


Figure 5. Qualitative Comparison of the proposed Zflow with baselines. Zflow achieve significant improvement across varying dimensions of geometric and textural integrity.