DM-VTON: Distilled Mobile Real-time Virtual Try-On

Category: Research

ABSTRACT

The fashion e-commerce industry has witnessed significant growth in recent years, prompting exploring image-based virtual try-on techniques to incorporate Augmented Reality (AR) experiences into online shopping platforms. However, existing research has primarily overlooked a crucial aspect - the runtime of the underlying machine-learning model. While existing methods prioritize enhancing output quality, they often disregard the execution time, which restricts their applications on a limited range of devices. To address this gap, we propose Distilled Mobile Real-time Virtual Try-On (DM-VTON), a novel virtual try-on framework designed to achieve simplicity and efficiency. Our approach is based on a knowledge distillation scheme that leverages a strong Teacher network as supervision to guide a Student network without relying on human parsing. Notably, we introduce an efficient Mobile Generative Module within the Student network, significantly reducing the runtime while ensuring high-quality output. Additionally, we propose Virtual Try-on-guided Pose for Data Synthesis to address the limited pose variation observed in training images. Experimental results show that the proposed method can achieve 40 frames per second on a single Nvidia Tesla T4 GPU and only take up 37 MB of memory while producing almost the same output quality as other state-of-the-art methods. It also demonstrates the potential of integrating image-based virtual try-on into real-time AR applications.

Index Terms: Mixed/augmented reality—Real-time systems—Virtual try-on—Knowledge distillation

1 INTRODUCTION

In recent years, the fashion industry, particularly fashion e-commerce, has witnessed significant advancements. Despite these improvements, customers still face the limitation of having to physically visit stores to try on their wanted clothes. As a result, there is a growing interest in virtual try-on [4, 9, 25, 29], which demonstrates the potential to enhance shopping experiences by integrating Augmented Reality (AR).

To the best of our knowledge, existing works on image-based virtual try-on do not put their concern about the complexity of their models. They depend on many information like the human semantic segmentation map (body-parser map) and the human pose, denoted as *human representation*, to enhance the output quality. As a result, they take too much time for inference, preventing them from being applied in real-time applications (ACGPN [27], SDAFN [1] and C-VTON [3] as in Fig. 1). Recent methods [4, 12] have removed the dependency on human representation, thus, reducing the run time considerably. However, they still suffer from using large memory footprints, impacting the requirements for AR devices to run them (PF-AFN [4] and FS-VTON [9] as in Fig. 1).

Addressing those issues, we propose a novel framework, **D**istilled **M**obile real-time **V**irtual **T**ry-**ON (DM-VTON)**, to achieve faster run time and less memory consumption while also producing results of the same quality. The framework consists of two networks: a Teacher network and a Student network. The Teacher network serves as the source of information and guides the Student network through Knowledge Distillation [11]. Specifically, the Teacher is trained with the objective of generating the try-on result from the person image, target garment, and human representation (parser-based approach). Because we only use the Teacher during the training phase, we build it using the state-of-the-art virtual try-on architecture [9]. Taking advantage of the synthetic result and the original garment, the Student can learn to reconstruct the original realistic image without the need for human representation (parser-free approach). As we use the Student network for inference, we consider the trade-off between its runtime and output quality. We introduce two components in the Student network: the Mobile Feature Pyramid Network (MFPN) and Mobile Generative Module (MGM). Both components are based on the lightweight MobileNet [23] architecture, which has been proven to achieve higher throughput and smaller memory usage while producing comparable results to other architectures in the same work.

