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

Aiyu Cui Jay Mahajan Viraj Shah Preeti Gomathinayagam Chang Liu Svetlana Lazebnik
University of Illinois Urbana-Champaign
{aiyucui2, jaym2, vjshah3, preetig3, changl25, slazebni}@illinois.edu

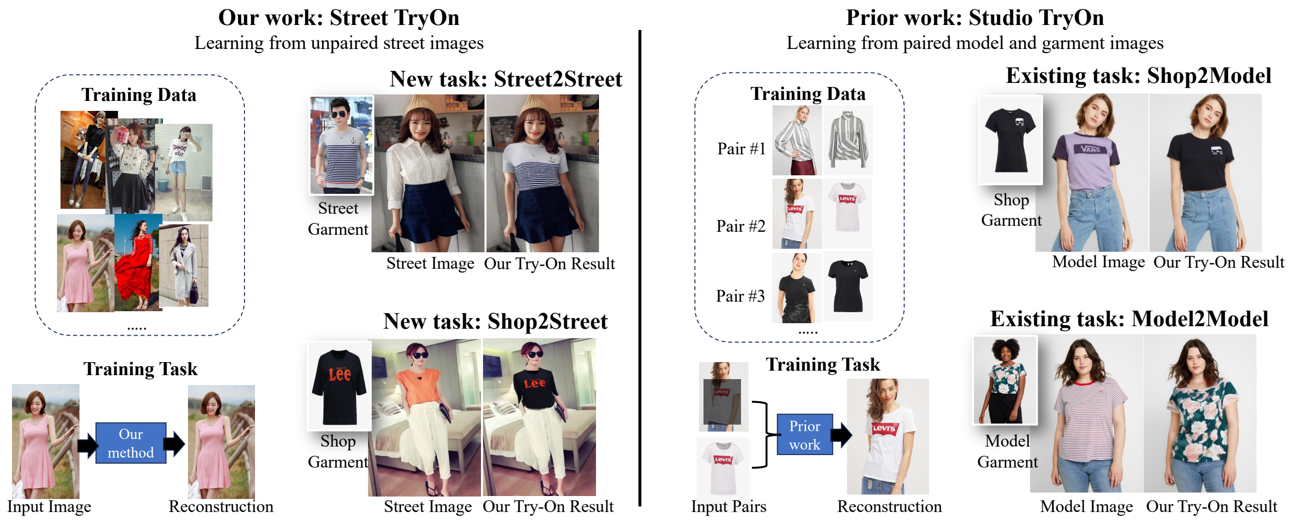

Figure 1. Our proposed Street TryOn benchmark and method, contrasted with existing work focusing on studio images and paired training.

Abstract

Most existing methods for virtual try-on focus on studio person images with a limited range of poses and clean backgrounds. They can achieve plausible results for this studio try-on setting by learning to warp a garment image to fit a person's body from paired training data, i.e., garment images paired with images of people wearing the same garment. Such data is often collected from commercial websites, where each garment is demonstrated both by itself and on several models. By contrast, it is hard to collect paired data for in-the-wild scenes, and therefore, virtual try-on for casual images of people with more diverse poses against cluttered backgrounds is rarely studied.

*In this work, we fill the gap by introducing a **StreetTryOn** benchmark to evaluate in-the-wild virtual try-on performance and proposing a novel method that can learn it without paired data, from a set of in-the-wild person images directly. Our method achieves robust performance across shop and street domains using a novel DensePose warping correction method combined with diffusion-based conditional inpainting. Our experiments show competitive performance for standard studio try-on tasks and SOTA performance for street try-on and cross-domain try-on tasks.*

1. Introduction

Virtual try-on methods have advanced rapidly and reached high levels of performance for transferring garments from shop images to model person images [9, 14, 22, 24] or from one model image to another [1, 4, 20, 21]. By contrast, transferring garments to and from in-the-wild images is rarely studied. Although the dominant Shop2Model benchmark, VITON-HD [3], is getting saturated, virtual try-on research is still far from robust enough to enable the general population to visualize how a garment would look on their own bodies by taking photos in a casual environment.

