Supplemental Materials of Street TryOn: Learning In-the-Wild Virtual Try-On from Unpaired Person Images

1. Street TryOn Benchmark Details

The proposed Street TryOn benchmark is derived from the Fashion Retrieval Dataset DeepFashion2 [7]. DeepFashion2 releases $191,961$ and $32,153$ in-the-wild fashion images for training and validation. These images feature models wearing an assortment of clothing items belonging to 13 popular clothing categories. For each image, a comprehensive set of annotations is available, encompassing information on scale, occlusion, zoom-in level, viewpoint, category, style, bounding box coordinates, dense landmarks, and per-pixel masks.

However, DeepFashion2 images cannot be directly used for Virtual Try-On, which requires a frontal-view person with at least the upper body fully present and without large occlusion in relatively bright lighting conditions. Therefore, a two-stage data filtering process was employed for both the training and validation sets.

1.1. Data Filtering Processes

At the first stage of filtering, we use DeepFashion2 [7]'s provided annotations to keep only the images with labels "frontal viewpoint", "no zoom in" and "slight occlusion". Besides, we filtered out the image sourced from customers, which contains a large number of selfies and images taken in dark rooms, as shown in Fig. 1. After the initial filtering stage, we get $15,556$ and $2,401$ training and test images.

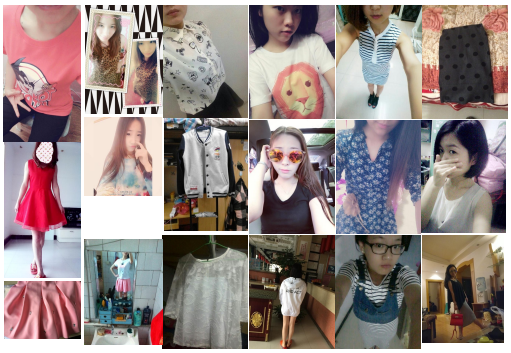


Figure 1. Example DeepFashion2 images that are not feasible for virtual try-on tasks and filtered out.

Subsequently, in the second stage of filtering, the focus was on identifying images that portrayed the entire upper body of the models. We run the DensePose detection [8], which also detects human bounding boxes on these images. We discarded images without human bounding boxes detected and images with the person present horizontally (e.g., lying down) or images with bounding boxes in an aspect ratio larger than $5:8$. We then pad the human bounding box to make its aspect ratio $5:16$, cropped the person from the images, and resized it to 512×320.

Following these two stages of meticulous filtering, the dataset was refined to encompass $12,364$ images for the training set and $2,089$ for the validation set. The examples of the selected and processed images can be found in Fig. 2.

1.2. Data Statistics

Because each image in the DeepFashion2 dataset has rich annotation provided, we further investigate the annotations of the Street TryOn benchmark derived from DeepFashion2, and we obtain a detailed data analysis for the data distributions. As shown in Fig. 3, the proposed Street TryOn benchmark contains a diverse set of garments in various categories.

Next, we looked into the images in the test set and

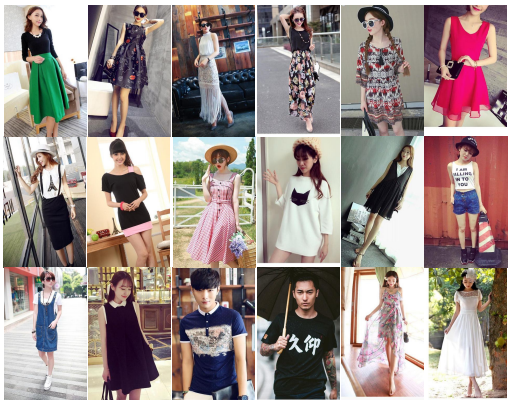


Figure 2. Examples of the selected DeepFashion2 images after cropping, which are included in the proposed Street TryOn Benchmark.