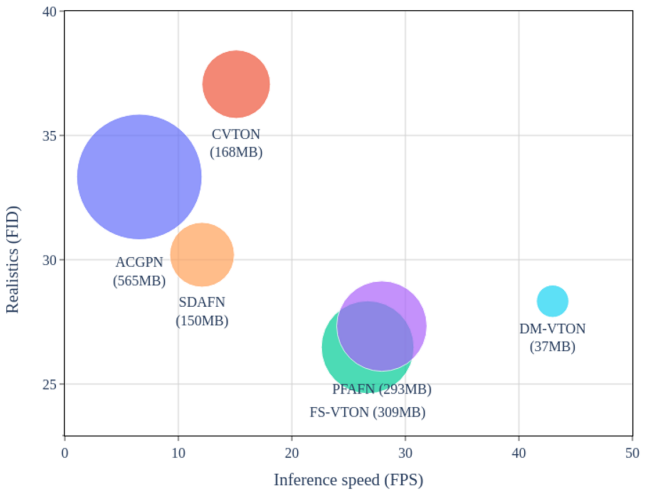


Figure 1: The comparison of our method (DM-VTON) and SOTA methods on VITON test set [8] in terms of realistic results (FID [10], lower is better), inference speed (FPS, higher is better), and memory usage. The size of each bubble represents the memory footprint. FPS is measured using a single Nvidia Tesla T4 GPU.

Besides, common virtual try-on datasets [8, 20] only contain a limited range of pose variations. That makes the models trained on those datasets suffer from overfitting. As a result, they perform poorly in real-life scenes. To overcome this problem, we introduce **V**irtual **T**ry-on-guided **P**ose for **D**ata **S**ynthesis (VTP-DS), an automatic pipeline to enrich the diversity of the poses in the mentioned datasets. The pipeline has two key ideas: self-checking the results by calculating the Object Keypoint Similarity (OKS) [18], and synthesizing new images by using a diffusion network [2]. Given a virtual try-on framework, the pipeline can automatically identify input images where the framework generates results with incorrect poses and extract the corresponding pose information. Then those poses are used as guidance to synthesize new person images from a single image of that person. By utilizing the VTP-DS pipeline, we can enrich the datasets with a wider range of pose variations, thus enhancing the training process and improving the performance of virtual try-on models in real-life scenarios.

We conducted experiments comparing our DM-VTON framework with other state-of-the-art (SOTA) methods in terms of inference speed, memory usage, and the realisticness of the output. During the experimentation, we carefully evaluated the trade-off between those factors. As depicted in Fig. 1, our DM-VTON framework

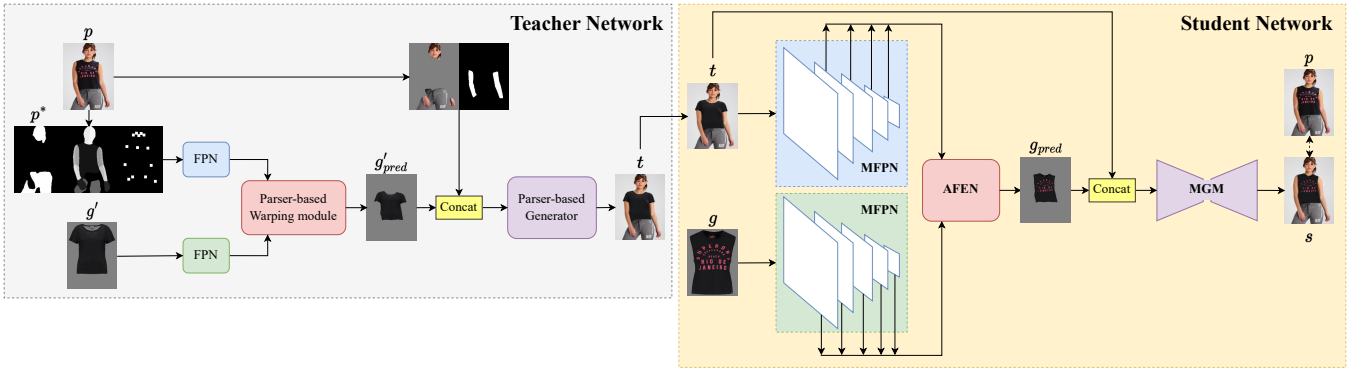


Figure 2: Overview architecture of the proposed Distilled Mobile Real-time Virtual Try-On (DM-VTON) framework. The parser-based Teacher network generates a synthetic image as the input for training the Student network.

outperforms all existing state-of-the-art methods regarding inference speed and memory usage while maintaining an equal quality of results. In summary, our contributions are as follows:

- To address the limitation of image-based virtual try-on that concerns inference time and memory consumption, we propose the DM-VTON framework based on knowledge distillation learning. By developing a lightweight Student network, we can reduce the number of floating point operations (FLOPs) and parameters of the model, thus making it easier to deploy and operate on AR devices.
- We introduce VTP-DS, an automatic fashion-pose data generation pipeline designed to enrich existing fashion datasets by synthesizing new person poses from a single image of that person. The pipeline utilizes a virtual try-on framework to identify challenging poses and subsequently generates additional images for those specific poses. These generated images are then utilized in the training process, improving the framework's performance in real-world scenarios.
- Experiment results show that DM-VTON achieves faster inference and more efficient resource utilization while producing comparable result quality with existing state-of-the-art methods, which proves the efficiency of our proposed framework.

