ISSN: 2710-0588

MG-VTON: A Lightweight Virtual Try-on Model with Knowledge Distillation for Real-time Performance

Xuan Yu[1,*]

[1]Yunnan Normal University, Jvxian Street, Kunming, China

[*] Corresponding Author: xuanyuscience@163.com

Abstract

In recent years, virtual try-on technology has seen a continuous surge in public visibility and has become a key tool for many companies to boost sales and enhance user experience. Existing virtual try-on methods are mainly divided into two categories: those based on Generative Adversarial Networks (GANs) and those based on diffusion models. GAN-based methods have been widely applied due to their compact model structures and fast execution speed, but there is still room for improvement in image quality and detail fidelity. In contrast, diffusion model-based methods excel in generating high-quality and realistic images, but their high computational complexity and slow inference speed limit their practicality in real-time applications. To address these issues, this paper proposes a lightweight and efficient virtual try-on model called MG-VTON that does not require human parsing. By introducing knowledge distillation techniques, we have streamlined the model to significantly improve computational efficiency and inference speed. Moreover, MG-VTON can still generate high-quality and realistic try-on effects without relying on human parsing. This work offers new insights for the further development of virtual try-on technology, enhancing the user experience and providing companies with more competitive solutions in digital apparel presentation and marketing.

Keywords

Virtual try-on, Generative Adversarial Networks (GANs), knowledge distillation, real-time performance.

1. Introduction

With the rapid advancement of information technology, online shopping has become an integral part of daily life, particularly in the context of clothing purchases. However, because online shopping cannot provide the physical try-on experience that offline shopping offers, many consumers often face challenges related to sizing, style, and fit when selecting clothes, which in turn affects their purchasing decisions. To address this limitation, virtual try-on (VITON) technology has emerged. It aims to simulate a near-realistic try-on experience using digital technologies, thereby enhancing users' shopping experiences and encouraging purchase behavior. This development not only improves the convenience of online shopping but also brings new sales models and business opportunities to the fashion retail industry.

Traditional virtual try-on technologies primarily rely on 3D reconstructions of real human bodies [1] and clothing models [2], employing complex physical simulations to achieve dynamic dressing effects. However, these methods require processing numerous parameters, leading to high computational costs for model training and demanding substantial hardware resources, which typical users may not have. Additionally, the collection and processing of these parameters involve significant time and labor, raising concerns about user privacy. These approaches also fall short in terms of simulation realism and privacy protection. Therefore,

ISSN: 2710-0588

developing a low-cost, high-quality, and user-friendly virtual try-on solution has become a shared goal of both academic and industry efforts.

In 2017, VITON [3] introduced the first 2D image-based virtual try-on method, which leveraged human pose estimation and segmentation techniques to effectively address the challenge of limited supervised training data, significantly reducing the reliance on 3D reconstructions. This method breathed new life into the virtual try-on field, marking the beginning of a new era in 2D virtual try-on technology.

As virtual try-on technology continued to evolve, researchers proposed various innovative methods to tackle emerging challenges. For instance, knowledge distillation techniques were introduced to minimize the impact of manual parsing during model training. In this approach, a teacher network guides the student network's learning process, leading to more efficient model training, as demonstrated by PF-AFN [4]. With improvements in computer hardware, technologies for generating high-resolution and high-quality images have also progressed, such as VITON-HD [5] and HR-VTON [6], further enhancing the realism of virtual try-on results. Inspired by Generative Adversarial Networks (GANs) and diffusion models, single-stage network architectures [7] and diffusion-based networks [8] have emerged, providing new momentum for virtual try-on research. Currently, GANs play a critical role in virtual try-on technology due to their compact structure and training efficiency, making them widely used in generating clothing images.

In recent years, diffusion models have garnered increasing attention due to their unique forward and reverse diffusion learning processes. These models gradually introduce and remove noise step by step, in combination with conditional control, achieving significant progress in image realism and detail restoration. Specifically, diffusion models introduce Gaussian noise incrementally during forward diffusion and then reverse the process step by step to denoise, ultimately generating a clear, detailed image. This process brings new possibilities to the virtual try-on field, especially for generating intricate clothing details that do not exist in the original images. While diffusion models still face challenges, particularly in preserving the original characteristics of the clothing, recent studies have significantly improved the controllability of these models [9], laying a strong foundation for their application in virtual try-on systems.

However, the computational resources and time required for image generation using diffusion models remain major bottlenecks. With the rise of the mobile era, the demand for faster response times has increased dramatically, yet diffusion models are still too slow during inference. In contrast, GANs hold a significant advantage in terms of speed and model size. Consequently, in this paper, we adopt knowledge distillation techniques in GAN-based methods to ensure high-quality image generation while substantially reducing model size, making them more suitable for real-time applications. In diffusion model-based methods, we explore their application in image restoration, improving existing models and retraining them to achieve higher-quality image generation.