Existing virtual try-on methods [1, 4, 9, 14, 20–22, 24] are typically designed to dress up studio models (Fig. 1-right) where the models demonstrate garments with a limited range of poses against clean backgrounds. Such methods train on image pairs showing the same garment in a shop view and worn by a model, which enables it to learn garment warping via a reconstruction loss. While they can yield high-quality results in the studio setting, these methods do not transfer well to in-the-wild images (Fig. 1-left), in which body poses and camera angles are less constrained, and lighting and backgrounds are more variable. As we will prove later, existing methods struggle with limb reconstruction, warping, and background rendering in such images.

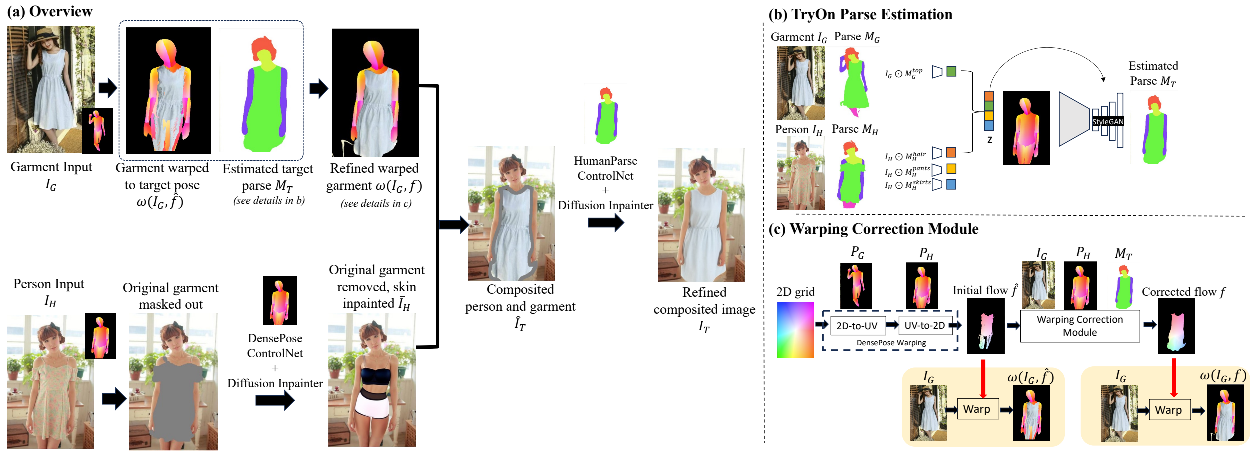

Figure 2. Overview of our proposed virtual try-on method (see text for details).

Since there is no existing dataset to evaluate virtual try-on in the wild, we introduce a new benchmark in Section 3, **StreetTryOn**, derived from the large in-the-wild fashion retrieval dataset DeepFashion2 [8] by filtering out the images that are infeasible for try-on tasks, resulting in a set of $12K$ training and $2K$ test images. Combining with the garment and person images in VITON-HD dataset [3], we obtain a suite of try-on tasks with garment and person inputs from various sources, as shown in Fig. 1. Benchmarking methods across all these tasks can give a comprehensive idea of the robustness and cross-domain generalization ability of different models (i.e., generalization of models trained on "studio" images to "street" images, and vice versa).

To obtain robust performance on the challenging in-the-wild try-on tasks, Shop2Street and Street2Street (Fig 1), we introduce in Section 4 a novel approach for learning virtual try-on from unpaired in-the-wild person images. An overview of our method is shown in Fig. 2. Comprehensive evaluation in Section 5 will show that our method outperforms all existing methods on our StreetTryOn benchmark and is competitive on the much more mature VITON-HD benchmark [3]. Our method is remarkably robust for the hardest try-on setting, Street2Street, achieving similar results whether trained on in-domain or out-of-domain data.