2 Related Work

Image-based virtual try-on techniques can be classified into two categories: parser-based and parser-free approaches. Both of them typically involve three steps: extracting the intrinsic input features, warping the input garment to fit the clothing area of the person image, and performing the replacement using a generative model.

As for parser-based virtual try-on methods, they require human representation, including the body-parser map and human pose, to calculate the warping transformation matrix. The very first methods that pave for this approach are VITON [8] and CP-VTON [27]. To improve the output quality, ACGPN [30] also warps the body-parser map along the target garment. SDAFN [1] reduces the need for the parser map but still needs the human pose, though. Recently, ClothFlow [6] is the first method that uses the appearance flow to guide the warping procedure, and this approach is still used in current SOTA methods [4, 9].

On the other hand, the parser-free approach only requires an input garment and a person image for inference. Thus, this makes the inference process much faster and more independent from other intermediate models. WUTON [12], the pioneering parser-free method, produces noticeable artifacts due to using the same input-output pairs for both Teacher and Student networks [4]. Addressing that issue, PF-AFN [4] introduces a new Knowledge Distillation-based training pipeline, in which the Student network takes the Teacher output as its input and has its output supervised by the original images. This training methodology has become the standard for subsequent parser-free methods. RMGN [16] improves the generation part by using SPADE blocks [21], while FS-VTON [9] improves the warping part by using StyleGan blocks [14].

There are also methods that aim to perform virtual try-on on a sequence of frames. These methods use techniques like memory-based [32] or optical flow [15] to keep the temporary consistency between the frames. However, these methods are still based on the parser-based approach, which takes considerable time to calculate the human representation.

In this paper, we adopt the parser-free approach to prioritize speed, as calculating human representation is a time bottleneck in the try-on process. However, we take a distinct step from existing methods by modifying the Student network for improved speed and reduced memory consumption while keeping the parser-based approach in the Teacher network to preserve the output quality.

3 Method

3.1 Overview

Our objective is to generate an image of a person wearing a specific garment while preserving the rest of the image. To achieve this goal, we adopt the knowledge distillation training pipeline [4, 9, 11, 12, 16] to develop a Distilled Mobile Real-time Virtual Try-On (DM-VTON) framework (see Fig. 2). Our proposed DM-VTON consists of two networks: Teacher and Student networks. Both include three main components: feature extractor, clothes-warping module, and generator.

The Teacher network aims to produce the virtual try-on result using the parser-based training process. The Student network then utilizes the Teacher network to generate synthetic input images, enabling the Student network to be supervised by the original images without relying on human representation. To ensure high-quality output, the Teacher network is built upon SOTA virtual try-on models. As we focus on inference speed, we propose lightweight components for the Student network.

3.2 Teacher Network

The main purpose of this network is to generate a synthetic person image that serves as the input for the Student training process. Furthermore, the Teacher also helps this process through a knowledge distillation scheme. In particular, we take advantage of the SOTA method of virtual try-on task: FS-VTON [9]. As shown in Fig. 2, it incorporates two feature pyramid networks (FPN) [17] constructed from residual blocks, enabling the extraction of features from the human representation p^* and garment image g'. To achieve the