In summary, the main contributions of this paper are as follows:

We introduce knowledge distillation techniques to build a novel, lightweight GAN-based model.

The new student network offers faster inference speed and a smaller model size, making it suitable for resource-constrained environments.

2. Related Work

Since the introduction of the innovative two-stage strategy proposed by VITON [3], virtual try-on has entered a new era. Currently, virtual try-on approaches can be broadly divided into two categories based on their underlying network architectures: GAN-based virtual try-on and diffusion model-based virtual try-on.

ISSN: 2710-0588

Most GAN-based virtual try-on methods follow a two-stage strategy: (1) First, the clothing is deformed to fit the pose of the target person as closely as possible. (2) Then, a GAN is used to fuse the deformed clothing image with the person's body, excluding the clothing, to generate the final image.

In the first stage, various methods like TPS [10], STN [11], Flow [12], and Implicit transformations are commonly used to deform the clothing so it fits the target pose. The quality of this deformation largely determines the final outcome. Thin Plate Spline (TPS) [10] treats a 2D image as a thin plate and deforms it by moving control points, causing the plate to bend or stretch. TPS minimizes the second derivative of the deformation, smoothing the image and avoiding unnatural folds or sharp angles. While TPS is effective for global transformations, it is limited in controlling fine local details. Spatial Transformer Networks (STN) [11] is a deep learning framework consisting of three parts: the Localization Network, which adjusts the deformation location; the Grid Generator, which generates the target grid; and the Sampler, which samples the input image and outputs the transformed image. STN is highly adaptable and can be improved by better training data, though the quality of the dataset directly affects performance. Optical Flow (Flow) [12] assumes that each pixel moves and calculates its next position, providing pixel-level control for detailed transformations. Implicit methods, instead of using explicit spatial transformations, align clothing and body features in the feature space through deep neural networks. O-VITON [13] aligns clothing features in the body region, while TryOnGAN [7] utilizes StyleGAN's [14] feature alignment, ensuring clothing is aligned with the target pose. While implicit methods enhance clothing deformation, controlling the finer details of the clothing remains challenging. Finally, a GAN is often used to generate the final deformed clothing image.

In the second stage, a GAN is directly applied to generate the image, significantly affecting the final quality. Numerous methods have been explored to address unnatural artifacts. Early works like VITON [3] and CP-VITON [15] focused on the try-on process, achieving good results for simple clothing in central regions but causing blurring in non-clothing areas. CP-VITON+ [16] improved upon this by preserving non-clothing regions, mitigating some of these issues. Later methods like PFAFN [4] and Flow-Style-VTON [17] made significant improvements in the deformation module, resulting in clearer boundaries and more natural-looking images. Additionally, networks specifically designed for high-resolution image generation emerged, such as VITON-HD [5], which introduced misalignment-aware normalization to better align the clothing with the person, reducing misalignment and occlusion. HR-VITON [6] emphasized the coupling between body features and clothing, generating more natural results. Both methods enhanced image details after synthesis.

Recently, diffusion models have also been applied to this field. Originally, diffusion models aimed to eliminate Gaussian noise [18], but with the introduction of DDPM [19], their potential for image generation became apparent. Diffusion models consist of a forward process and a reverse process. In the forward process, noise is progressively added to the original image, represented as $p_\theta(x0) \coloneqq \int p\theta(x0{:}T)\,d\,x1{:}T$, where x_0 is the clean image and $x_1, \ldots, x_T$ are latent variables. The joint distribution $p_\theta(x0{:}T)$ is learned from the Gaussian distribution$p(xT) = \mathcal{N}(xT; 0, I)$, and noise $\beta 1, \ldots, \beta_T$ is added step by step in a Markov chain. The distribution at time step t after adding noise to the previous step $xt - 1$ is:

$$q(\,xt \mid x_{t-1}\,) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{1}$$

The denoising process is represented by the retention ratio $\alpha_t = 1 - \beta_t$, and the cumulative ratio $\bar{\alpha}t \coloneqq \prod s = 1^t\alpha_s$ represents the total change from the initial data to time step t. The distribution of x_t at time step t, given the initial data x_0, is:

ISSN: 2710-0588

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}}tx0, (1 - \bar{\alpha}_t)I\right) \quad (2)$$

The reverse process restores the noisy image to its original form. If the forward process can be reversed, i.e., by sampling from $q(x_t \mid x_{t-1})$, the original image can be restored from a random Gaussian sample $\mathcal{N}(0, I)$. Using $\mu_\theta(x_t, t)$, the mean predicted by the neural network θ, and σ_t^2, the variance at time step t, the reverse process formula is as follows:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (3)$$

Latent Diffusion Models (LDM) [20] introduced a pioneering cross-attention layer that allows flexible control of image generation through different modalities, laying a solid foundation for subsequent diffusion models.

