UMFuse: Unified Multi View Fusion for Human Editing applications Supplementary

Rishabh Jain
MDSR Adobe

Mayur Hemani
MDSR Adobe

Duygu Ceylan
Adobe Research

Krishna Kumar Singh
Adobe Research

Jingwan Lu
Adobe Research

Mausoom Sarkar
MDSR Adobe

Balaji Krishnamurthy
MDSR Adobe

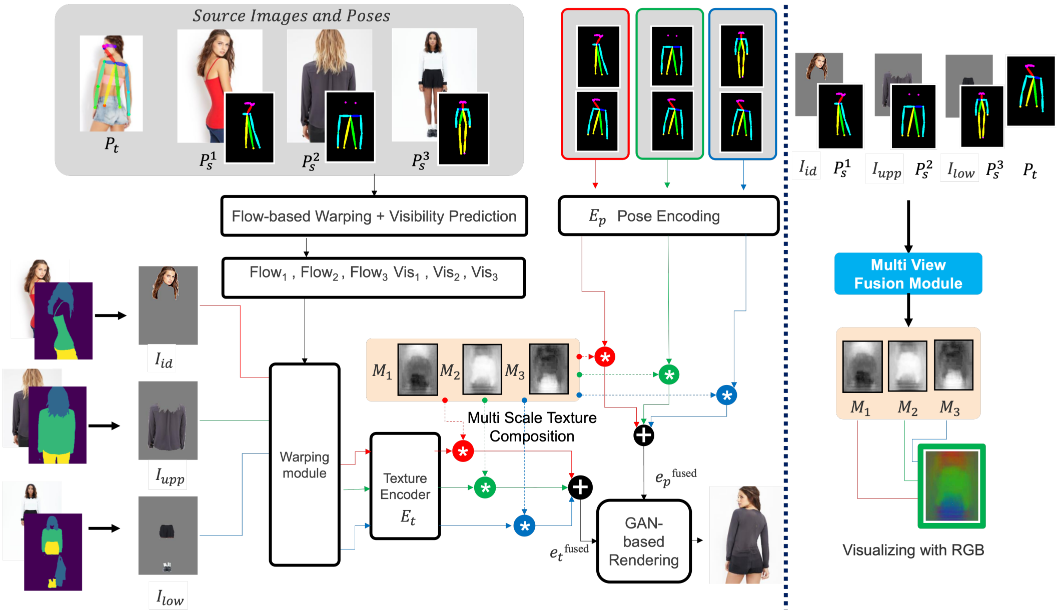

Figure 1: The Mix-and-Match inference task uses an off-the-shelf human body parsing solution[5] to prepare different fashion components(I_{id}, I_{upp}, I_{low}). These components are then fed to the UMFuse pipeline to generate the composition of these in a novel view(guided by P_t). The training is done with 4-tuples of the same person (Fig 3)

Mix and Match Training and inference: We modify Multi View Human Reposing pipeline so that the network automatically conforms to perform the MMHIG task. For the MMHIG task(Fig 1), the single-view PHIG network works normally to generate the flow fields, VisMap and pose encodings. For the texture encoding, the input source images are first segmented into the desired region by using an off-the-shelf human body parser[5]. These segmented regions and then warped with the predicted flow field. These segmented warped images goes into the texture encoder to produce the final texture encodings at multiple scales($e_{t,l}$). For the Multi-View Fusion module, only the segmented region of source images in different poses are given as input to produce the desired appearance retrieval map(ArMap). The model is trained to generate the combined output in the final target pose with an end to end training objective(Fig 3). We impose the same losses as the Multi-View reposing task for the MMHIG task also.

$$\begin{aligned} L_{MMHIG}(I_p, I_{gt}) =& \alpha_{rec}\|I_p, I_{gt}\|_1 + \alpha_{per}L_{vgg}(I_p, I_{gt}) + \\ & \alpha_{sty}L_{sty}(I_p, I_{gt}) + \alpha_{adv}L_{adv}(I_p, I_{gt}) \end{aligned}$$

For both the learning tasks, the hyper-parameters are set to $\alpha_{rec} = 2$, $\alpha_{per} = 0.5$, $\alpha_{sty} = 2$, $\alpha_{adv} = 1$.

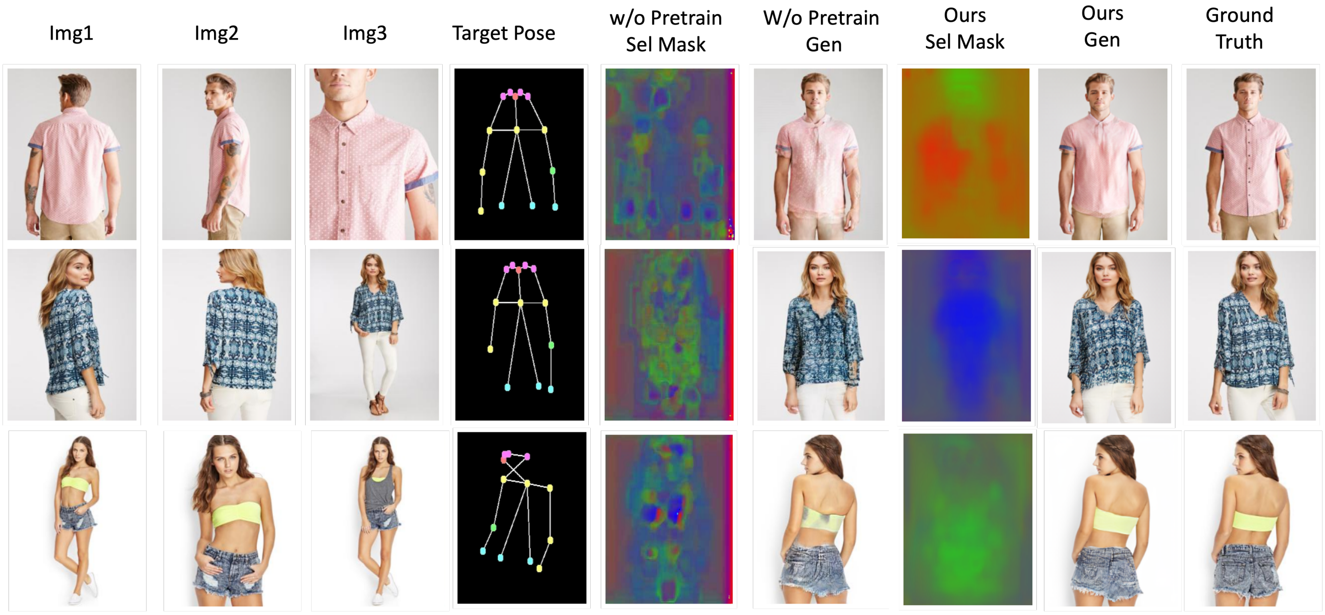

Figure 2: Here, we see the qualitative difference between (Swin+Uper) architecture trained with and without pretraining. Without pretraining, our model architecture was not able to learn the visibility correspondence by itself and produced a random mask at output. Hence, the generated result produced was also of an inferior quality. In contrast, after initializing with the pretrained checkpoint, the selection mask becomes interpretable and the generated output quality also increases significantly

We compare our generated output quantitatively and qualitatively with the DiOr[2] architecture, which performs sequential edits on a person image, as it closely matches the Mix-and-Match task. Accordingly, we use the official implementation [1], providing DiOr's network with multiple source views in the order recommended by the authors (target reposing(P_t) → hair (I_{id}) → top clothing (I_{upp}) → lower clothing (I_{low})).

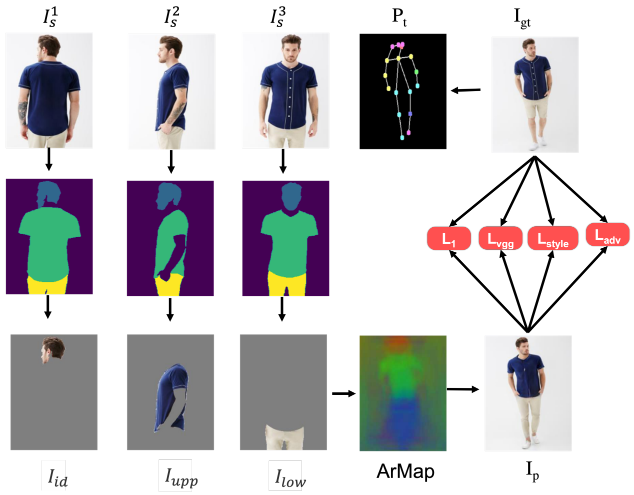

Figure 3: Training pipeline for MMHIG task

Effect of visibility informed preTraining As mentioned in the main paper(Sec 3), we first train the network to improve its attention to visible regions in each input image (as per the target pose). The visibility-informed training is important as it is reasonable to assume that the network will perform better if it retrieves the appearance from a specific source image for the regions which were visible in the corresponding target pose. Hence, the Multi-View Fusion module is first trained to predict the VisMap3D instead of the full appearance retrieval map. VisMap for a single source image is obtained by fitting densepose [3] on the image and matching its UV coordinates with the target image UV coordinates. We combine the visible regions(green in Fig 3, main paper) of all the 3 source images by concatenating them in 3 RGB channels. After concatenation, we normalize the weights across channels to make it compatible with the softmax probability function(For ex, if 1 pixel in target pose is visible in 2 sources images, it's weight will be 0.5 in those 2 channels and 0 in the third). The network is trained with L1 loss for this training and the qualitative benefits of preTraining are shown in Fig 2. We also show quantitative benefits in the main paper(Tab 2) which shows that pre-training improves SSIM(0.732→0.737), PSNR(18.05→18.32), LPIPS(0.178→0.168) and FID(12.77→12.00) metrics for the Swin+Uper[4] combination.