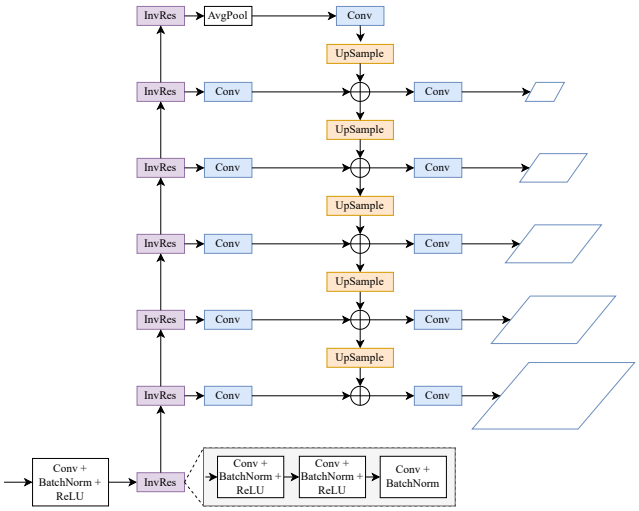


Figure 3: Mobile Feature Pyramid Network architecture

garment deformation functionality, the Teacher network utilizes a style-based global appearance flow estimation network that uses modulated convolution [14]. This network first predicts a coarse appearance flow via extracted global style vector and then refines it locally. The last flow thus can capture the global and local correspondence between the garment and the target person. This makes the Teacher network more robust against the problems of detail-preserving and large misalignment. Finally, the warped clothes and the preserved region on the human body are concatenated as the generator input for try-on result generation. The generator of our Teacher network follows the encoder-decoder architecture with skip connections, which have been proven effective in detail preservation.

Because the inputs of the parser-based model (i.e., human representation) contain more semantic information when compared to those in the parser-free model, we employ an adjustable knowledge distillation learning scheme [4] with a distillation loss to guide the feature extractor of the Student network:

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

$$\psi = \begin{cases} 1, \ if \ \|t-p\|_1 < \|s-p\|_1 \\ 0, \ otherwise \end{cases}, \tag{2}$$

where t_{p_i} and s_{p_i} denote the feature maps at the i-th scale extracted from the human representation p^* within the Teacher network and the synthetic image t within the Student network; t, s are the try-on result of the Teacher and Student, respectively; p is the person image ground truth.

3.3 Student Network

We propose a parser-based approach for synthesizing try-on images with increased speed compared to previous methods while ensuring accuracy. As shown in Fig. 2, our Student network consists of three key components: Mobile Feature Pyramid Network (MFPN), Appearance Flow Estimation Network (AFEN), and Mobile Generative Module (MGM). These components synergistically collaborate to extract features, manipulate garments through deformation, and generate try-on images. The AFEN introduced by Ge et al. [4] proved effective in deforming garments by using appearance flow estimates from pyramid features. The MFPN and MGM are built upon the architecture of MobileNetV2 [23] with Inverted Residual blocks specifically designed to optimize computational efficiency and model size.

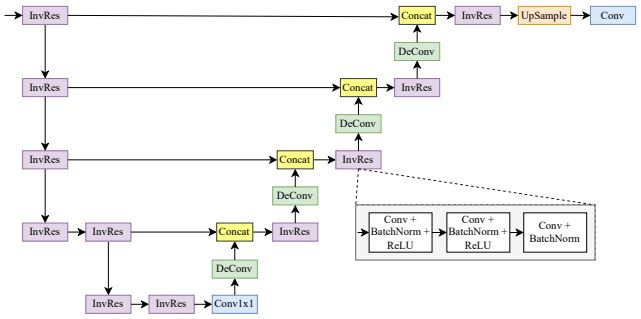


Figure 4: Mobile Generative Module architecture

3.3.1 Mobile Feature Pyramid Network

As shown in Fig. 3, MFPN incorporates the architecture of MobileNetV2 with Inverted Residual blocks [23] to a Feature Pyramid Network. By leveraging the capabilities of two MFPN blocks, we extract two-branch N-level feature maps from person and garment images within a parser-free network. These features are fed into the Appearance Flow Estimation Network (AFEN) to predict the appearance flow map for garment deformation.

