Received: 10 January 2024 | Revised: 31 July 2024 | Accepted: 22 October 2024 | Published online: 13 March 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 3(3) 275–283
DOI: 10.47852/bonviewAIA52022458

BON VIEW PUBLISHING

FashionFlow: A Lightweight Pix2Pix-Based Approach to Virtual Try-On

Samar Pratap[1], **Alston Richard Aranha**[1], **Divyanshu Kumar**[1,*] and **P. Preethi**[1]

[1]*Department of Computer Science, PES University, India*

Abstract: The proposed virtual try-on model works on synthesizing naturalistic images by superimposition of target clothing item on a person. The proposed model offers a significant advancement in virtual try-on technology, by introducing a lightweight architecture that reduces the system requirements required to run the application. Many different preprocessing methods have been utilized to account for different poses and body shapes. Great efficiency has been achieved by employing the Pix2Pix generative adversarial networks (GAN) architecture with a minimal number of training epochs. The proposed model's performance shows great results without a discriminator, a change from traditional GAN setups, thus emphasizing the need for simple and effective virtual try-on solutions. While the performance is characterized as good, this novel approach encourages further exploration in lightweight model design for practical and efficient try-on applications. Additionally, this paper showcases results from other GAN architectures providing a comprehensive overview of our research contributions in virtual try-on technology.

Keywords: generative adversarial networks, Pix2Pix GAN, image superimposition, fashion

1. Introduction

A massive increase in online fashion shopping has been noticed recently. The sales cycle of 2022 indicated that there were more than 1.6 billion user visits in the online fashion market from 150 of the world's biggest and most prestigious fashion e-commerce brands. By the end of 2023, the global fashion e-commerce market is forecast to reach a value of over 820 billion U.S. dollars. By 2027, it could reach just over 1.2 trillion U.S. dollars.[1] A study by Forrester Research shows the growth of the fashion industry's e-commerce share of total retail sales from 14% in 2014 to 38% in 2022. Since noticeable growth is witnessed in the online fashion market, it is crucial that appropriate steps should be taken to enhance user's interaction with fashion e-commerce markets. A major limitation of online fashion e-commerce includes enabling users to visualize themselves in different clothing items. Thus, virtual try-on holds an important role in covering this gap, resulting in a better user experience. Currently, just a handful of the online marketplaces have been focusing on usage of 3D modeling of customers using in-depth cameras for image superposition. This technique has the ability to generate realistic results but requires high usage of resources.

Image-based try-on methods have been noticed to give highly superior performance under different conditions by preserving the original body shape and other facial features. A majority of the existing approaches face common issues such as inappropriate wrapping of clothes, changes in sleeve length and not being able to catch body shapes like in cases involving pregnant women.

Thus, we propose FashionFlow, a lightweight solution for efficient imposition of target cloth upon the given input image while conserving all the features of the person. To accomplish better accuracy of cloth placement over the person, different body representations of the human body are considered as features for the generative adversarial networks (GAN) architecture. For body representation, the features included are skeletons of the human, agnostic representation, and face-hair segmentation. These features improve the ability of the system to adapt to different body shapes. A GAN network is then used to fit the cloth and generate the super-imposed output image. The proposed architecture of GAN includes Pix2Pix GAN without a discriminator in the model, and fine-tuning of loss functions is performed. It was reported that this kind of architecture gives equally good results and even superior results in cases where body shape is different from a general body shape. FashionFlow utilizes popular annotated datasets such as VITON-HD [1] for the development and evaluation of models across various scenarios and use cases. The results demonstrate that FashionFlow generates realistic virtual try-on results. Thus, the main contributions of the paper are summarized below:

1) A novel GAN-based approach is proposed as an efficient and accurate virtual try-on method incorporating an architecture without discriminator. The proposed method in this paper is noted to provide more accurate outcomes in contrast to the actual Pix2Pix GAN [2] proposed.
2) Diverse range of preprocessing techniques are employed that enable the analysis of input image characteristics, including body shape, cloth wrapping, and preservation of facial features.
3) The proposed approach's performance is evaluated on the basis of comparison with existing methods on sample images. The

[1] https://web.archive.org/web/20240223163827/https://www.salecycle.com/blog/featured/online-fashion-retail-11-essential-statistics/

*Corresponding author: Divyanshu Kumar, Department of Computer Science, PES University, India. Email: pes1202101503@pesu.pes.edu

remainder of the paper is organized into sections. Section 2 summarizes the Related Work. Section 3 elaborates on the proposed method and its implementation. Finally, Section 4 concludes the paper and outlines future work.

2. Literature Review

1) GANs in fashion

Fashion and trends play a significant role in shaping society by influencing individuals' self-expression, values, and lifestyle choices, reflecting the societal effect of the ever-changing world of fashion [3]. These sectors can be influenced by technology drastically, with countless innovations going on and immense potential for future application, as aptly pointed out by Akram et al. [4]. In this domain, Virtual Try-On-based systems have been a subject of great discussion and research because of their applicability in the online shopping ecosystem. Such systems are greatly aided by GANs. GANs have been utilized extensively for image-to-image translation for years now. From the most basic image-to-image translation tasks to more sophisticated ones like virtual try-on that encompass a variety of facets, including but not limited to body shapes, sizes, and preferences, allowing for a tailored and personalized try-on experience. Approaches [5, 6], such as VITON [7], have demonstrated the effectiveness of GANs in the realistic generation of virtual try-on results.