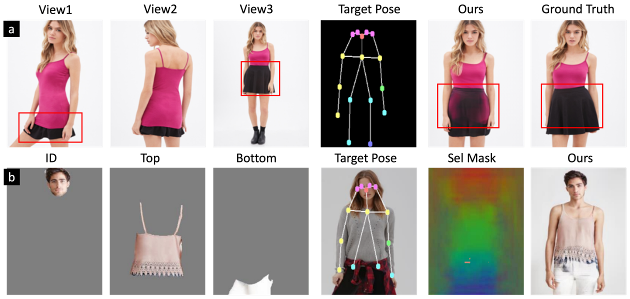

Figure 4: Limitations of our model caused due to different layering in images(a) and gender mismatch for clothing(b)

Limitations Even though UMFuse creates highly accurate images, there are still limitations. In Fig 13 row 1, the network was not able to preserve the collar even after taking all the 3 images as input. In Fig 4 (a), the network produces an interpolation between View1 and View3 due to their different cloth layering style in input. In (b), the Mix&Match output looks unrealistic due to a gender mismatch of male identity and female clothing.

Additional results We provide additional qualitative examples to highlight UMFuse superiority for the Multi View reposing task(Page 4-8), advantages of using 3 views(Page 9-13) and Mix-and-Match task(Page 14-19) in the subsequent pages.

References

[1] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order (github). https://github.com/cuiaiyu/dressing-in-order. 2

[2] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14638–14647, October 2021. 2