3.3.2 Appearance Flow Estimation Network

This component aims to deform the garment to fit the human pose while preserving the texture. Following the work of Ge et al. [4], we adopt an appearance flow estimation network (AFEN) comprising subnetworks equipped with varying sizes of convolution layers. These subnetworks are responsible for estimating flows based on extracted multi-level feature maps. The outcome of this network can capture the long-range correspondence between the garment image and the person image, effectively minimizing issues related to misalignment. To enhance the preservation of clothing characteristics, this module is optimized with the second-order constraint:

$$L_{sec} = \sum_{i=1}^{N} \sum_{t} \sum_{\pi \in N_t} CharLoss\left(f_i^{t-\pi} + f_i^{t+\pi} - 2f_i^t\right), \tag{3}$$

where f_i^t denotes the t-th point on the i-th scale flow map; N_t is the set of horizontal, vertical, and diagonal neighborhoods around the t-th point; and *CharLoss* denotes generalized Charbonnier loss [26].

3.3.3 Mobile Generative Module

To synthesize the entire try-on image from the warped image and target person image, we develop a Mobile Generative Module, the integration of the architectural principles of UNet [22] and MobileNetV2 [23] as illustrated in Fig. 4. The primary objective behind the design of this generator is to reduce both the computational burden and the model's overall size.

3.3.4 Loss Function

During training, we optimize the warping module separately in the first stage and then train together with the generator in the last stage. The loss function used in the first stage is defined as:

$$L^{warp} = \lambda_l^{warp} L_l^{warp} + \lambda_{per}^{warp} L_{per}^{warp} + \lambda_{sec} L_{sec} + \lambda_{dis} L_{dis}, \tag{4}$$

where $L_l^{warp} = \|g_{pred} - p \odot m_{gt}\|$ denotes pixel-wise L1 loss, $L_{per}^{warp} = \sum_i \|\Phi_i(g_{pred}) - \Phi_i(p \odot m_{gt})\|$ is the perceptual loss [13], L_{sec}^{warp} is the smooth loss (second-order constrain), L_{dis}^{warp} is the distillation loss, g_{pred} is warped garment. p is the person image ground truth with the garment mask m_{gt}; Φ_i denotes the i-th block of pre-trained VGG19 [24].

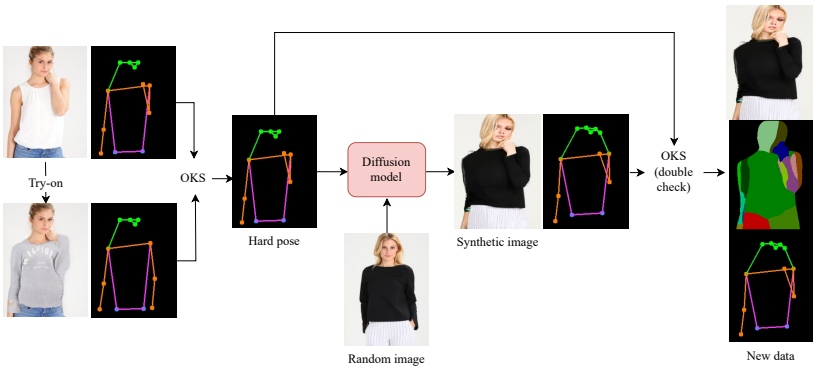


Figure 5: Overview of Virtual Try-on-guided Pose for Data Synthesis pipeline.

With the generative module, we also apply L1 loss perceptual loss [13] between the synthesized image and the ground truth image to supervise the training process of MGM:

$$L^{gen} = \lambda_l^{gen} L_l^{gen} + \lambda_{per}^{gen} L_{per}^{gen}, \tag{5}$$

where $L_l^{gen} = \|s - p\|$ is L1 loss and $L_{per}^{gen} = \sum_i \|\Phi_i(s) - \Phi_i(p)\|$ is the perceptual loss [13]. s and p are the generated output of the Student network and the person image ground truth, respectively.