There have been multiple approaches where a high-resolution output has been obtained without focusing on the shape variations. A latest approach by Li et al. [8] employed state-of-the-art warping methods and a learned rendering procedure, the proposed POVNet approach performs accurate texture preservation and fine detail representation. Though the paper did not address body shape variations, its scalability and real-time responsiveness enhance user engagement for fashion industry applications.

Size-aware and pose-guided implementations [9, 10] particularly focus on problems like translation between images with sleeves and without them and improper pose detection. There have been significant developments in specific use-case-based implementations for this task. An approach by Yan et al. [11] semantically associates landmarks to account for issues like sleeve differences and length of clothes. Another paper by Zhu et al. [12] proposed a pose transfer method using a sequence of transfer blocks defined by the authors as Pose-Attentional Transfer Blocks, which utilizes a GAN model to progressively generate pose-transferred images for non-rigid objects. We approach these problems using Pix2Pix GAN and propose a lightweight approach to fix such sleeve and pose-based issues, also accounting for cases where body shapes are inconsistent such as pregnant women.

2) Pix2Pix GAN in use

It is a conditional GAN that uses real data, noise, and labels to generate images. Extensive research has explored its effectiveness in image-to-image translation tasks, demonstrating its utility in this domain [13]. A noteworthy paper by Dong et al. [14] introduced a soft-gated-warping technique that compared the outcomes of the conventional Pix2Pix model, originally presented at CVPR 2017 with respect to virtual clothing try-on systems. The merit of Pix2Pix-based systems has been further highlighted in papers such as NVIDIA corporation's paper on high-resolution conditional GANs for picture synthesis [15]. For virtual try-on-based systems, the results of the traditional Pix2Pix model proposed by Isola et al. [2] revealed significant distortions in the resulting generated images. In our paper, we aim to build upon this system, trying to produce superior, more accurate, and practical images. Our approach has demonstrated substantial improvements over the earlier Pix2Pix-based implementation, the outcomes of which were demonstrated in Dong's paper and showed inconsistencies in recognizing arms, faces, and sleeves. We have fixed these problems using efficient preprocessing techniques and changing the model to suit our needs. The utilization of Pix2Pix GAN has led to the creation of a more lightweight system while also reducing computational expenses. This approach exhibits an ability to distinguish sleeves, poses, and body shapes even with fewer training epochs.

Consequently, our paper's approach not only produces comparatively superior results with respect to the previous Pix2Pix-based implementation but also aids in the creation of a more computationally efficient system. Our discriminator-less approach greatly helps this objective by cutting down on the training time and resource-intensive nature of such applications. This lightweight nature, along with the system's ability to distinguish details highlights its practicality for real-world applications. The improvements and findings of this study can also be generalized to other image generation tasks.

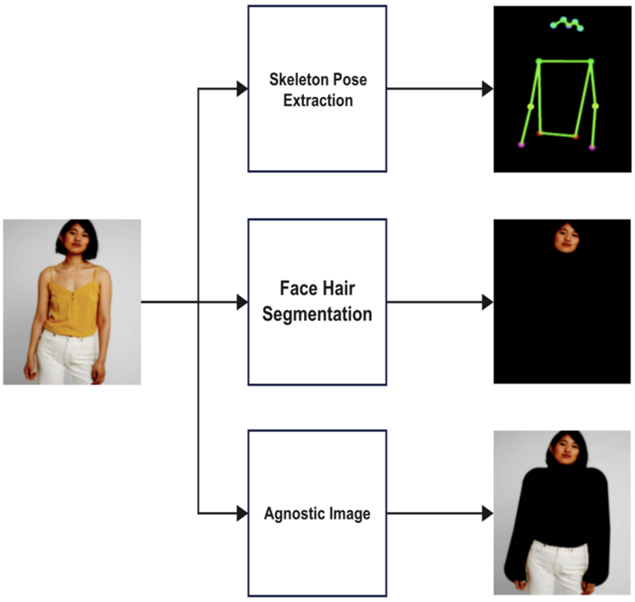
2.1. Theoretical framework

In a practical case, only a reference image along with a target cloth will be passed as inputs to the FashionFlow model. For a given input image, our proposed model aims to achieve smooth superimposition for virtual try-on by following certain preprocessing steps such as pose extraction. Stage 1, namely Pose extraction aids the model in understanding the body's orientation. Stage 2 includes face and hair segmentation helping in the refinement of input details for a more accurate overlay. Stage 3 performs steps to generate an agnostic representation providing a further detailed overview of the original image having destroyed the details of previous cloth. The detailed description of each stage is explained in Sections 3.1–3.3. By the inclusion of these essential stages, it helps the U-Net-based GAN architecture to generate an accurate virtual try-on output. The U-Net architecture then superimposes target cloth over the human body. Its encoder-decoder structure with skip connections facilitates the conservation of details and spatial information. Figure 1 shows the different steps of preprocessing performed on the input image before generation of output image.

3. Research Methodology

Our paper follows a structured research methodology that includes principles from computer vision, machine learning, and image processing. The preprocessing is divided into steps like skeleton pose extraction, face-hair segmentation, and agnostic image generation. These steps help in preparing the images and fixing recurring issues such as distortion of the face and hair, incorrectly changing parts of the image other than the region of interest and inability of the systems to adapt to different body shapes and poses. As part of our study, we explored two approaches for training the Pix2Pix GAN, one with the discriminator and the other without it. The results of both the approaches were later compared to state-of-the-art systems using metrics like FID score and Learned Perceptual Image Patch Similarity (LPIPS) score.