[3] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1

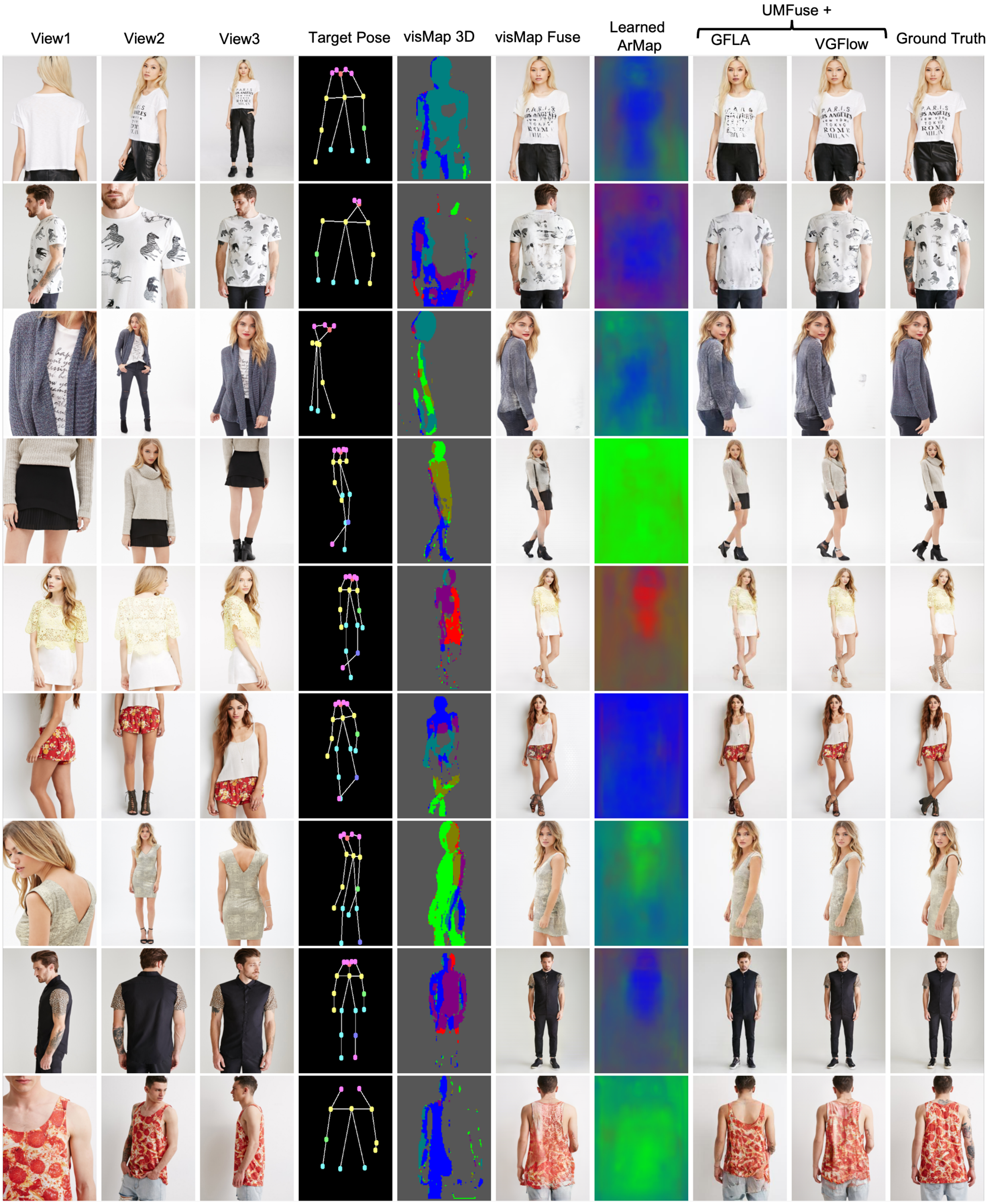

Figure 5: Multi View reposing qualitative examples

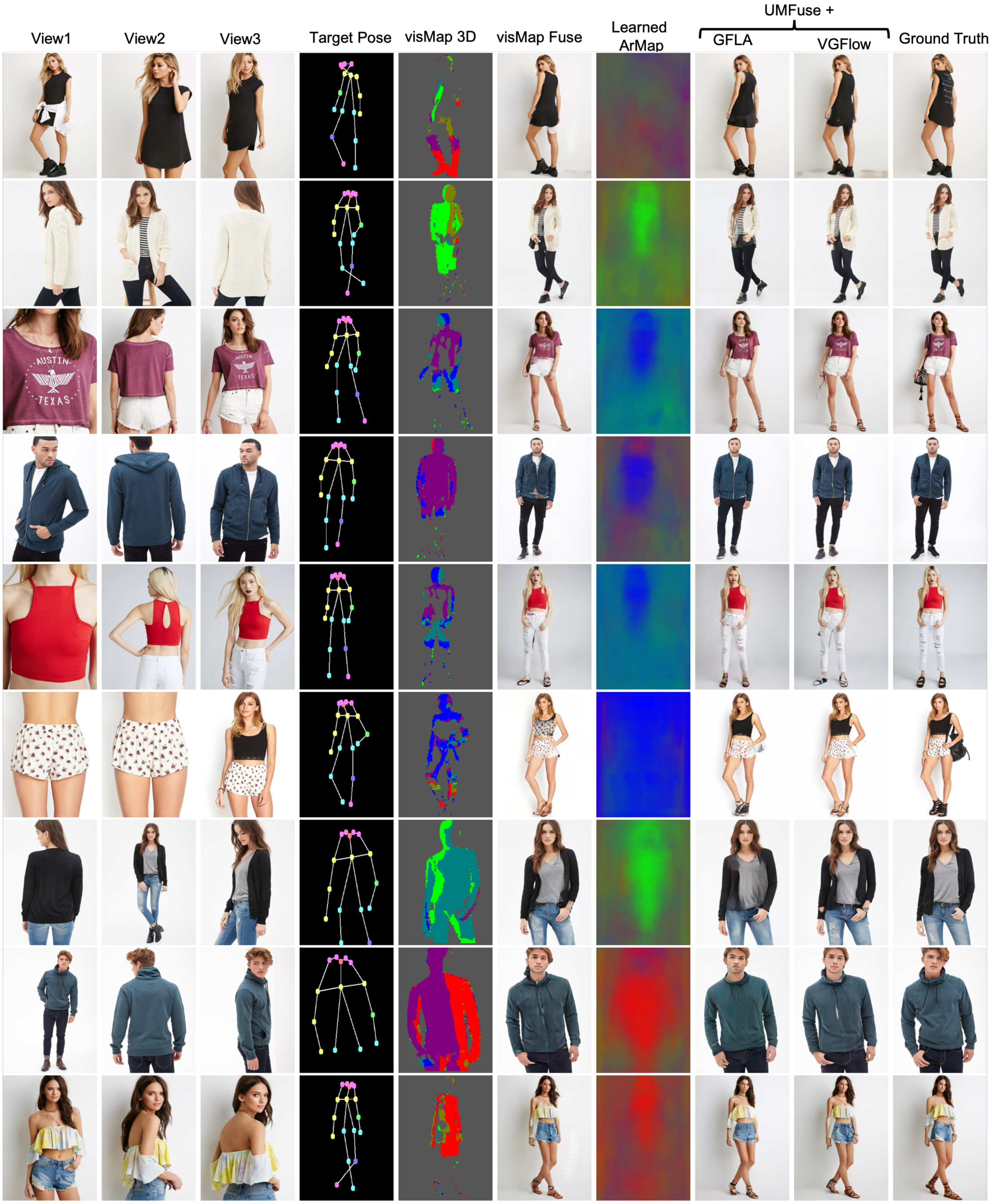

Figure 6: Multi View reposing qualitative examples