In practice, we empirically set $\lambda_l^{warp} = 1$, $\lambda_{per}^{warp} = 0.2$, $L_{sec}^{warp} = 6$, $L_{dis}^{warp} = 0.04$, $\lambda_l^{gen} = 5$, $\lambda_{per}^{gen} = 1$. The overall loss function when training the whole model in the last stage is:

$$L = 0.25 * L^{warp} + L^{gen}. \tag{6}$$

3.4 Virtual Try-on-guided Pose for Data Synthesis

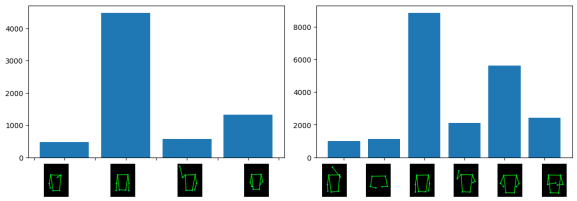
By using the K-Means clustering algorithm, we observe that the original VITON dataset [8] is mainly composed of images with straight-arm poses (as in Fig. 6(a)). This bias creates a challenge as models trained on such data are prone to overfit and perform poorly on images with different upper-body poses. To tackle this problem, we propose the Virtual Try-on-guided Pose for Data Synthesis (VTP-DS) pipeline. With the objective of improving the existing virtual try-on framework, the pipeline incorporates two key ideas: automatically detecting poorly performed poses using the Object Keypoint Similarity (OKS) metric [18] and synthesizing new training data specifically targeting those poses. The overview of the VTP-DS pipeline is illustrated in Fig. 5.

Given an input image containing a person, we extract that person's pose by using the YOLOv7 pose estimation method [28]. Then, we utilize our trained DM-VTON model to perform virtual try-on on the input image. The extracted pose from the resulting image is compared with the pose of the input image using a customized OKS metric Equation 7.

$$\frac{1}{|P|} \sum_{i \in P} \exp\left(\frac{-d_i^2}{2s^2 k_i^2}\right), \tag{7}$$

where P denotes the set of arm and hand keypoints, while the original formula uses all keypoints; d_i denotes the Euclidean distance between the keypoint i of two poses; s denotes the total area containing the pose; k_i is the constant provided by Lin et al. [18] to represent the standard deviation for keypoint i. As we focus on distinguishing different upper-body poses, only the arm and hand keypoints contribute to the formula. If the OKS score falls below a specified threshold $t = 0.9$, it is identified as a hard pose.

Once identifying a hard pose, we randomly pick a person image from the VITON dataset. Then it leverages Bhunia's Diffusion model [2] to synthesize a new image of the person in the corresponding pose. To ensure the accuracy of the synthesized image, we perform a double-check using the OKS metric to verify the correctness of the output pose. Finally, DensePose [5] is utilized to generate the body-parser map of the synthesized image.



(a) VITON (b) VITON + synthesized data

Figure 6: Pose distribution in VITON dataset [8].

To synthesize additional data for training networks, we initially collected videos from Youtube, specifically focusing on posing or catwalk videos. These videos had varying durations, ranging from 1 to 10 minutes. Subsequently, we extracted individual frames from these videos, which served as the input for our VTP-DS pipeline. After that, we manually removed low-quality results, resulting in 14,314 high-quality synthesized images for training the networks.

To access the quality of synthesized images, we use the K-means algorithm combined with our modified OKS metric. As details of the pose clustering results shown in Fig. 6, data in the VITON training set mainly focuses on poses with simple poses (i.e. arms are less covered, low rotation amplitude). Meanwhile, when combined with our synthesized images, new pose clusters appear and the concentration of data in groups is less imbalanced, which can help to train robust virtual try-on models.

4 Experiments

4.1 Detailed Implementation

The Teacher and Student network training process follows the same strategy with two stages: the first stage only trains the warping module, while the latter trains the entire network. Both were under the same setting and carried out on a single Nvidia A100 GPU. We trained the model for 100 epochs with the initial learning rate is 5×10^{-5}, which decays linearly after the first 50 epochs.

4.2 Experimental Settings

VITON [8], the most popular dataset for evaluating virtual try-on, was used to evaluate methods. It contains 16,253 frontal-view upper-body woman and top clothing image pairs with 256×192 resolution. However, we followed the work of Han et al. [6] to filter out duplicates and ensure no data leakage happens, remaining 6,824 training image pairs and 416 testing image pairs in the cleaned VITON

Table 1: Quantitative results between DM-VTON and SOTA virtual try-on methods. The † marker indicates the results measured by the generated images provided by the authors. The speed was evaluated on a single Nvidia T4 GPU.