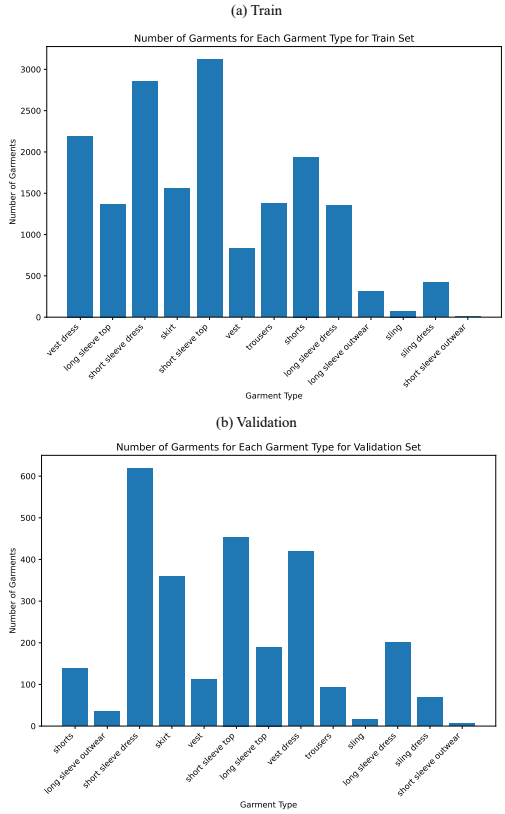


Figure 3. The number of garments in each garment type.

manually labeled six attributes for each image, which are "is_full_body", "is_frontal_view", "has_arm_around_torso", "has_large_occlusion_torso", "has_watermark" and "has_padding". These labels can be used to divide the validation set into different difficulty levels in the future. The attribute distributions can be found in Fig. 4.

Since DeepFashion2 is a fashion retrieval dataset that originally contains multiple images for the same garments

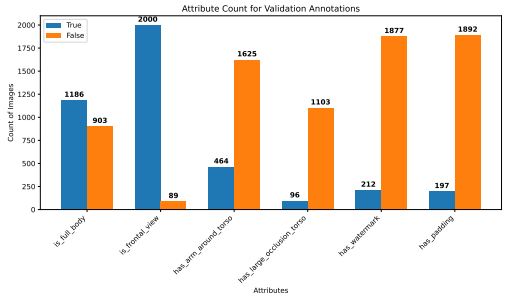


Figure 4. The distribution of manually labeled attributes in the test set.

Figure 5. The images that share the same 'garment_id' in the DeepFashion2 Dataset. Color variations cannot be distinguished.

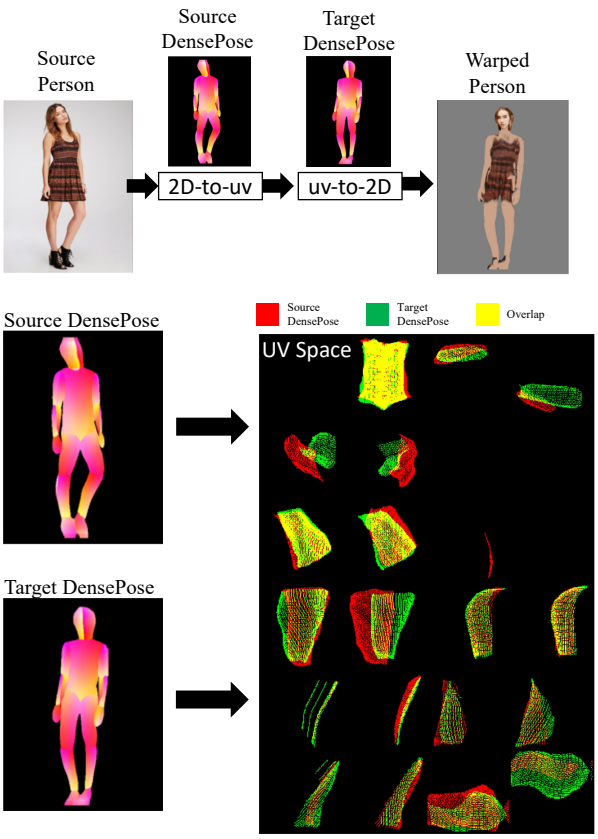
(by sharing the same 'garment_id'), we further investigate if we can build image pairs to enable paired training and validation for virtual try-on tasks. After the filtering processes, in the Street TryOn benchmark, there are 14% unique images without any other images labeled as the same garments. For the rest of the images, although the DeepFashion2 annotation suggests they have potentially paired images, the annotation is too noisy to be directly used to construct pairs, because the color variations of garments are not distinguished by the DeepFashion2's garment_ids, as shown in Fig. 5. Therefore, if one wants to build paired images for the Street TryOn benchmark in the future, additional manual annotation would be necessary.

