

DH-VTON: Deep Text-Driven Virtual Try-On via Hybrid Attention Learning

Jiabao Wei

School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
weijiabao@bit.edu.cn

Zhiyuan Ma

Department of Electronic Engineering
Tsinghua University
Beijing, China
mzyth@tsinghua.edu.cn

Abstract—Virtual Try-ON (VTON) aims to synthesis specific person images dressed in given garments, which recently receives numerous attention in online shopping scenarios. Currently, the core challenges of the VTON task mainly lie in the fine-grained semantic extraction (*i.e.*, *deep semantics*) of the given reference garments during depth estimation and effective texture preservation when the garments are synthesized and warped onto human body. To cope with these issues, we propose DH-VTON, a deep text-driven virtual try-on model featuring a special hybrid attention learning strategy and deep garment semantic preservation module. By standing on the shoulder of a well-built pre-trained paint-by-example (*abbr.* *PBE*) approach, we present our DH-VTON pipeline in this work. Specifically, to extract the deep semantics of the garments, we first introduce InternViT-6B as fine-grained feature learner, which can be trained to align with the large-scale intrinsic knowledge with deep text semantics (*e.g.*, “neckline” or “girdle”) to make up for the deficiency of the commonly adopted CLIP encoder. Based on this, to enhance the customized dressing abilities, we further introduce Garment-Feature ControlNet Plus (*abbr.* *GFC+*) module and propose to leverage a fresh hybrid attention strategy for training, which can adaptively integrate fine-grained characteristics of the garments into the different layers of the VTON model, so as to achieve multi-scale features preservation effects. Extensive experiments on several representative datasets demonstrate that our method outperforms previous diffusion-based and GAN-based approaches, showing competitive performance in preserving garment details and generating authentic human images.

Index Terms—Virtual try-on, Stable diffusion, Hybrid attention learning

I. INTRODUCTION

Image-based virtual try-on (VTON) has recently attracted significant interests in generative research community [1] with the increasing popularity of online shopping [2]–[10]. Despite significant progress having been witnessed, the existing VTON models still face several critical issues. One key issue lies in that the given garments must be naturally deformed to fit the target person’s pose and body shape. Based on this, the other key issue is the patterns and texture details of the deformed garments need to be fine-grained preserved.

To address the above two critical issues, existing image-based VTON approaches generally can be categorized into two categories: warping-based and warping-free approaches. **a)** The former [8]–[20] typically perform garment warping before image synthesis via GANs [21] or LDMs [22]. Early

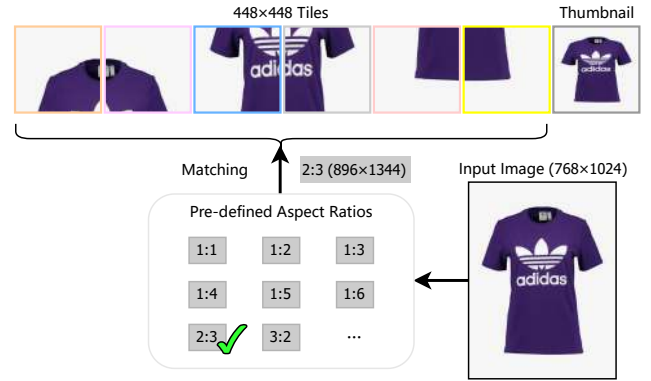


Fig. 1. **Demonstration of dynamic high-resolution capabilities.** InternViT-6B-448px-V1-5 [26] dynamically match an optimal aspect ratio from pre-defined ratios, dividing the image into tiles of 448×448 pixels and creating a thumbnail for global context.