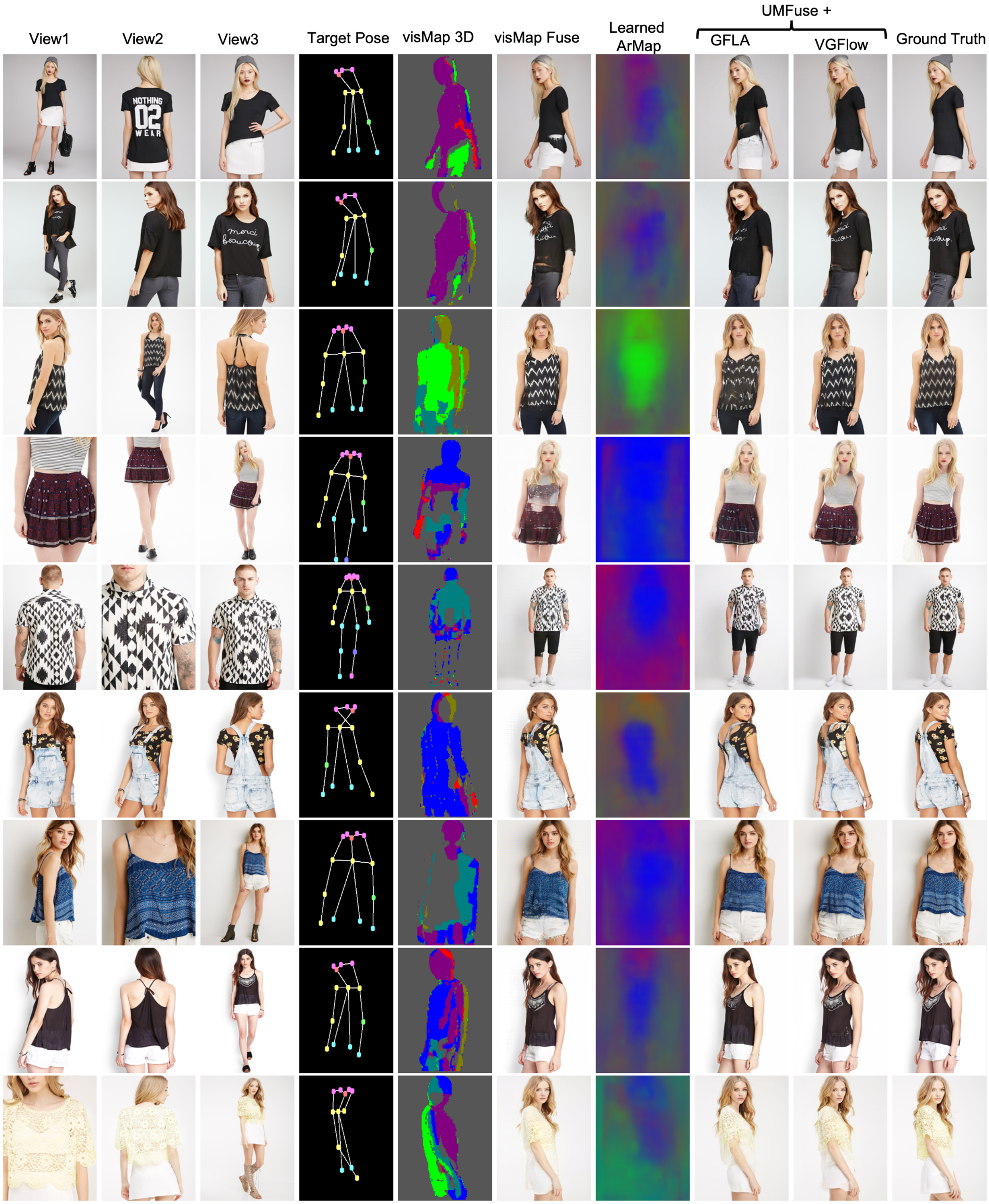

Figure 7: Multi View reposing qualitative examples

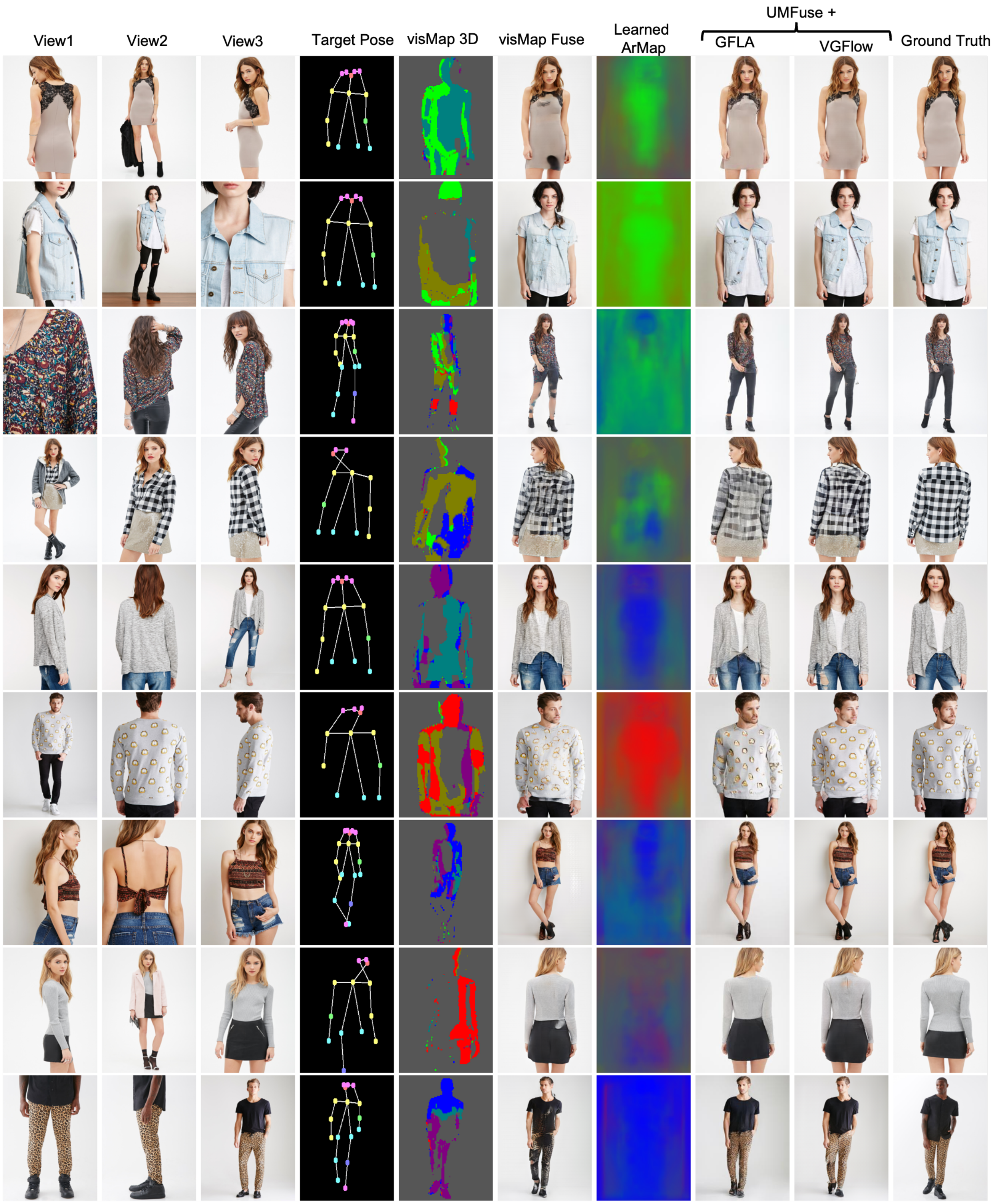

Figure 8: Multi View reposing qualitative examples

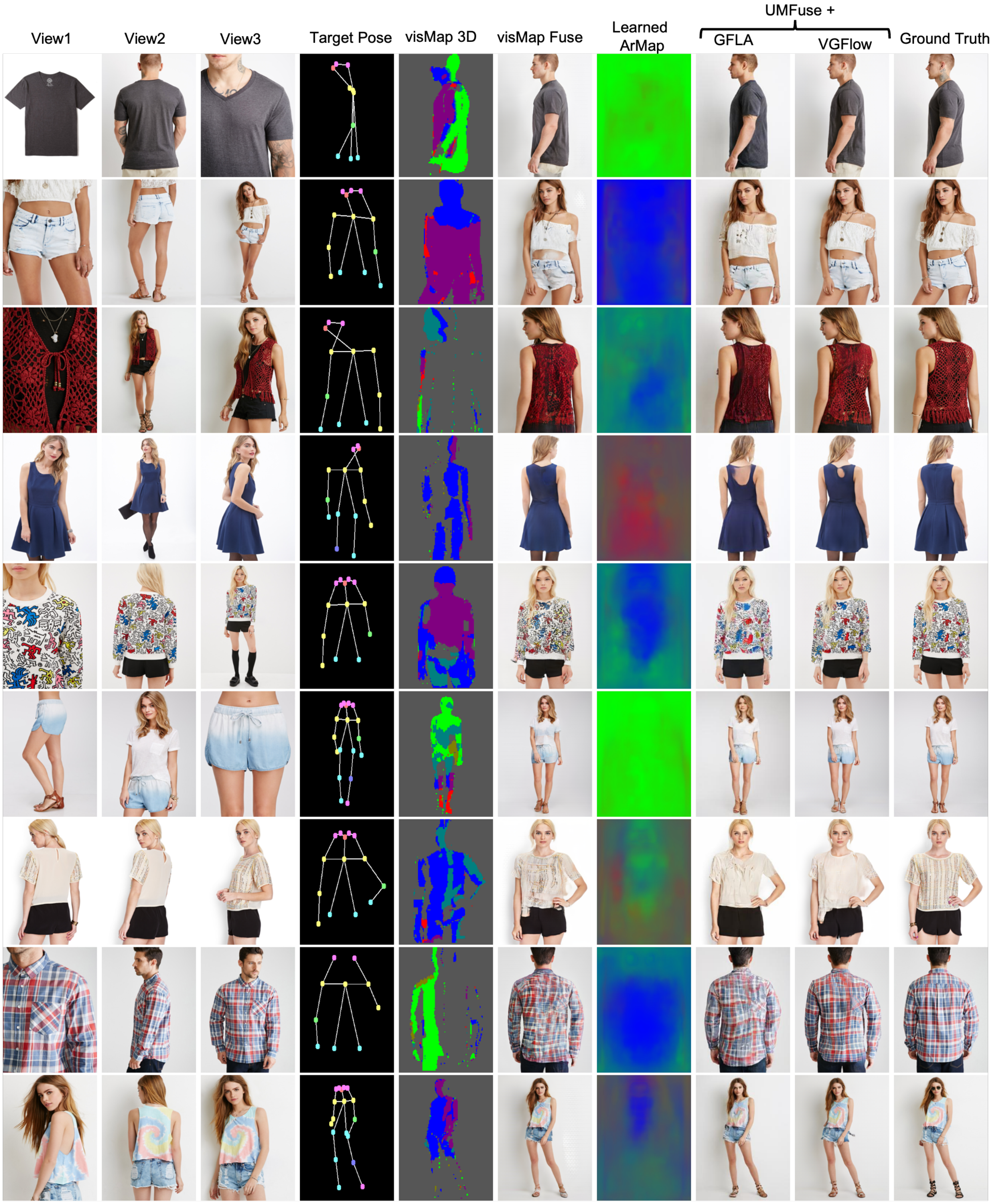

