International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multipose Attaire-fit-in using Deep Neural Network Frame Works

Chandrashekhara K T*[1], Gireesh Babu C N*[2], Vijaykumar Gurani[3], Ashwini N[4], Bhavya G[5],Sumith S[6]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: Creating an image-based virtual try-on system that seamlessly fits in-shop clothing onto a reference person across various poses poses a significant challenge. Previous efforts have concentrated on preserving intricate clothing details such as textures, logos, and patterns during the transfer of desired garments onto a target person in a fixed pose. However, when extending these methods to accommodate multi-pose scenarios, the performance of existing approaches has notably declined. This paper introduces an innovative end-to-end solution, the Multi-Pose Virtual Try-On Network, designed to effectively adapt desired clothing onto a reference person in arbitrary poses. The virtual try-on process involves three key sub-modules. Firstly, a Semantic Prediction Module (SPM) is employed to generate a comprehensive semantic map of the desired clothing. This predicted semantic map enhances guidance for locating the desired clothing region, leading to the creation of a preliminary try-on image. Secondly, a module is introduced to warp the clothing to the desired shape based on the predicted semantic map and pose. To address misalignment issues during the clothing warping process, a conductible cycle consistency loss is incorporated. Lastly, a Try-on Module combines the coarse result and the warped clothes to produce the final virtual try-on image, ensuring the preservation of intricate details and alignment with the desired pose.Additionally, a face identity loss is introduced to refine facial appearance while maintaining the identity of the virtual try-on result. The proposed method is evaluated on a vast multi-pose dataset, demonstrating its superiority over state-of-the-art methods. The qualitative and quantitative experiments indicate that the virtual try-on system exhibits robustness to data noise, including changes in background and accessories such as hats and handbags. This showcases its scalability and effectiveness in real-world scenarios.

Keywords: *Deep Learning, Virtual Try- on, Apparel Try-on, Post Estimation*

1. Introduction

The surge in online shopping has propelled the need for virtual dress fitting or virtual try-on experiences, where customers can visualize how a garment will look on them before making a purchase. Image-based virtual fitting involves the generation of images that transform a person's clothing to match another specified garment in a product image. While this concept enhances customer satisfaction, ensuring a satisfactory virtual try-on experience requires attention to key aspects

1) Preservation of the client's pose, body shape, and identity.

(2)Natural deformation of clothing based on the desired region, considering posture and skeleton structure.

(3)Attention to the intricacies of the apparel material.

(4)Accurate rendering of body parts originally covered by the reference individual's costume in the primary image.

Various methods have been proposed for virtual try-on, including approaches that distort clothing images to fit the target body, followed by pixel-level refinement. Some methods use segmentation maps to determine the client's confirmation from the final image. However, existing methodologies face challenges such as low-resolution synthesis due to misalignment between contorted apparels and the human figure during cloth image warping. The 'thin-plate spline' (TPS) transformation and ClothFlow are examples of these approaches, with the latter involving higher processing costs due to pixel-level clothing movement prediction.

Additionally, current techniques employing a simple U-Net architecture for high-resolution image synthesis encounter issues with unbalanced training and poor-quality images, especially when attempting to recreate originally obstructed human body regions. Recognizing these challenges, recent research introduces clothing-agnostic person representation by removing clothing information and utilizing posture information and segmentation maps. The Alignment Aware Segment (ALIAS) Normalization is employed to enhance image generation, excluding irrelevant information in displaced areas and propagating semantic information throughout the network. The ALIAS generator uses normalization to synthesize images of customers wearing the target garment, preserving garment texture and details even in misaligned areas through multi-scale improvements at the feature level [2]

2. Existing System

Here is a summary of the introduction and proposed method sections of the Paper: The paper discusses the challenges and benefits of using immersive virtual

[1] *BMS Institute of Technology and Management – 560064, India*
ORCID ID : 0000-0002-6716-4961
[2] *BMS Institute of Technology and Management – 560064, India*
ORCID ID : 0000-0002-1017-2573
[3]*Karnatak University, Dharwad – 580003, India*
ORCID ID : 0000-0002-2516-9561
** Corresponding Author Email:chandru@bmsit.in, gireeshbabu@bmsit.in*

platforms (IVPs) for education and training, and proposes a framework for integrating and evaluating IVPs in different contexts.

Proposed Method: The paper presents a four-step method for designing, implementing, assessing, and improving IVPs, based on a literature review and a case study of an IVP for teaching English as a foreign language.

The method involves: - Defining the learning objectives and outcomes of the IVP, and aligning them with the curriculum and the pedagogical approach. Selecting the appropriate IVP features and tools to support the learning objectives and outcomes, such as avatars, interactions, scenarios, feedback, and analytics. Evaluating the effectiveness and usability of the IVP, using both quantitative and qualitative methods, such as surveys, interviews, observations, and learning analytics.