approaches primarily rely on GANs [21] as image generator and tentatively reduce the mismatch between the warped garment and the target person such as in VITON-HD [8], HR-VITON [9] and GP-VTON [17]. Afterwards, researchers have considered leveraging LDMs [22] instead of GANs [21] as image generator due to their impressive generation capabilities [10], [16]. Specifically, DCI-VTON [10] and LaDI-VTON [16] are two representative works by utilizing LDMs [22] to merge the warped garment onto the target person. However, the main disadvantage of the warping-based approaches is the artifacts produced by the garment warping process, which may be difficult to eliminate during image synthesis. Furthermore, existing garment deformation methods such as TPS [23], STN [24], and FlowNet [25] basically all lack well customized dressing abilities under giving the various postures as conditions.

b) In contrast, another type of warping-free approaches [2]–[7], [10], [16], [27], [28] usually adopt LDMs [22] as image generator because of their strong intrinsic generation capabilities. To avoid generating artifacts, they generally bypass garment warping and utilize a feature extractor and several cross-attention blocks to capture and transfer the textures of the given garments. For instance, CLIP [29] has emerged as a robust image encoder and is frequently employed as feature extractor in various VTON methods, including MGD [3] and PBE [27]. However, CLIP [29] is pre-trained to align with the

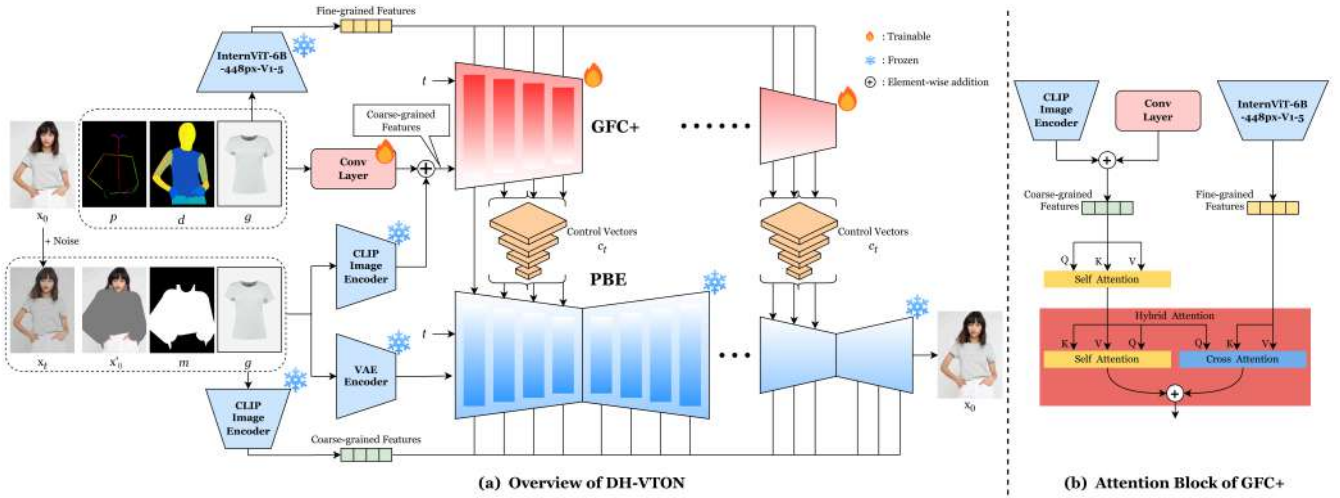


Fig. 2. **Overview of DH-VTON.** We demonstrate the training pipeline of our DH-VTON and details of the attention block. **(Left)** DH-VTON comprises a fixed-parameter PBE [27] and a trainable GFC+. Apart from the given noisy image \mathbf{x}_t , mask m , masked image \mathbf{x}'_0 , garment image g , time steps t , GFC+ generates a set of control vectors c_t by incorporating additional control conditions, such as pose p and densepose d . Control vectors are integrated into PBE [27] to enhance the model’s controllability while preserving PBE’s generation capabilities. **(Right)** We introduce a hybrid attention strategy in GFC+ to ensemble different layers of fine-grained characteristics for multi-scale features preservation.

holistic features of coarse textual captions [30], [31]. Therefore, the extracted features are usually also coarse-grained, which may lead to undesirable effects [32]. Recent methods have improved the garment feature extraction abilities by utilizing a tale of two UNet modules, such as TryOnDiffusion [28], OOTDiffusion [5], and IDM-VTON [6]. However, these methods still suffer from preserving meticulous details of garments, dampening their applications to real-world scenarios.