Figure 9: Multi View reposing qualitative examples

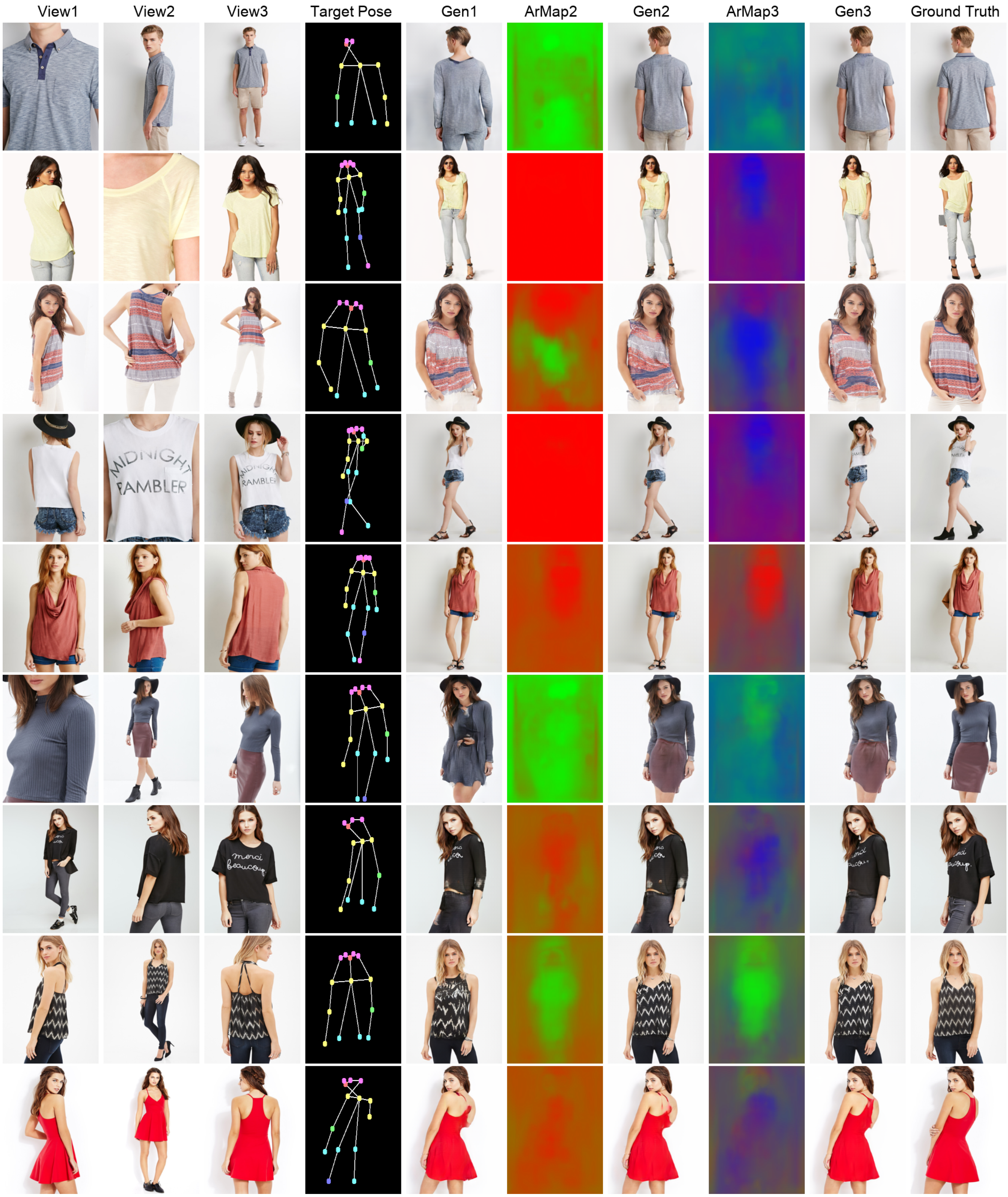

Figure 10: Benefits of using multiple views

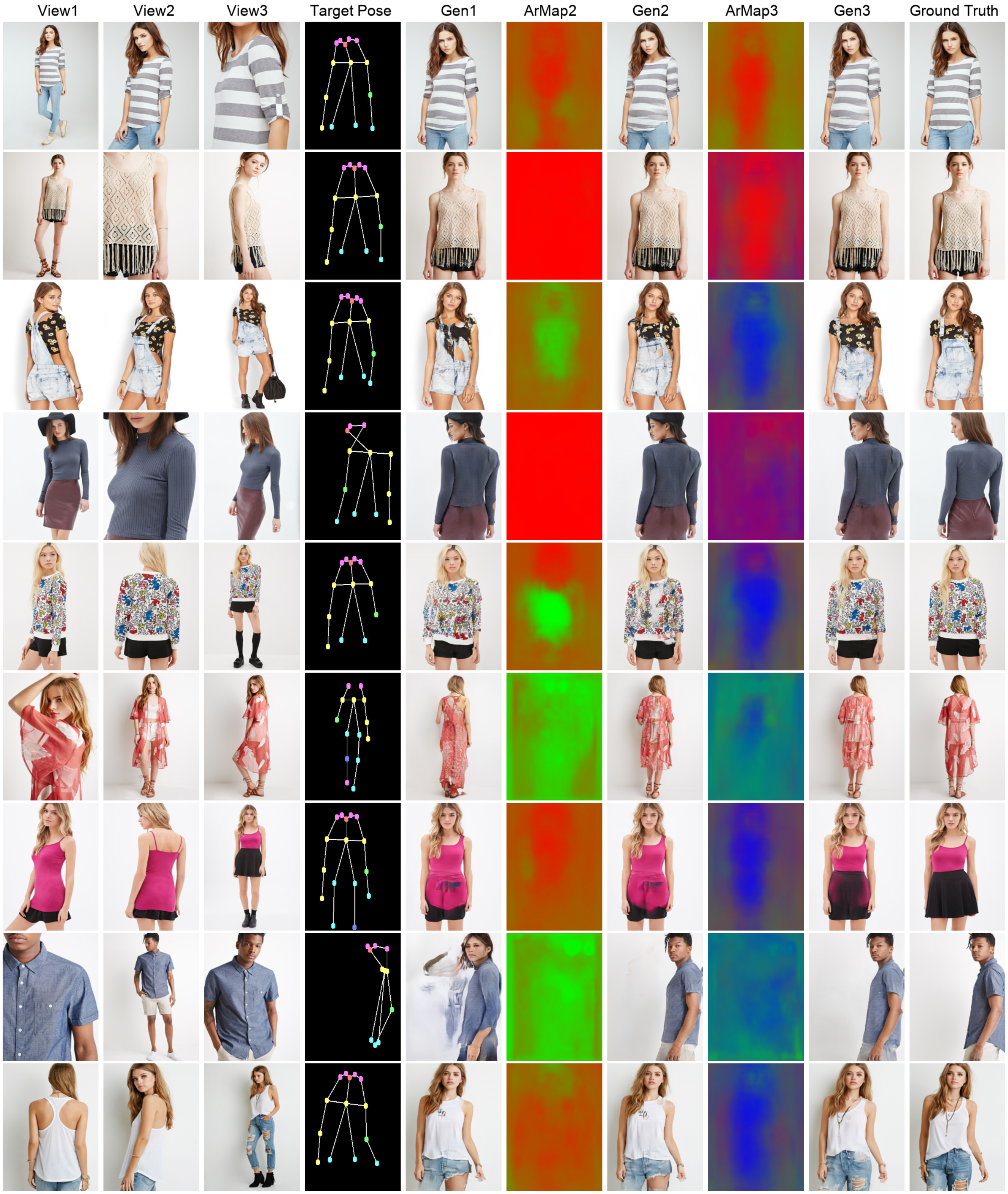

Figure 11: Benefits of using multiple views

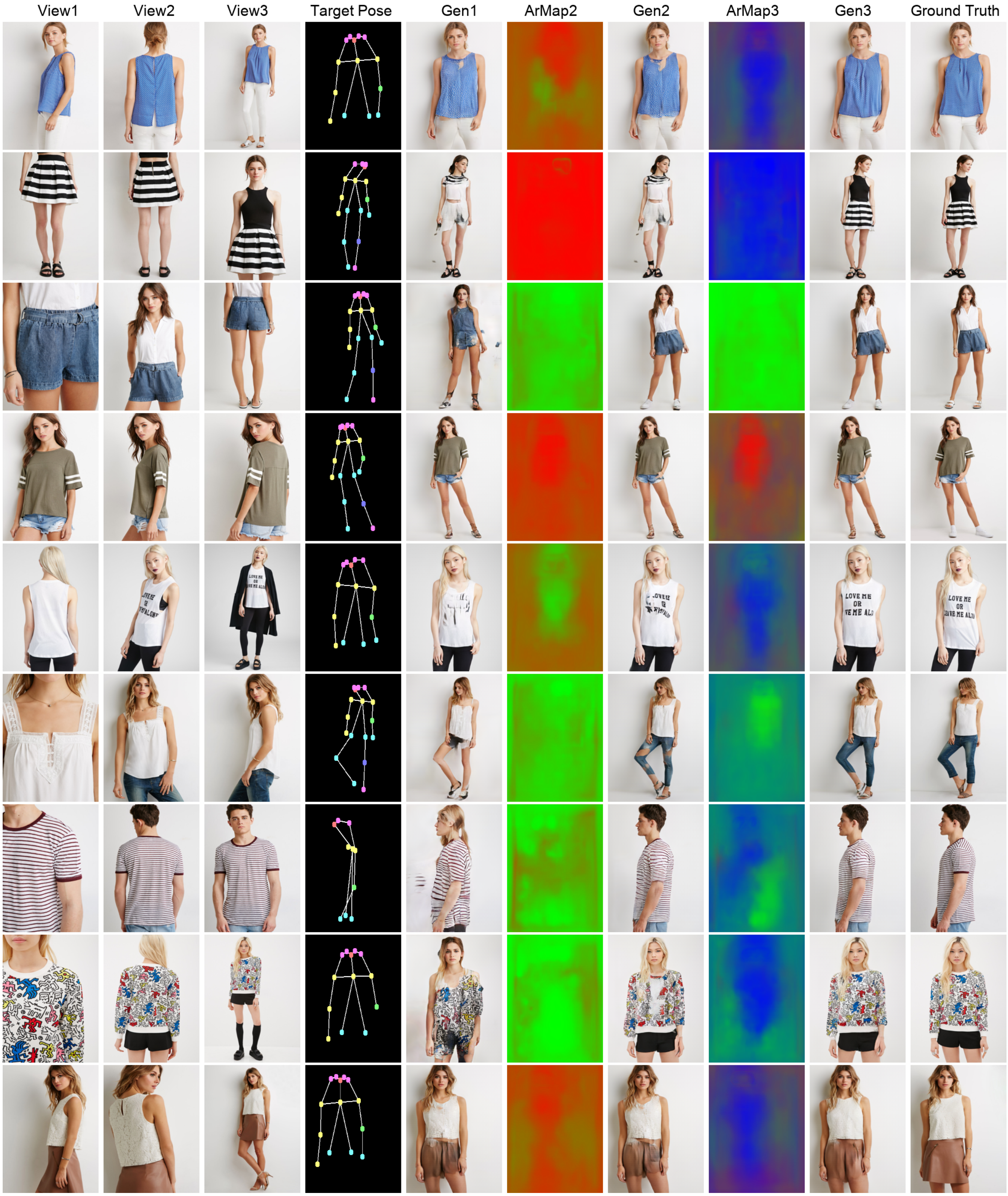

