ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

Virtual Try-on: A comprehensive Study of Different Methodologies and the Architectures

[1] Neha Nageswaran, [2] Bastian Scharnagl, [3] Prof. Christian Groth

[1] [2] [3] Hof University of Applied Science
Corresponding Author Email: [1] neha.nageswaran@hof-university.de, [2] bastian.scharnagl@hof-university.de,
[3] christian.groth@hof-university.de

Abstract— ***In the midst of the growth in fashion retail industry, fitting of clothing remotely can be utilized to reduce return rates and unnecessary shipping. Due to recent advancements in Generative AI, models for generating a virtual try-on experience have also been able to develop further. This paper evaluates different models and techniques which are built for the application of virtual try-on. Additionally, this paper discusses about different models - ClothFlow model, Contextual-VTON model or FitGAN along with a deep insight into algorithms and methodologies namely Residual Networks, U-Net and Generative Adversarial Networks. Additionally, the paper compares these state-of-the-art models against the various evaluation metrics such as Frechet Inception Distance (FID), Inception Score (IS), Structural Similarity index (SSIM). It also explains, how the generated image outputs vary based on the changes made in hyper-parameters such as learning rate, batch size etc., and how these changes impact on the generated output.***

Index Terms— ***Virtual try-on, artificial intelligence, deep learning, neural networks, performance metrics, ClothFlow, C-VTON, FID, SSIM, inception score, machine learning, CNN, GAN, encoder, decoder.***

I. STATE-OF-THE-ART

A. Virtual Try-on: Future of Shopping

Virtual try-on for clothes is created to provide a revolutionary experience in shopping. In this world, which is purely driven by e-commerce, this innovative technology act as a game changer to the everlasting problem of buying cloth online without trying them on [7]. This platform attracts more customers because it produces an option of virtually trying out different cloths even before purchasing. Users could see how the cloth suits them and how it would look on the[1]m based on color, complexion and style. One of the major advantages of this technology is that it saves time and effort, eliminating the requirement of visiting a physical store to try cloths on. Moreover, it gives consumers confidence in their purchases as they already are aware about the style and look of the cloth on them. Additionally, it would become an integral part of the digital shopping world which brings the trial room to one's own home and could further help reducing the return rate.

B. Residual Neural Network (ResNet)

ResNet is a type of Convolutional Neural Network (CNN) model usually used for computer vision applications such as object detection, image classification, or semantic segmentation [28]. Context-driven image-based virtual fitting room network (C-VTON) [5] represents a state-of-the-art system characterised by realistic virtual try-on results. The modules used in the processing pipeline of C-VTON is built on ResNet-like blocks which consist of convolutional layers, a ReLU activation function and a trainable shortcut connection [11]. The two main components of C-VTON are the Context-Aware Generator (CAG) and the Body-Part Geometric Matcher (BPGM). Although the CAG and the BPGM contain ResNet blocks, the CAG includes context-aware normalization added before each individual convolutional layer, while the BPGM uses two encoders with stacked convolutional layers that also look like ResNet. In addition, the paper provides insights into the main architecture, implementation and evaluation based on the performance metrics.

C-VTON harness the capabilities of deep learning to extract the advanced features from input images by incorporating ResNet as a foundational framework of the generator network. This consequently enhances the quality of the generated virtual try-on results. To amplify the efficiency of the training process, the end-to-end training of the generator network is facilitated by back-propagation, which is another advantage of using ResNet. By embedding context-aware normalization (CAN) operations that can adapt to the image context, C-VTON introduces a change to the ResNet blocks and also refines the normalization procedure. This adaption relinquishes further increase in the performance of the generator network, enabling it to produce virtual try-on results that present more convincing visual results that are also semantically appealing.

1) Architecture

Figure 1 shows the two main components CAG and BPGM helps the model to generate a visually appealing try-on result. BPGM and CAG use convolutional neural networks (CNNs) together with ResNet like blocks for their

Manuscript received June 13, 2024. This work has been supported and funded by the European Regional Development Fund (ERDF).

processing pipeline. The problem of vanishing gradient in deep neural networks is addressed in the ResNet model by introducing residual connections [11]. The residual connections allow information to flow directly from one layer to another without being transformed, which in turn helps to mitigate the problem of information loss that could potentially occur in deep networks and allows much deeper models.