Figure 1
Preprocessing steps performed on the original image



3.1. Research design

As described above, we have followed a multi-stage approach for virtual try-on, comprising of steps the steps outlined in Section 2.1 like skeleton pose extraction, face-hair segmentation, and agnostic image generation. The rationale behind these steps and details about their implementation are elaborated in the following subsections.

3.1.1. Skeleton pose extraction

Skeleton pose extraction involves identifying key points or joints in the human body and creating a digital model of the body's pose based on those points. These points are computed with the help of certain labels. 20 essential key landmarks have been identified which include features such as eyes, ears, shoulders, wrists, elbows, and hips. They are then further linked to form a skeleton-like model of the human body which helps in efficient pose extraction from the input image. Additionally, this process makes the model resilient to variations in poses and body sizes.

3.1.2. Face and hair segmentation

To perform face and hair segmentation, U-Net architecture (made using RESNET-5) is put into use. The state-of-the-art RTNet model [16] used to execute the task performs face parsing by introducing "RoI Tanh-polar Transform" [17] for in-the-wild face analysis, preserving context and enabling rotation-equivariant learning. "Hybrid Residual Representation Learning Blocks" [18] helps in extracting hybrid representations which boost performance. iBugMask dataset has been utilized for training this model, containing over 22,000 diverse images. With the iBugMask dataset and extensive experiments, it achieves state-of-the-art performance helping in the preparation of preprocessed images.

3.1.3. Agnostic image representation

For generating an agnostic image for a target input, this step uses a U-Net-based generator equipped with skip connections for the generation of an agnostic representation from an input image. This agnostic representation offers a comprehensive overview of the actual image while intentionally removing any information regarding the previous clothing. A series of transformations and neural network operations are included to modify the image, enhancing its features and details while performing the elimination of specific clothing-related information existing in the initial image.

3.1.4. U-Net: GAN framework

Figure 2 demonstrates the GAN U-Net Architecture implemented in our study to ensure image superimposition. Given all the preprocessed images corresponding to a particular input, a modified Pix2Pix-based GAN framework from the originally proposed Pix2Pix GAN is proposed which synthesizes the output

Figure 2
GAN architecture with U-Net and data flow: illustrates the preprocessed image inputs, the U-Net structure, and the resultant output generation

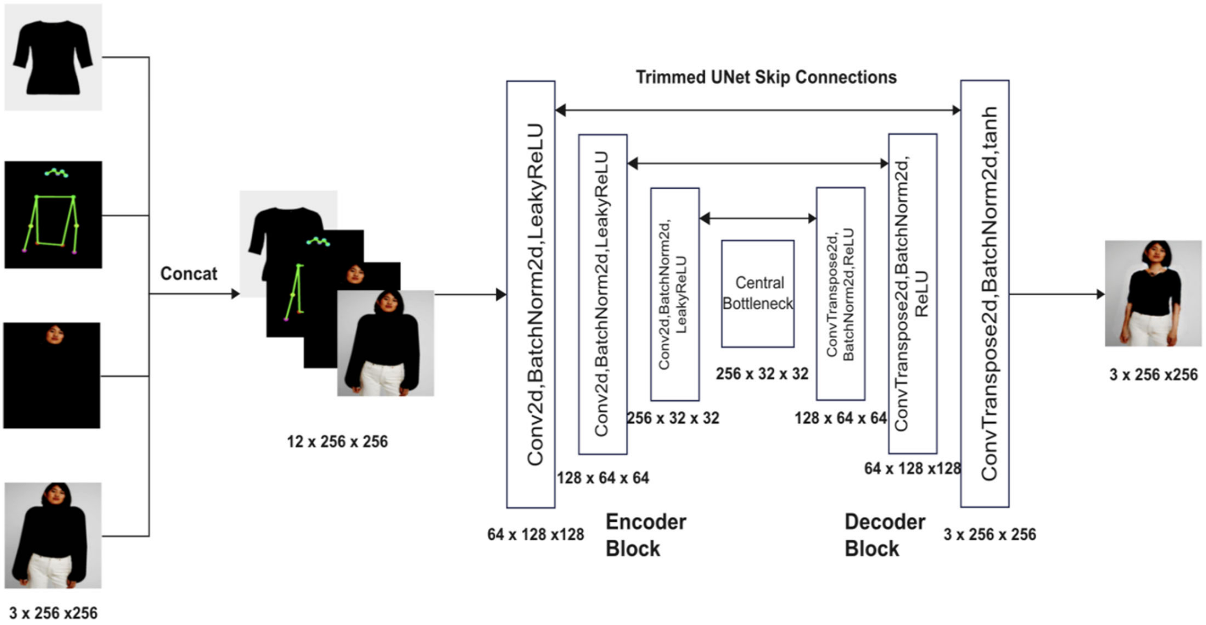

Concat
12 x 256 x 256
3 x 256 x256
Trimmed UNet Skip Connections
Conv2d,BatchNorm2d,LeakyReLU
Conv2d,BatchNorm2d,LeakyReLU
Conv2d,BatchNorm2d, LeakyReLU
Central Bottleneck
ConvTranspose2d, BatchNorm2d,ReLU
ConvTranspose2d,BatchNorm2d, ReLU
ConvTranspose2d,BatchNorm2d,tanh
256 x 32 x 32
256 x 32 x 32
128 x 64 x 64
128 x 64 x 64
64 x 128 x128
64 x 128 x128
3 x 256 x 256
Encoder Block
Decoder Block
3 x 256 x256

