C-VTON Supplementary Material

In the main part of the paper we presented a wide variety of results to highlight the benefits of the proposed C-VTON virtual try-on approach. In this *Supplementary material* we now show additional visual result to further demonstrate the capabilities of C-VTON. Specifically, we: (i) show comparisons with competing models from the literature using a wider range of input images, (ii) present additional try-on results in different settings, e.g., under pose variations, with varying sleeve lengths, or with differently textured garments, (iii) investigate the performance of C-VTON on an additional high-resolution dataset, i.e., VITON-HD, (iv) study the impact of body-part segmentations on the overall performance, (v) compare results between the Geometric Matching Module from [34] and our Body-Part Geometric Matcher, and (vi) provide a more in-depth analysis of the limitations of the proposed (virtual try-on) model.

A. Additional Comparisons

Figure 10 shows additional comparisons between C-VTON and competing approaches from the literature. Here, the same models as in the main part of the paper are again used for the comparison. As can be seen, C-VITON consistently outperforms others in arm and texture generation, both on the VITON as well as on the MPV dataset. Additionally, it can be seen that on MPV, S-WUTON seems to produce images with extreme color saturation in certain cases, e.g. see results in rows 8 and 10. C-VTON does not suffer from such issues.

B. Additional Try-On Examples

To further demonstrate the capabilities of C-VTON, we present additional try-on results on the VITON and MPV datasets in Figures 11 – 16. All presented images are generated based on the test set (consisting of people and target clothing images) used throughout other parts of paper.

Figures 11 and 12 show virtual try-on results in a basic setting, i.e. when transferring a garment without a particular texture to what we subjectively deem a simple pose[1]. Here, different sleeve lengths (from shortest to longest) are shown down the rows. Figures 13 and 14 feature the same simple pose scenario, but with textured target garments. The figures show transfer results with dotted patterns, followed by striped garments, smaller and larger repetitive patterns, graphics and text-based logotypes. Last but not least, Figures 15 and 16 show sample results in more challenging poses, ranging from poses, where arms slightly occlude the body, or are in an upward position, to poses, where the arms and hands cover a large portion of the body and corresponding clothing.

Note how C-VTON is able to generate realistic results in all settings described above. Specifically, observe the quality of the warped graphics, textures and texts on the transferred clothing across all presented results and especially in Figures 15 and 16, where synthesis results with difficult poses are displayed.

Next, we conduct a study involving multiple subjects and target garments. The aim of the study is not only to analyze the capabilities of C-VTON when transferring the same target clothing to multiple subjects, but to also demonstrate clothing transfers from short to long sleeves (Figures 17 and 18) and vice versa (Figures 19 and 20). For the study, we carefully chose a subset of test images that feature a wide range of variation with respect to arm positions, clothing that people are initially wearing and a variety of target garments with short and long sleeves. It can be seen from the presented results that C-VTON consistently generates high-quality results regardless of the target clothing and that differing sleeve lengths have little effect on the quality of the generated results. V-neck and sleeveless clothes are also largely synthesized correctly.

C. Results on the VITON-HD Dataset

The results presented in the main part of the paper were generated on the VITON and MPV datasets, which contain images of size 256×192 pixels. In this section, we now use a higher resolution dataset, VITON-HD, to illustrate how C-VTON performs with larger sized images. Note that VITON-HD is high-resolution version of the original VITON dataset featuring images of 512×368 pixels in size and split between a training set of $14,221$ images and a testing set of 2032 images. The dataset has only recently become popular due to being more difficult to train on. Similarly as in the main part of the paper, we again exclude all images that are present both in the training and testing data from the experimental assessment and report results over a clean test set of 416 test images. To accommodate the higher resolution images, we slightly change the topology of the discriminators needed when training C-VTON and add 1 additional ResNet block to the matching and patch discriminators, D_{mth} and D_{ptc}, and 2 additional ResNet blocks to the segmentation discriminator, D_{seg}.

The results, presented in Figure 21, show that C-VTON can generate crisp results with minimal artefacts and generally comparable image and try-on quality as observed with models trained on the VITON and MPV datasets. In terms of quantitative scores, C-VTON achieves a FID score of 21.929 and and LPIPS score of 0.139 ± 0.035 (LPIPS) over the cleaned VITON-HD test set, which is slightly worse than results for the C-VTON model trained on the VITON

[1] From the perspective of virtual try-on models, simple poses include arm configurations that do not substantially cover the initial clothing, so as to interfere with the segmentation/parsing procedure needed by most try-on models.

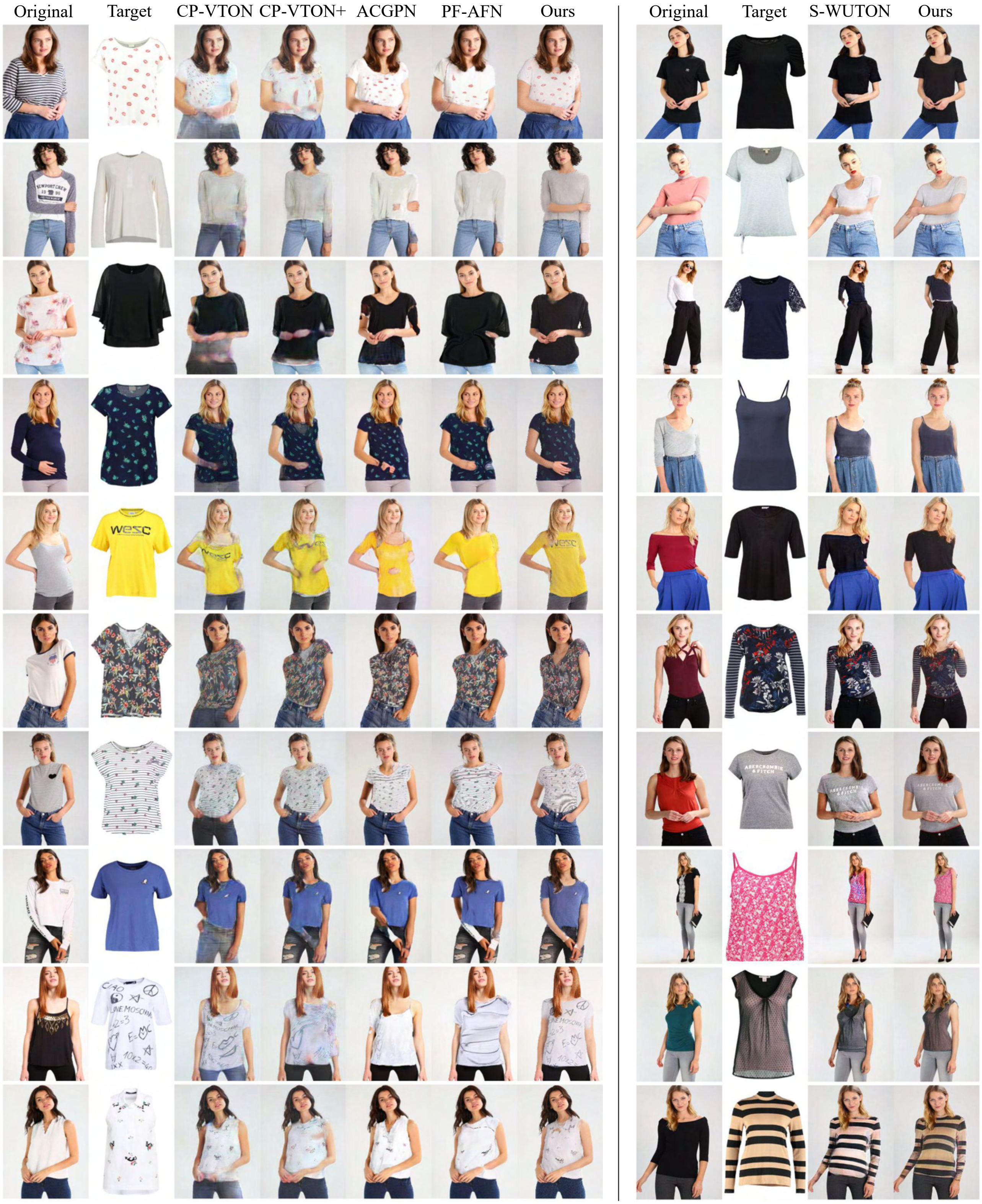

Figure 10. Additional comparisons on the VITON (left) and MPV (right) datasets. The proposed C-VTON model fares exceedingly well when synthesizing people in difficult positions and with detailed on on-garment graphics. The figure is best viewed electronically and zoomed-in for details.