Method	**Published**	Parser	Pose	**FID** ↓	**LPIPS** ↓	**Runtime (ms)** ↓	**FLOPs (B)** ↓	**Memory usage (MB)** ↓
ACGPN [30]	CVPR 2020	✓	✓	33.33	0.231	153.64	399.08	565.86
PF-AFN [4]	CVPR 2021			27.33	0.216	35.80	137.85	293.25
C-VTON† [3]	CVPRW 2022	✓		37.06	0.241	66.90	108.47	168.60
SDAFN [1]	ECCV 2022		✓	30.20	0.245	83.42	149.40	150.87
FS-VTON [9]	CVPR 2022			26.48	0.200	37.49	132.98	309.25
DM-VTON	Ours			28.33	0.215	23.27	69.82	37.79

Figure 7: Qualitative comparison on VITON-Clean dataset [8].

dataset, denoted by VTION-Clean. We combine the VTION-Clean training set and our synthesized images to train our proposed method.

Fréchet Inception Distance (FID) [10] and Learned Perceptual Image Patch Similarities (LPIPS) [31] metrics were used to evaluate the similarity of try-on results to real images.

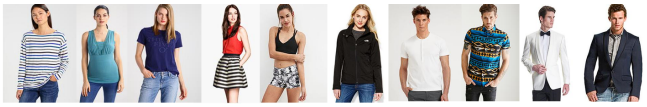
4.3 Experimental Results

We compared the performance of our proposed MD-VTON with SOTA methods in virtual try-on, such as ACGPN [30], PF-AFN [4], C-VTON [3], SDAFN [1], FS-VTON [9]. Comparison of MD-VTON against those methods in terms of image quality (i.e., FID and LPIPS), inference speed (i.e., ms), FLOPs (Floating point operations), and memory usage (MB) is shown in Table 1. Our proposed method outperforms all other SOTAs in terms of runtime, FLOPs, and memory consumption. On the other hand, our DM-VTON achieves slightly higher FID and LPIPS scores than those of PF-AFN [4] and FS-VTON [9]. The experimental results prove that the proposed DM-VTON can run in real-time (i.e., 43 frames per second) with small memory consumption but still retains high-quality virtual try-on results. The visualization of compared methods is illustrated in Fig. 7.

5 Pilot Study

We invited 12 participants who are university students and researchers in the 18-44 age range. We let the users experience our DM-VTON and collected their feedback. The experimental scenario was that while shopping online at home, the users come across a particular garment that catches their eyes, but they are unsure whether it looks good on them. We offered them an application to try on such garments before making the purchasing decision. Specifically, we prepared a collection of 20 garments taken from the VITON [8] and PolyvoreOutfits [7] datasets (see Fig. 8). These garments were carefully selected to represent a variety of colors, shapes, and textures.

Each participant took part in a 10-minute session in which the participant was asked to perform a virtual try-on using our provided model images and virtual try-on on themselves directly captured



(a) Human model samples

(b) Garment samples

Figure 8: Examples of data used in our pilot study.

from our camera. In terms of human models, we used 10 person images in the VITON [8] and DeepFashion [19] datasets (see Fig. 8).

Upon completing the trial, we interviewed participants and asked for their feedback. Our primary objective was to evaluate how our system influenced their purchasing decisions. We also gathered their feedback about the output quality and whether they prefer using the model images or their own images to enhance the overall user experience in the future.

Most of the participants agreed that trying on clothes in various poses helped them visualize the suitability of the garments before making a purchase decision. Specifically, 66.7% of the participants felt confident enough to make the purchasing decision after using our system, while the remaining 33.3% had doubts about the truthiness of the models. Moreover, 83.3% of the participants preferred using their images for try-on, as it provided a more realistic experience for them. On the other hand, the remaining 16.7% considered both options, as the provided models allowed them to see the best representation of the garments, such as with appropriate brightness and poses. Fig. 9 illustrates some virtual try-on results on our provided models. Due to privacy issues, we did not capture the virtual try-on results on the participants' images.

Figure 9: Results obtained when users performed virtual try-on on our provided human models in the pilot study.