Improving the IVP design and implementation based on the evaluation results, and iterating the process until the desired outcomes are achieved [3].

This addresses the problem of virtual try-on, which is to generate realistic images of a person wearing a desired clothing item. The paper identifies two main challenges: occlusion and misalignment. The paper proposes a novel method called OccluMix, which leverages a semantically-guided mixup technique to de-occlude the person and the clothing, and a multi-scale alignment module to align the clothing to the person.

Proposed Method: The paper presents the details of the OccluMix method, which consists of four main components: a segmentation network, a mix-up network, a generation network, and a discriminator network. The paper explains the role and function of each component, and how they work together to achieve the virtual try-on task. The paper also describes the loss functions and the training procedure of the OccluMix method.

This paper introduces a novel virtual try-on system leveraging deep learning technologies, specifically style transfer algorithms and generative adversarial networks (GANs), to enhance the online shopping experience. The system predicts the semantic layout of reference images and generates detailed clothing representations, aiming to provide users with a realistic and visually appealing virtual try-on experience. The system's architecture is based on Python Web, and its performance has been validated through experiments, showcasing its potential in the e-commerce sector [5].

The evolution of virtual try-on technology in e-commerce has been significant, yet it faces challenges such as inefficiency and unrealistic outcomes. Previous methods, including video streaming-based, 3D modeling-based, and image processing-based approaches, have attempted to address these issues with limited success. This study identifies the need for a more effective solution that can deliver high-quality virtual try-on experiences, overcoming the limitations of existing technologies[6].

The proposed system utilizes a combination of style transfer algorithms and GANs to apply the user's input image style onto a virtual scene and generate content while maintaining image details. The system architecture comprises a front-end for user interaction and a back-end for data processing and model computation. The back-end system is structured into four layers: data, service, algorithm, and application, with the ACGPN algorithm selected for virtual try-on applications. The system is trained on extensive clothing datasets to ensure accurate predictions and high-quality results.

Experimental analyses demonstrate that the system achieves high accuracy, authenticity, and efficiency in simulating clothing changes within a virtual environment. The system outperforms previous virtual try-on systems in terms of both accuracy and realism, offering a comprehensive and high-quality virtual dressing service. It effectively handles challenging scenarios and is capable of accommodating various types of clothing, making it a robust solution for fashion applications

The study presents a significant advancement in virtual try-on technology, offering a deep learning-based system that surpasses existing solutions in terms of performance and user experience. The system's ability to generate realistic and detailed clothing representations in a virtual setting has the potential to revolutionize the e-commerce industry. Future research will focus on further optimizing the system's efficacy and scalability, making it suitable for deployment on a larger scale[4].

Deep learning, virtual try-on, convolutional neural network, generative adversarial network, style transfer, e-commerce, ACGPN algorithm. RESUNET refers to Deep Residual UNET. It’s an encoder-decoder architecture developed by Zhengxin Zhang et al. for semantic segmentation. It was initially used for the road extraction from the high-resolution aerial images in the field of remote sensing image analysis. Later, it was adopted by researchers for multiple other applications such as polyp segmentation, brain tumour segmentation, human image segmentation, and many more. The RESUNET consists of an encoding network, decoding network and a bridge connecting both these networks, just like a U-Net. The U-Net uses two 3 x 3 convolution, where each is followed by a ReLU activation function. In the case of RESUNET, these layers are replaced by a pre-activated residual block.

Fig 1. Incorrect fit in ResNet Architecture

Fig 2. Cloth fitting in ResNet Architecture

Incorporates residual blocks from ResNet architectures, which increases the network's depth and complexity. This can lead to better performance on large and complex datasets but might cause overfitting on smaller or simpler datasets due to its increased capacity and parameters.This is demonstrated in the image above.As shown in figure 1 and figure 2, you can see in the above image that fitting is near perfect but not perfect as the cloth is not fitting on the shoulder perfectly having an improper fitting and in the second image the cloth has been tried to fit upon the base image of the model without removing her existing cloth[12]

3. PROPOSED SYSTEM

A significant issue known as pixel-squeezing artifacts remains. These artifacts occur due to the excessive warping of clothing near occluded regions within the HR-VITON framework, which is designed to handle misalignment and occlusion conditions. The root of this problem lies in the disconnection between the warping and segmentation map generation modules, which prevents proper information exchange and restricts the range of possible poses for the person images. This limitation poses a challenge for applying virtual try-on technology in practical scenarios [19].