TryOnDiffusion [9] employed two parallel Unets for training, achieving impressive results, although this approach requires significant computational resources and challenging datasets. Consequently, many recent methods have focused on fine-tuning large pre-trained models [21] or guiding pre-trained models [22]. LaDI-VTON [8] used LDM to convert clothing image features into CLIP tokens to guide image generation. DCI-VITON [21] utilized a two-stage strategy, first deforming the clothing and then using a diffusion model to restore the masked region. CAT-DM [22] and StableVITON [23] adopted ideas from ControlNet [24], using part of a Unet to enhance the detail guidance for clothing. OOTDiffusion [25] encoded both clothing images and text descriptions and input them into a Unet for guidance. IDM-VTON [26] used IP-Adapter [27] to fine-tune TryOnNet, enhancing the background and target clothing generation. These studies show that diffusion-based methods produce more natural and realistic images, though due to the stochastic nature of diffusion models, inconsistencies in clothing detail generation may still arise.

In conclusion, both approaches have their strengths. Diffusion models excel in generating fine details but require large-scale models and significant computational resources, leading to slower inference times. Additionally, while diffusion models generate high-quality details, there is still room for improvement in maintaining consistency. In contrast, GAN-based methods offer faster inference with smaller model sizes, making them more suitable for resource-constrained environments requiring quick results.

3. Preliminaries

In terms of clothing deformation, this paper primarily focuses on the method introduced by Flow [12]. This method captures the displacement of pixels or features before and after transformation. Let (u_x, u_y) represent the displacement. At the target position (x, y), sampling is performed from $(x - u_x, y - u_y)$ in the original distribution, with bilinear interpolation used for non-integer coordinates. This enables varying levels of prediction complexity, depending on whether the prediction involves pixel-level details or higher-level features, and whether it is handled on a single layer or across multiple layers.

Single-layer optical flow networks perform sampling at a single feature level, while multi-layer optical flow networks sample at multiple feature levels, progressing from coarse to fine. To achieve more natural and realistic clothing deformations, modern multi-layer optical flow methods are more refined compared to traditional optical flow networks. These methods estimate various factors, including body posture and key points, using multiple optical flow maps, resulting in more detailed deformation results.

ISSN: 2710-0588

Since the introduction of GANs [28], they have become a cornerstone of image generation. The goal of GANs is to learn the data distribution and generate images that closely resemble those in the training dataset. The fundamental idea of GANs is based on adversarial training between a generator network and a discriminator network. The generator strives to maximize the probability that its generated images are classified as real by the discriminator, while the discriminator aims to maximize its ability to distinguish real images from generated ones.

The process begins by defining a prior on the input noise variable $p_z(z)$. To learn the generator's data distribution $p_g(x)$, the generator maps the noise variable z to the data space through a differentiable function $G(z;\theta_g)$, where G is parameterized by θ_g. A multilayer perceptron $D(x;\theta_d)$ is then used as the discriminator, outputting a scalar value representing the probability that x comes from the real data distribution rather than from p_g.

By jointly training both the discriminator D and the generator G, the discriminator learns to distinguish real from generated data, maximizing the probability of correct classification. Meanwhile, the generator minimizes the objective function $\log\left(1-D\left(G(z)\right)\right)$, learning to generate more realistic data that makes it harder for the discriminator to differentiate. In essence, D and G engage in a minimax game, with the value function $V(G,D)$ expressed as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{dt}(x)}[\log D(x)]\backslash notag + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad (4)$$

4. Methodology

To fully leverage the advantages of GANs, this paper employs a knowledge distillation approach to minimize model size and reduce input parameters. One of the most advanced GAN models in this field, GP-VTON[29], is selected as the teacher network for this study, and a new student network is constructed. The overall network architecture is shown below Figure 1:

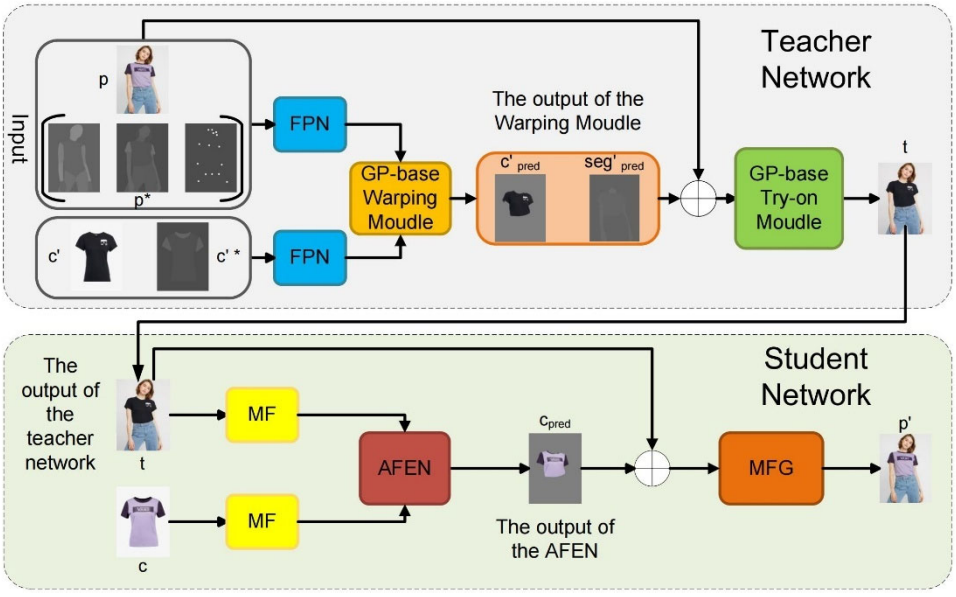

Figure 1. MG-VTON Overall Network Structure.

ISSN: 2710-0588

In this figure, p represents the original person image; p^* is the parsed person image, including various auxiliary images such as dense pose maps, human keypoint estimations, and segmentation maps; c′ is a randomly selected garment from the training set that differs from the person's original garment; c'^* represents auxiliary images for the garment, such as segmentation maps; seg'_{pred} is the predicted segmentation structure; c'_{pred} is the warped garment output from the teacher network; and t is the teacher network's final output. Similarly, c is the image of the garment originally worn by the person, c_{pred} is the warped garment output from the student network, and p′ is the final output of the student network.

First, the person image p is selected, and a different garment c′ is randomly chosen. In the Warp stage of the teacher network, garment warping is performed based on parsed information from the person image p^* and the garment image c'^*, resulting in a warped garment image. This image is then fed into the Try-on stage of the teacher network to generate the try-on result t. The try-on result t and the original garment image c are then passed through the Warp stage of the student network to obtain the warped original garment, which is subsequently input into the Try-on stage of the student network, generating the reconstructed person image p′.

In the teacher network GP-VTON[29], two Feature Pyramid Networks are initially applied to extract five multi-scale features from both the condition and input images. These features are then fed into a warping module to obtain both local and global transformations. The garment is decomposed into three parts: left, middle, and right. Local transformations are applied to each part individually, and the decomposed, warped garment images are then combined via global parsing to produce the complete warped garment image.

In the student network, a lightweight structure is achieved by skipping garment decomposition and performing holistic transformations. The main components are: an MF network for feature extraction, an AFEN network for flow-based warping, and an MFG network for generating the try-on image.

The structure of the MF network is shown in Figure 2. Its primary architecture is similar to the Feature Pyramid Network in the teacher network but is made lighter by introducing the UIB module from MobileNetv4[30]. This module is used to construct the feature pyramid, with additional expansion layers to enhance accuracy. The UIB module builds on the concept of Inverted Residuals, adding two optional depthwise convolutions (Starting depthwise conv and Expansion conv) to achieve four different instantiations, including ConvNext. Spatial mixing is performed before expansion, allowing for a larger receptive field as needed. Inverted Residuals enable spatial mixing for expanded features, allowing efficient and accurate feature extraction with a smaller network structure. ExtraDW can cheaply increase the network's depth and field of view, combining the advantages of ConvNext and Inverted Residuals, while FFN accelerates operations to maximize computational efficiency. Similar to the teacher network, a five-layer feature pyramid is constructed, with the person and garment images input separately to extract features. Since the teacher network has more inputs than the student network, a tunable knowledge distillation scheme[4] is used, employing a knowledge distillation loss to guide feature extraction in the student network.

$$L_{\text{dis}} = \psi \sum_{i=1}^{N} ||t_{p_i} - s_{p_i}||_2 \tag{5}$$

$$\psi = \begin{cases} 1, & if \|t-p\|_1 < \|s-p\|_1, \\ 0, & otherwise. \end{cases} \tag{6}$$

ISSN: 2710-0588

Here, ψ is the adjustment factor, and t_{p_i} and s_{p_i} represent the features extracted at the i-th layer of the feature pyramid in the teacher and student networks, respectively.