Figure 12: Benefits of using multiple views

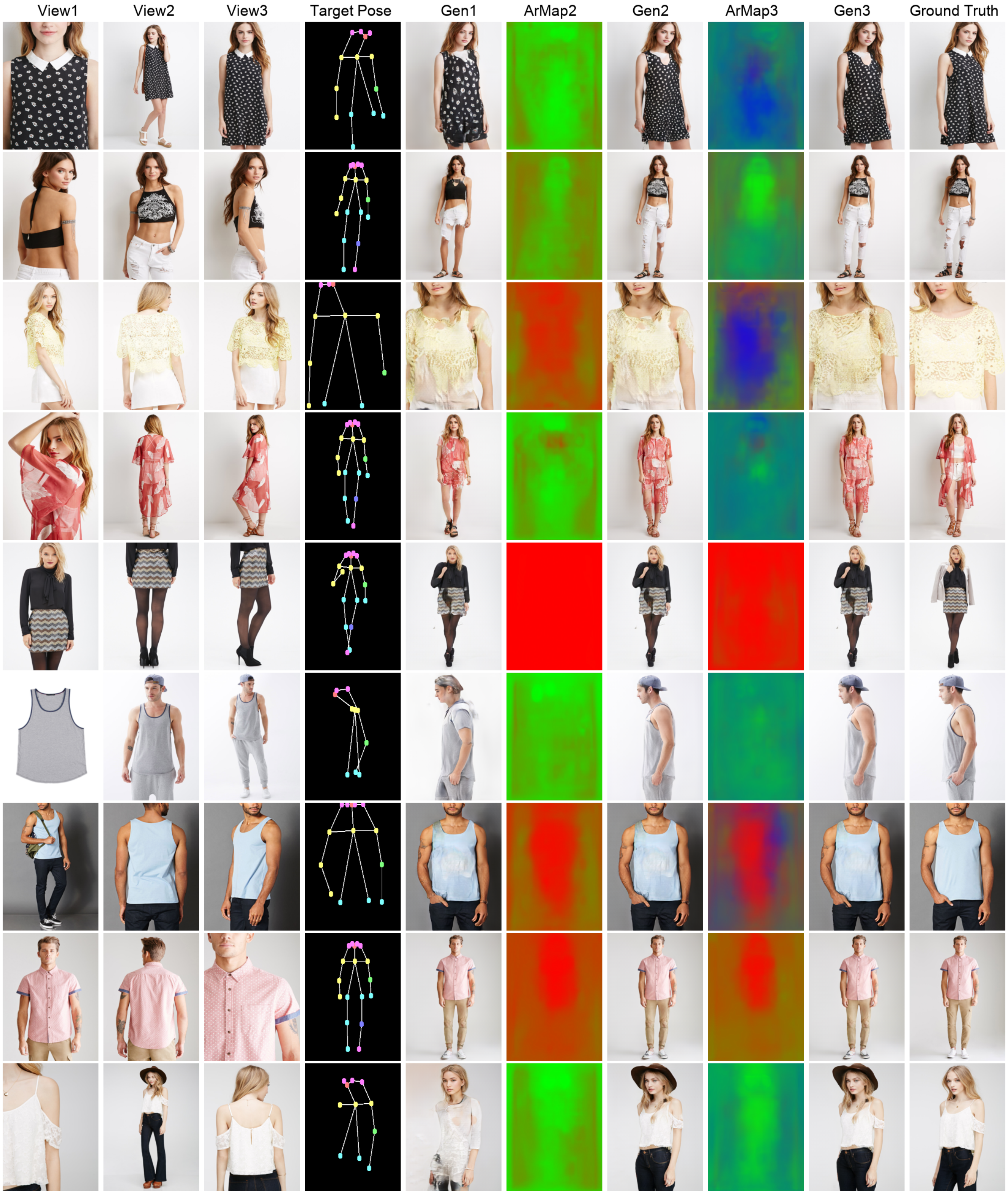

Figure 13: Benefits of using multiple views

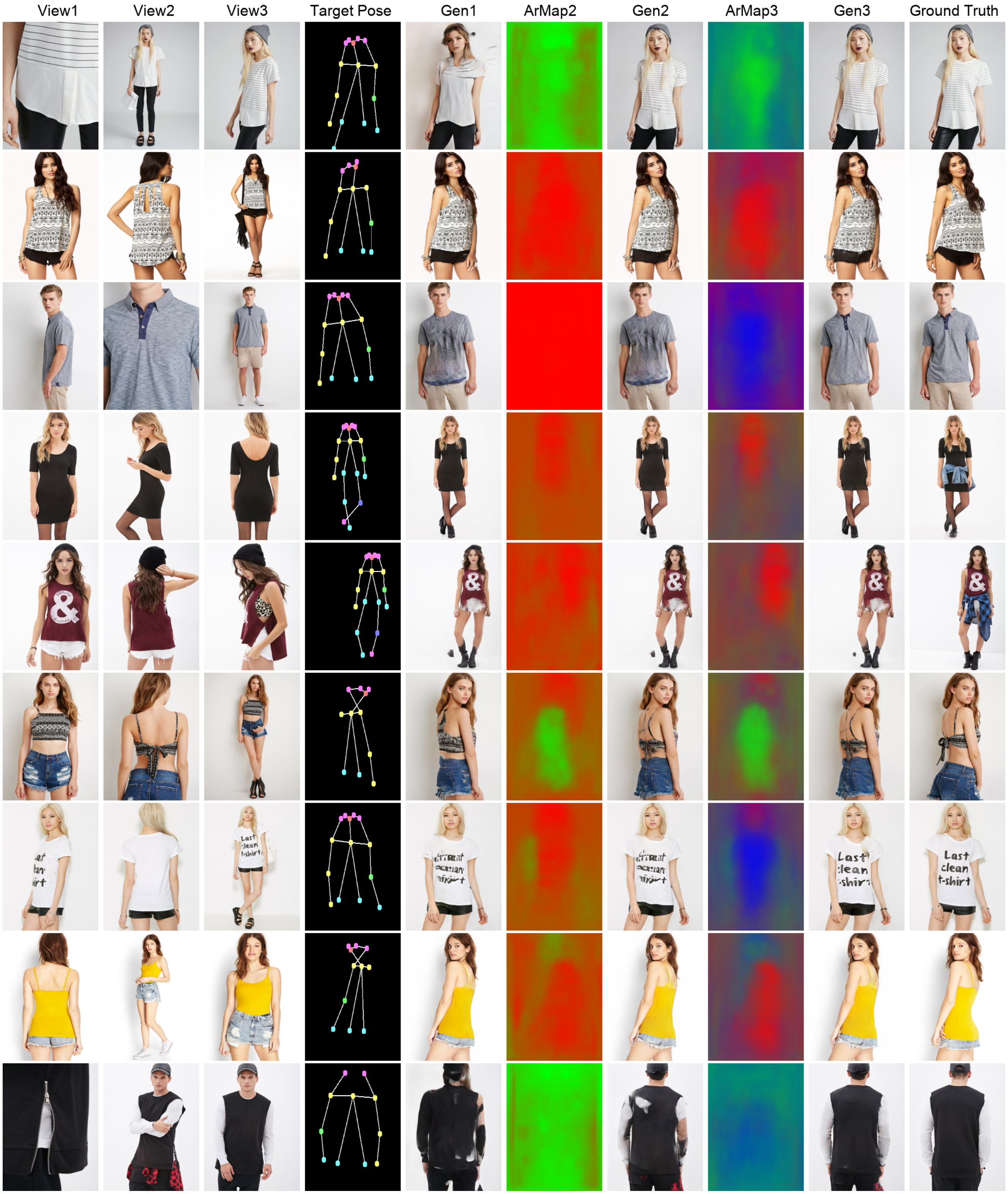

Figure 14: Benefits of using multiple views

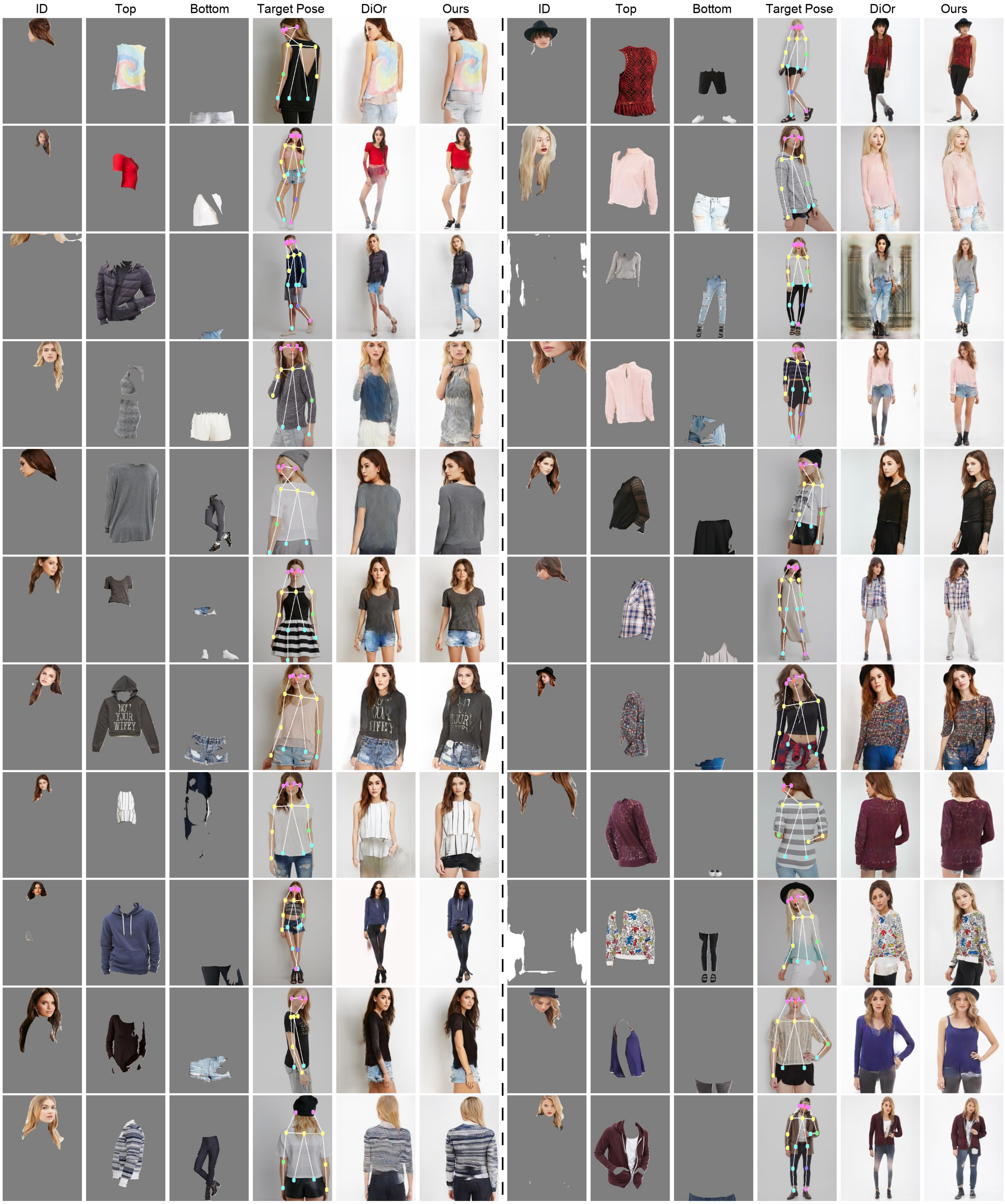