2. DensePose Perturbation with Cosine Noise

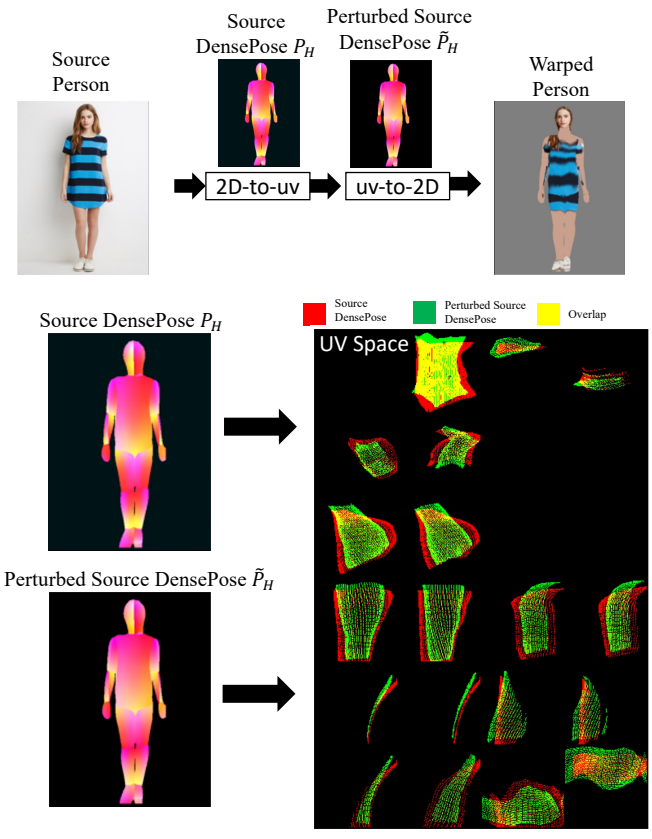
As mentioned in the main paper, when we train the DensePose correction module, a cosine noise is added to the DensePose's pixel values to mimic the imperfect DensePose alignment at the test time. This way, we can train it without paired data but still achieve robust performance at the test time.

Fig. 6(A) shows how the imperfect DensePose prediction causes warping distortion when we transform a garment from one pose to another. Clearly, in the visualized UV overlapping, the source DensePose and the target DensePose are not perfectly matched.

After observing the misalignment patterns, we propose to use a cosine perturbation to mimic this misalignment. Given a person's DensePose P_H, in which each pixel (i, j) contains a (u, v) coordinate to map the person into UV space as $P_H[i, j] = (u, v)$, we add a cosine noise to its



(A) DensePose Misalignment at the Inference Time



(B) Mimicked DensePose Misalignment at the Training Time

Figure 6. **DensePose Perturbation with Cosine Noise.** The example full-body person images are from the DeepFashion dataset [10] and are only used in this figure for illustration purposes. In this illustration, the full person is warped to demonstrate how the 24 UV maps in DensePose (for 24 different body parts) are misaligned. In the final warped image, we fill the skin with the average skin color for visualization.

pixel values as

$$\tilde{P}_H[i,j] = \Big(u + k_1 \cos(\alpha_1 u + \beta_1), v + k_2 \cos(\alpha_2 v + \beta_2)\Big), \tag{1}$$

where $k_1, k_2, \alpha_1, \alpha_2, \beta_1$ and β_2 are randomly sampled coefficients for each UV map in DensePose P_H. As shown in Fig. 6(B), the cosine perturbation can effectively simulate the DensePose misalignment at the inference time. Therefore, with the cosine perturbation, we can mimic the inference scenarios during the training time without paired data.

Figure 7. The detected garment DensePose.

3. Additional Implementation Details

3.1. Pre-trained Diffusion Models

Here, we provide additional implementation details of the diffusion models used in this work.

We set the inference steps for both the skin and refinement diffusion inpainters as 20. The guidance scale for both is 7.5. The skin inpainter has the negative prompt set as "art, clothes, garments, long-sleeves, sleeves, cloak, loose, thick clothes, loose clothes, pants, shirts, skirts, dresses, long jackets, jackets, cloth between legs, cloth around the body, cloth around arms." The refinement diffusion inpainter uses a negative prompt as "blurry, cracks on skins, poor shirts, poor pants, strange holes, bad legs, missing legs, bad arms, missing arms, bad anatomy, poorly drawn face, bad face, fused face, cloned face, worst face, three crus, extra crus,

fused crus, worst feet, three feet, fused feet, fused thigh, three thighs, fused thigh, extra thigh, worst thigh, missing fingers, extra fingers, ugly fingers, long fingers, horn, extra eyes, huge eyes, 2girl, amputation, disconnected limbs, cartoon, cg, 3d, unreal, animate".