image on the basis of target clothing item. The U-Net architecture, defined as UnetGenerator, comprises a series of interconnected blocks, starting with an innermost block operating at 512 filters and progressively expanding to the further intermediate blocks, eventually concluding in an outermost block with reduced filter counts. The structure involves skip connections, enabling feature concatenation between different layers of the model. The outermost block responsible for the final output operates through an order of these interconnected blocks to generate the desired output. The Pix2Pix GAN framework employs a paired group of input and output images to train the generator model by using differences between images generated from the inputs and the desired outputs. In a typical GAN framework, the generator is supposed to evade the discriminator perception while reaching near the actual output in an L2 sense. But Pix2Pix focuses using L1 distance rather than L2 as L1 encourages less blurring:

$$\mathcal{L}_{L1}(\mathrm{G}) = \mathbb{E}x,y,z[||y - G(x,z)||_1] \tag{1}$$

The input configuration to the framework consists of a concatenated stack of images produced from the preprocessing stages alongside the target clothing image. During training, the output image is the original representation passed to the generator. This input-output setup is important as the previous preprocessing stages ensure that the original cloth information is indiscernible. It ensures that the model learns to synthesize clothing on a person rather than merely memorizing the specific image of the clothed person.

The original Pix2Pix GAN proposed by Isola et al. [2] includes a discriminator model working with the generator to improve the realism of the generated images. The discriminator, often a convolutional neural network, assesses the authenticity of the generated images by distinguishing between the synthetic image from the UNET generator and the actual image. This adversarial relationship enables the generator to refine its output based on the feedback received from the discriminator, thus increasing the similarity between the generated images and the actual images.

Originally, within a GAN model, a discriminator's role is to help in generating images resembling the actual image when a target image was absent from the model. This highlights the discriminator's significance within the model, ensuring the generation of desired images.

The Pix2Pix GAN, similar to the previously described GAN model, integrates both generator and the discriminator for image generation. Unlike a conventional DCGAN model, the Pix2Pix GAN already factors in pixel-wise loss in the generated image, with a target output image provided within the GAN framework. Hence, the necessity of a separate discriminator is brought into question.

The potential presence of noise in a given image dataset raises concerns about the need for a separate discriminator. Its usage can potentially direct the learning process towards false, non-existent patterns in the original input image, leading the generator away from generating the intended image.

Therefore, the FashionFlow, introduced in this paper, diverges from conventional approaches by removing the discriminator from the system. It selectively works with the integration of pixel-wise L2 loss during the training process of the generator to facilitate image generation. It has been observed that this framework offers several advantages, ensuring stability in the synthesis of images given the target cloth. By following this architecture, FashionFlow fundamentally simplifies its training process, focusing only on optimizing the generator's output in every iteration by usage of pixel-wise L2 loss. By emphasizing the pixel-wise loss function, FashionFlow maintains a more stable and light-weighted approach towards generating high-quality images.

3.2. Proposed algorithm

FashionFlow Algorithm

```
Input: Image of the person, I_person
       Image of the new cloth, I_cloth

Output: Image of the person wearing the new cloth, I_output
procedure FashionFlow(I_person, I_cloth)
  skeleton ← DetectSkeleton(I_person)
  faceHair ← SegmentFaceAndHair(I_person)
  agnostic ← GenerateAgnostic(I_person)
  I_output ← FeedToGAN(skeleton, faceHair, agnostic, I_cloth)
  return I_output
end procedure

procedure DetectSkeleton(I_person)
  Load pre-trained keypoint detection model
  output ←ModelForwardPass(I_person)
  skeleton ← DrawSkeleton(I_person, output)
  return skeleton
end procedure

procedure SegmentFaceAndHair(I_person)
  Load pre-trained face and hair segmentation model
  faceHair ← ModelForwardPass(I_person)
  return faceHair
end procedure

procedure GenerateAgnostic(I_person)
  Load pre-trained agnostic model
  agnostic ← ModelForwardPass(I_person)
  return agnostic
end procedure

procedure FeedToGAN(skeleton, faceHair, agnostic, I_cloth)
  combinedImage ← CombineImages(skeleton, faceHair,
  agnostic)
  Initialize generator model
  generated_image ← GeneratorForwardPass(combinedImage)
  return generated_image
end procedure
```

This algorithm shows the steps involved in the virtual try-on process using a Pix2Pix GAN-based approach with a U-Net generator where I_{output} signifies the final output from the algorithm. For getting the skeleton pose of the person in the image, we have used a pre-trained key point detection model from the torchvision library. It defines a set of key points and extracts limbs from them, forming the skeletal structure. It then connects the key points using a threshold score. The resulting image provides a visual representation of the human poses. For the face and hair segmentation, a collection of pre-trained face detectors has been used, for example S3FD and RetinaFace with weights trained on the WIDER dataset. We trained a U-Net-based model to generate an agnostic image representation from the input image. These images are then stacked with the target cloth and passed as input to the GAN. To train the GAN we have used both the conventional approach and the approach where the discriminator is not taken into consideration. This study mainly focuses on the improvement of the final results just by introducing a few simple preprocessing steps and the comparable performance of the Pix2Pix GAN even when trained without the discriminator. The model is tested on the previously preprocessed input and evaluated on metrics like FID and LPIPS score. The limitations of the system include the inability of the system to generate complex patterns and to generate high-resolution images.