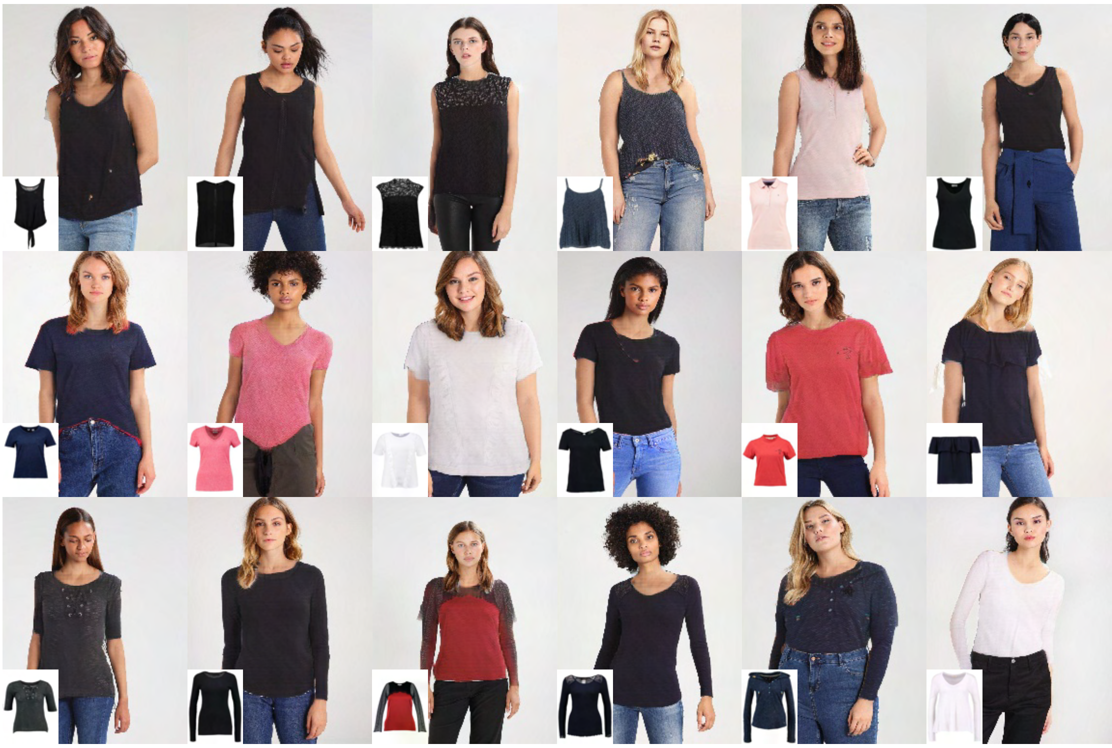

Figure 11. Example results from the VITON test dataset containing simple garments and people in simple poses ordered by sleeve length. The image in the bottom left of each example shows the target garment. The figure is best viewed in color.

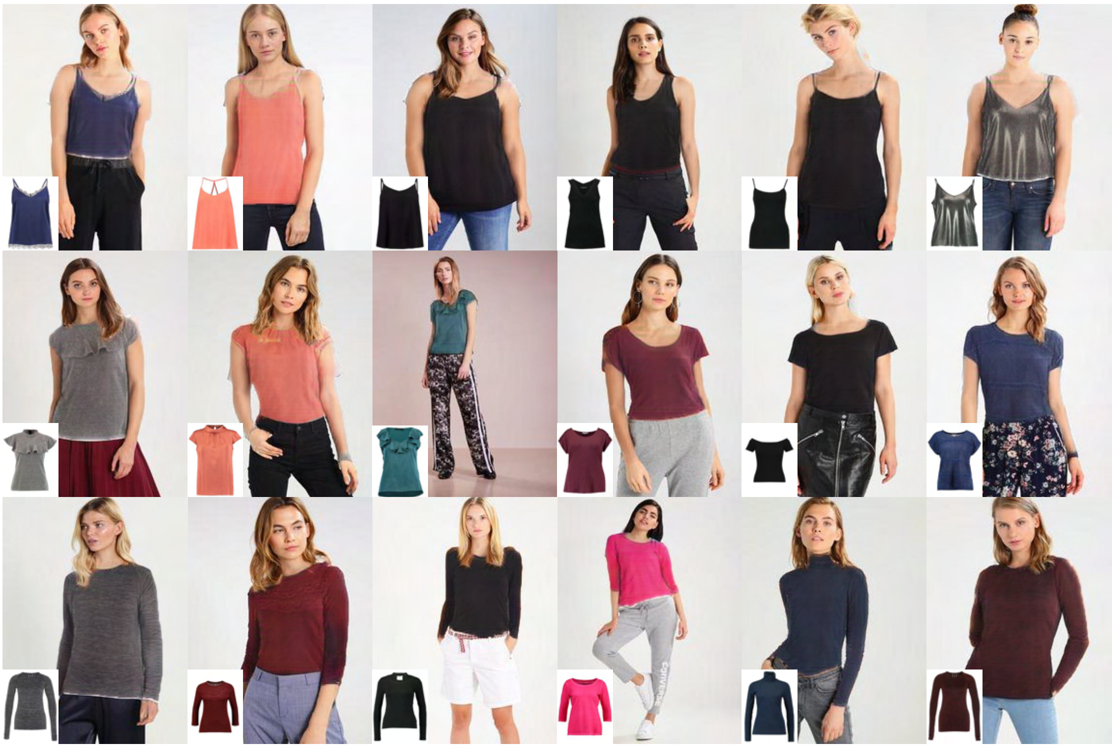

Figure 12. Example results from the MPV test dataset containing simple garments and people in simple poses ordered by sleeve length. The image in the bottom left of each example shows the target garment. The figure is best viewed in color.

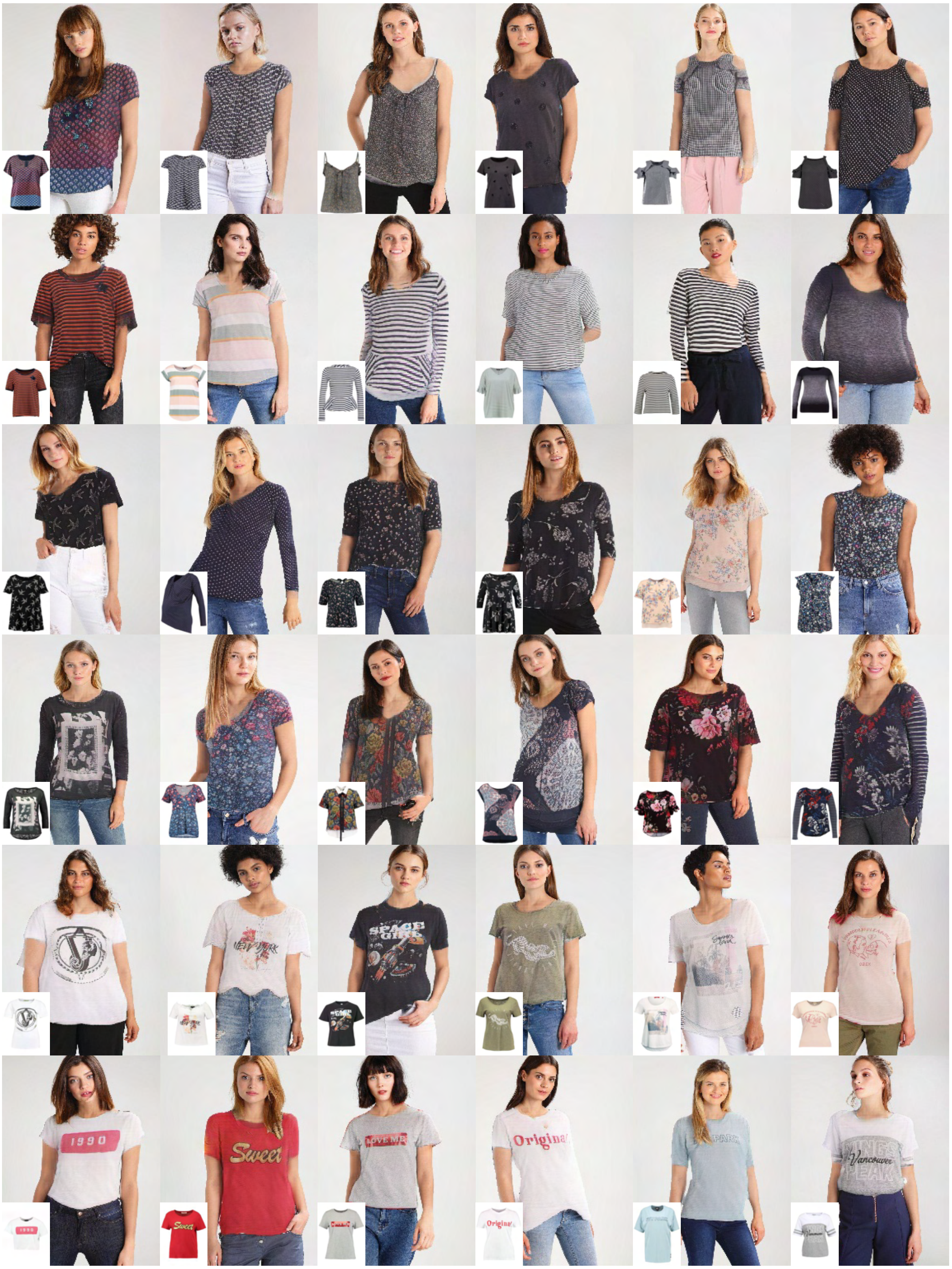

Figure 13. Example results from the VITON test dataset containing textured garments and people in simple poses. The type of on-garment texture displayed differs from row to row to show our model's ability to synthesize a wide array of graphics. The image in the bottom left of each example shows the target garment. Zoom in for details.