Driven by the above issues, we propose DH-VTON, a deep text-driven virtual try-on model featuring a special hybrid attention learning strategy and deep garment semantic preservation module. Specifically, inspired by the success of PBE [27], we present our DH-VTON pipeline in this work. Furthermore, for extracting the deep semantics of the garments, we are the first to introduce InternViT-6B [33] into VTON tasks as fine-grained feature learner, which can be trained to align with the large-scale intrinsic knowledge with deep textual semantics to compensate for the deficiency of the commonly adopted CLIP [29] encoder. On this basis, to enhance the customized dressing capabilities, we further design GFC+ module and propose to utilize a novel hybrid attention strategy for training, which can adaptively integrate fine-grained characteristics of the garments into the different layers of the VTON model, so as to achieve multi-scale features preservation.

Experiments on two representative datasets VITON-HD [8] and DressCode [34] demonstrate the effectiveness of the proposed DH-VTON model, showing that it achieves competitive performance against previous warping-based or warping-free models. DH-VTON not only significantly enhances the fine-grained semantic extraction of the given garments but also effectively captures and preserves the texture details.

II. METHOD AND METHODOLOGY

To effectively improve the fine-grained semantic extraction abilities and accurately preserve the texture details, we

propose a novel deep text-driven virtual try-on (DH-VTON) model, which integrates a special hybrid attention strategy and deep garment semantic preservation module, as depicted in Fig. 2(a). DH-VTON mainly contains two parts: a fixed-parameter PBE [27] and a trainable GFC+. The former aims to ensure high realism of generated images, while the latter aims to further enhance the customized dressing abilities.

A. ControlNet Architecture

Given a target person image \mathbf{x}_0 , DH-VTON gradually adds noise to \mathbf{x}_0 , receiving a noisy image \mathbf{x}_t , with t representing the frequency of noise addition. And given a group of conditions including noisy image \mathbf{x}_t , mask m , masked image \mathbf{x}'_0 , given garment image g , time steps t as well as additional control conditions (e.g., pose p and densepose d), GFC+ generates a suite of control vectors c_t . Then these vectors are incorporated into the SD Middle Block and the skip-connections of PBE’s UNet, consequently guiding the generation process of PBE [27]. Similar to LDM [22], DH-VTON learns a network ϵ_θ to predict the noise added to the noisy image \mathbf{x}_t with:

$$\mathcal{L}_{DH-VTON} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{x}'_0, m, g, p, d, t)\|_2^2], \quad (1)$$

where $t \in \{1, \dots, T\}$ denotes the time step of the forward diffusion process, \mathbf{x}_0 is the target person image and \mathbf{x}_t is \mathbf{x}_0 with the added standard Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$.

B. Garment Feature Extraction

To make up for the deficiency of the commonly adopted CLIP [29] encoder, we are the first to introduce InternViT-6B [33] into VTON tasks as fine-grained feature learner to extract the deep semantics of the garments.

Specifically, InternViT-6B [33] first dynamically matches the image to the optimal aspect ratio from a set of pre-defined aspect ratios. Once the appropriate aspect ratio is determined, the image is resized to the corresponding resolution. For