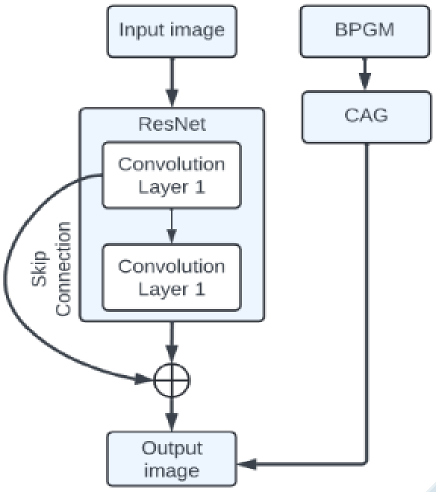


Fig. 1. ResNet architecture [22]

Overall, the architecture of C-VTON is designed to utilize these benefits by adapting ResNet blocks to the ContextAware Generator (CAG).

- BPGM: BPGM, which follows the design of the Geometric Matching Module (GMM) [23], takes the image of a person and clothing as input and in turn generates a warped clothing image aligned with the person’s pose. The BPGM uses a simplified geometric matching module that can generate accurate challenging poses and arm configurations. It contains two encoders with five stacked convolutional layers, followed by a downsampling operation, a ReLu activation and a batch normalization, where two encoders, one for the person’s image and one for the clothing, generate feature representations. These feature representations are then used to compute a dense correspondence field between the two images, which is ultimately used to warp the clothing image to match the pose of a person. BPGM is mainly suitable for handling complex poses and on-shirt graphics by dividing the body into several parts, each of which is aligned separately. This is useful when the target garment is partially occluded or has a very complex pattern. The training here is done using the combination of target shape loss, appearance loss and perceptual loss.
- CAG: Here the warped image and the image of the person is given as input to the CAG, and it generates a photo-realistic and contextually consistent final try on image. Although CAG, which uses a U-Net-like [20] architecture, is designed as a standard residual network, it relies on context-dependent (conditional) normalization operations to ensure that contextual information is sufficient when generating virtual tryon results. It synthesizes high-quality try-on results by using different sources of contextual information. Training here is done with a combination of perceptual losses and adversarial losses.

Target shape loss render the target clothing into a shape that matches the pose of a person, whereas appearance loss forces the visual appearance of the warped clothing in the body area to resemble the input image. Of the losses in CAG, perceptual loss measures the similarity between the generated image and ground truth image in terms of high-level features extracted from a pre-trained network, while adversarial loss ensures that the generated image cannot be distinguished from real images. In short, C-VTON handles challenging situations and generates visually appealing and contextually consistent results by considering both the context of the input image and the orientation of the target garment.

2) Performance Metrics

The paper delves into the quantitative evaluation of the performance of C-VTON by comparing it with other approaches such as CP-VTON [23], CP-VTON+ [18], ACGPN [24] and PF-AFN [6]. Both Fréchet Inception Distance (FID) [12] and Learned Perceptual Image Patch Similarity (LPIPS) [26] scores are used to evaluate the performance by accessing image quality and visual similarity. C-VTON consistently surmounts other approaches in both VITON [10] and Multi-Pose Virtual try on (MPV) datasets [3]. Compared to the closest runner-up, C-VTON 28.2% reduction in FID score and a 53.6% reduction in LPIPS score and similar performance is shown in the MPV dataset as well. C-VTON differs from other models in various aspects. For example, BPGM can accurately transform the garments under challenging poses and complex arm configurations. On the other hand, CAG leverages contextual information from various sources to synthesize high-quality virtual try-on results. Additionally, CAG implements conditional normalization operations to ensure that the contextual cues are effectively integrated when generating the results. Figure 4(a) shows how accurately the cloth is fit for the subject. C-VTON is versatile when it comes to handling diverse input image, making it a more adaptable solution. The evaluation of performance metrics confirms the excellence of C-VTON as a virtual try-on model that consistently outperforms state-of-the-art methods on both VITON and MPV datasets. The combination of BPGM and CAG and its ability to handle diverse input images make C-VTON a promising solution for virtual try-on applications.