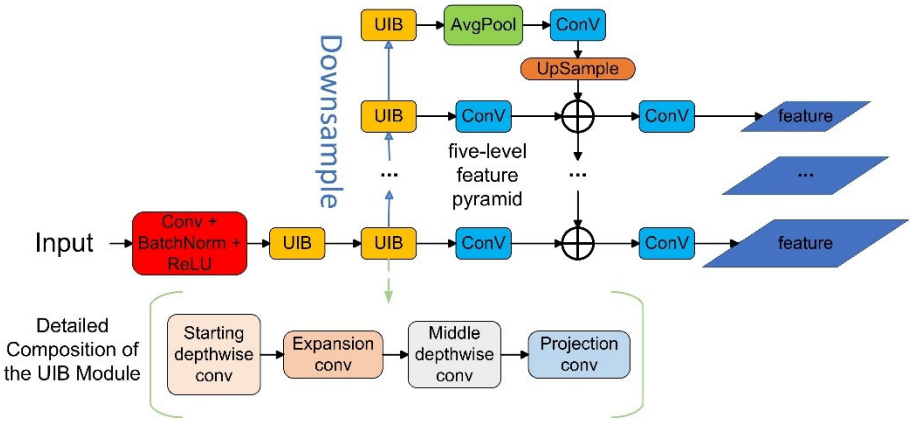

Figure 2. MF Network Architecture.

Next, the extracted features are input into the warping network. Following the method by Ge et al.[4], the AFEN network is used for synthesis. This module aims to warp the garment image according to the human pose image while preserving the garment's original texture, aligning as closely as possible with the person's posture and any occlusions. The module consists of multiple convolutional networks of varying sizes, which estimate multi-level feature maps from the MF module to produce an overall deformation estimation map for the garment. This map is then applied to the garment image to obtain the warped garment output. To enhance the preservation of garment features, the module is optimized using a second-order smoothness loss:

$$L_{sec} = \sum_{i=1}^{N} \sum_{t} \sum_{\pi \in N_t} \text{CharLoss}(f_i^{t-\pi} + f_i^{t+\pi} - 2f_i^t) \tag{7}$$

Here, f_i^t is the t-th point in the flow map at the i-th layer of the feature pyramid; N_t represents the set of horizontal, vertical, and diagonal neighboring points around point t; and CharLoss denotes the generalized Charbonnier loss[31].

The final component, MFG, is the generation network, as shown in Figure 3. To achieve a compact yet accurate network structure, it incorporates the UIB module from MobileNetv4[30] and the Unet architecture[32]. In the Unet architecture, the encoder (downsampling path) and decoder (upsampling path) are implemented using UIB modules with different hyperparameters, such as varying strides and expansion rates.

Since the training is divided into two stages, the loss functions are also split accordingly. In the Warp stage, the loss function is defined as follows:

$$L_{warp} = \lambda_{warp} L_{warp} + \lambda_{per} L_{per} + \lambda_{sec} L_{sec} + \lambda_{dis} L_{dis} \tag{8}$$

Here, $L_l^{warp} = ||c_{pred} - p \odot m_{gt}||$; $p \odot m_{gt}$ represents the element-wise product of the person image p and its corresponding garment mask m_{gt} , retaining only the garment region and masking out other areas. L_l^{warp} calculates the pixel-level $L1$ loss between the generated image

ISSN: 2710-0588

and the real image. $L_{\text{per}}^{\text{warp}} = \sum_i || \Phi_i(c_{\text{pred}}) - \Phi_i(p \odot m_{\text{gt}}) ||$ is the perceptual loss[33]; $L_{\text{sec}}$is the second-order smoothness loss; and L_{dis} represents the distillation loss. These four types of losses together constitute the loss function for the Warp stage.

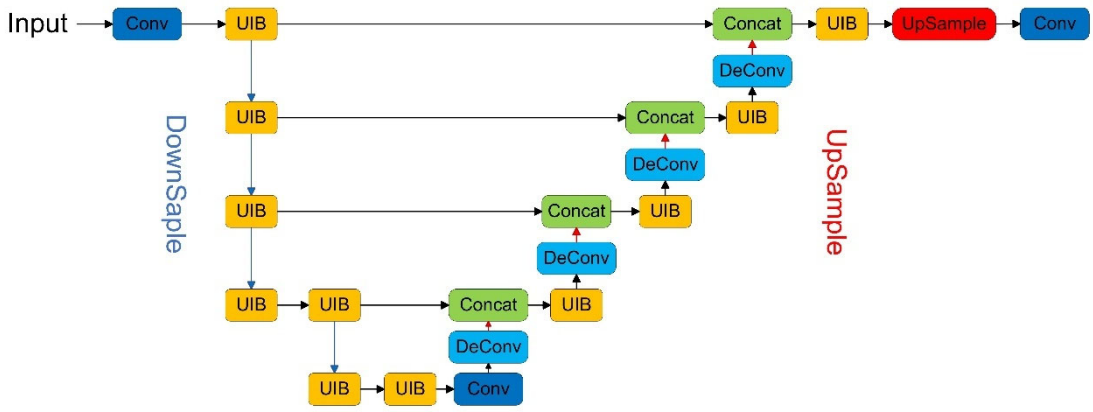

Figure 3. MFG Network Architecture.