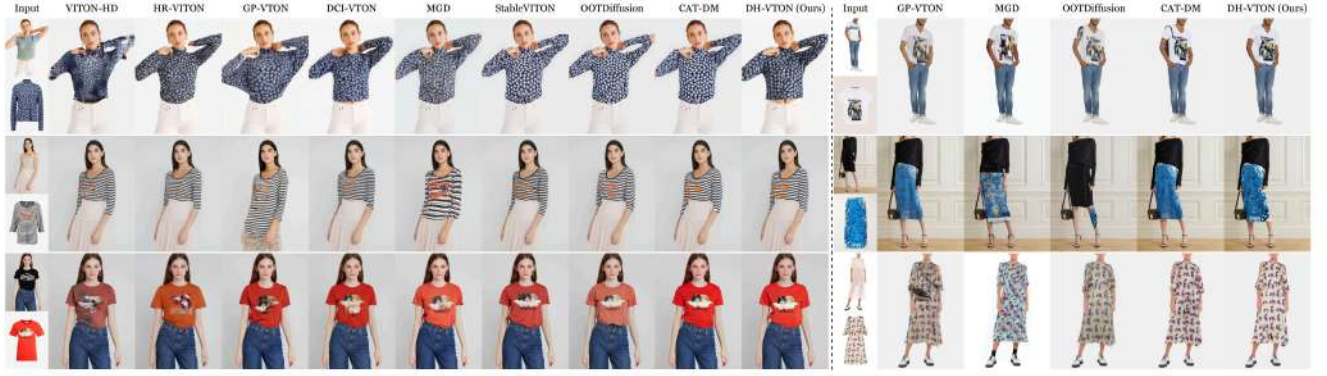


Fig. 3. Qualitative results on VITON-HD and DressCode test datasets. Please zoom in for more details.

TABLE I

ABLATION STUDY OF HYBRID ATTENTION AND DIFFERENT VALUES OF λ ON VITON-HD [8]. THE BEST AND SECOND BEST RESULTS ARE REPORTED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Hybrid Attention	λ	Paired				Unpaired	
		SSIM \uparrow	FID \downarrow	KID \downarrow	LPIPS \downarrow	FID \downarrow	KID \downarrow
\times	-	0.795	7.63	4.21	0.1304	14.46	5.25
\checkmark	0.25	0.826	6.54	2.36	0.1072	12.47	3.88
\checkmark	0.5	0.863	5.92	0.98	0.0861	10.92	1.62
\checkmark	0.75	<u>0.871</u>	<u>5.58</u>	<u>0.71</u>	<u>0.0589</u>	<u>9.31</u>	1.07
\checkmark	1.0	0.874	5.53	0.67	0.0562	9.02	<u>1.08</u>
\checkmark	1.25	0.859	6.12	1.96	0.0938	11.29	2.76
\checkmark	1.5	0.801	8.01	4.83	0.1395	15.03	6.09

TABLE II

ABLATION STUDY OF DH-VTON WITHOUT OR WITH GFC+ ON VITON-HD [8].

GFC+	Paired				Unpaired	
	SSIM \uparrow	FID \downarrow	KID \downarrow	LPIPS \downarrow	FID \downarrow	KID \downarrow
\times	0.763	14.32	5.44	0.2254	15.77	6.22
\checkmark	0.874	5.53	0.67	0.0562	9.02	1.08

example, as shown in Fig. 1, an 768×1024 image is resized to 896×1344 . After that, the resized image is divided into 6 tiles of 448×448 pixels and each tile is processed independently. In addition to these tiles, a 448×448 thumbnail of the entire image is also included to capture the global context for comprehending the overall features.

C. Hybrid Attention Learning

In order to adaptively integrate fine-grained characteristics of the garments into the different layers of the VTON model, as shown in Fig. 2(b), we propose to leverage a fresh hybrid attention strategy for training, accordingly achieving multi-scale features preservation. Here, assuming \mathbf{O}_s represents the output of self attention and \mathbf{I}_g represents the fine-grained features from InternViT-6B [33] at corresponding positions, the output of hybrid attention \mathbf{O}_h can be defined as follows:

$$\mathbf{O}_h = \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)}_{\text{Self Attention}} \mathbf{V} + \lambda \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}}\right)}_{\text{Cross Attention}} \mathbf{V}', \quad (2)$$

where $\lambda \in [0, 1.5]$ is a hyper-parameter to control the scale of fine-grained features. Note that we share a query matrix

TABLE III

ABLATION STUDY OF DH-VTON WITH DIFFERENT FEATURE EXTRACTORS ON VITON-HD [8].