As defined by the pretrained diffusion inpainter [5], at the inference time, it takes a concatenation of noise, the latent masked image, and the mask as the input to start denoising. Instead of directly using the Gaussian noise, our noise is built by first broadcasting the mean features into each segmentation class for the latent masked image based on the predicted tryon human parse M_T and adding noises to it with timestep as 999.

3.2. Garment DensePose Detection

For preprocessing, we implement and train the garment DensePose Detection algorithm proposed in the prior work of Cui et al. [4] at resolution 256×192. Then, we resize the detected DensePose to the desired resolutions. The examples of detected garment DensePose can be found in Fig. 7.

4. Additional Experiments for PASTAGAN++

PASTAGAN++ [13] is the state-of-the-art method among the prior work that learns virtual try-on from unpaired images. The main paper shows that the PASTAGAN++ trained with model images is not robust enough to generalize to street2street try-on task and cross-domain try-on tasks.

Here, we run an additional experiment that trains PASTAGAN++ from scratch with the street images in the proposed Street TryOn benchmark. The results in Tab 1 and Fig. 8 show that PASTAGAN++ cannot deal with street images (i.e., casual images of people against cluttered backgrounds). As a StyleGAN2-based method, PASTAGAN++ learns a latent space to encode the data distribution for image patches. Although such a distribution can be learned for relatively structural data like person images with a clean background, it is hardly possible to learn a highly diverse distribution [11] to represent the complex background for in-the-wild images. Therefore, the PASTAGAN++ trained with street images is struggling with the background reconstructions and cannot generate the street try-on images as faithfully as our method does.

5. Additional Ablation Study

We report an ablation study that separately analyzes the effects of our DensePose warping and inpainting-based compositing. We take FS-VTON [9], a non-diffusion-based method with separate warping and refinement modules, whose warping module is trained using paired studio images. We individually replace our warping and refinement modules with theirs. Here, our method is trained with paired data on VITON-HD for a fair comparison. As shown in Fig.9, both our proposed warping module and the diffusion

6. More Visual Results

In this section, more visual results are presented to verify the effectiveness of our approach.

6.1. Shop2Model Test

More results can be found in the Shop2Model test on VITON-HD in Fig 10 and Fig. 11, comparing with PFAFN [6], FS-VTON [9], SDAFN [2] and GP-VTON [12].

6.2. Street2Street Test

We finally show more results for our Street2Street try-on with intermediate outputs in Fig. 12 and Fig. 13.

6.3. Ablation Study

Here, we also provide more examples of ablation studies to validate the effectiveness of each component in the design of our approach in Fig. 14.

Figure 8. Visual Comparisons with PWS [1] and PASTAGAN++ [13].

	Training Settings			Model2Model	Model2Street	Street2Street
	paired	garment src	person src	FID ↓	FID ↓	FID ↓
PastaGAN++ [13] (released)	×	model	model	13.848	71.090	67.016
PastaGAN++ [13] (street)	×	street	street	40.841	70.461	67.088
ours (1)	✓	shop	model	10.961	**34.050**	33.165
ours (2)	×	model	model	11.040	34.434	33.742
ours (3)	×	street	street	**10.214**	34.191	**33.039**

Table 1. **Comparsions with prior work in same-domain and cross-domain tests.** PASTAGAN++ (released) is the officially released model trained on UPT dataset [14], which contains unpaired model images. We train PASTAGAN++ (street) on the training split of the Street TryOn dataset with the default settings.

Warping Method	Fusion Method	Shop2Street (FID)
FS-VTON	FS-VTON	90.972
FS-VTON	SD-inpainting	77.456
ours	FS-VTON	64.046
ours	SD-inpainting	**33.819**

Figure 9. Comparison with warping and compositing (fusion) modules of FS-VITON [9] in FID. **Top:** Visual Results. **Bottom:** Quantitative Results.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with Style: Detail-Preserving Pose-Guided Image Synthesis with Conditional StyleGAN. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 5

[2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single Stage Virtual Try-on via Deformable Attention Flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 4

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021. 7, 8

[4] Aiyu Cui, Sen He, Tao Xiang, and Antoine Toisoul. Learning Garment Densepose for Robust Warping in Virtual Try-On. *arXiv preprint arXiv:2303.17688*, 2023. 4

[5] Hugging Face. Stable diffison inpainting. https://huggingface.co/runwayml/stable-diffusion-inpainting. 4