However, it has a few limitations when it comes to generating accurate and realistic size output images. Missing of different sizes of input images can indeed contribute to the model’s inability to generate realistic size output images.

Challenges with loose clothing occur when the model struggles with accurately fitting oversized or loose-fitting clothing into person, which in turn impacts the realism in the generated images. By addressing these issues and by refining the model, there is a possibility to improve upon model's performance.

3) Hyperparameters

Various hyperparameters including learning rate, batch size, number of epochs, loss functions, number of ResNet blocks, and contextual normalization were used to fine-tune and optimize the performance of the CVTON model. For both BPGM and CAG, an initial learning rate of 0.0002 was used, which was eventually reduced by a factor of 10 after 100 epochs. As for the batch size, both components use a batch size of 4 and are trained for 200 epochs. In addition, the C-VTON model uses a variety of loss functions and tunes the number of ResNet blocks to optimize the performance of the model. To ensure the contextual information, the contextual normalization parameters were fine-tuned. In this way, the hyperparameters used in the C-VTON model are fine-tuned to optimize the performance of the model on both VITON and MPV datasets.

C. U-Net

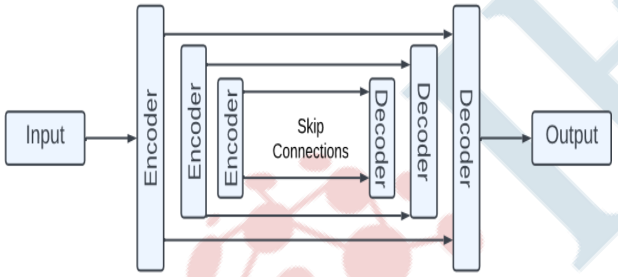


Fig. 1. U-Net based architecture [13]

U-Net is a type of CNN architecture usually used or image segmentation, feature extraction and translation [1]. ClothFlow [9] highlights the encoder-decoder structure with skip connections. It is mainly employed for generating good fitting clothing output and to enhance the appearance of the cloth in the output [9]. StableVITON [14] consists of the same encoder-decoder network with skip connections. Here, the encoder part of the U-Net down samples the input features in order to capture high level representation while the decoder unsamples these feature in order to generate the output images. U-Net in [14] plays a crucial role in generating high fidelity images for various try-on applications. This is done by learning the semantic correspondence between the clothing and the human body by preserving clothing details during the generation process [14]. As for the VTNFP [25] model, it consists of an attention gated U-Net, which is mainly used for feature extraction. This helps the model to focus on relevant regions of the input data and discards the irrelevant information [25]. Here, the U-Net is utilized for extraction of features from the person representation and the warped clothing thus generating a realistic and detailed try-on results [25].

1) Architecture

The U-Net [21] architecture is a type of CNN most used for image segmentation, image-to-image translation, etc. Encoder and decoder are two symmetric parts of the U-Net, where the encoder part helps to extract high-level features from the given input image, while the decoder part reconstructs the output image from these important features.

Figure 2, inspired from [20] shows the structure of U-Net, which is used for generating within the latent space of the autoencoder. When the given input image gets encoded into latent space, the U-Net generates a new latent space with this information [21]. The decoder uses this new latent space to decode the new information back to the image. The U-Net is trained to learn the relationship between clothing and the human body by using intermediate feature maps from a spatial encoder [21]. This helps the model align the clothing with the body and preserve clothing details during image generation. Models such as StableVITON[14], ClothFlow[9], Virtual Try-on Network with Body and Clothing Feature Preservation (VTNFP)[25] etc make use of U-Net for its efficient way of preserving the clothing details.

StableVITON [14] comprises of U-Net architecture for feature extraction, denoising and to condition intermediate feature maps to align the clothing to the human body to achieve a visually convincing result [8].

The use of the U-Net architecture in StableVITON helps to improve the quality of the generated image by increasing the sharpness of the images, reducing artifacts and increasing the diversity of the generated images. It enhances the image sharpness by learning the fine details of the images, which helps to make the generated image more realistic. Artifacts caused by the diffusion model can be reduced by smoothing, which in turn makes the generated image more visually appealing [8]. The ability of the U-Net to learn a variety of styles from the training data helps to make the generated images more diverse and visually interesting [14]. In addition to the U-Net, StableVITON uses other techniques to improve the quality of the generated images.