Extractor	Paired				Unpaired	
	SSIM \uparrow	FID \downarrow	KID \downarrow	LPIPS \downarrow	FID \downarrow	KID \downarrow
CLIP [29]	0.853	7.90	1.38	0.1111	10.21	1.77
IP-Adapter [35]	0.847	8.13	2.86	0.1127	11.23	3.90
DINO-V2 [36]	<u>0.862</u>	<u>7.11</u>	<u>1.12</u>	<u>0.0988</u>	<u>9.67</u>	<u>1.36</u>
InternViT-6B [33]	0.874	5.53	0.67	0.0562	9.02	1.08

\mathbf{Q} for both self attention and cross attention. $\mathbf{Q} = \mathbf{O}_s \mathbf{W}_q$, $\mathbf{K} = \mathbf{O}_s \mathbf{W}_k$, $\mathbf{V} = \mathbf{O}_s \mathbf{W}_v$, $\mathbf{K}' = \mathbf{I}_g \mathbf{W}'_k$, and $\mathbf{V}' = \mathbf{I}_g \mathbf{W}'_v$. Here, \mathbf{W}_q , \mathbf{W}_k , \mathbf{W}_v , \mathbf{W}'_k , and \mathbf{W}'_v are the weight matrices of the trainable linear projection layers.

III. EXPERIMENTS

A. Experimental Setup

a) *Datasets and Metrics*: Our experiments are performed on two high-resolution (768×1024) VTON datasets, i.e., VITON-HD [8] and DressCode [34]. And test experiments are conducted under both paired and unpaired settings. In the paired and unpaired settings, we employ FID [37] and KID [38] for realism and fidelity assessment. Furthermore, in the paired setting with available ground truth, we additionally employ LPIPS [39] and SSIM [40] to evaluate the coherence of VTON images.

b) *Baselines*: For more holistic comparisons, we compare DH-VTON with the two categories of baseline models: 1) warping-based models, including VITON-HD [8], HR-VITON [9], GP-VTON [17], and DCI-VTON [10]; 2) warping-free models, including MGD [3], StableVITON [4], OOTDiffusion [5], and CAT-DM [7].

c) *Implementation Details*: During the experiments, we use an end-to-end training process. All experiments are conducted on four NVIDIA RTX A6000 GPUs with a batch size of 2. We utilize the AdamW optimizer and set the learning rate to 3×10^{-5} . Moreover, the hyper-parameter λ is searched from $\{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$.

B. Ablation Studies

a) *Hyper-parameter λ* : We investigate the effect of hybrid attention learning as well as the different values of the

TABLE IV
QUANTITATIVE RESULTS UNDER BOTH PAIRED AND UNPAIRED SETTINGS ON VITON-HD [8] AND DRESSCODE [34] TEST DATASETS.

Methods	VITON-HD						DressCode					
	Paired				Unpaired		Paired				Unpaired	
	SSIM \uparrow	FID \downarrow	KID \downarrow	LPIPS \downarrow	FID \downarrow	KID \downarrow	SSIM \uparrow	FID \downarrow	KID \downarrow	LPIPS \downarrow	FID \downarrow	KID \downarrow
VITON-HD [8]	0.848	12.81	5.52	0.1216	14.64	6.10	-	-	-	-	-	-
HR-VITON [9]	0.860	9.92	3.06	0.1038	12.15	3.42	-	-	-	-	-	-
DCI-VTON [10]	0.862	9.41	4.55	0.0606	12.53	5.25	-	-	-	-	-	-
StableVITON [4]	0.854	6.44	0.94	0.0905	11.05	3.91	-	-	-	-	-	-
GP-VTON [17]	0.871	8.73	3.94	<u>0.0585</u>	11.84	4.31	0.771	9.93	4.61	0.1801	12.79	6.63
MGD [3]	0.827	11.12	3.38	0.1280	13.34	3.93	0.786	8.24	3.27	0.1078	10.19	5.84
OOTDiffusion [5]	0.819	9.31	4.09	0.0876	12.41	4.69	0.885	4.61	<u>0.96</u>	<u>0.0533</u>	12.57	6.63
CAT-DM [7]	0.877	<u>5.60</u>	<u>0.83</u>	0.0803	8.93	<u>1.37</u>	0.866	4.17	1.21	0.0674	8.22	<u>1.98</u>
DH-VTON (Ours)	<u>0.874</u>	5.53	0.67	0.0562	<u>9.02</u>	1.08	<u>0.881</u>	3.79	0.84	0.0435	6.31	1.36