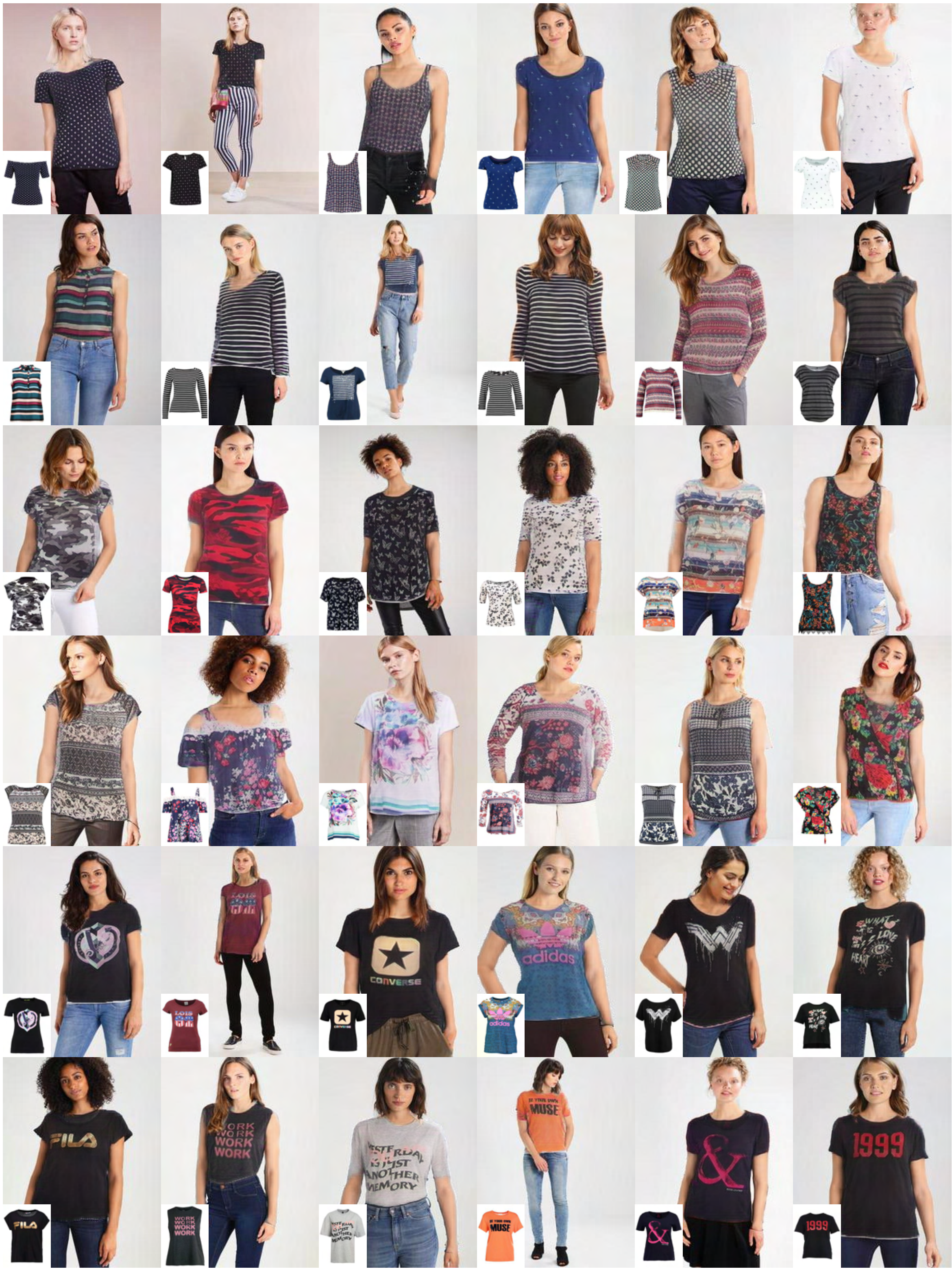

Figure 14. Examples results from the MPV test dataset containing textured garments and people in simple poses. The type of on-garment texture displayed differs from row to row to show our model's ability to synthesize wide array of graphics. The image in the bottom left of each example shows the target garment. Zoom in for details.

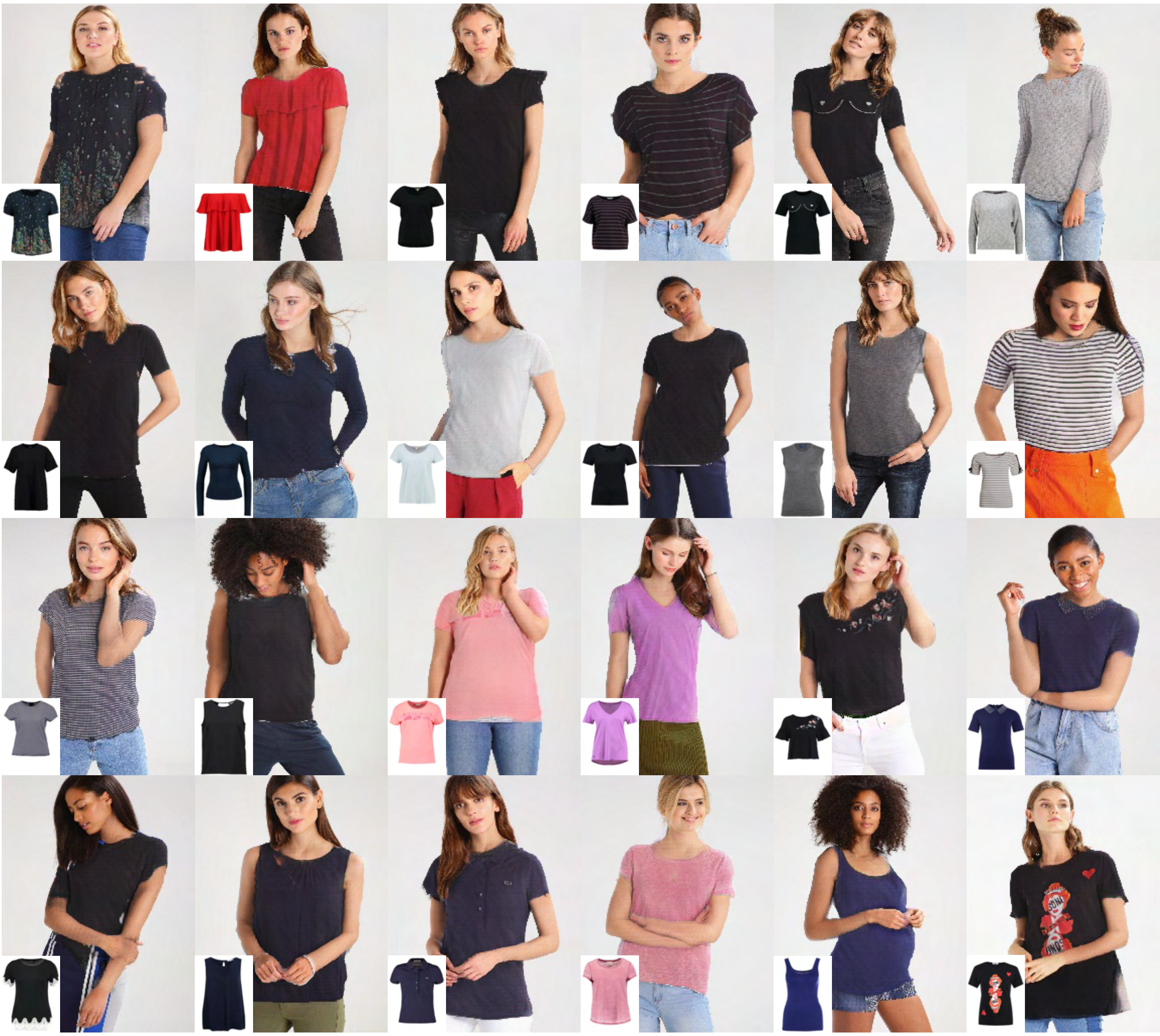

Figure 15. Examples results from the VITON test set containing people in what we consider difficult poses. The generated images are arranged into 4 distinct poses to show the consistency of the synthesis procedure in a wide variety of conditions. The image in the bottom left of each example, shows the target garment. The figure is best viewed in color and electronically.

dataset.