3.3. Experiments

To prepare and train our proposed model, we utilize the popular and publicly available VITON-HD dataset which has 13,679 pairs of garment and person images. Our model was trained on this dataset for 30 epochs, using the preprocessing steps executed beforehand. The preprocessed images are noted to have contributed to preserving essential facial features and body shapes, ensuring better results. To confirm the model's efficacy, the paired virtual try-on experimental setup from VITON-HD is used where the segment of the person's image representing the garment is exchanged with the associated paired garment. The expected outcome for the generated try-on image is supposed to be a close resemblance to the original paired image.

3.3.1. Qualitative analysis

Figure 3 demonstrate a few of the outputs from the proposed model also showing the layers of preprocessed images before generation of final image. With reference to both the figures:

1) First column represents the original input image to be given
2) Second column represents the target input cloth

Figure 3
FashionFlow results I: illustration of steps involved along with the final image

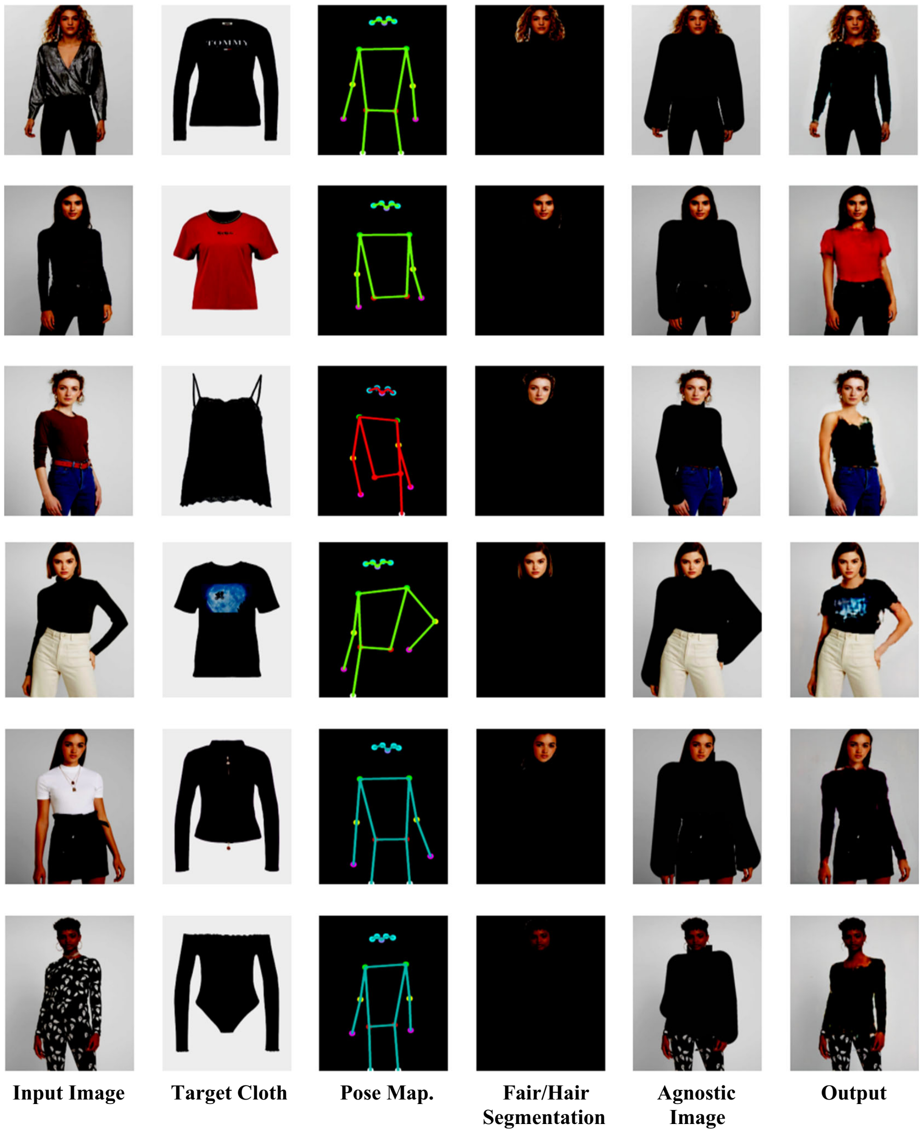
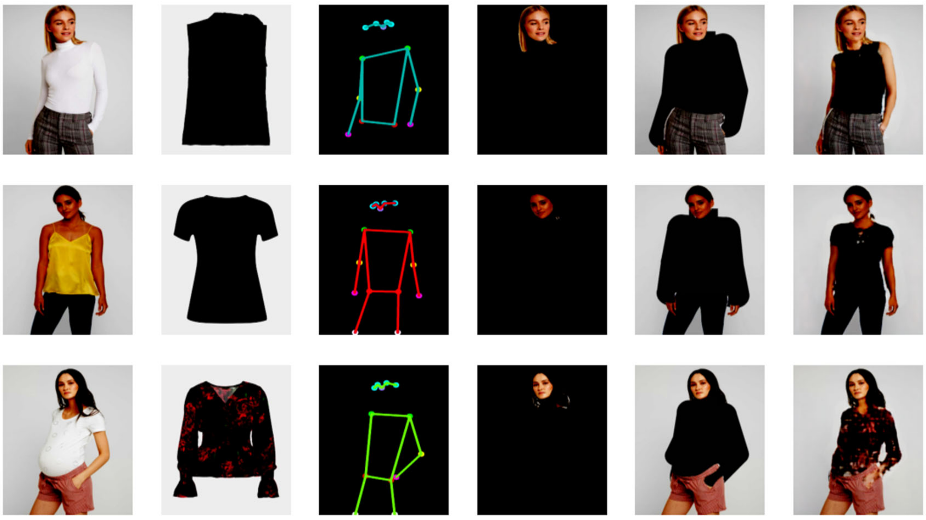


Figure 4
FashionFlow results II: demonstrating results of the proposed model in different scenarios



3) Third, Fourth, and Fifth columns represent the different representation techniques in the proposed model as a preprocessing step
4) The last column shows the output image generated by our model

With reference to Figure 4, the first and second row of outputs clearly demonstrates that the sleeves have been properly removed and added as per the given input cloth without any visible errors. This proves that the preprocessing techniques have been beneficial in successful generation of images. Our model has been capable in generation of accurate images, taking into account diverse body shapes for instance, a pregnant woman as shown in the third row.