Fig. 4. **Effect of λ .** We compare the results of DH-VTON trained without/with hybrid attention strategy and using different values of λ .



Fig. 5. **Effect of InternViT-6B.** We compare the results of DH-VTON when using different feature extractors.

guidance scale λ on VITON-HD [8]. Experimental results are presented in Fig. 4 qualitatively and Tab. I quantitatively. We can find that the optimal λ value is around 1.0 on VITON-HD [8]. Meanwhile, for more complicated dress images of DressCode [34], a larger λ is needed to match more complex and detailed garment features. According to this study, we consistently conduct hybrid attention learning for DH-VTON, and empirically set $\lambda = 1.0$ for VITON-HD [8] and $\lambda = 1.25$ for DressCode [34] in the following experiments.

b) *Effect of InternViT-6B:* We conduct a series of ablation studies to investigate the effect of InternViT-6B [33]. Experimental results on how different feature extractors affect the performance of DH-VTON are illustrated in Fig. 5 qualitatively and Tab. III quantitatively. With the integration of InternViT-6B [33], DH-VTON obtains the most realistic and natural VTON results and has shown great progress and improvement across all metrics on VITON-HD [8].

c) *Effect of GFC+:* We also investigate the effect of GFC+ on VITON-HD [8]. Experimental results are shown in



Fig. 6. **Effect of GFC+.** We compare the results of DH-VTON trained without/with GFC+.

Fig. 6 qualitatively and Tab. II quantitatively. GFC+ visually enhances the customized dressing abilities of preserving the textures and patterns of the given garments (e.g., the texts and graphics of t-shirts) and quantitatively improves all evaluation metrics, which consistently shows the superior of our model.

C. Qualitative Results

Some test results of DH-VTON compared to other VTON methods on VITON-HD [8] and DressCode [34] datasets are visually shown in Fig. 3. Compared with other methods, we can observe that our DH-VTON significantly achieves the best try-on effect for various kinds of garments. Moreover, our DH-VTON not only generates realistic images but also preserves most of the fine-grained garment details.

D. Quantitative Results

The quantitative comparisons between DH-VTON and other methods are minutely reported in Tab. IV. DH-VTON outperforms other methods on the majority of metrics, particularly in KID [38] and LPIPS [39], demonstrating its effectiveness in image generation quality on both paired and unpaired tasks.

IV. CONCLUSION

In this paper, we present DH-VTON, a deep text-driven virtual try-on model. We are the first to introduce InternViT-6B [33] into VTON tasks as fine-grained feature learner, which significantly improves the deep semantic extraction abilities. Besides, we make full use of inherent power within PBE [27] and design an additional GFC+ module to enhance the customized dressing abilities. Based on this, we further propose a fresh hybrid attention strategy to ensemble different layers of fine-grained characteristics for multi-scale features preservation. Experiments on the two representative datasets demonstrate the effectiveness and superior of our approach.