D. Impact of Design Choices

We attribute a considerable part of the capabilities of C-VTON for realistic image synthesis to the use of the proposed Body-Part Geometric Matcher and reliance on body-part segmentations. To additionally showcase the importance of this contributions, we implement and train two additional C-VTON variants, i.e.,

- The first uses the Geometric Matching Module (GMM) from [34] instead of the proposed Body-Part Geometric Matcher (BPGM). The change in the matching module results in differently warped clothing in the first stage of the pipeline and, therefore, has a significant impact on the final results.

- The second variant uses our BPGM, but relies on clothing segmentations rather than body parts (similarly in spirit to [35, 36]). Because of the different correspondence, this change is again significantly affecting virtual try-on performance.

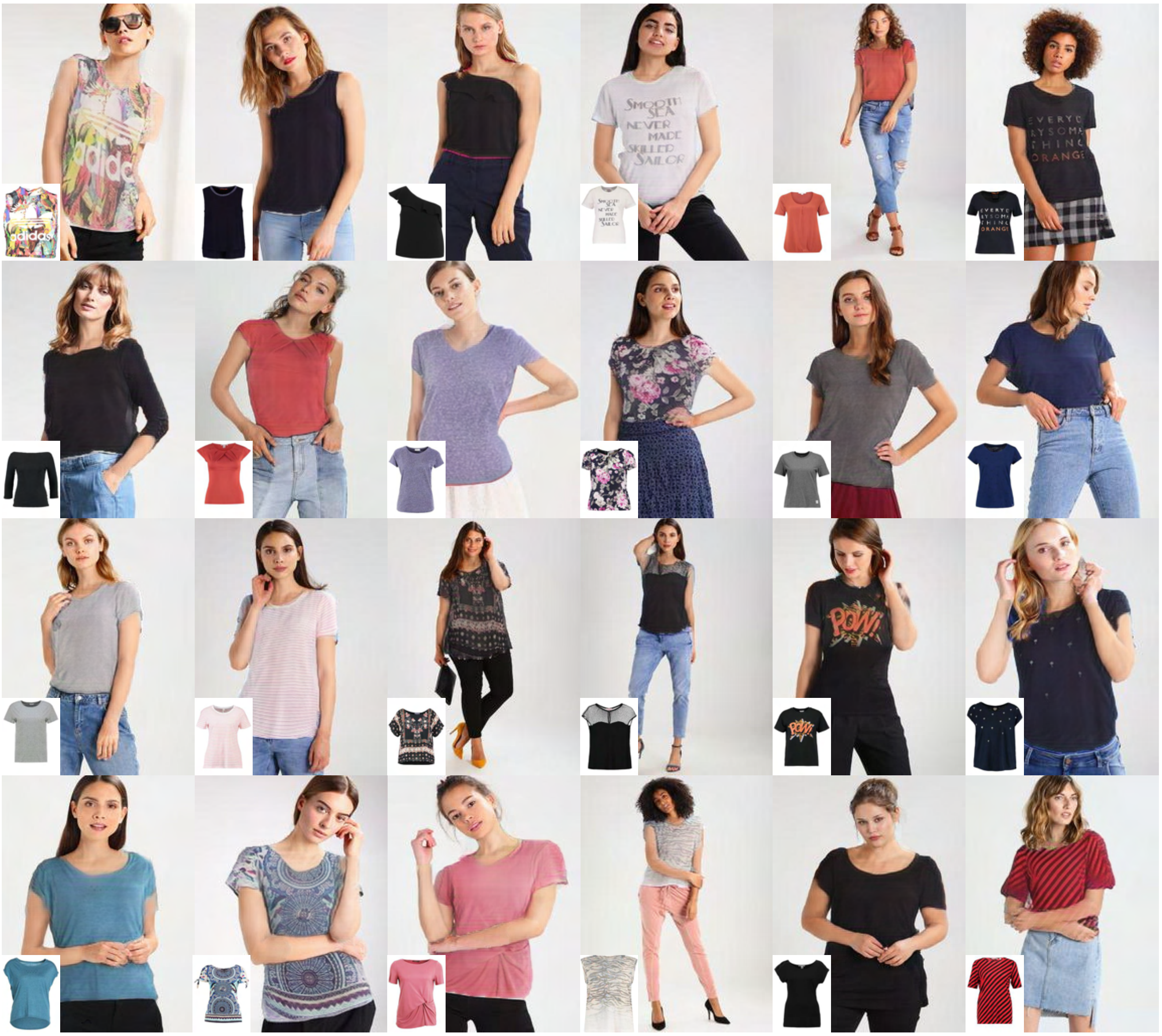

Figure 16. Example C-VTON results from the MPV test set containing people in what we consider difficult poses. The generated images are arranged into 4 distinct poses to show the consistency of the synthesis procedure in a wide variety of conditions. The figure in the bottom left of each example, shows the target garment. The image is best viewed in color and electronically.

Except for the changes mentioned above, the rest of the model, including the network architecture and hyperparameters used, are kept the same. Example results for the GMM versus BPGM comparison are shown in Figures 22 and 23, and for the clothing versus body-part segmentations in Figures 24 and 25.

GMM vs. BPGM. When comparing warping results, it is easy to see that the GMM tries to match the garment shape precisely, leading to artefacts in the neck and arm areas. Additionally, the GMM exhibits problems with preserving on-garment textures and is susceptible to various issues, such as hair occluding the clothing the module is trying to match. Problems with on-garment texture warping can, for example, be seen in the last row of Figure 23, where the stripes on the garment do not match the pose of the subject well after warping. Such issues are not present in the proposed BPGM, where garments are fitted to the general body-area and not to the particular shape of the initial garment/clothing. Quantitative performance scores also point to a deterioration of results when using the GMM instead of the proposed BPGM. On VITON, the GMM-based version of C-VTON results in FID and LPIPS scores of 24.613