ClothFlow [9] uses U-Net because of its effectiveness in capturing high-level features and sharing information across the network. In conditional layout generation and image-based rendering in virtual try-on, U-Net performs well due to its ability to handle complex spatial relationships and capture detailed information [13]. In conditional layout generation, the U-Net architecture is used to generate the conditional map from the semantic layout map, where the encoder part gradually reduces the spatial dimensions while capturing the high-level features. The decoder, on the other hand, extracts the feature maps and combines them with skip connections to recover the spatial dimensions and generate a conditional map. In this way, the network can effectively capture the structural constraints for good appearance and clothing details [13]. During the image rendering phase, the final image is synthesized from warped source clothing regions of the source and other input conditions are being synthesized by the U-Net.

The U-Net's ability to capture high-level features and pass information via skip connections enables the generation of high-quality images that are realistic and preserve the details of the clothing [4].

Yu, Ruiyun, et al. [25] has explained how the U-Net is used in the Virtual Try-on Network with Body and Clothing Feature Preservation (VTNFP). U-Net is being employed in the try-on synthesis module (M3) of the VTNFP model to extract features of the person in the image and the garment after warping. In addition, it captures and fuses the information it obtains from the predicted body part segmentation map, the warped clothing and additional auxiliary information about the body to synthesize the final image, preserving the original trouser and arm features in it preserving clothing and body parts [27].

Figure 2 shows how U-Net is used for virtual try-on tasks. There is an encoder that extracts features from the input person image and the warped input clothing image. By processing these features, the U-Net generates a segmentation map that displays different body parts and a conditional map that contains information about warped clothing features and additional auxiliary body information. [25] The decoder uses the information encoded in the feature maps to generate a detailed and refined output [13].

2) Performance Metrics

Han, Xintong et al. [9] explains the performance of the model in ClothFlow in terms of performance metrics, Structured Similarity Index Metric (SSIM) and Inception Score (IS). ClothFlow achieves comparable quantitative performance to other state-of-the-art methods in terms of SSIM and IS. IS measures the diversity and quality of the synthesized result, while the SSIM metric provides information about the structural similarity and perceptual quality of the generated images [9]. The results of these evaluation metrics show that the method produces images with higher structural similarity and diversity, which are comparable to other advanced methods in the field. In addition to the quantitative metrics, a user study is conducted to evaluate the perceptual quality and to see to what extent the generated image is real. Humans are tasked with selecting more realistic imagefrom pairs of generated images, providing subjective feedback on the visual fidelity and realism of the results [9]. Additionally, quantitative comparisons in terms of Warp-SSIM and Mask-SSIM are conducted on VITON dataset showcasing the improvement in warping accuracy achieved by the ClothFlow model. It also proves that the model excels in synthesizing desired clothing, as evidenced by its higher scores compared to other methods. These metrics highlight ClothFlow's capability to accurately warp clothing regions and reconstruct clothing details, contributing to the overall quality of the synthesized clothed person images [9].

In [25], the evaluation metrics such as FID, SSIM score etc are not explicitly mentioned, instead the performance of the model is measured based on a user perception study, which is a qualitative evaluation method that rely on human feedback to assess the quality of the synthesized images. This study involved A/B test to compare the image quality synthesized by VTNFP against those synthesized by VITON/CP-VTON. The results of A/B test proved that 67.87% of the images generated by VTNFP have better quality than those generated by VITON. 77.38% of the images generated by VTNFP have better quality than those generated by CP-VTON.

Kim,Jeongho et al. explains [14] highlights the performance of the model in terms of the evaluation metrics, which include Learned Perceptual Image Patch Similarity (LPIPS), Kernel Inception Distance (KID), FID, and SSIM, which are commonly used to evaluate the quality and fidelity of the generated images in the context of virtual try-on applications. When evaluating the realism and variety of the generated images, FID and KID play a major role, with lower FID and KID values indicating that the generated images are more realistic and varied, thus reflecting a better quality of the virtual try-on results [14]. LPIPS measures the perceptual similarity between images. The lower the LPIPS values, the closer the generated images are to the real images, which means a higher quality of the virtual try-on results in terms of visual similarity [14]. SSIM also plays an important role in evaluating the performance of the model by measuring the similarity of structure between images. The higher the SSIM values, the greater the structural similarity of the generated images to the ground truth, suggesting better preservation of structural details. StableVITON uses the above performance metrics to evaluate the model's performance in terms of image quality, fidelity and perceptual similarity.