In the GEN stage, the loss function is defined as follows:

$$\mathrm{L}^{\text{gen}} = \lambda_{\mathrm{l}}^{\text{gen}} \mathrm{L}_{\mathrm{l}}^{\text{gen}} + \lambda_{\text{per}}^{\text{gen}} \mathrm{L}_{\text{per}}^{\text{gen}} \tag{8}$$

Here, only the $L1$ loss and perceptual loss[33] are used to supervise the training of MFG by comparing the pixel-wise differences between the generated image and the original image.

5. Experiments

5.1. Experimental Setup

Dataset: In this study, we utilize the VITON-HD dataset[5], focusing primarily on images with a resolution of 512 × 384 pixels. The dataset comprises a total of 13,679 pairs of person and clothing images, which are further divided into 11,647 training pairs and 2,032 testing pairs. We process the dataset following the methodology proposed by Zhenyu Xie et al.[30], employing techniques from[38] and[39] to obtain 2D human pose images, dense pose images, and parsing maps for both persons and garments.

Benchmark and Evaluation Metrics: We compare our proposed MG-VTON model with several state-of-the-art parser-free virtual try-on methods, including PF-AFN[4], FS-VTON[40], and DM-VTON[32]. All these methods are retrained on the VITON-HD dataset[5] using the official code provided by the authors to produce results at a resolution of 512 × 384 pixels. The retrained models are then evaluated using the original test images to ensure a fair comparison.

Our evaluation focuses on two main aspects:

Quality of Generated Images: Assessed using the Fréchet Inception Distance (FID) and the Learned Perceptual Image Patch Similarity (LPIPS) metrics, which measure the similarity between the generated images and the ground truth.

Model Efficiency: Evaluated in terms of inference speed (frames per second, FPS), computational complexity (measured in gigaflops, GFLOPs), memory usage (in megabytes, MB), and the number of model parameters (in millions, M). These metrics provide insights into the practicality of deploying the models in real-world applications.

ISSN: 2710-0588

5.2. Qualitative Results

Table 1 presents the quantitative results of our experiments. As shown, our proposed MG-VTON model significantly reduces the number of parameters and memory consumption compared to existing methods, while still maintaining high-quality virtual try-on results. Specifically, MG-VTON achieves a favorable balance between performance and efficiency, making it suitable for scenarios where computational resources are limited.

Table 1. Quantitative Comparison of Different Methods on the VITON-HD Dataset.

Model	FID↓	LPIPS↓	FPS↑	Memory(MB)↓	GFLOPs↓	Parameters(M)↓
PF-AFN	9.36	0.197	16.53	279.30	275.71	73.20
FS-VTON	11.41	0.180	18.06	330.63	265.95	85.66
DM-VTON	11.18	0.223	24.69	35.82	139.65	9.35
MG-VTON	10.29	0.210	45.98	35.88	140.00	9.36

Figure 4 provides a visual comparison of the results produced by different methods. It illustrates how our MG-VTON model performs in comparison to other state-of-the-art methods on the VITON-HD dataset. As observed, MG-VTON generates images with realistic details and accurate garment fitting, demonstrating its superiority in visual quality.