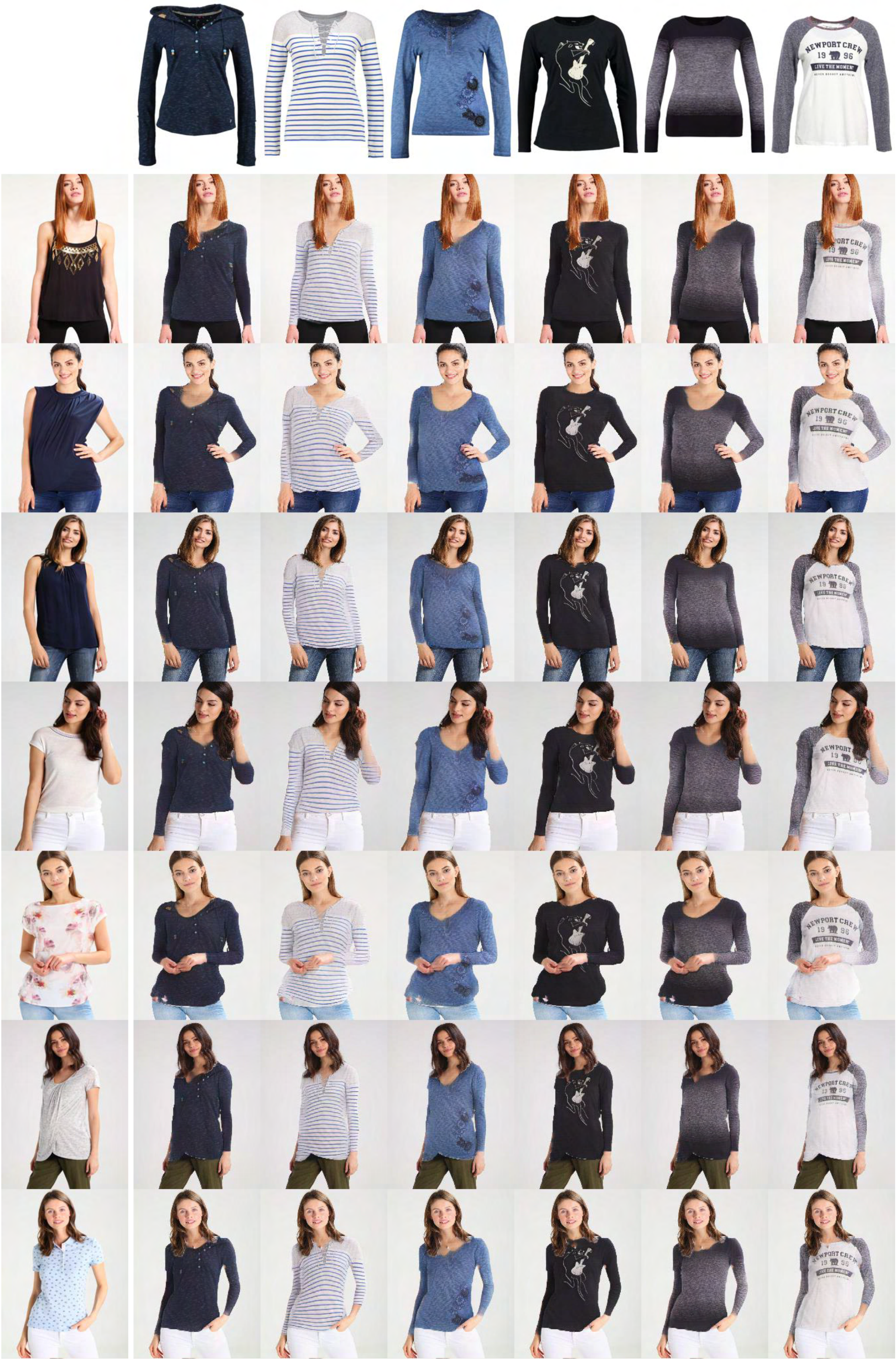

Figure 17. Examples of transfer from short to long-sleeved garments on the VITON dataset. The results show that realistic try-on results are achieved even with images with varying pose and/or different target clothing types. Best viewed electronically and zoomed-in.

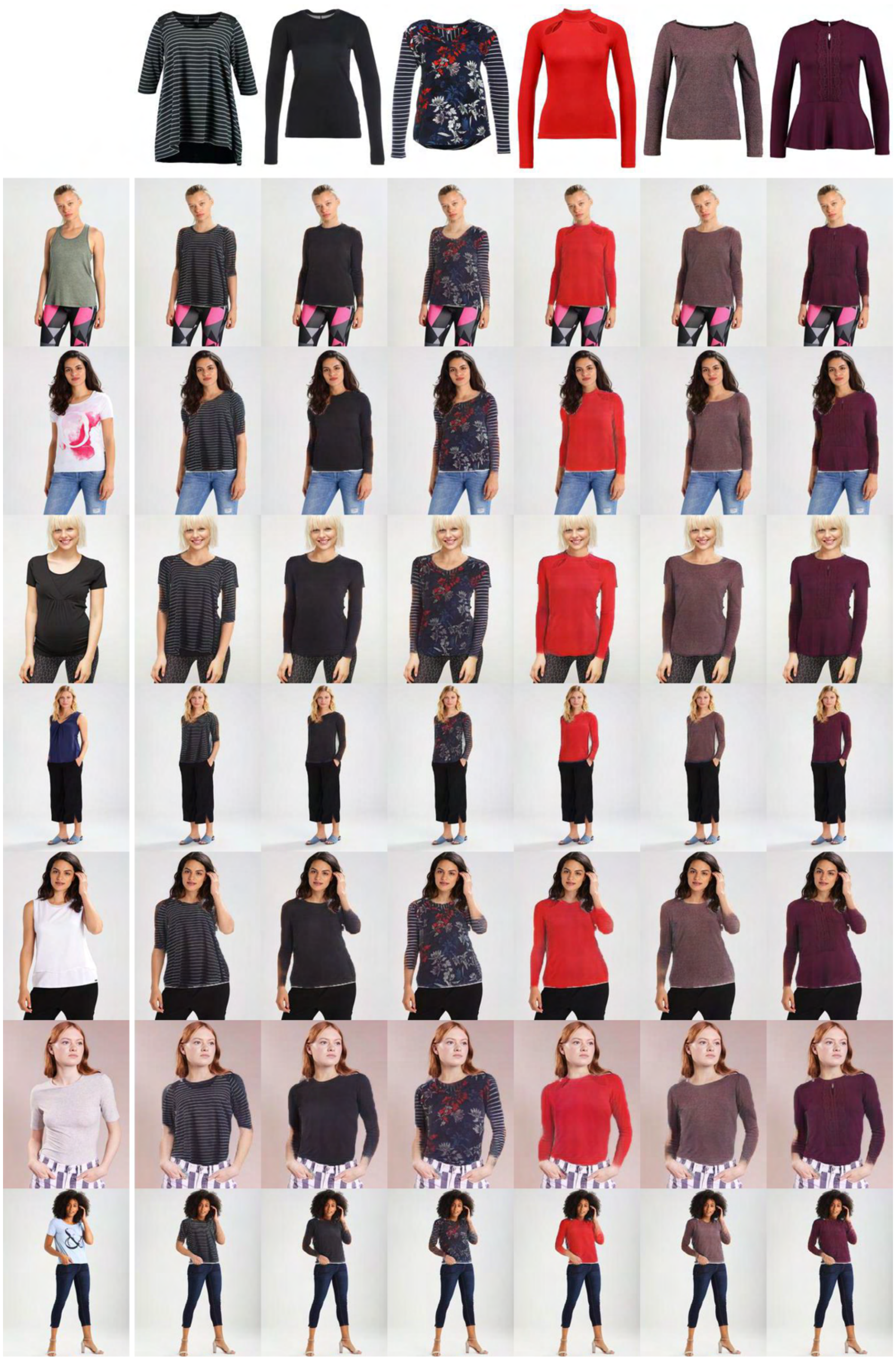

Figure 18. Examples of transfer from short to long-sleeved garments on the MPV dataset. The results show that realistic try-on results are achieved even with images with varying pose and/or different target clothing types. Best viewed electronically and zoomed-in.

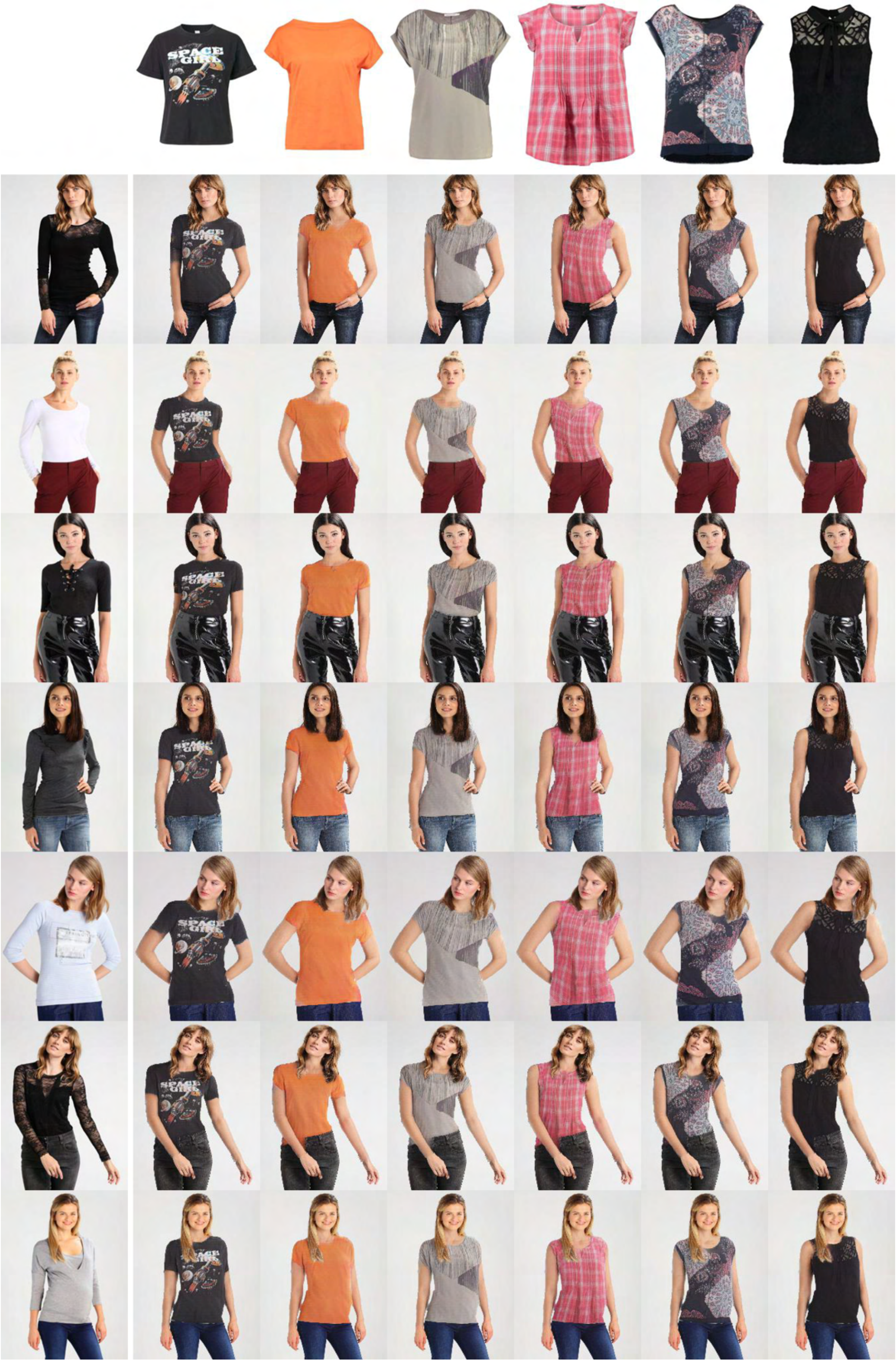

Figure 19. Examples of transfer from long to short-sleeved garments on the VITON dataset. The results show that realistic try-on results are achieved even with images with varying pose and/or different target clothing types. Best viewed electronically and zoomed-in.