To address these challenges, we propose a novel try-on condition generator that integrates the warping and segmentation generation modules. This integrated module predicts both the warped garment and the segmentation map simultaneously, ensuring perfect alignment between the two. Our approach completely eliminates misalignment and effectively manages occlusions caused by body parts. Through extensive experiments, we demonstrate that our framework achieves state-of-the-art results on high-resolution datasets (e.g., 1024×768) both quantitatively and qualitatively [21].

Several methodologies have been suggested for virtual try-on systems, which typically involve distorting clothing images to fit the target body and then refining the image at the pixel level. Among these, some methods utilize segmentation maps to finalize the client's virtual appearance. However, these existing techniques often struggle with low-resolution synthesis. This issue arises primarily due to the misalignment between the contorted clothing and the human figure during the image warping process. Techniques like the 'thin-plate spline' (TPS) transformation and ClothFlow exemplify these methods, though they suffer from high processing costs due to the pixel-level prediction of clothing movement.

Furthermore, the proposed methods employ a simple U-Net architecture for high-resolution image synthesis and frequently face problems like unbalanced training and poor-quality images. These issues become particularly evident when the systems attempt to reconstruct body regions that were originally obstructed by clothing [9].

The U-Net architecture, first published in the year 2015, has been a revolution in the field of deep learning. The architecture won the International Symposium on Biomedical Imaging (ISBI) cell tracking challenge of 2015 in numerous categories by a large margin. Some of their works include the segmentation of neuronal structures in electron microscopic stacks and transmitted light microscopy images. With this U-Net architecture, the segmentation of images of sizes 512X512 can be computed with a modern GPU within small amounts of time. There have been many variants and modifications of this architecture due to its phenomenal success. Some of them include LadderNet, U-Net with attention, the recurrent and residual convolutional U-Net (R2-UNet), and U-Net with residual blocks or blocks with dense connections.

Although U-Net is a significant accomplishment in the field of deep learning, it is equally essential to understand the previous methods that were employed for solving such similar tasks. One of the primary examples that comes to mind was the sliding window approach, which won the EM segmentation challenge at ISBI in the year 2012 by a large margin. The sliding window approach was able to generate a wide array of sample patches apart from the original training dataset [17].

This result was because it used the method of setting up the network of sliding window architecture by making the class label of each pixel as separate units by providing a local region (patch) around that pixel. Another achievement of this architecture was the fact that it could localize quite easily on any given training dataset for the respective tasks.

However, the sliding window approach suffered two main drawbacks that were countered by the U-Net architecture. Since each pixel was considered separately, the resulting patches overlapped a lot. Hence, a lot of overall redundancy was produced. Another limitation was that the overall training procedure was quite slow and consumed a lot of time and resources. The feasibility of the working of the network is questionable due to the following reasons [16].

The U-Net is an elegant architecture that solves most of the occurring issues. It uses the concept of fully convolutional networks for this approach. The intent of the U-Net is to capture both the features of the context as well as the localization. This process is completed successfully by the type of architecture built. The main idea of the implementation is to utilize successive contracting layers, which are immediately followed by the up sampling operators for achieving higher resolution outputs on the input images.

Additionally, we introduce a discriminator rejection mechanism designed to filter out incorrect segmentation map predictions, which can lead to unnatural final results. This rejection mechanism significantly enhances the performance of virtual try-on frameworks, making it a critical feature for real-world applications.

In summary, our contributions are as follows:

We propose an innovative architecture that simultaneously performs warping and segmentation map generation, inherently eliminating misalignment and naturally handling occlusions caused by body parts.

We adapt a discriminator rejection mechanism to improve the accuracy of segmentation map predictions

We achieve state-of-the-art performance on high-resolution datasets, setting a new benchmark for virtual try-on technology.

4. METHODOLOGY

The goal of Attire Fit-In is to generate a synthetic image of individual wearing target clothing, while preserving the person's original stance and body shape from the reference image and maintaining the features of the garment. Given a reference image of a person and an image of the garment, the following steps are involved:

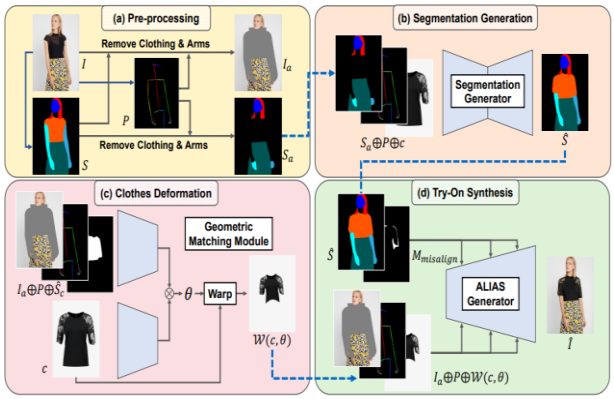

Fig 3. Overview of Virtual Try-On

1. Garment-Agnostic Person Representation: We create a garment-agnostic representation of the person by removing the clothing information from. This is achieved by utilizing both the pose map and the segmentation map of the individual.