[6] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-Free Virtual Try-on via Distilling Appearance Flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2021. 4

[7] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019. 1

[8] 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. 1

[9] Sen He, Yi-Zhe Song, and Tao Xiang. Style-Based Global Appearance Flow for Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 4, 6

[10] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 3

[11] Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling Stylegan to Large Diverse Datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 4

[12] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. GP-VTON: Towards General Purpose Virtual Try-on via Collaborative Local-Flow Global-Parsing Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023. 4

[13] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, Xin Dong, Feida Zhu, and Xiaodan Liang. Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. *arXiv preprint arXiv:2207.13475*, 2022. 4, 5, 6

[14] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 34:2598–2610, 2021. 6

Figure 10. **More examples for Shop2Model test** on VITON-HD benchmark [3] 1.

Figure 11. **More examples for Shop2Model test** on VITON-HD benchmark [3] 2.

Figure 12. **More examples for Street2Street test** for top try-on.

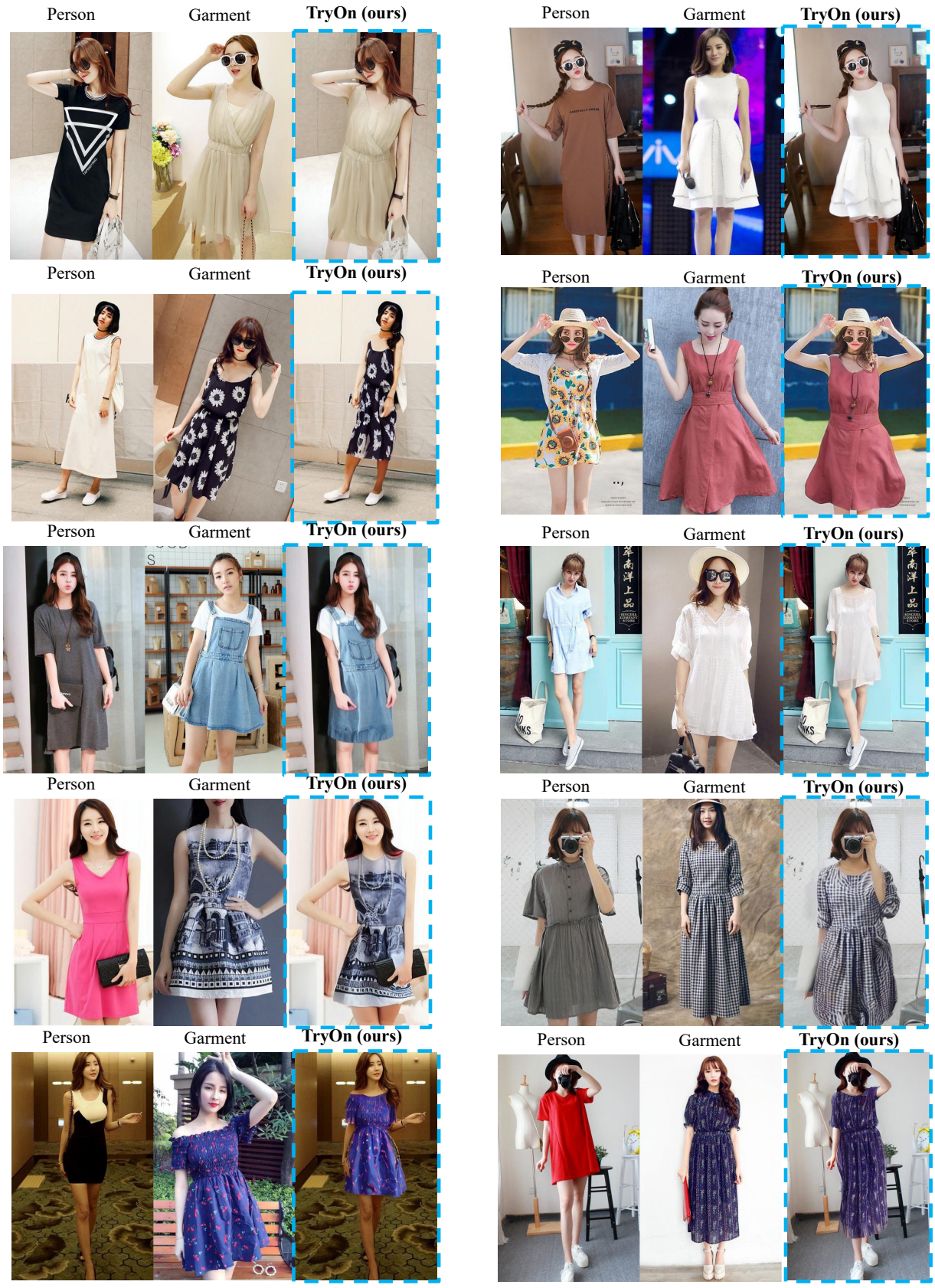


Figure 13. **More examples for Street2Street test** for dress try-on.