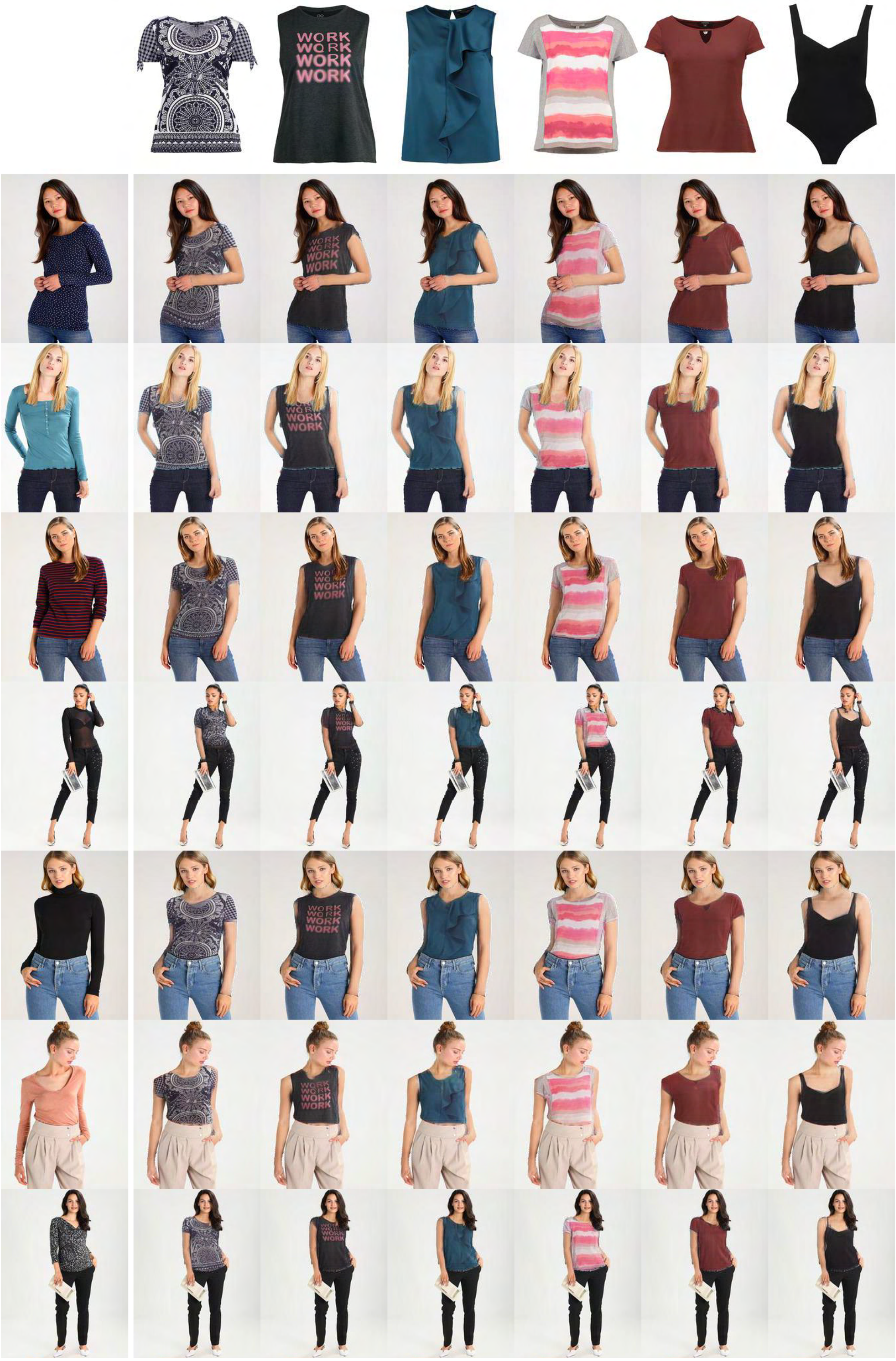

Figure 20. Examples of transfer from long to short-sleeved garments on MPV dataset. The results show that realistic try-on results are achieved even with images with varying pose and/or different target clothing types. Best viewed electronically and zoomed-in.

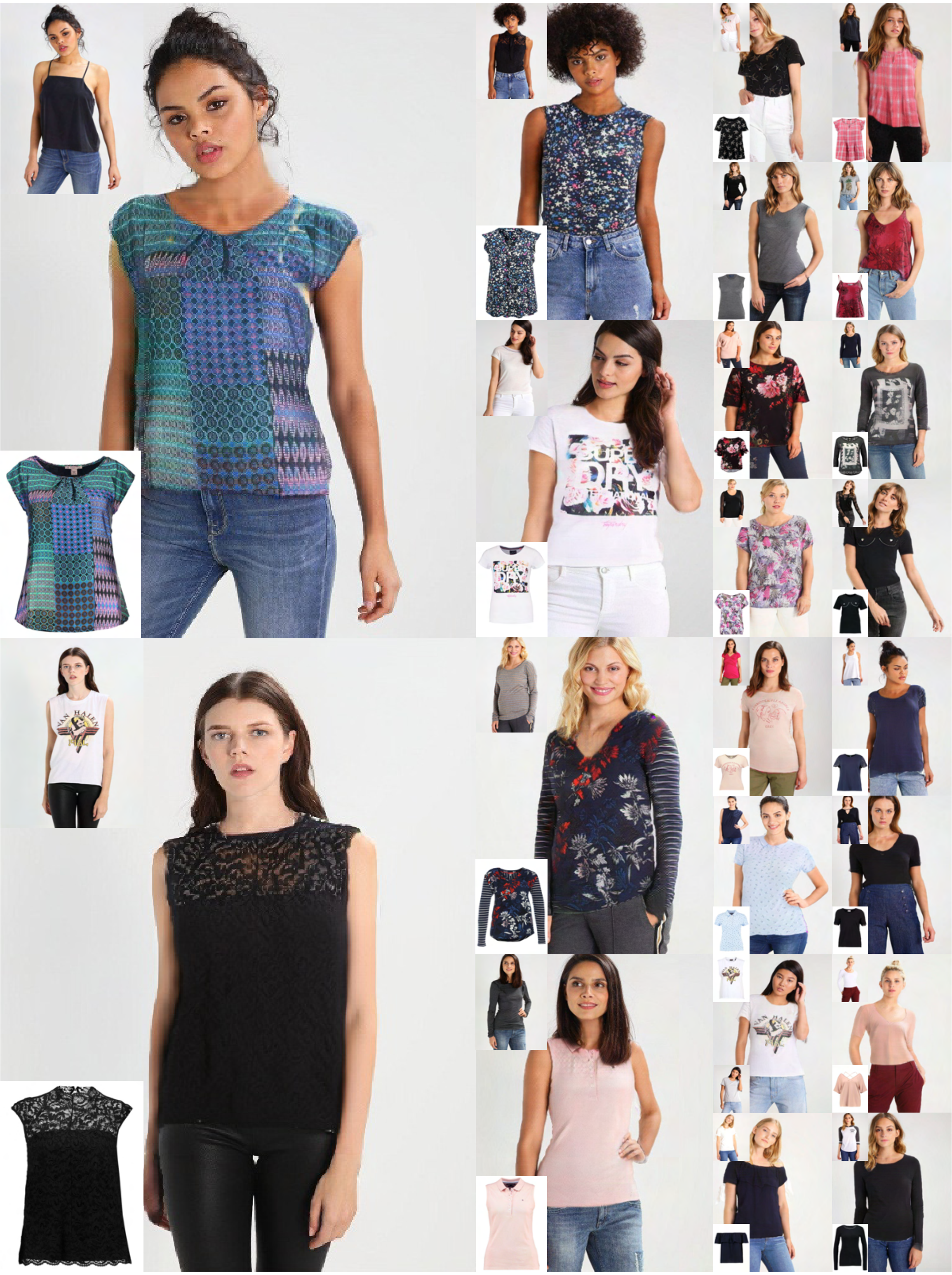

Figure 21. Sample results generated by C-VTON on the VITON-HD dataset. Even when synthesizing higher resolution images, C-VTON produces sharp images with comparable transfer quality to models trained on regular resolution. For each sample result presented, the image in the upper left shows the original input image, and the image in the bottom left shows the target clothing. Results are best viewed electronically.