2. Segmentation Map Generation: The simulation constructs a segmentation map based on the garment-agnostic person representation. This segmentation map serves as a blueprint for placing the new garment.

3. Garment Distortion: To approximately align the garment with the human body, we apply a distortion process to the garment image.

4. ALIAS Normalization: After distorting, we use the ALIgnment-Aware Segment (ALIAS) normalization method to remove any misleading data from misaligned areas. The ALIAS generator fills in these areas by preserving the garment texture and features, ensuring a seamless integration with the person's body.

Through these steps, Attire Fit-In successfully synthesizes an image where the individual appears naturally dressed in the target clothing, maintaining both the person's body shape and the garment's characteristics.

PRE-PROCESSING

To train the model for the virtual try-on task using pairs of garment images C and corresponding reference images I (where I depicts a person wearing C), we need to create a person representation that excludes the original clothing data in I. These representations must fulfill the following requirements:

- Remove the original garment to be replaced.

- Retain sufficient information to predict the person’s stance and body shape.

- Preserve areas that are not to be altered (such as the face, hands, etc.) while maintaining the individual's identity.

To address the challenge of accurately reproducing body parts, our approach utilizes a garment-agnostic image Ia and a garment-agnostic segmentation map Sa as key

stages. These components effectively remove the outline of the apparel while preserving the body parts that need to be replicated.

Our method begins by predicting a segmentation map S and a pose map P for the image I using pre-trained networks. The segmentation map S is used to eliminate the region corresponding to the garment to be changed, while keeping the rest of the image intact. Given the difficulty of replicating hands, the pose map P is employed to eliminate the arms without affecting the hands. Based on S and P, we create a cloth-independent image Ia and a cloth-independent segmentation map Sa. This process allows the model to remove the original clothing data and retain the rest of the image, ensuring an accurate and detailed representation of the individual [23].

SEGMENTATION GENERATOR

The segmentation generator is tasked with predicting the segmentation map ^S of an individual in the reference image, now dressed in the target garment C. This process relies on the garment-agnostic person representation (Sa, P) and the target garment image C. The primary objective is to train the segmentation generator Gs to transform the inputs (Sa, P, C) into the segmentation map ^S, effectively excluding any information about the original clothing item from the image.

For the architecture of Gs, we utilize the U-Net model, as introduced by Ronneberger et al. in 2015. U-Net is a refined architecture that addresses numerous challenges using a fully convolutional network approach. The fundamental concept involves employing a series of down-sampling layers followed by up-sampling layers to produce a high-resolution output from the input image.

The process begins with the input image being passed through the initial layers of the model. These layers comprise several convolutional layers activated by ReLU functions. The image size reduces due to unpadded convolutions, a sequence of convolution operations without zero-padding, leading to a smaller output.

On the left side of the U-Net, known as the encoder block, max-pooling layers are used to progressively reduce the image size. The encoder also includes multiple convolutional layers where the number of filters increases with each layer, capturing complex features at various abstraction levels. This detailed feature extraction is crucial for the subsequent reconstruction phase.

The right side of the U-Net, referred to as the decoder block, gradually reconstructs the image to its original size. In the decoder block, the number of filters decreases, and up-sampling layers are utilized to increase the resolution of the feature maps. The up-sampling operation involves layers that expand the spatial dimensions of the feature maps, counteracting the down-sampling performed by the encoder.

A distinctive feature of the U-Net architecture is the use of skip connections. These connections link the output of each layer in the encoder block directly to the corresponding layer in the decoder block. Skip connections help retain essential spatial information that might be lost during the down-sampling process, ensuring that the network can recover fine-grained details. This capability leads to more accurate and higher-quality segmentation results as shown in figure 4[15].

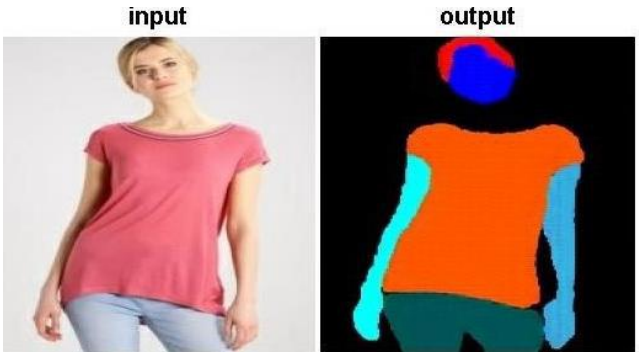

Fig 4. Generation of the segmentation map of image

In summary, the U-Net architecture processes an input image through multiple convolutional layers, down-samples and then up-samples it while preserving crucial spatial information through skip connections. This design enables the segmentation generator Gs to produce high-resolution segmentation maps effectively, ensuring that the final synthetic image accurately depicts the individual wearing the target garment C.