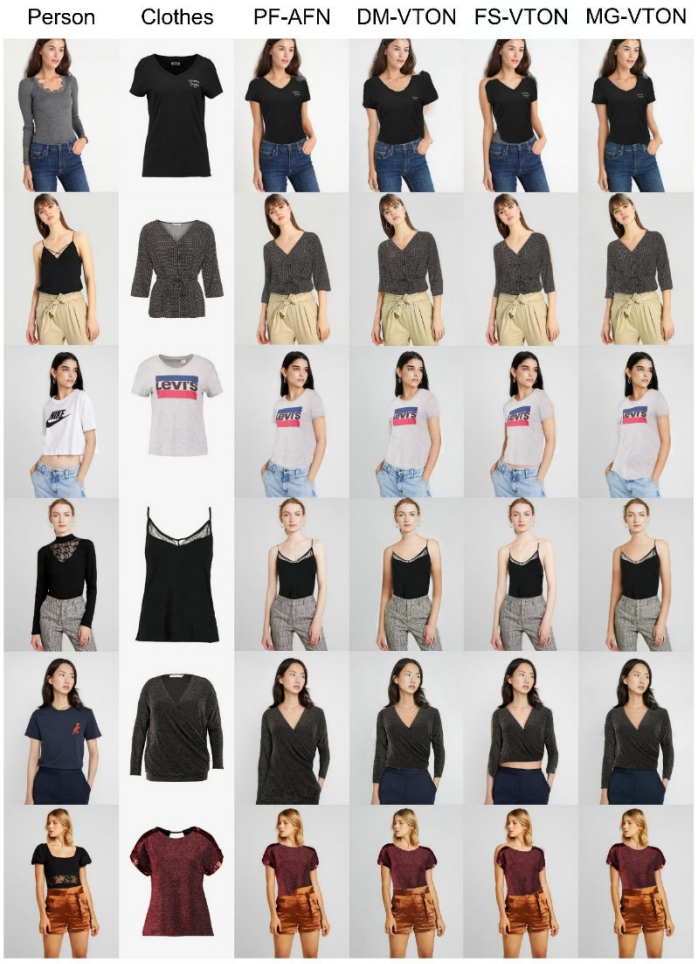

Figure 4. MFG Network Architecture

ISSN: 2710-0588

5.3. Ablation Study

Our approach primarily builds upon the methodology of DM-VTON[32], which serves as our baseline for conducting ablation studies to verify the effectiveness of our proposed modifications. Specifically, we assess the impact of replacing the teacher network, integrating the MF feature extraction module, and implementing the MG generation module.

Effect of Replacing the Teacher Network: We first keep the student network of DM-VTON unchanged but replace the teacher network with GP-VTON, resulting in a model we refer to as GPDM-VTON. We retrain the student network accordingly to observe the effect of the new teacher network on the student's learning process.

Integration of the MF Feature Extraction Module: Next, we introduce the MF module into the feature extraction network and conduct qualitative comparisons. Under the condition of using GP-VTON as the teacher network, we adopt the MF module for feature extraction, denoted as MG-VTON(MF). This allows us to evaluate the impact of the MF module on the model's ability to capture detailed features.

Implementation of the MG Generation Module: Finally, with the feature extraction modules all utilizing MF, we incorporate the MG generation module, resulting in our complete model, MG-VTON. This step assesses the combined effect of both the MF and MG modules on the overall performance.

Table 2. Performance Comparison in Ablation Study.

Model	FID↓	LPIPS↓	FPS↑	Memory(MB)↓	GFLOPs↓	Parameters(M)↓
DM-VTON	11.18	0.223	24.69	35.82	139.65	9.35
GPDM-VTON	10.86	0.234	38.35	35.82	139.65	9.35
MG-VTON(MF)	10.80	0.228	32.67	35.87	139.94	9.36
MG-VTON(Ours)	10.29	0.210	45.98	35.88	140.00	9.36

From the results presented in Table 2, we observe that each modification contributes to an improvement in the quality of the generated images. Specifically, replacing the teacher network with GP-VTON enhances the FID score, indicating better alignment with real images. The integration of the MF module further improves the LPIPS metric, reflecting enhanced perceptual similarity. The implementation of the MG module leads to the best overall performance, demonstrating the effectiveness of our proposed modules in enhancing the virtual try-on results.

These results confirm that each component of our proposed method contributes to enhancing the model's performance, both in terms of image quality and computational efficiency.

6. Conclusion

This paper presents a lightweight and efficient network called MG-VTON. By utilizing a mobile device framework, the network achieves strong real-time computational capabilities with relatively low resource consumption. Specifically, an advanced teacher network is employed to guide the student network's learning process, enabling the student network to generate high-quality images without relying on human parsing.

However, some challenges remain. When the person's pose is complex, the flow network struggles to effectively handle overlapping regions. This issue could be mitigated by improving the deformation network, AFEN. Additionally, due to the small size of the network model and the large size of the processed images, instability can occur during training. This problem could be addressed by slightly increasing the network's capacity.

ISSN: 2710-0588

Overall, the experimental results demonstrate the potential of the proposed method, which can be applied on mobile devices for rapid virtual try-on image generation, rather than relying on centralized large servers. Furthermore, its real-time capabilities make it well-suited for applications such as augmented reality (AR) and other virtual experiences.

References

[1] Tang M , Wang H , Tang L ,et al.CAMA: Contact-Aware Matrix Assembly with Unified Collision Handling for GPU-based Cloth Simulation[J].Computer Graphics Forum, 2016, 35(2):511-521.DOI:10.1111/cgf.12851.

[2] Cao C , Wu H , Weng Y ,et al.Real-time Facial Animation with Image-based Dynamic Avatars[J].ACM Transactions on Graphics, 2016, 35(4):1-12.DOI:10.1145/2897824.2925873.

[3] Han X , Wu Z , Wu Z ,et al.VITON: An Image-based Virtual Try-on Network[J]. 2017.DOI:10.1109/CVPR.2018.00787.