Figure 15: Mix and Match qualitative examples

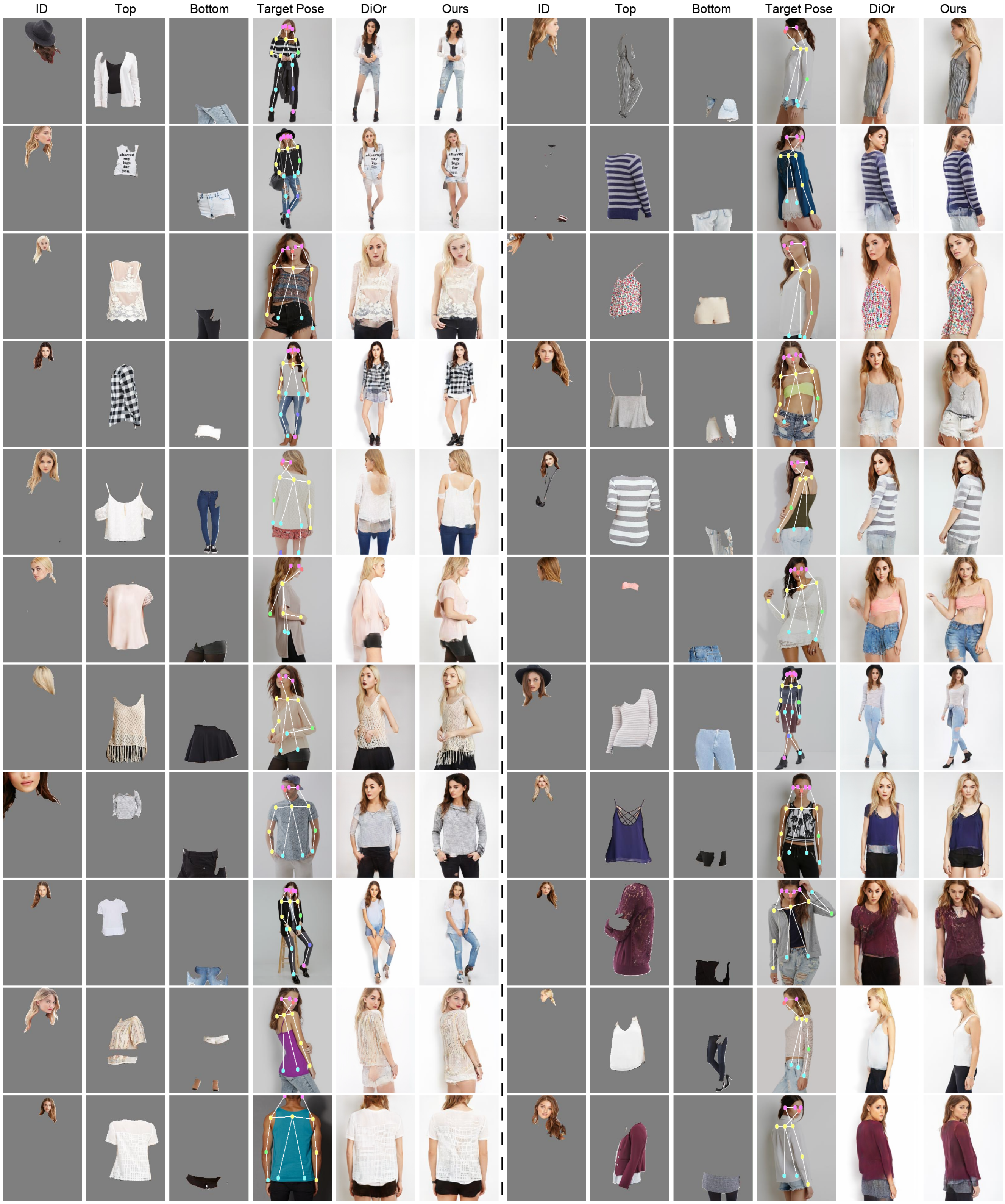

Figure 16: Mix and Match qualitative examples

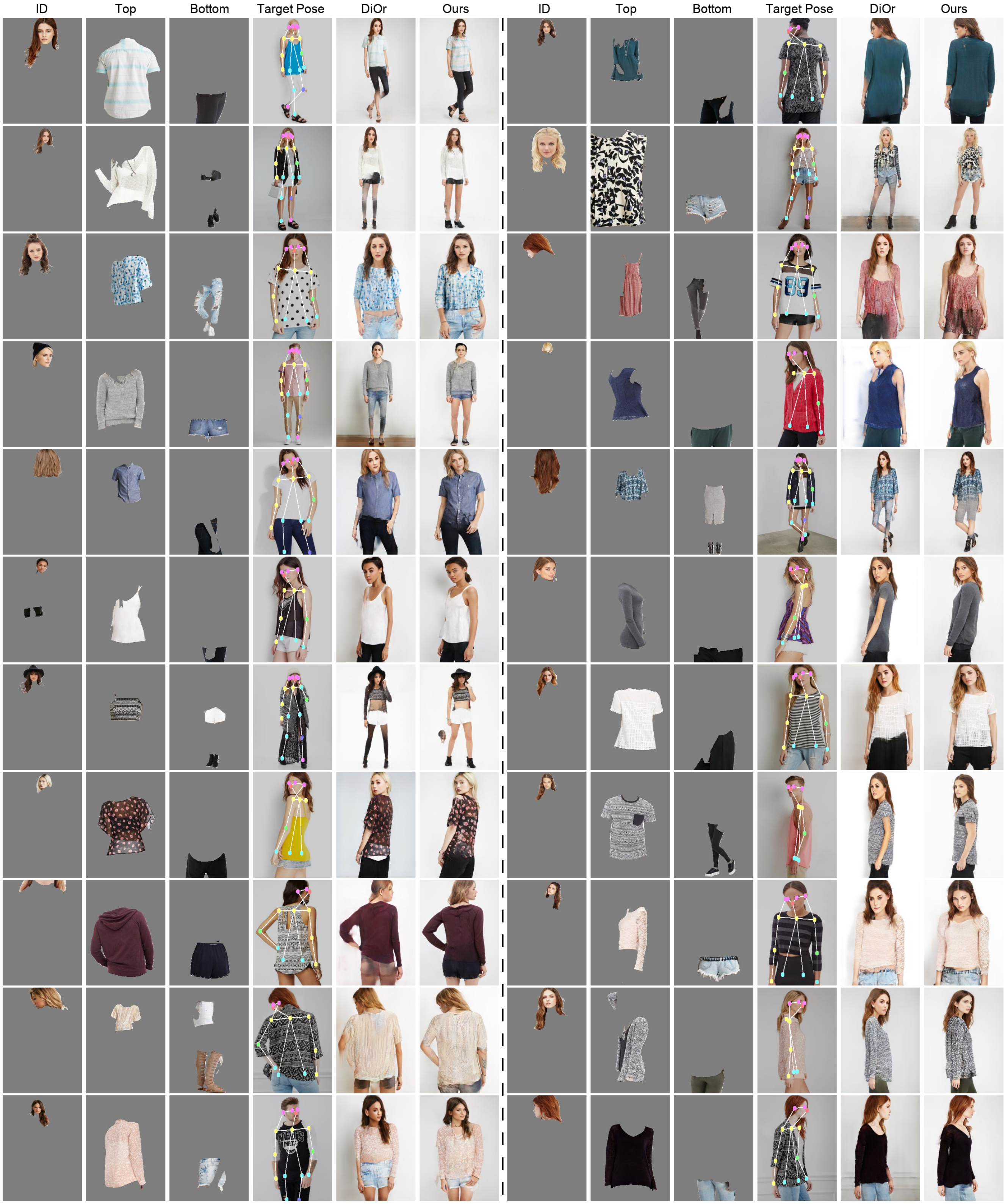

Figure 17: Mix and Match qualitative examples