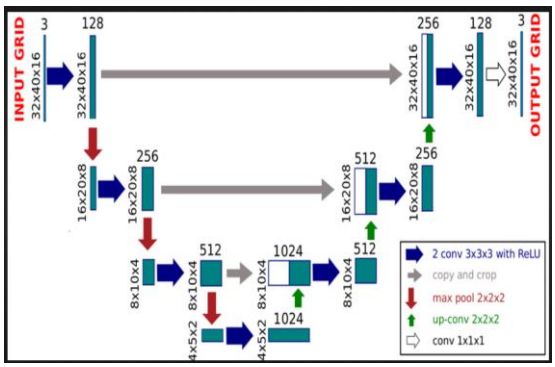

Fig 5. U-net architecture

CLOTHING IMAGE DEFORMATION

At this stage, the selected garment C is transformed to align with ^Sc, which represents the garment region in the segmentation map ^S. The clothing-agnostic person representation (Ia, P) and Sc are provided as inputs to the geometric matching module, as described in the CPVTON framework (Wang T. C., 2018). According to the cited paper, the geometric matching module (GMM) consists of four main components.

First, the GMM utilizes two feature extractors to extract significant characteristics from Ia and C. These feature extractors play a crucial role in identifying and isolating key attributes of the person and the garment, respectively. The features obtained from these extractors are then combined into a single tensor using a correlation layer. This tensor serves as the input for the regressor network, which processes the combined features.

The next component involves predicting the spatial transformation parameters, denoted as θ, using the regression network. The regression network uses the correlation matrix generated from the combined features to make these predictions. The transformation parameters θ are essential for determining how the garment C needs to be warped to fit the target segmentation map ^Sc.

Finally, the garment C is warped to produce c^ = Tθ(c) using the Thin Plate Spline (TPS) transformation module T. The TPS transformation ensures that the garment is deformed smoothly and accurately to match the target body shape and pose.

The process begins with generating a correlation matrix between the features obtained from (Ia, P) and C. This correlation matrix is then used by the regression network to predict the TPS transformation parameters. During the training phase, the model uses Sc derived from S instead of ^Sc. The training involves minimizing the L1 loss between the warped garment and the garment Ic recovered from Ic. Additionally, a quadratic difference constraint is applied to reduce visible distortions in the deformed clothing image caused by the transformation process.

By following these steps, the geometric matching module effectively transforms the target garment to align with the segmentation map, ensuring that the final output maintains a natural and realistic appearance. This method allows the virtual try-on system to produce high-quality images of the person wearing the target garment, accurately reflecting the intended fit and style.

Building on the outputs from the previous stages, the final synthetic image ^I is generated at this phase. The process integrates the garment-agnostic human representation (Ia, P) with the warped garment image W(c,θ) using the segmentation map S. Each layer of the generator receives Ia P, and W(c,θ) as inputs to ensure that all relevant information is utilized throughout the synthesis process.

To achieve this integration, we employ the Alignment-Aware Segment (ALIAS) normalization as a novel conditional normalization technique. Introduced by Wang in 2018, ALIAS normalization is designed to handle the unique challenges posed by virtual try-on systems. By using the segmentation map S and the corresponding mask of these regions, ALIAS normalization ensures that semantic information is preserved while simultaneously eliminating any misleading data that may arise from misaligned regions.

In practical terms, ALIAS normalization operates by adapting the normalization process based on the alignment of the input data. This approach allows the model to maintain the integrity of the garment's texture and details, even in areas where misalignment might otherwise cause distortions. The conditional normalization provided by ALIAS ensures that each part of the image receives appropriate adjustments, leading to a more coherent and realistic final output.

Throughout the generation process, the combination of Ia, P, and W(c,θ) is refined layer by layer. The ALIAS normalization method leverages the segmentation map to guide this refinement, ensuring that the synthesized image accurately reflects the intended appearance of the person wearing the target garment. This step is crucial for achieving high-quality results that preserve the individual's identity and the garment's characteristics.

By integrating these techniques, the final synthetic image ^I not only aligns well with the original body shape and pose of the individual but also maintains the aesthetic and functional attributes of the target garment. This comprehensive approach ensures that the virtual try-on system produces realistic and visually appealing results, enhancing the overall user experience.