Figure 5 shows the comparison between the results of the Pix2Pix GAN model with and without the discriminator, displaying that even when the model is trained without using the discriminator, the results are comparable to the model trained with the discriminator.

Figure 6 demonstrates the results from CVPR model proposed by Isola et al. [2]. The figure is adapted from Dong et al. [14]. From the figure, we can observe that the facial features are not saved and most of the images look distorted due to the same reason. The model proposed in this paper saves the facial features during preprocessing to ensure stable output image.

3.3.2. Quantitative analysis

To ensure a fair comparison, we evaluate FashionFlow using two popular metrics: Frechet inception distance (FID) [19] and LPIPS [20]. FID is superior to Inception Score in capturing how similar produced images are to the original ones. A two-time-scale update rule (TTUR) has been put into use to train GANs with stochastic gradient descent on any loss function in order to determine the FID score. Each generator and discriminator in TTUR has a unique learning rate. The update rule is signified as:

$$w_{n+1} = w_n + b(n)\left(g(\theta_n, w_n) + M_n^{(w)}\right) \tag{2}$$

$$\theta_{n+1} = \theta_\mathrm{n} + \alpha(n)\left(h(\theta_n, w_n) + M_n^{(\theta)}\right) \tag{3}$$

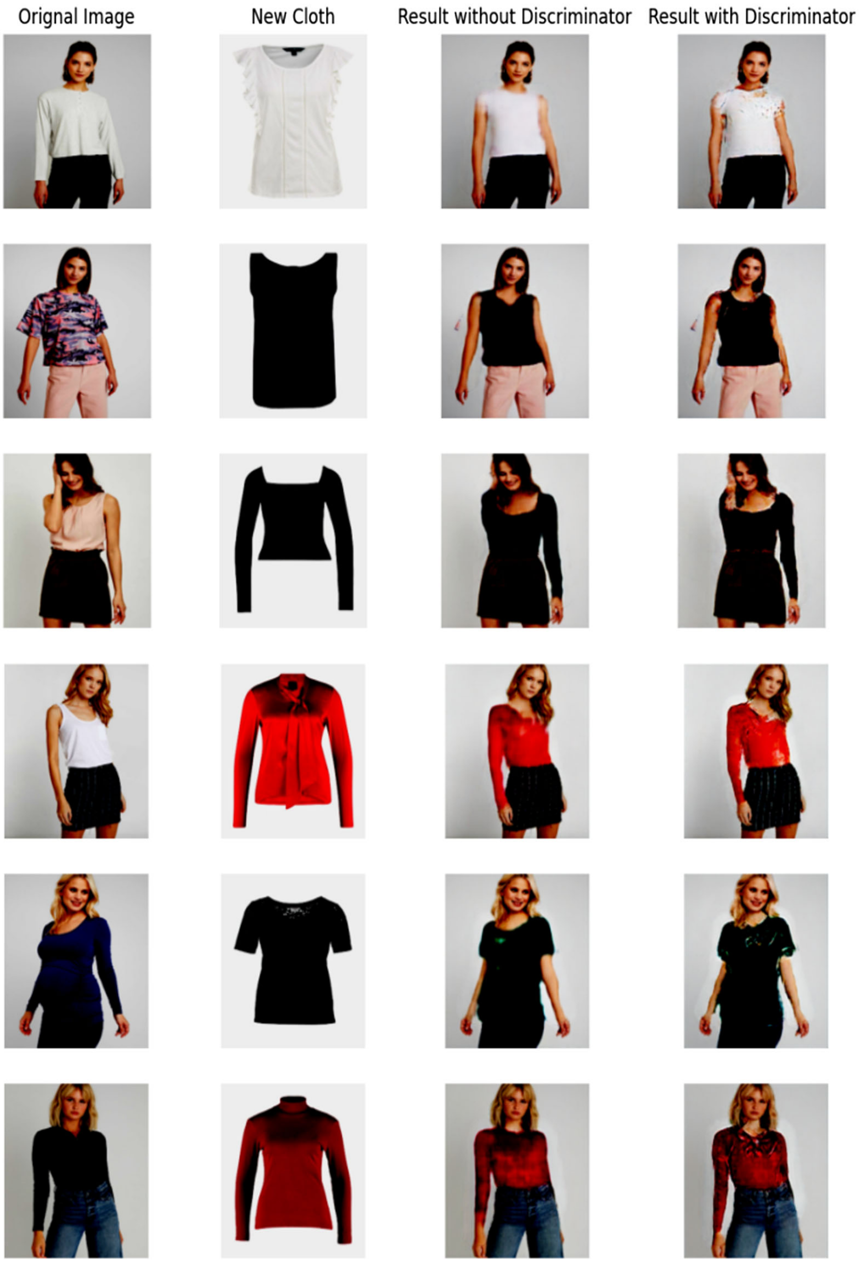
where w and θ signify the vector of parameters for generator and discriminator, respectively. Furthermore, $a(n)$ and $b(n)$ are the learning rates for discriminator and generator. ($g(\theta_n,\ w_n)$ and $h(\theta_n,\ w_n)$ represent the stochastic gradient for both loss functions. LPIPS is a metric used to quantify the perceptual similarity between two images. It is based on a deep-trained neural network which can evaluate the visual similarity between patches of images. The model not only considers low-level features like color and texture but also higher-level features related to object and scene semantics. Quantitatively, it calculates the distance between reference and distorted patches with network F. Let I1 and I2 be two images, and Pi and Fi be the patches and feature representations of those patches for images I1 and I2 respectively. The LPIPS score L between I1 and I2 can be indicated as the average L2 distance between corresponding feature representations:

$$\mathrm{L}(\mathrm{I}_1, \mathrm{I}_2) = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} ||\mathrm{F}_{1\mathrm{i}} - \mathrm{F}_{2\mathrm{i}}||^2 \tag{4}$$

Table 1 demonstrates FID and LPIPS Score of different models calculated over public datasets. It is noted that the proposed model FashionFlow performs reasonably well comparatively.

Additionally, we evaluate our models with existing models on the basis of the number of parameters along with the model size. Results from Table 2 demonstrate that FashionFlow not just becomes a highlight with fewer parameters and a smaller size but also maintains great accuracy in its generated outputs. It is also worth noting that the preprocessing steps implemented in these

Figure 5
Comparing results between "with" and "without" discriminator



studies are like the ones used in FashionFlow, thus exhibiting comparable complexity with respect to the proposed approach. This demonstrates the novelty of the proposed approach in maintaining the image quality with a nearly same complexity of preprocessing steps with an even smaller model with lesser parameters.

3.3.3. Discussion

In the qualitative analysis, it is proven that the proposed model demonstrates its superior ability to generate realistic and precise images. The model can be seen to effectively handle complex clothes and diverse body shapes. Thus, it is proven that the preprocessing techniques used to enhance the model's capability have contributed to the overall effectiveness of the model.

The appropriate results from the qualitative analysis section are further justified from the quantitative analysis of the model. The evaluation performed using the popular FID and LPIPS metrics shows that FashionFlow achieves a great performance. The lower FID score confirms the model's ability to generate images that are close to the original in terms of quality. Similarly, the LPIPS scores affirm the perceptual similarity of the generated images to the reference images, reflecting the

Figure 6
Results from the previous Pix2Pix implementation

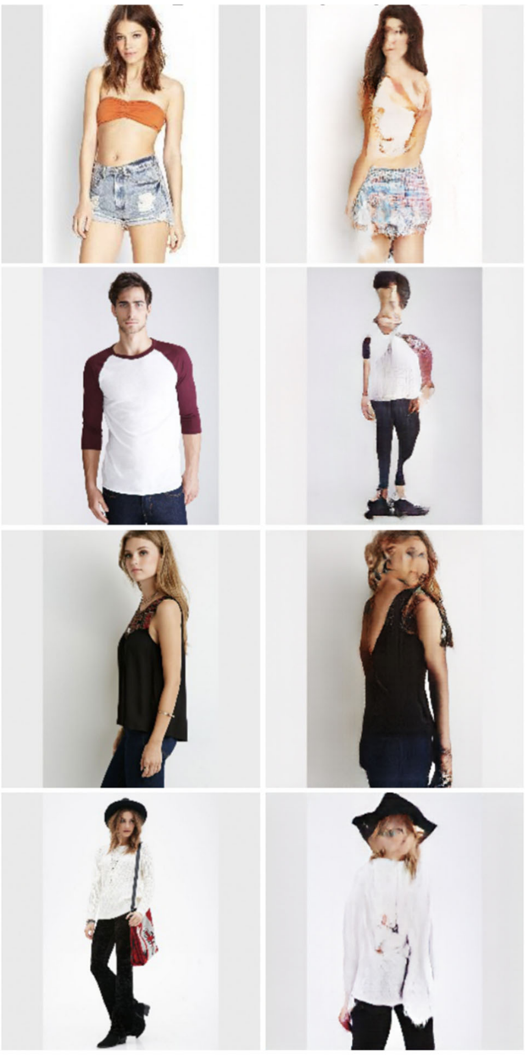


Table 1
Comparative results of FashionFlow with respect to FID and LPIPS scores

	Metrics	
Table models	FID	LPIPS
FashionFlow (Ours)	51.83	0.043
ACGPN [21]	43.29	0.112
CPVTOn	43.28	0.159
PASTA-GAN [22]	29.43	0.1215
VITON-HD	11.59	0.077

Table 2
Evaluation of models based on its size and number of parameters

Model name	Model size (in MB)	Total number of parameters (in millions)
FashionFlow (Ours)	207.61	54.42
HR-ViTOn	564.25	147.91
VITON-HD	587.74	154.07

model's proficiency in capturing both low-level and high-level features effectively.

Overall, these qualitative and quantitative findings suggest that the implemented techniques in FashionFlow, including the robust preprocessing steps and the unique architecture, offer significant benefits over existing models. The same has been successfully validated in the results, highlighting its potential for practical implementation in cloth superimposition applications.

4. Conclusion and Future Work

In our paper, we propose a novel virtual try-on method, called FashionFlow that replaces a person's garment by superimposing it with a target cloth. A modified version of Pix2Pix-based GAN has been utilized. In this version, the discriminator is not taken into consideration. The removal of discriminator in Pix2Pix GAN has been justified by the prior presence of L1 or pixel-wise error.