2. Related Work

Virtual Try-On Benchmarks. Existing Shop2Model virtual try-on benchmarks include VITON [12], VITON-HD [3], MPV [5], and DressCode [19], all of which have paired person and garment images with studio model as person source and ghost mannequin images as garment source. DeepFashion[18], and UPT [21] datasets have also been used for Model2Model try-on. However, none of the existing datasets are representative of in-the-wild try-on settings. The SHHQ-1.0 dataset [7] has previously been proposed to evaluate in-the-wild try-on performance. However, at least 25% images in SHHQ-1.0 are studio model images aggregated from the DeepFashion dataset [18] and the African fashion dataset [11]. Therefore, our proposed StreetTryOn benchmark is a necessary addition to the literature.

Virtual Try-On Methods. Most of the top-performing methods for the Shop2Model try-on [9, 13, 14, 16, 22, 24] are trained on paired datasets mentioned above, like VITON-HD [3] and DressCode [19]. Such methods can achieve high-quality results on in-domain images, but do not transfer well to in-the-wild data. Several other works [1, 4, 24] can achieve Model2Model try-on by training on paired data (people wearing the same outfits in multiple poses). PASTAGAN [20] and PASTAGAN++ [21] are the only prior works for Model2Model try-on trained without paired training data on the UPT dataset [20], but all of them suffer from the warping for free-form pose and complex backgrounds of street images.

3. StreetTryOn Benchmark

To explore in-the-wild and cross-domain try-on, we introduce a new benchmark called **StreetTryOn**, derived from the existing fashion retrieval dataset DeepFashion2 [8]. DeepFashion2 contains $191,961$ training and $32,153$ test images of people with diverse poses, outfits and backgrounds, but unfortunately, most of them cannot directly be used for virtual try-on since they only show portions of the body, have large occlusions, non-frontal views, or dark lighting conditions. To remove such unsuitable images, we apply a multi-step filtering process using a combination of provided DeepFashion2 annotations, person detection, and manual selection, resulting in a clean set of $12,364$ training and $2,089$ test images.

Benchmark Tasks. The try-on tasks of greatest interest to us are **Street2Street**, **Shop2Street**, and **Model2Street** (Fig. 1). For the latter two cross-domain tasks, we obtain the needed shop and model test images from VITON-HD [3]. For **Street2Street**, we use the $2,089$ test street images in StreetTryOn, which are partitioned into two subsets of 909 "top" images and $1,190$ "dresses." Then we construct 909 and $1,190$ unpaired (person, garment) test tuples by random shuffling. For **Shop2Street** and **Model2Street** try-on, we randomly sample 909 garment ghost mannequin images and 909 model images from VITON-HD to construct two sets of 909 cross-domain (person, garment) test tuples. Combining the above test sets with existing **Shop2Model** and **Model2Model** test sets from VITON-HD gives us a comprehensive suite of scenarios for evaluation.

4. Our Try-On Method

Task Definition & Our pipline. Given a person image I_H and a garment image I_G[1], our goal is to generate the try-on image I_T with person I_H wearing I_G. We preprocess I_H and I_G to obtain semantic segmentations or parses M_H and M_G, as well as DensePose [10] estimates P_H and P_G.

An shown in the overview in Fig. 2, our try-on inference pipeline starts by predicting the semantic parse M_T for the try-on output image using a TryOn Parse Estimator. Next, we predict a flow field f to warp the garment I_G to the output pose P_H using DensePose correspondence followed by a trained Warping Correction Module. At the same time, for the person image I_H, we remove the original garment and inpaint skin regions by a pre-trained diffusion inpainter with a DensePose ControlNet conditioned on P_H. Then, we combine the warped garment $\omega(I_G, f)$ and the inpainted person $\bar{I}_H$ to get the composited person I'_T. Finally, we use the pre-trained diffusion inpainter with a Human Parse ControlNet conditioned on M_T to inpaint a masked garment boundary to get the final try-on output I_T.