ALIAS NORMALIZATION

ALIAS normalization (Choi, 2021) has two inputs: the synthetic segmentation map ˆS; the misalignment binary mask Misalign.

L(H x W) , which excludes the warped mask of the target clothing image W(Mc, θ) from Sˆc (Mc denotes the target clothing mask), i.e.,

Malign = Sˆc ∩ W(Mc, θ) (3)

M-misalign = Sˆc – Malig (4)

First, Malign and M-misalign from the formula (3) and formula (4) Obtained. The reconstructed edition of ˆS is, defined as ˆSdiv, and ˆSc is separated into ˆS to become Malign and M-misalign. Furthermore, the regions of M-misalign and the other regions in hi are standardized separately with the help of ALIAS normalization. This is followed by standardized activation modulation using the affine transformation parameters derived from ˆS div. ALIAS Normalisation is basically utilized for the removal of the misleading information in the misaligned regions

ALIAS GENERATOR

The ALIAS generator is structured using a series of residual blocks, each incorporating an up-sampling layer. Within each ALIAS residual block, three convolutional layers and three ALIAS normalization layers are utilized

The purpose of these blocks is to operate at different resolutions, necessitating the adjustment of inputs to the normalization layer, S and M-misalign, before integrating them into each layer. Similarly, the generator input (Ia, P, W(c,θ)) is scaled to various resolutions to ensure compatibility across layers

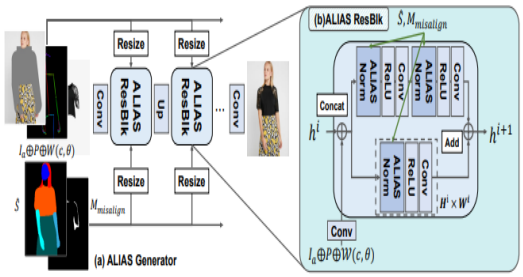

Fig 6. Flowchart of ALIAS generator

As the inputs (Ia, P, W(c,θ)) pass through the convolutional layer, they are resized and concatenated with the activation of the preceding layer before entering each residual block. This concatenation process allows each residual block to enhance the activations with the newly combined inputs, promoting more effective multi-scale functional level improvements. This approach helps retain garment details more effectively than improvements made solely at the single-pixel level.

The ALIAS generator's design is inspired by the principles of SPADE (Park, 2019) and pix2pixHD (Wang T. C., 2018). The training process for the ALIAS generators involves conditional adversarial losses, feature matching losses, and perceptual (or sensory) losses. Conditional adversarial losses help the generator create outputs that are indistinguishable from real images when conditioned on specific inputs. Feature matching losses ensure that the generated images maintain a high level of detail and realism by matching the feature distributions between the generated and real images. Perceptual losses further refine the image quality by ensuring that the high-level content and style of the generated images are consistent with the reference images.

By integrating these training techniques, the ALIAS generator is able to produce high-quality synthetic images that accurately reflect the intended appearance of the person wearing the target garment. The multi-scale approach and advanced normalization techniques ensure that the final images are not only realistic but also maintain the intricate details of the clothing and the person's features. This results in a robust virtual try-on system capable of delivering visually appealing and highly accurate representations.

5. Results and Discussion

Table 1: Accuracy Metrics

Model	Jaccard	F1	Recall	Precision	Accuracy
U-Net	**0.7392**	**0.8242**	**0.8689**	**0.8231**	**0.9313**
ResU-Net	**0.6256**	**0.7404**	**0.8021**	**0.7422**	**0.8963**

As shown in Table 1, Comparing UNet and ResU-Net based on the provided metrics reveals notable differences in their performance in image segmentation tasks. Beginning with the Jaccard score, a metric quantifying the overlap between predicted and ground truth segmentation masks, UNet achieves a substantially higher score of 0.73928 compared to ResU-Net's 0.62564. This suggests that UNet produces segmentation outputs that closely align with the true segmentation areas, indicating superior segmentation accuracy and efficacy in delineating object boundaries within images.

Moving on to the F1 score, which balances precision and recall, UNet again demonstrates its superiority with a score of 0.82429, surpassing ResU-Net's score of 0.74074. The higher F1 score of UNet implies a better trade-off between correctly identifying positive cases (precision) and capturing all positive cases (recall), reinforcing its efficacy in producing accurate and comprehensive segmentation results. Moreover, the recall score further highlights UNet's advantage, as it achieves a higher recall of 0.86891 compared to ResU-Net's 0.80218. This suggests that UNet is better at capturing a larger proportion of true positive cases, thus minimizing the risk of missing important segmentation details within images.