[4] Ge Y , Song Y , Zhang R ,et al.Parser-Free Virtual Try-on via Distilling Appearance Flows[J]. 2021.DOI:10.48550/arXiv.2103.04559.

[5] Choi S , Park S , Lee M ,et al.VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization[J]. 2021.DOI:10.48550/arXiv.2103.16874.

[6] Lee S , Gu G , Park S ,et al.High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions[J]. 2022.DOI:10.48550/arXiv.2206.14180.

[7] Lewis K M , Varadharajan S , Kemelmacher-Shlizerman I .TryOnGAN: body-aware try-on via layered interpolation[J].ACM Transactions on Graphics, 2021, 40(4):1-10.DOI:10.1145/3450626.3459884.

[8] Zhu L , Yang D , Zhu T L ,et al.TryOnDiffusion: A Tale of Two UNets[J].ArXiv, 2023, abs/2306.08276.DOI:10.48550/arXiv.2306.08276.

[9] Morelli D , Baldrati A , Cartella G ,et al.LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On[J].ArXiv, 2023, abs/2305.13501.DOI:10.48550/arXiv.2305.13501.

[10] Liu G , Song D , Tong R ,et al.Toward Realistic Virtual Try-on Through Landmark Guided Shape Matching[J]. 2021.DOI:10.1609/aaai.v35i3.16309.

[11] Wang B, Zheng H, Liang X, et al. Toward characteristic-preserving image-based virtual try-on network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 589-604.

[12] Cui A , Mckee D , Lazebnik S .Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-on and Outfit Editing[J]. 2021.DOI:10.48550/arXiv.2104.07021.

[13] Neuberger A , Borenstein E , Hilleli B ,et al.Image Based Virtual Try-On Network From Unpaired Data[J].IEEE, 2020.DOI:10.1109/CVPR42600.2020.00523.

[14] Gou J, Sun S, Zhang J, et al. Taming the power of diffusion models for high-quality virtual try-on with appearance flow[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 7599-7607.

[15] Zeng J, Song D, Nie W, et al. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8372-8382.

[16] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International conference on machine learning. PMLR, 2015: 2256-2265.

[17] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[18] Kim J, Gu G, Park M, et al. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8176-8185.

ISSN: 2710-0588

[19] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 3836-3847.

[20] Xu Y, Gu T, Chen W, et al. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on[J]. arXiv preprint arXiv:2403.01779, 2024.

[21] Choi Y, Kwak S, Lee K, et al. Improving diffusion models for virtual try-on[J]. arXiv preprint arXiv:2403.05139, 2024.

[22] Ye H, Zhang J, Liu S, et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models[J]. arXiv preprint arXiv:2308.06721, 2023.

[23] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

[24] Karras T. A Style-Based Generator Architecture for Generative Adversarial Networks[J]. arXiv preprint arXiv:1812.04948, 2019.

[25] Yang H, Zhang R, Guo X, et al. Towards photo-realistic virtual try-on by adaptively generating-preserving image content[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7850-7859.

[26] Minar M R, Tuan T T, Ahn H, et al. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on[C]//CVPR workshops. 2020, 3: 10-14.

[27] He S, Song Y Z, Xiang T. Style-based global appearance flow for virtual try-on[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3470-3479.

[28] Xie Z, Huang Z, Dong X, et al. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 23550-23559.

[29] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[30] Xie Z, Huang Z, Dong X, et al. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 23550-23559.

[31] Qin D, Leichner C, Delakis M, et al. Mobilenetv4-universal models for the mobile ecosystem. arXiv 2024[J]. arXiv preprint arXiv:2404.10518.

[32] Nguyen-Ngoc K N, Phan-Nguyen T T, Le K D, et al. DM-VTON: Distilled mobile real-time virtual try-on[C]//2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE, 2023: 695-700.

[33] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.

[34] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 694-711.

[35] Sun D, Roth S, Black M J. A quantitative analysis of current practices in optical flow estimation and the principles behind them[J]. International Journal of Computer Vision, 2014, 106: 115-137.

[36] Yang B, Gu S, Zhang B, et al. Paint by example: Exemplar-based image editing with diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18381-18391.

[37] Zhu Z, Feng X, Chen D, et al. Designing a better asymmetric vqgan for stablediffusion[J]. arXiv preprint arXiv:2306.04632, 2023.

ISSN: 2710-0588

[38] Qiao S, Wang Y, Li J. Real-time human gesture grading based on OpenPose[C]//2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2017: 1-6.

[39] Güler R A, Neverova N, Kokkinos I. Densepose: Dense human pose estimation in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7297-7306.

[40] He S, Song Y Z, Xiang T. Style-based global appearance flow for virtual try-on[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3470-3479.