TryOn Parse Estimator. The architecture of our parse estimator is shown in Fig. 2-b. We encode the target DensePose P_H into a 16×16 feature map by a trainable encoder as $\mathbf{E}_{dp}(P_H)$, and set it as the initial feature map of the StyleGAN. While the initial feature map controls the pose of predicted parse, we use the style code z of StyleGAN to control the contents of human parse. In more detail, the style code z is a concatenation of four segment style codes $\{z^{top}, z^{hair}, z^{pants}, z^{skirt}\}$. Each of the segments $i \in \{top, hair, pants, skirt\}$ is encoded by a segment encoder $\mathbf{E}_{seg}$ as $z^i_H = \mathbf{E}(I_H \odot M^i_H)$ with the mask M^i_H of the segment i from the source person image I_H. At inference time, the top segment will come from the garment image, and the rest will come from the person input, so we predict the human parse as

$$M_T = \mathbf{G}(\{z^{top}_G, z^{hair}_H, z^{pants}_H, z^{skirt}_H\}|\mathbf{E}_{dp}(P_H)) \quad (1)$$

where $\mathbf{G}$ is the StyleGAN. During training, all segment codes will come from the same person image, and the model is trained to reconstruct the original human parse using cross-entropy loss.

[1]For simplicity, we use I_G to denote the garment image with everything except for the try-on garment masked out.

Warping Correction Module DensePose is a mapping from a person image to the coordinate system (UV space) of a parametrized 3D human model, which can be used for garment warping directly. In practice, DensePose estimation are far from perfect, especially for loose garments, and direct warping results in missing or misaligned areas. Thus, we apply a trained correction after the initial DensePose warping. As shown in the top of Fig. 2-c, we obtain an initial flow field $\hat{f}$ by projecting a mesh grid to the UV space using the garment's DensePose P_G, and then warping it back to the person's pose in image space via P_H. Next, we train a correction module that takes in the naive flow $\hat{f}$ and adjusts it to obtain the final flow $f = \mathbf{C}(\hat{f}|I_G, P_H, M_T)$.

To train the correction module without paired data, we attempt to reconstruct the person image I_H from a perturbed version $\tilde{I}_H$. For $\tilde{I}_H$, we apply a cosine perturbation (a noise that is a cosine function of the grid) to the pixel values of DensePose P_H, which mimics imperfect registration at inference time. Given this synthetic data, we train the corrector $\mathbf{C}$ with the same objectives as in prior work [9, 14] with total variation loss, L1 loss and VGG loss [15].

Garment Removal and Skin Inpainting. To prevent information leakage from the mask used to remove the old garment, before rendering the new warped garment, we introduce a separate step of removing the original garment and inpainting it with as much skin as possible.

Refining the composited image. Finally, to compose the warped garment $\omega(I_G, f)$ and the processed person image $\bar{I}_H$ together, we first create a naive composite image $\hat{I}_T$ as $\hat{I}_T = \bar{I}_H \odot e[1 - M^{top}_T] + \omega(I_G, f) \odot e[M^{top}_T]$, where e is an erosion function and M^{top}_T is the predicted try-on garment mask. Then, we obtain the final try-on output I_T by applying the second diffusion inpainter on the composite image $\hat{I}_T$ to inpaint the erased gaps and refine the details.

Both of the above steps are accomplished by a pre-trained Stable Diffusion inpainter [6] combined with ControlNets [23] trained on our own data. Specifically, for skin inpainting, we train a ControlNet using DensePose P_H as conditioning information, and for the final compositing, we train a ControlNet with predicted parse M_T as conditioning.

5. Experiments

We report performance on the proposed StreetTryOn benchmark by running experiments at 512×320 for Street2Street, Model2Street, and Model2model tests, and 512×384 resolution for Shop2Street and Shop2Model.

To evaluate the proposed method on the proposed benchmarks, we compare three training settings for our method: (1) training with the standard paired VITON-HD training data; (2) unpaired training with VITON-HD person images only; (3) unpaired training with StreetTryOn person images.

Garment Transfer from Person Images. Tab. 5, Fig. 3 and Fig. 4 reports results for in-the-wild try-on task

Figure 3. Street2Street Try-On examples for our method.

(a) Comparisons with methods trained on the **Model2Model** setting

(b) Comparison with methods trained on the **Shop2Model** setting

Figure 4. (a)-top: trained on Model2Street. (a)-bottom: Model2Model. (b)-top: Shop2Street. (b)-bottom: Shop2Model.