Precision, which measures the proportion of positive cases predicted by the model that are actually positive, also favors UNet with a score of 0.82315, slightly higher than ResU-Net's score of 0.74228. This indicates that UNet generates fewer false positives in its segmentation predictions, leading to more accurate delineation of object boundaries and reduced instances of misclassification. Finally, considering the accuracy score, which evaluates the overall correctness of the model's predictions, UNet outperforms ResU-Net with a score of 0.93131 compared to 0.89632. This reaffirms UNet's superiority in producing segmentation outputs that closely match the ground truth labels across all classes, indicating its robustness and reliability in various image segmentation applications.

In summary, the comparison of UNet and ResU-Net based on the provided metrics underscores UNet's dominance in image segmentation tasks, characterized by higher Jaccard, F1, recall, precision, and accuracy scores. These findings suggest that UNet exhibits superior performance in

accurately delineating object boundaries within images, minimizing false positives, and maximizing segmentation accuracy, thereby making it a compelling choice for applications requiring precise and reliable image segmentation.

In the figure 7, a model is posing gracefully, showcasing an elegant outfit. Beside her, there's an image of the same garment displayed separately. Using a virtual try-on method implemented with a U-Net architecture, the garment is perfectly fitted onto the model, highlighting a seamless and realistic integration. The virtual try-on process enhances the visual experience, making the cloth look as if it naturally drapes over her. The overall effect demonstrates the precision and effectiveness of U-Net in virtual clothing applications.

Fig 7. Cloth fitting generated image

In the Figure 8, a model strikes a complex and dynamic pose, adding an element of movement to the scene. Next to her is an image of a garment displayed independently. Using a virtual try-on method with a U-Net architecture, the garment is flawlessly fitted onto the model, perfectly adapting to her intricate posture. The virtual try-on seamlessly merges the cloth with her pose, demonstrating remarkable precision and realism. This showcases the advanced capabilities of U-Net in handling complex poses for virtual clothing applications.

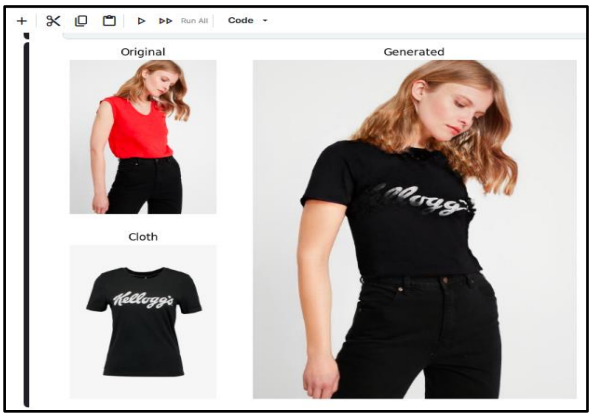

Fig 8. Perfect fitting in complex posture

In the figure 9, a model poses with one arm extended and the other folded, creating a dynamic and engaging stance. Next to her, there is an image of a garment displayed separately. Using a virtual try-on method with a U-Net architecture, the garment fits her perfectly, conforming seamlessly to her unique pose. The virtual try-on method ensures the cloth drapes naturally over her body, adjusting to her extended and folded arms. This highlights the effectiveness of U-Net in achieving a realistic and precise fit for virtual clothing applications.

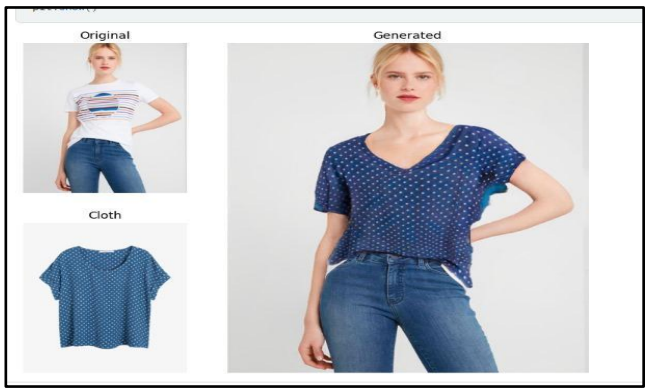

Fig 9. Perfect Cloth Fitting for a model posing with one arm extended and another folded

In the figure 10, a model is posing confidently, showcasing her stance. Beside her, there is an image of a garment displayed separately. Using a virtual try-on method with U-Net architecture, the garment is flawlessly fitted onto her, aligning perfectly with her pose. The virtual try-on process ensures the cloth looks natural and well-draped on her figure. This demonstrates the precision and realism achieved by the U-Net in virtual clothing applications