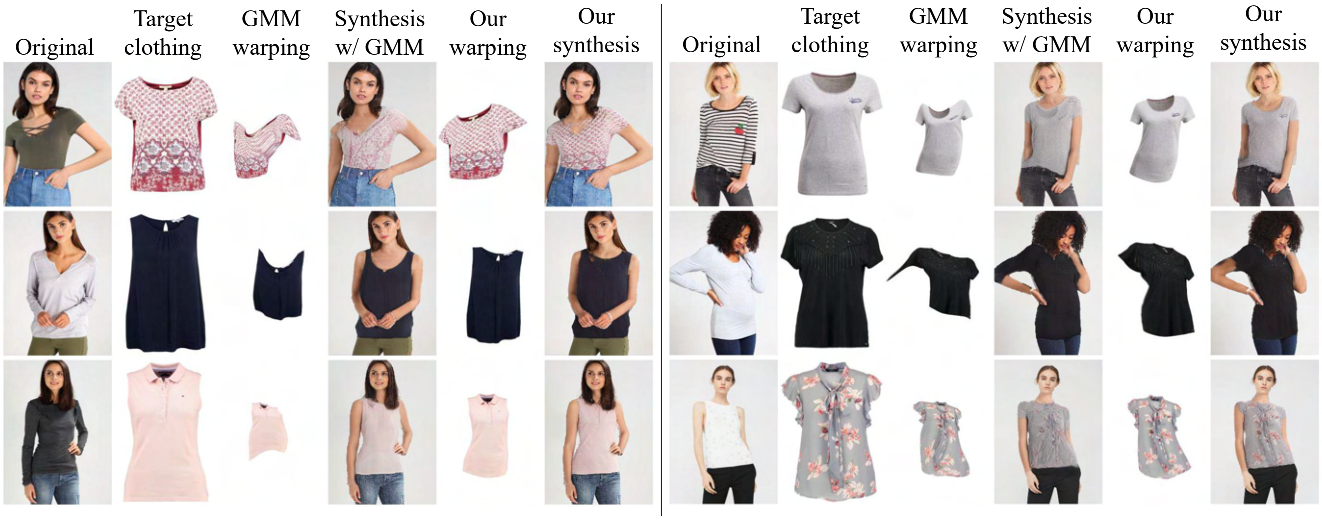

Figure 22. Comparison between the Geometric Matching Module from [34] and the proposed Body-Part Geometric Matcher (BPGM - marked *Our warping*) on the VITON dataset. Even though C-VTON focuses on matching the body-area only, the generated results are better aligned with the person's pose.

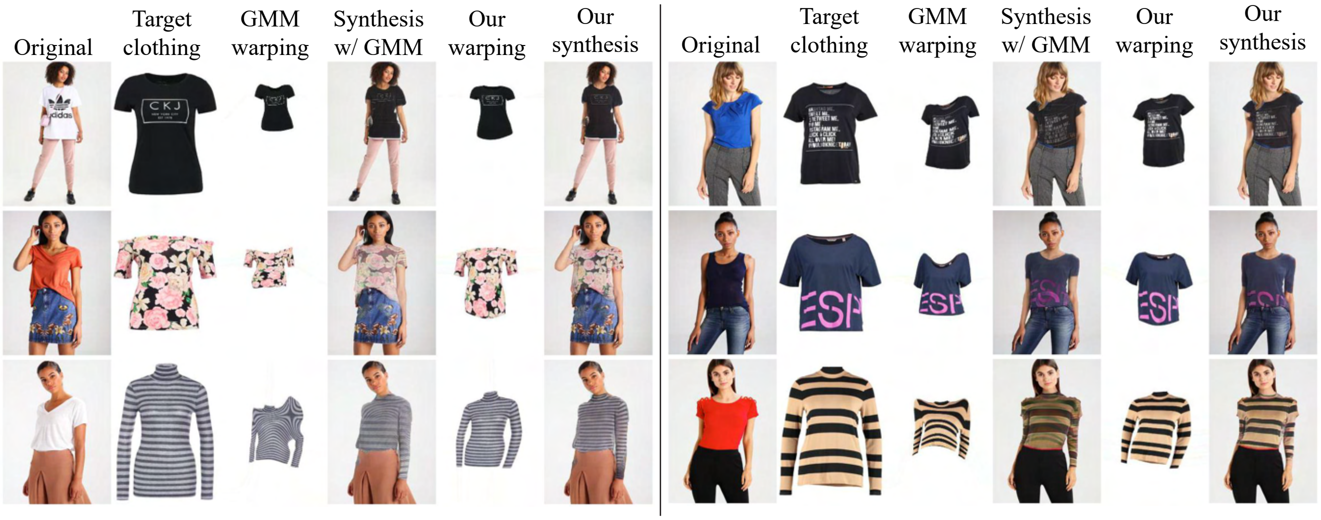

Figure 23. Comparison between the Geometric Matching Module from [34] and the proposed Body-Part Geometric Matcher (BPGM - marked *Our warping*) on the MPV dataset. Even though C-VTON focuses on matching the body-area only, the generated results are better aligned with the person's pose.

and 0.136, respectively (down by 5.078 and 0.028), while on MPV, FID and LPIPS scores of 6.014 and 0.078 are achieved (down by 1.097 and 0.005), respectively.

Body-part vs. Clothing Segmentation. A comparison between the use of body-part and clothing segmentations and their effect on the C-VTON model is shown in Figures 24 and 25. It is evident from the presented results that clothing segmentations are much noisier compared to body-part segmentations which has an adverse effect on the end results. When using clothing segmentations, the synthesized images are more susceptible to incorrect garment-swapping areas and are negatively affected by incorrect information on arm and hand areas that mostly impact image generation in difficult poses. Using body-part segmentations improves upon these issues by providing consistent information to the generator. Interestingly, quantitative performance scores show only minor improvements when comparing the two configurations. On VITON, FID and LPIPS scores are 20.766 and 0.116 when using clothing segmentations compared to 19.535 and 0.108 with body-part segmentations, while on MPV these scores are 5.092 and 0.071 compared to 4.846 and 0.73 with body-part seg-

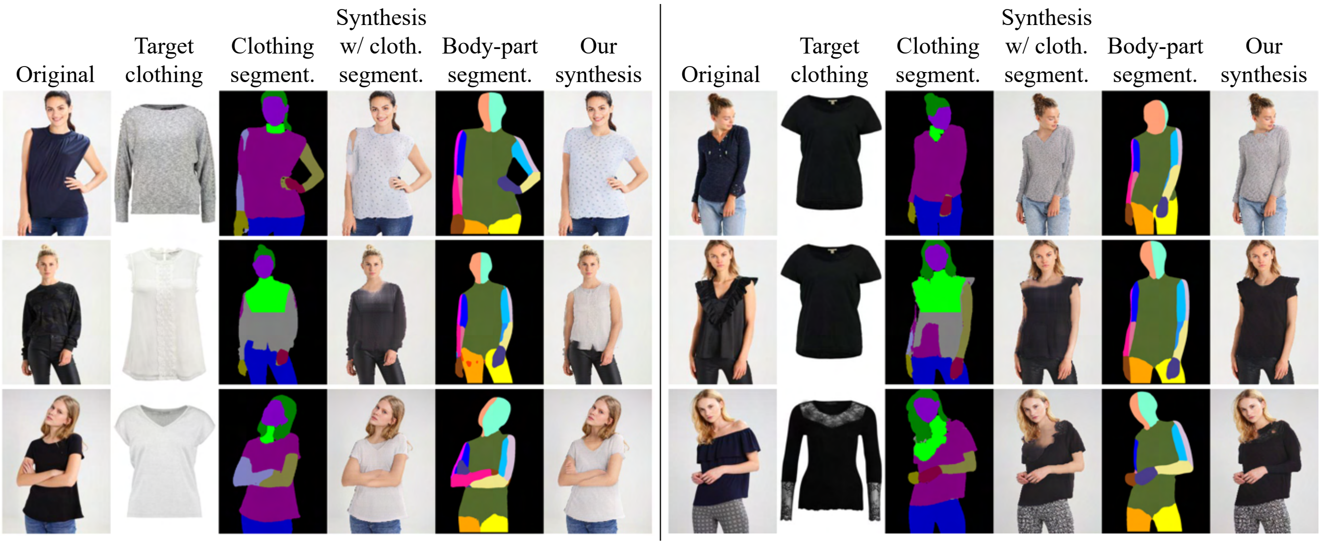

Figure 24. Comparison of clothing and body-part segmentations and their effect on the syntehsis quality of C-VTON on the VITON dataset. The use of clothing segmentations affects sleeve generation, garment edges and arm generation in difficult poses. The figure is best viewed in color and zoomed in.