This helps the method to achieve higher computational efficiency along with accurate results. The different preprocessing methods give an insight into the input image of the person by studying its pose and by preserving the facial features. In order to perform the process of training and experimentation, we have utilized the VITON-HD dataset. The model performs fairly well, according to the qualitative and quantitative data, compared to the models with discriminator's inclusion and the original Pix2Pix model. In this work, the formation of complex patterns is not taken into account to generate a lightweight model with the goal of proving better results without using a discriminator. There exists a potential to enhance the model's capability for generation of images for abstract garments as a future work. Additionally, there is room for extension for superimposing clothes on multiple people in a video. This paper demonstrates that 2D image generation pipeline can be successfully used as an alternative to expensive 3D-based image generation methods. Overall, the proposed model approach shows great opportunities in the improvement of online fashion shopping experience without having to physically try the clothes on.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/shadow2496/VITON-HD.git; in Kaggle at https://www.kaggle.com/datasets/marquis03/high-resolution-viton-zalando-dataset.

Author Contribution Statement

Samar Pratap: Conceptualization, Formal analysis, Data curation. **Alston Richard Aranha:** Methodology, Validation, Visualization. **Divyanshu Kumar:** Software, Investigation, Writing – original draft. **P. Preethi:** Writing – review & editing, Supervision, Project administration.

References

[1] Choi, S., Park, S., Lee, M., & Choo, J. (2021). VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14131–14140.

[2] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[3] Saravanan, D., & Nithyaprakash, V. (2015). Fashion trends and its impact on society. In *International Conference on Textile, Apparel & Fashion 2015*.

[4] Akram, S. V., Malik, P. K., Singh, R., Gehlot, A., Juyal, A., Ghafoor, K. Z., & Shrestha, S. (2022). Implementation of digitalized technologies for fashion industry 4.0: Opportunities and challenges. *Scientific Programming*, *2022*(1), 7523246. https://doi.org/10.1155/2022/7523246

[5] Lata, K., Dave, M., & Nishanth, K. N. (2019). Image-to-image translation using generative adversarial network. In *3rd International Conference on Electronics, Communication and Aerospace Technology*, 186–189. https://doi.org/10.1109/ICECA.2019.8822195

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . , & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

[7] Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). VITON: An image-based virtual try-on network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7543–7552. https://doi.org/10.1109/CVPR.2018.00787

[8] Li, K., Zhang, J., & Forsyth, D. (2023). POVNet: Image-based virtual try-on through accurate warping and residual. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 12222–12235. https://doi.org/10.1109/TPAMI.2023.3283302

[9] Chen, C. Y., Chen, Y. C., Shuai, H. H., & Cheng, W. H. (2023). Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *IEEE/CVF International Conference on Computer Vision*, 7513–7522. https://doi.org/10.1109/ICCV51070.2023.00691

[10] Li, Z., Wei, P., Yin, X., Ma, Z., & Kot, A. C. (2023). Virtual try-on with pose-garment keypoints guided inpainting. In *IEEE/CVF International Conference on Computer Vision*, 22731–22740. https://doi.org/10.1109/ICCV51070.2023.02083

[11] Yan, K., Gao, T., Zhang, H., & Xie, C. (2023). Linking garment with person via semantically associated landmarks for virtual try-on. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17194–17204. https://doi.org/10.1109/CVPR52729.2023.01649

[12] Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2342–2351. https://doi.org/10.1109/CVPR.2019.00245

[13] Henry, J., Natalie, T., & Madsen, D. (2021). Pix2pix GAN for image-to-image translation. *Research Gate Publication*, 1–2. https://doi.org/10.13140/RG.2.2.32286.66887

[14] Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., & Yin, J. (2018). Soft-gated warping-gan for pose-guided person image synthesis. *Advances in Neural Information Processing Systems*, *31*.

[15] Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8798–8807. https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00917

[16] Rafiei, F., Shekhar, M., & Rahnev, D. (2024). RTNet neural network exhibits the signatures of human perceptual decision making. *BioRxiv Preprint*. https://doi.org/10.1101/2022.08.23.505015

[17] Lin, Y., Shen, J., Wang, Y., & Pantic, M. (2021). RoI Tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, *112*, 10419. https://doi.org/10.1016/j.imavis.2021.104190

[18] Yang, B., Wu, J., Ikeda, K., Hattori, G., Sugano, M., Iwasawa, Y., & Matsuo, Y. (2022). Face-mask-aware facial expression recognition based on face parsing and vision transformer. *Pattern Recognition Letters*, *164*, 173–182. https://doi.org/10.1016/j.patrec.2022.11.004

[19] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6629–6640.

[20] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595. https://doi.org/10.1109/CVPR.2018.00068

[21] Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., & Luo, P. (2020). Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7847–7856. https://doi.org/10.1109/CVPR42600.2020.00787

[22] Xie, Z., Huang, Z., Zhao, F., Dong, H., Kampffmeyer, M., Dong, X., . . . , & Liang, X. (2022). PASTA-GAN++: A versatile framework for high-resolution unpaired virtual try-on. *arXiv Preprint:2207.1347*.

How to Cite: Pratap, S., Aranha, A. R., Kumar, D., & Preethi, P. (2025). FashionFlow: A Lightweight Pix2Pix-Based Approach to Virtual Try-On. *Artificial Intelligence and Applications*, *3*(3), 275–283. https://doi.org/10.47852/bonviewAIA52022458