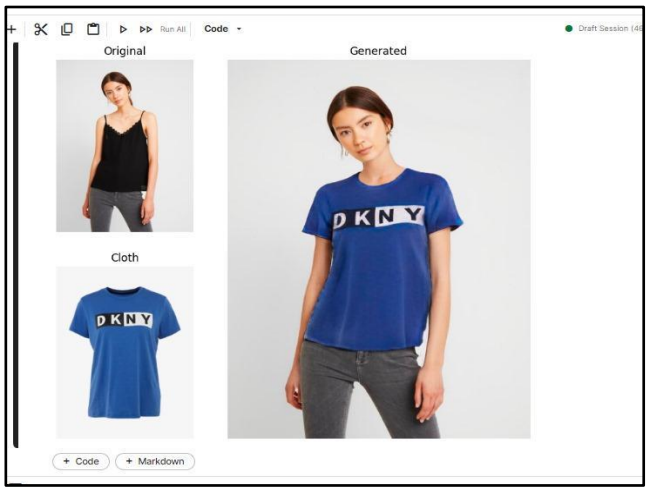

Fig 10. Perfect Cloth fitting generated image

6. CONCLUSION

As we have approached the end of the work, we can successfully wrap the images of clothing garments onto the target person's body by deforming it following the client's body shape and posture, thereby aiding the customer to visualize the attire's fitting virtually.

To synthesize the photo-realistically, we have used the Attire Fit-In for the virtual try-on. The ALIAS generator

can normalize and process staggered areas, propagate semantic information all through the ALIAS generator, and retain clothing facts through multiscale improvements.

In this Model of diffusion models for authentic virtual try-on, particularly in the wild scenario. We incorporate two separate modules to encode the garment image, i.e., visual encoder and parallel UNet, which effectively encode high-level semantics and low-level features to the base UNet, respectively. In order to improve the virtual try-on on real world scenarios, we propose to customize our model by fine-tuning the decoder layers of UNet given a pair of garment-person images.

Potential negative impact, this work introduces a method that enhances the performance of virtual try-on using generative diffusion models. The virtual try-on technology comes with benefits and pitfalls - the tool could be helpful for users to effectively visualize their look with a given garment. As we have shown in results our work is supporting multiple poses with most clothes in the dataset proving the fitting accuracy.

References

[1] Lewis KM, Varadharajan S, Kemelmacher-Shlizerman I, Tryon-gan: body-aware try-on via layered interpolation. ACM Trans Graph.2021

[2] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, "CP-VTON+: Clothing shape and texture preserving image-based virtual try-on," inProc. IEEE/CVF Conf. Compute. Vis. Pattern Recognit. Workshops, Jun. 2021

[3] Gong K, Liang X, Zhang D, Shen X, Lin L Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition,2021

[4] Choi S, Park S, Lee M, Choo J Viton-hd: high-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2021.

[5] Ghodhbani H, Neji M, Razzak I, Alimi AM .You can try without visiting: a comprehensive survey on virtually try-on outfits.,2022

[6] B. Akinkunmi and P. C. Bassey. A Qualitative Approach for Spatial Qualification Logic. International Journal of Artificial Intelligence & Applications, 2017.

[7] Jetchev N, Bergmann U, The conditional analogy gan: swapping fashion articles on people images. In: Proceedings of the IEEE international conference on computer vision workshops,2022.

[8] Sarkar K, Golyanik V, Liu L, Theobalt C ,Style and pose control for image synthesis of humans from a single monocular view,2021.

[9] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in Proc. Eur. Conf. Comput. Vis., Cham, Switzerland, 2020..

[10] Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2021).

[11] Han,X., Hu,X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proc. of the IEEE international conference on computer vision (ICCV). pp. 10471–10480 (2019).

[12] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proc. the Advances in Neural Information Processing Systems (NeurIPS) (2017).

[13] Issenhuth, T., Mary, J., Calauz` enes, C.: Do not mask what you do not need to mask: a parser-free virtual try-on. In: European Conference on Computer Vision. pp. 619–635. Springer (2020).

[14] Jandial, S., Chopra, A., Ayush, K., Hemani, M., Krishnamurthy, B., Halwai, A.: Sievenet: A unified framework for robust image-based virtual try-on. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2182–2190 (2020).

[15] Li, K. Chong, M.J., Zhang, J., Liu, J.: Toward accurate and realistic outfits visualization with attention to details. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15546–15555 (2021)

[16] Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017).

[17] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017).

[18] Minar, M.R., Ahn, H.: Cloth-vton: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. In: Proceedings of the Asian Conference on Computer Vision (2020).

[19] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning

Representations (2018).

[20] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[21] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp.

[22] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21.

[23] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1(2), 3 (2022)

[24] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[25] Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241, Springer (2018).