Figure 4(c), 4(d) and 4(b) show how the models fit the cloths accurately. However, there are a few drawbacks which could potentially affect the generated images. ClothFlow model [9] performs well in generating realistic clothed person images but faces challenges with intricate clothing deformations and occlusions which in turn can lead to distortions in the output. The training data could be augmented, and the chosen hyper parameters could be fine-tuned. Additional constraints which push the model in handling challenging scenarios. On the other hand, StableVITON model faces challenges in preserving fine details of clothing and facial features and in handling the occlusions in images. The model struggles with intricate details due to the complexity of accurately representing fine details of image and handling additional elements beyond clothing during the image generation process. [25] struggle in preserving the body and cloth details when it comes to complex poses which leads to blurry or deformed regions at the point where the cloth and body interact. It is due to the issue in retaining the cloth and body information where they both intersect which generates blurry pictures or incorrect clothing placement output.

3) Hyperparameters

This publication [14] highlights the performance of the model by adjusting the learning rate using the Adam optimizer with a fixed learning rate of 1e-4 for 360k iterations.

The autoencoder's decoder is fine-tuned on the DressCode and VITON-HD training datasets with a learning rate of 5e-5 for 10k iterations on each dataset.

On the other hand, [25] regulates the model for its VTNFP model by using the Adam optimizer and initially setting the learning rate to 0.0001 and then linearly reducing it to zero over a certain number of epochs. The decoder of the autoencoder is fine-tuned on training datasets DressCode and VITON-HD with a learning rate of 5e-5 for 10k iterations on each dataset. [9] sets the learning rate to 0.0002 and is kept fixed throughout the training process and later being fine-tuned to ensure stable convergence and optimal performance.

When it comes to batch size, [25] batch size of four for the clothing deformation module and batch size of five for the segmentation map generation module, however, [14] uses a batch size of 32 for the training model. [9] sets the batch size based on the computational resource availability and size of training dataset to twelve.

The author of this paper [14] makes use of various augmentation techniques like random shift, horizontal flip, etc,. These techniques are applied to the clothing and the input conditions during training. Pseudo Linear Multi-Step (PLMS) sampler is set to fifty sampling steps for inference and a novel loss function, attention total variation loss is introduced in order to improve the sharpness of attention map thus preserving the clothing details. The author in paper [25] uses a combination of L1 loss, adversarial loss and focal loss which are used to optimize different aspects of the model, such as pixel-wise intensity differences, perceptual differences, and adversarial training. The author in paper [9] uses two regularization parameters in order to balance the contributions of the structural loss and loss of smoothness. They are set to certain values based on empirical observations and they are fine tuned to achieve optimal performance.

D. Generative Adversarial Network (GAN) [2]

GAN [2], a type of feed forward neural network architecture is mainly chosen for its exemplary performance in generating realistic data. They work with variety of data including text, image, audio, video, etc,. which includes work with unlabelled data as well [2]. Generator and discriminator are two main parts of GAN where the generator creates new data while the discriminator evaluates its authenticity.

In this paper [17], the author highlights the advantages of using a type of GAN, which is StyleGAN2 architecture that generates high quality images and also preserves the body shape, skin color and identity features while transferring the garment from a different person. This is achieved by incorporating the network with information on 2D human body pose and the segmentation of the clothing. In order to determine the most effective interpolation coefficients for each layer, an optimization method is being employed [17].

On the other hand, S, Kumar et al. in TryOnGAN architecture, addresses the challenges of virtual try-on by leveraging conditional image generation and latent space manipulation [15].

The author in this paper [19] explains the GAN architecture which is used in FitGAN to learn disentangled item representations and generate realistic images reflecting the true fit and, shape properties of fashion articles, providing a fundamental baseline capable of achieving controllable generation and rich semantically meaningful article representations.