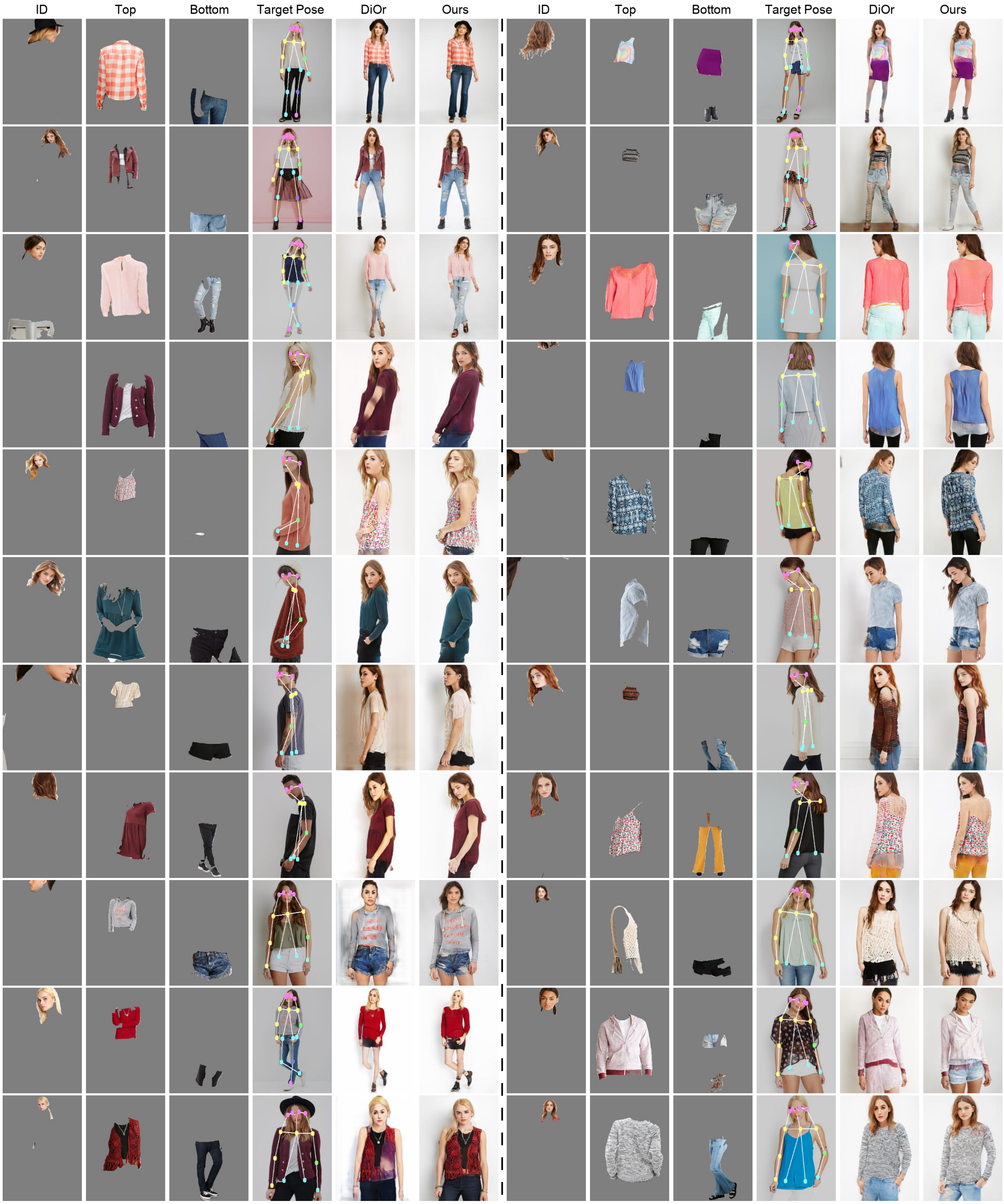

Figure 18: Mix and Match qualitative examples

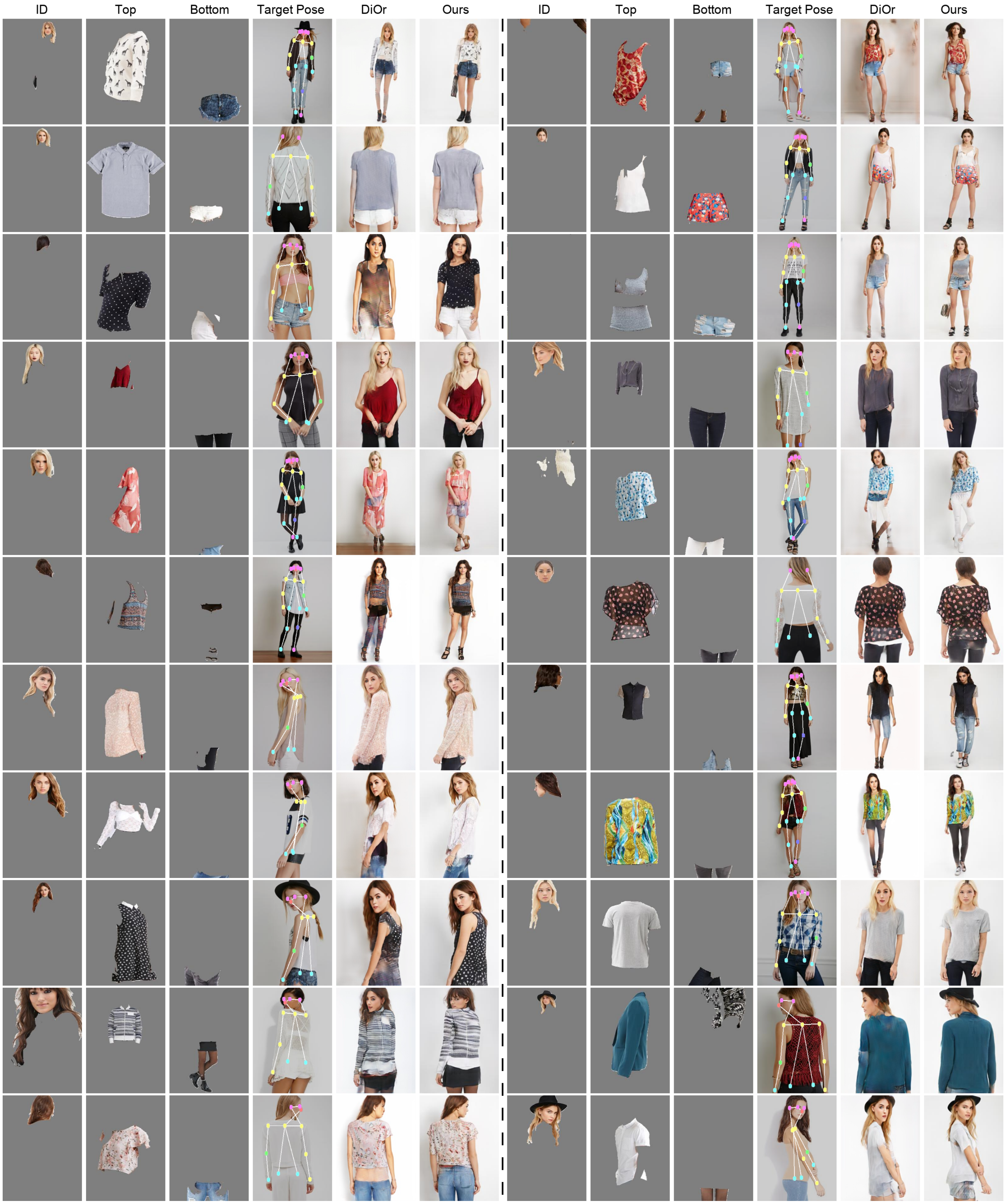

Figure 19: Mix and Match qualitative examples

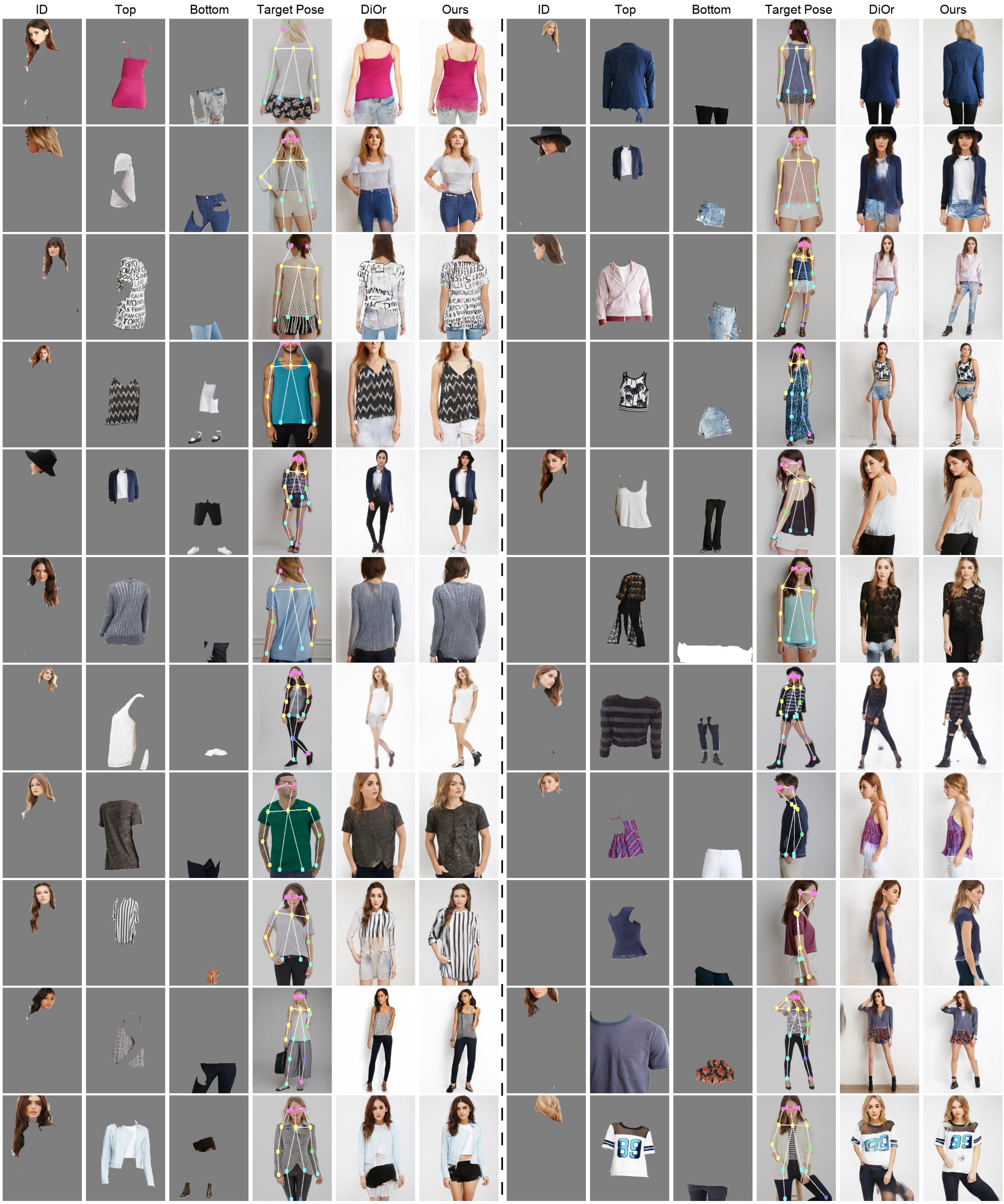

Figure 20: Mix and Match qualitative examples