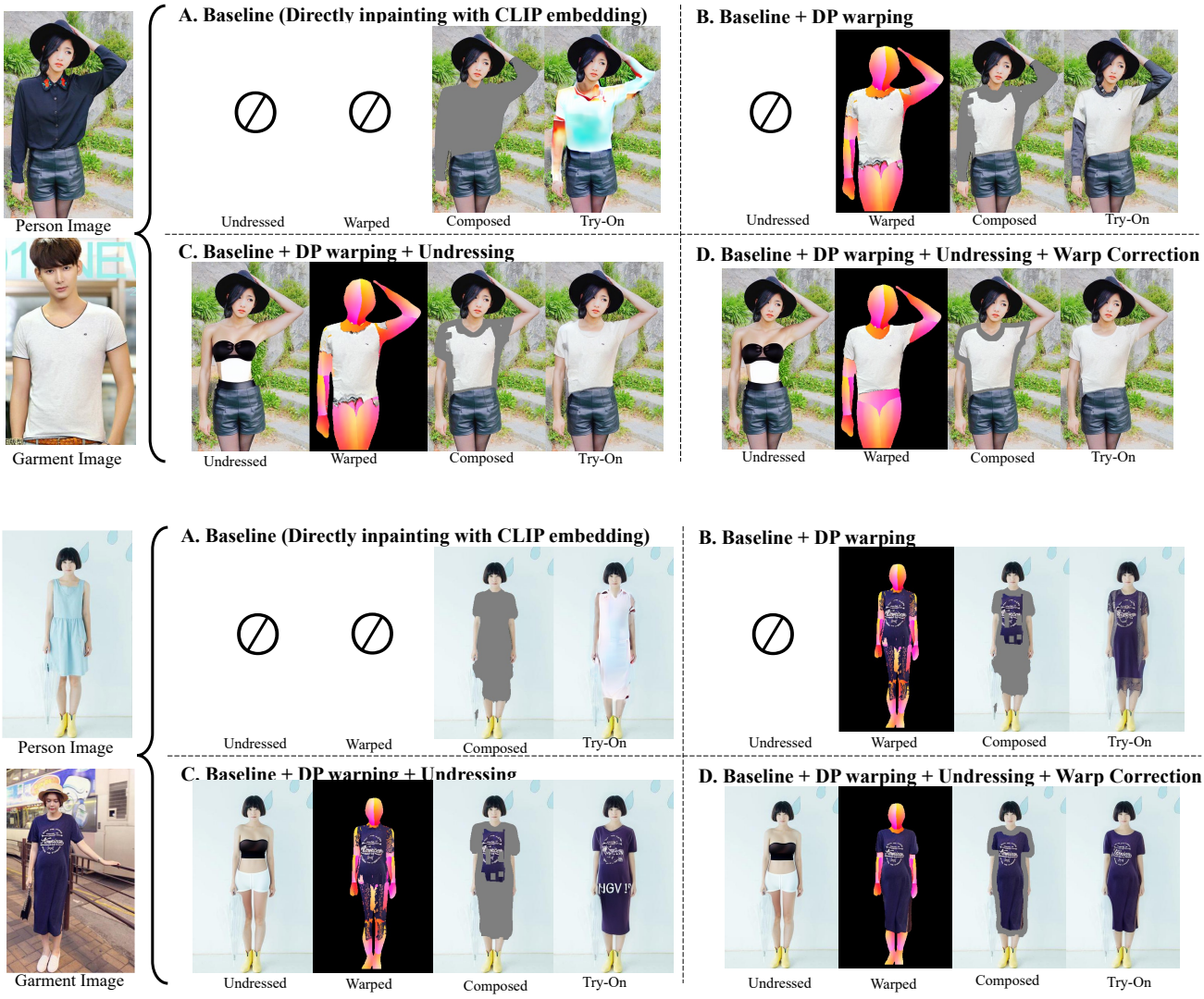


Figure 14. More examples for the ablation study to verify the effectiveness of each component in the proposed method.