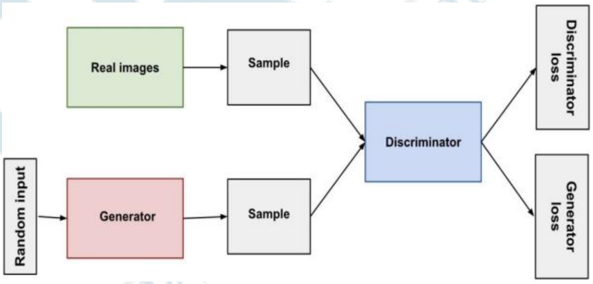


Fig. 2. GAN architecture [16]

1) Architecture

Figure 3 [16] inspired from [16] explains the general GAN architecture. The generator creates fake images from random noise while the discriminator has to differentiate between real and fake images [16]. Generator loss function encourages generating data that maximizes the probability of fooling the discriminator [16]. Discriminator loss function penalizes misclassifications, encouraging it to correctly identify the source of the data [16].

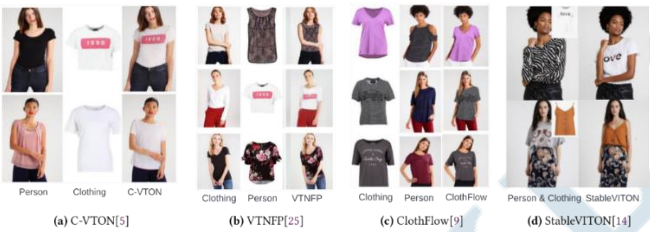
The pose conditioned StyleGAN2 model uses 2D human body pose and clothing segmentation to create. By incorporating these aspects into the network's conditioning, it maintains essential characteristics like identity, body contours, and skin tone while transferring attire across different individuals. Employing an optimization procedure, it determines the best interpolation coefficients per layer, leading to the realistic portrayal of a person adorned in a particular ensemble [15].

Try-on GAN adopts a customized StyleGAN2 structure to facilitate virtual try-on experiences with clothing items. By conditioning the model using pose embeddings and latent style embeddings provided by users, it generates highly lifelike depictions of a chosen garment on a target individual, even amidst pose variations between the source and target subjects. Through inversion optimization, it disentangles pose and style data, granting control over garment characteristics like shape, hue, and fit [17].

Within this paper [19], the author explains how FitGAN relies on a GAN framework to produce authentic representations embodying the fit and form attributes of fashion pieces. By conditioning the generator network on fit descriptors such as shape labels, FitGAN delivers visually realistic and varied interpretations of clothing items. GANs are selected for their capacity to generate lifelike imagery and their adaptability to diverse data formats and conditioning parameters, making them well-suited for addressing the varied aspects of fashion items handled by FitGAN.

2) Performance Metrics

Pecenakova, S et al. [19] explains about different performance metrics such as Inception Score, FID, precision and recall. Here, the results are described as demonstrating the ability of the model to produce visually realistic and diverse images for a lot of online garments.



(a) C-VTON[5] (b) VTNFP[25] (c) ClothFlow[9] (d) StableVITON[14]

Fig. 4. Comparison of generated examples

Lewis, K et al. [17] accentuates on evaluation metrics such as FID, Embedding Similarity (ES) score, and perceptual user study. The ES score evaluates the distance between embeddings of the original garment and the garment in the try-on image. When the ES score is higher, it indicates better similarity between the original and transferred garments [17]. Perceptual user study is being done in order to evaluate the quality of the try-on results based on the perceptive of humans. The participants for the study choose the best result among various pairs of try-on images generated by different methods [17].

Kumar, S et.al [15] focuses on using FID scores to evaluate the photo-realism of the image which are generated. Qualitative comparison of different model or training strategies is done by using of image grids, which helps comparing the images side by side by visually assessing the differences and similarities between different models.

The author further highlights the impact of transfer learning by showing how pretrained weights lead to a quick drop in FID scores initially with lesser training times. The pretrained models obtain lower FID scores post certain number of training images which demonstrate the effectiveness of transfer learning in improving the quality of generated images. Apart from this, the author explores the properties of the learned latent spaces in terms of style continuity and disentanglement of pose and style codes. These are being explored in order to highlight the ability of the model to generate highly realistic virtual try-on images.