REFERENCES

- [1] Z. Ma, L. Zhao, B. Qi, and B. Zhou, “Neural residual diffusion models for deep scalable vision generation,” *arXiv preprint arXiv:2406.13215*, 2024.
- [2] X. Yang, C. Ding, Z. Hong, J. Huang, J. Tao, and X. Xu, “Texture-preserving diffusion models for high-fidelity virtual try-on,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7017–7026.
- [3] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, “Multimodal garment designer: Human-centric latent diffusion models for fashion image editing,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 393–23 402.
- [4] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, “Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8176–8185.
- [5] Y. Xu, T. Gu, W. Chen, and C. Chen, “Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on,” *arXiv preprint arXiv:2403.01779*, 2024.
- [6] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, “Improving diffusion models for virtual try-on,” *arXiv preprint arXiv:2403.05139*, 2024.
- [7] J. Zeng, D. Song, W. Nie, H. Tian, T. Wang, and A.-A. Liu, “Catdm: Controllable accelerated virtual try-on with diffusion model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8372–8382.
- [8] S. Choi, S. Park, M. Lee, and J. Choo, “Vitton-hd: High-resolution virtual try-on via misalignment-aware normalization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 131–14 140.
- [9] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, “High-resolution virtual try-on with misalignment and occlusion-handled conditions,” in *European Conference on Computer Vision*. Springer, 2022, pp. 204–219.
- [10] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang, “Taming the power of diffusion models for high-quality virtual try-on with appearance flow,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7599–7607.
- [11] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, “Vitton: An image-based virtual try-on network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7543–7552.
- [12] B. Fele, A. Lampe, P. Peer, and V. Struc, “C-vton: Context-driven image-based virtual try-on network,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3144–3153.
- [13] X. Han, X. Hu, W. Huang, and M. R. Scott, “Clothflow: A flow-based model for clothed person generation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 471–10 480.
- [14] A. Chopra, R. Jain, M. Hemani, and B. Krishnamurthy, “Zflow: Gated appearance flow-based virtual try-on with 3d priors,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5433–5442.
- [15] C.-Y. Chen, Y.-C. Chen, H.-H. Shuai, and W.-H. Cheng, “Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7513–7522.
- [16] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, “Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8580–8589.
- [17] Z. Xie, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, F. Zhu, and X. 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*, 2023, pp. 23 550–23 559.
- [18] T. Issenhuth, J. Mary, and C. Calauzenes, “Do not mask what you do not need to mask: a parser-free virtual try-on,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 619–635.
- [19] H. Yang, X. Yu, and Z. Liu, “Full-range virtual try-on with recurrent tri-level transform,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3460–3469.
- [20] R. Yu, X. Wang, and X. Xie, “Vtnfp: An image-based virtual try-on network with body and clothing feature preservation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 511–10 520.
- [21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [23] J. Duchon, “Splines minimizing rotation-invariant semi-norms in sobolev spaces,” in *Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*. Springer, 1977, pp. 85–100.
- [24] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [25] Y. Li, C. Huang, and C. C. Loy, “Dense intrinsic appearance flow for human pose transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3693–3702.
- [26] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” *arXiv preprint arXiv:2404.16821*, 2024.
- [27] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, “Paint by example: Exemplar-based image editing with diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 381–18 391.
- [28] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, and I. Kemelmacher-Shlizerman, “Tryondiffusion: A tale of two unets,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4606–4615.
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [30] Z. Ma, J. Li, G. Li, and K. Huang, “Cmal: A novel cross-modal associative learning framework for vision-language pre-training,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4515–4524.
- [31] Z. Ma, J. Li, G. Li, and Y. Cheng, “Unitranser: A unified transformer semantic representation framework for multimodal task-oriented dialog system,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 103–114.
- [32] Z. Ma, Z. Yu, J. Li, and G. Li, “Hybridprompt: bridging language models and human priors in prompt tuning for visual question answering,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 371–13 379.
- [33] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” *arXiv preprint arXiv:2404.16821*, 2024.
- [34] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, “Dress code: High-resolution multi-category virtual try-on,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2231–2235.
- [35] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models,” *arXiv preprint arXiv:2308.06721*, 2023.
- [36] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
- [37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [38] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying mmd gans,” *arXiv preprint arXiv:1801.01401*, 2018.
- [39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.