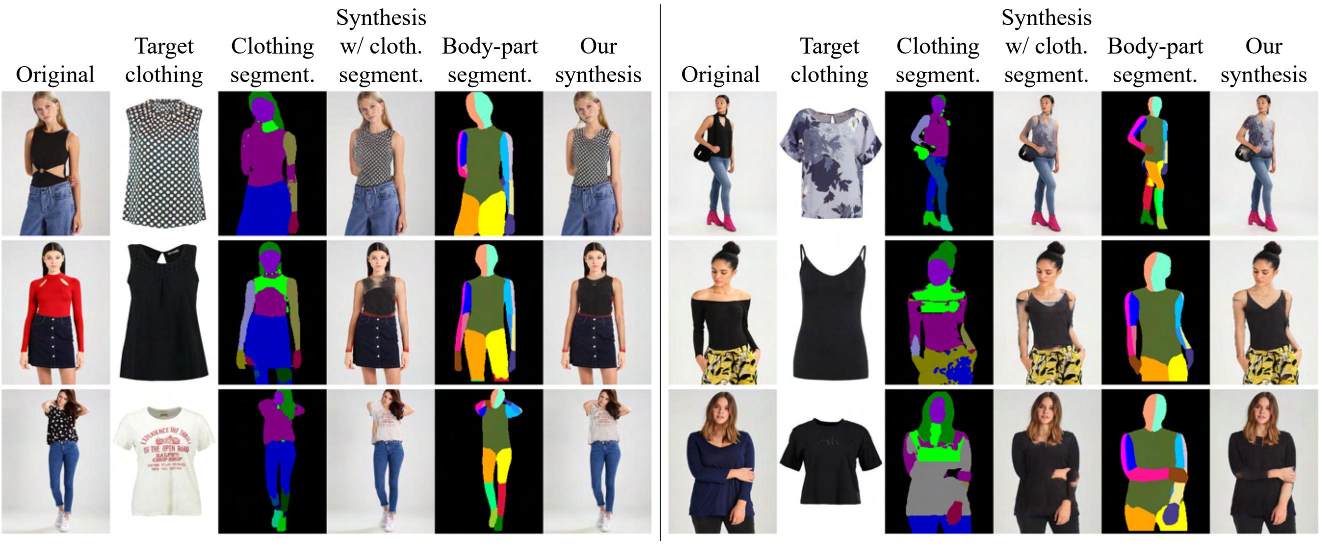

Figure 25. Comparisons of clothing and body-part segmentations and their effect on the syntehsis quality of C-VTON on the MPV dataset. The use of clothing segmentations affects sleeve generation, garment edges and arm generation in difficult poses. The figure is best viewed in color and zoomed in.

mentations, respectively.

E. Analysis of Limitations

Last but not least, we analyze the limitations of C-VTON and present a comparative evaluation with competing approaches in Figure 26. As can be seen, the main issues (on both datasets) can be categorized as stemming from three different causes:

- *Preprocessing*: incorrect clothing segmentations generated during the preprocessing stage used for generating the masked input image I_m that produce erroneous inputs for the context-aware generator (CAG);
- *Input Data Characteristics*: loose clothing on the subjects in the original input image that make it difficult to infer the correct body shape and transfer the target clothing without visual (shape-induced) artefacts, and
- *Target Clothing Characteristics*: the inability of C-VTON to identify the backside of the target garment

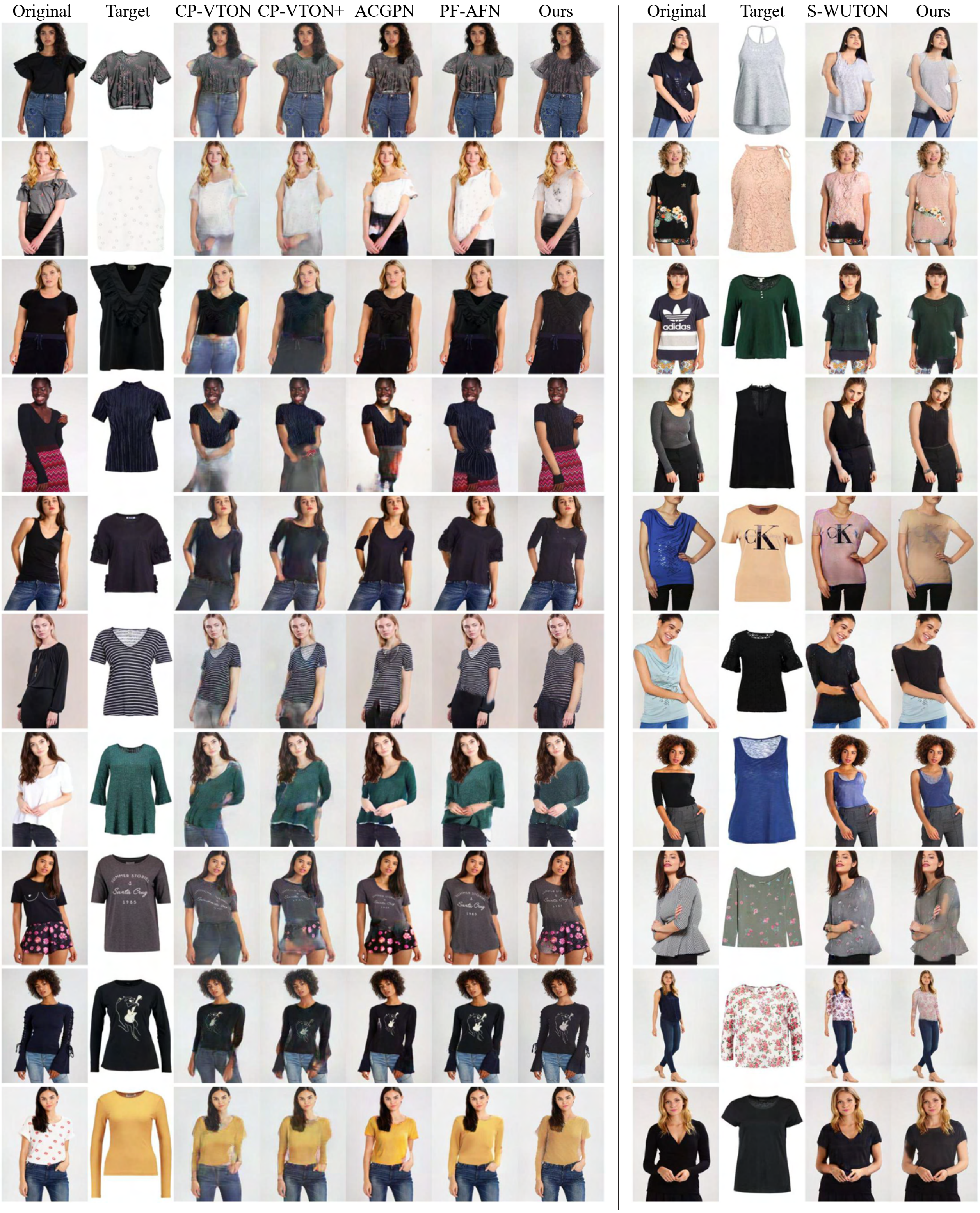

Figure 26. Examples of some of the limitations of C-VTON which entail blurry clothing edges and garments being only partially transferred. Note, however, that the competing methods included in the analysis have problems with realistic synthesis with the presented images as well.

  - especially with clothing with deeper necklines.

These causes lead to unrealistic and soft garment edges on the synthesized images or incorrectly synthesized clothing-types and V-neck areas improperly rendered. Additionally, arms of dark-skinned people, which are the minority in the dataset, are often synthesized with lighter skin tones. However, similar issues are also present with the considered competing techniques, which often result in even more unrealistic results, as illustrated by the examples in Figure 26. For instance, most competing models also exhibit issues caused by the inability to distinguish the backside of the target clothing C from the front and render sleeves inconsistently as well. The parser-free method, PF-AFN, does generate sharper clothing edges when our approach suffers from poor clothing segmentations, but nonetheless often estimates the edges incorrectly. Overall, we observe that even when C-VTON generates less convincing results, it still mostly outperforms or at least matches the quality of the results generated by the competing methods.