3) Hyperparameters

In this paper [19], parameters such as loss functions, learning rate, regularization techniques, batch size are adjusted in order to fine tune the model which helps with stabilizing training in StyleGAN-adaptive discriminator augmentation (ADA) with limited data.

Lewis K.M. et al. in [17] mentions about the hyperparameters which were used in the optimization of loss functions and are fine tuned in order to achieve good results in terms of preserving identity features, body shape, and skin color while transferring the garment from a different person.

Kumar, S. et.al [15], TryOnGAN model focuses on using transfer learning which helps in achieving faster convergence and improving the quality of the generated images.

II. SUMMARY AND CONCLUSION

The integration of AI on retail industry through applications like Virtual Try-on has revolutionized the new era of shopping and has enhanced the shopping experience for the customers. Through the introduction of AI technologies such as ResNet, U-Net, auto encoder, GAN, etc,. models like C-VTON, StableVITON, etc,. took shape in order to make the application more precise. By leveraging deep learning techniques, these systems can accurately warp clothing regions, synthesize high-quality try-on results, and preserve important visual and structural details of both the garments and the human body.

Performance metrics like FID, LPIPS, IS, SSIM, etc,. validate how effectively the models generate realistic and visually appealing output images. Additionally, user perception studies gives an affirmation on how good the AI generated try-on image is, when compared to traditional methods. Improved image quality and stable convergence can be achieved by fine tuning the hyper-parameters and incorporating advanced techniques such as transfer learning, normalization etc.

These models address enduring hurdles in online shopping such as trying cloths virtually, convenience for the user to see how different outfit fit on them. Leveraging deep learning methods, these systems accurately manipulate clothing regions, generating high quality outcomes retaining the visual and structural features of the garment.

Out of all models, StyleGAN is one model which poses a potential for rendering the realistic length of the clothing. As a future improvement, the size information of the cloths could be taken into consideration and potentially use this information to be rendering of realistic size clothing images.

III. APPENDIX

In this section, we try to provide overall summary of various research publications, datasets, models and metrics evaluated to compose this study.

	Paper	Dataset	Model	Objective	Performance Metrics	Hyperparameters considered	Base architecture
1	[5]	• MPV • VITON	C-VTON	Develop a method that generate realistic and visually convincing try-on images, even under challenging conditions.	• FID • LPIPS	• Number of layers • Learning rate • Batch size • Number of epochs • Loss functions • Number of • ResNet blocks • Contextual normalization	Resnet
2	[9]	• DeepFashion • VITON	ClothFlow (flow based model)	To develop a novel flow-based model for clothed person generation that can generate realistic and visually convincing images of people wearing clothes, even under challenging conditions such as complex poses and self occlusions.	LPIPS score, SSIM, FSR	No. of epochs Batch size Learning Rate	U-Net
3	[14]	• VITONHD	Stableviton	Preserving the clothing details by learning the semantic correspondence and synthesizing high-fidelity images by leveraging the pre-trained models' inherent knowledge about humans in the warping process.	• SSIM • LPIPS • FID • KID	• Learning rate • Adam optimizer	U-Net
4	[25]	• Zalando dataset		To propose a new image-based virtual try-on network, VTNFP designed to synthesize photorealistic images of a person wearing a target clothing item while preserving both body and clothing details and also to demonstrate that VTNFP outperforms existing state-of-the-art methods in terms of image quality and detail preservation.	User Perception Study	• AdamOptimizer • learning rate • batch size	U-Net
5	[18]	• Deep Fashion • VITON	CP-VTON+	To develop a new image-based virtual try-on method that can generate more realistic and visually convincing try-on images with preserved clothing shape and texture, even under challenging conditions such as complex poses.	• Intersection over Union • (IoU) • LPIPS • SSIM • IS	• Learning Rate • Batch Size • Adamoptimizer • Loss function	CNN
6	[19]	• VITON • Deca WVTO	FitGAN	To introduce FitGAN, a generative framework for realistically conveying physical fit and shape	• IS • FID • Precision	• loss functions • learning rate • regularization	GAN