We also received valuable feedback from participants on areas for improvement. In real-life conditions, the background, brightness, and quality of the captured images might not be suitable for trying on clothes, which is due to the fact that our train and test datasets only contain simple backgrounds and have proper brightness conditions. Thus, applying some pre-processing techniques such as segmentation and brightness equalization is necessary to address this issue. Additionally, participants also suggested that our system exhibited inconsistencies when dealing with complex poses such as half-turn poses or crossed-arm poses. Their feedback is useful in enhancing the user experience in the future.

6 Conclusion

This paper presents the simple yet efficient Distilled Mobile Real-time Virtual Try-On (DM-VTON) framework. By leveraging the knowledge distillation scheme, we developed a lightweight parser-free network to boost the processing speed. Our proposed network utilizes mobile-based architectures, resulting in achieving real-time virtual try-on capabilities while maintaining high-quality output and computational efficiency. In addition, the Teacher network, trained using a parser-based approach, provides supervision to the Student network, enabling it to learn without relying on the human representation of ground truth.

Additionally, to address the limited pose variation observed in the training images, we introduced the Virtual Try-on-guided Pose for Data Synthesis (VTP-DS) to enrich the diversity of poses in the training data. VTP-DS automatically identifies input images with incorrect poses generated by the framework and generates additional images for those specific poses.

Experimental results and the pilot study showcase the potential of our proposed framework. It can be applied to real-time augmented reality (AR) applications, paving the way for improved user experiences in a virtual fashion context.

References

[1] S. Bai, H. Zhou, Z. Li, C. Zhou, and H. Yang. Single stage virtual try-on via deformable attention flows. In *ECCV*, pp. 409–425, 2022.

[2] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan. Person image synthesis via denoising diffusion model. In *CVPR*, pp. 5968–5976, 2023.

[3] B. Fele, A. Lampe, P. Peer, and V. Struc. C-vton: Context-driven image-based virtual try-on network. In *WACV*, pp. 3144–3153, 2022.

[4] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pp. 8485–8493, 2021.

[5] R. Guler, N. Neverova, and I. DensePose. Dense human pose estimation in the wild. In *CVPR*, pp. 18–23, 2018.

[6] X. Han, X. Hu, W. Huang, and M. R. Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, pp. 10471–10480, 2019.

[7] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *ACM Multimedia*, pp. 1078–1086, 2017.

[8] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *CVPR*, pp. 7543–7552, 2018.

[9] S. He, Y.-Z. Song, and T. Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, pp. 3470–3479, 2022.

[10] 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. *NeurIPS*, 30, 2017.

[11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] T. Issenhuth, J. Mary, and C. Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, pp. 619–635, 2020.

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pp. 694–711, 2016.

[14] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4401–4410, 2019.

[15] G. Kuppa, A. Jong, X. Liu, Z. Liu, and T.-S. Moh. Shineon: Illuminating design choices for practical video-based virtual clothing try-on. In *WACV*, pp. 191–200, 2021.

[16] C. Lin, Z. Li, S. Zhou, S. Hu, J. Zhang, L. Luo, J. Zhang, L. li Huang, and Y. He. Rmgn: A regional mask guided network for parser-free virtual try-on. In *IJCAI*, pp. 1151–1158, 2022.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pp. 2117–2125, 2017.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

[19] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pp. 1096–1104, 2016.

[20] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, pp. 2231–2235, 2022.

[21] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pp. 2337–2346, 2019.

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pp. 4510–4520, 2018.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] W. Song, Y. Gong, and Y. Wang. Vtonshoes: Virtual try-on of shoes in augmented reality on a mobile device. In *ISMAR*, pp. 234–242, 2022.

[26] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, 106:115–137, 2014.

[27] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pp. 589–604, 2018.

[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, pp. 7464–7475, 2023.

[29] Y. Xu, S. Yang, W. Sun, L. Tan, K. Li, and H. Zhou. 3d virtual garment modeling from rgb images. In *ISMAR*, pp. 37–45, 2019.

[30] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, pp. 7850–7859, 2020.

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

[32] X. Zhong, Z. Wu, T. Tan, G. Lin, and Q. Wu. Mv-ton: Memory-based video virtual try-on network. In *ACM Multimedia*, pp. 908–916, 2021.