	Street2Street FID ↓	Model2Street FID ↓	Model2Model FID ↓
Ours (Paired, VITON-HD)	33.165	**34.050**	10.961
Ours (Unpaired, VITON-HD)	33.742	34.434	11.040
Ours (Unpaired, StreetTryOn)	**33.039**	34.191	**10.214**
FS-VTON [14]	67.009	77.273	13.926
HR-VITON [17]	63.539	55.172	20.404
SDAFN [2]	42.432	44.537	14.316
PWS [1]	84.326	76.889	34.224
PastaGAN++ [21]	67.016	71.090	13.848
PastaGAN++ (street)	67.088	70.461	40.841

Table 1. **Evaluation on Street2Street, Model2Street, and Model2Model tests.** We retrain FS-VTON, HR-VTON and SD-VTON on paired DeepFashion dataset for Model2Model try-on at 512×320. PWS is trained on paired DeepFashion [18], and PASTAGAN++ is trained on UPT dataset [20]. PASTAGAN++ (street) is trained on the proposed Street TryOn dataset.

	Shop2Street	Shop2Model (VITON-HD)		
	FID ↓	FID ↓	SSIM ↑	LPIPS ↓
Ours (Paired, VITON-HD)	**33.819**	9.671	0.840	0.113
Ours (Unpaired, VITON-HD)	35.135	11.675	0.826	0.128
Ours (Unpaired, StreetTryOn)	34.054	11.951	0.823	0.129
SDAFN [2]	62.735	9.400	0.882	0.092
FS-VTON [14]	77.843	9.552	0.883	0.091
HR-VITON [17]	63.516	16.21	0.862	0.109
GP-VTON [22]	n.a.	9.197	0.894	0.080
StableVITON [16]	37.085	**8.233**	**0.888**	**0.073**

Table 2. **Evaluation on Shop2Street and Shop2Model tests** at 512×320 and 512×384 respectively. The methods are retrained at 512×384 if their released models have a lower resolution. We resize output images to the resolution for these methods with released models at higher resolutions.

Street2Street, cross-domain task Model2Street, and studio task Model2Model. All of these take garments from person images (either studio or street). As shown, the prior methods perform much worse than ours on the three tasks, because all prior methods suffer from limb reconstruction, warping, and background rendering. Fig. 3 further proves that our method can well handle the diverse pose and backgrounds for Street2Street try-on.

Garment Transfer from Shop Images. Tab. 2 and Fig. 4 presents an evaluation on Shop2Street and Shop2Model tasks, in which a garment from a ghost mannequin image is transferred to a person image. Although the prior work shows better performance on their highly-tuned task, Shop2Model (VITON-HD task), our method gets the best performance on the cross-domain Shop2Street task, confirming both the robustness of our method and the challenging nature of our Street TryOn benchmark. For most prior methods, both the warping and rendering steps tend to fail on out-of-domain street images. Even though the concurrent work, StableVITON [16], shows significant improvement on Shop2Street try-on, it still struggles in limb reconstruction and background rendering. This proves that the prior work trained on studio images fails to fully capture the diverse distribution of in-the-wild person images.

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. 1, 2, 4

[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. 1, 2

[4] 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*, pages 14638–14647, 2021. 1, 2

[5] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards Multi-pose Guided Virtual Try-on Network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035, 2019. 2

[6] Hugging Face. Stable diffison inpainting. https://huggingface.co/runwayml/stable-diffusion-inpainting. 3

[7] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 2

[8] 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. 2

[9] 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. 1, 2, 3

[10] 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. 3

[11] Gilles Hacheme and Noureini Sayouti. Neural fashion image captioning: Accounting for data diversity. *arXiv preprint arXiv:2106.12154*, 2021. 2

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-on Network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2

[13] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. ClothFlow: A Flow-Based Model for Clothed Person Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019. 2

[14] 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. 1, 2, 3, 4

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Rerceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 3

[16] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. *arXiv preprint arXiv:2312.01725*, 2023. 2, 4

[17] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 4

[18] 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. 2, 4

[19] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-resolution Multi-Category Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022. 2

[20] 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. 1, 2, 4

[21] 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. 1, 2, 4

[22] 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. 1, 2, 4

[23] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[24] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. TryOnDiffusion: A Tale of Two UNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 1, 2