		N		attributes of garments at scale, with the aim of addressing the complex problem of remote fitting of fashion articles in e-commerce.	• Recall	techniques • batch size	
7	[15]	• Deep Fashion • Private Dataset	TryOnGAN	To explore how transfer learning and pose conditioning affect the TryOnGAN model and to investigate different properties of latent space interpolation.	• • FID	Use of transfer learning	GAN
8	[17]	—	Pose conditioned StyleGAN2	To introduce a novel image-based try-on algorithm, TryOnGAN, which seamlessly integrates person-specific components with garment shape and details to generate high-quality, photorealistic try-on images.	• FID • ES score • perceptual user study	Loss function	GAN

REFERENCES

[1] Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M Taha, and Vijayan K Asari. 2019. Recurrent residual U-Net for medical image segmentation. Journal of medical imaging 6, 1 (2019), 014006–014006.

[2] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. IEEE signal processing magazine 35, 1 (2018), 53–65.

[3] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. 2019. Towards multi-pose guided virtual try-on network. In Proceedings of the IEEE/CVF international conference on computer vision. 9026–9035.

[4] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. 2016. The importance of skip connections in biomedical image segmentation. In International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, 179–187.

[5] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. 2022. CVTON: Context-driven image-based virtual try-on network. In Proceedings of the IEEE/CVF winter conference on applications of computer vision. 3144–3153.

[6] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. 2021. Parser-free virtual try-on via distilling appearance flows. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 8485–8493.

[7] google. 2023. https://blog.google/products/shopping/ai-virtual-try-ongoogle-shopping/. In virtual try-on.

[8] Sangkwon Han, Seungbin Ji, and Jongtae Rhee. 2023. Diffusion-

[9] Denoising Process with Gated U-Net for High-Quality Document Binarization. Applied Sciences 13, 20 (2023), 11141.

[10] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. 2019. Clothflow: A flow-based model for clothed person generation. In Proceedings of the IEEE/CVF international conference on computer vision. 10471–10480.

[11] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. Viton: An image-based virtual try-on network. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7543–7552.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings´ of the IEEE conference on computer vision and pattern recognition. 770–778.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017).

[14] Xinrong Hu, Junyu Zhang, Jin Huang, JinXing Liang, Feng Yu, and Tao Peng. 2022. Virtual try-on based on attention U-Net. The Visual Computer 38, 9-10 (2022), 3365–3376.

[15] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2023. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. arXiv preprint arXiv:2312.01725 (2023).

[16] Saurabh Kumar and Nishant Sinha. 2022. Probing TryOnGAN. In 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD). 300–301.

[17] Maria Elena Laino, Pierandrea Cancian, Letterio Salvatore Politi, Matteo Giovanni Della Porta, Luca Saba, and Victor Savevski. 2022. Generative adversarial networks in brain

imaging: A narrative review. Journal of imaging 8, 4 (2022), 83.

[18] Kathleen M Lewis, Srivatsan Varadharajan, and Ira KemelmacherShlizerman. 2021. Tryongan: Body-aware try-on via layered interpolation. ACM Transactions on Graphics (TOG) 40, 4 (2021), 1–10.

[19] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. 2020. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In CVPR Workshops, Vol. 3. 10–14.

[20] Sonia Pecenakova, Nour Karessli, and Reza Shirvany. 2022. Fitgan: fit-and shape-realistic generative adversarial networks for fashion. In 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 3097–3104.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 234–241.

[22] Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. 2021. U-net and its variants for medical image segmentation: A review of theory and applications. Ieee Access 9 (2021), 82031–82057.

[23] Sasha Targ, Diogo Almeida, and Kevin Lyman. 2016. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029 (2016).

[24] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In Proceedings of the European conference on computer vision (ECCV). 589–604.

[25] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 7850–7859.

[26] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. 2019. Vtnfp: An imagebased virtual try-on network with body and clothing feature preservation. In Proceedings of the IEEE/CVF international conference on computer vision. 10511–10520.

[27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 586–595.

[28] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. 2019. Virtually trying on new clothing with arbitrary poses. In Proceedings of the 27th ACM international conference on multimedia. 266–274.

[29] Tao ZHOU, Bing-qiang HUO, Hui-ling LU, and Hai-ling REN. 2020. Research on residual neural network and its application on medical image processing. ACTA ELECTONICA SINICA 48, 7 (2020), 1436