Neural Style Transfer for Image-Based Garment Interchange Through Multi-Person Human Views

Hajer Ghodhbani[1][a], Mohamed Neji[1,2][b] and Adel M. Alimi[1,3][c]

[1]*Research Groups in Intelligent Machines (REGIM Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia*

[2]*National School of Electronics and Telecommunications of Sfax Technopark, BP 1163, CP 3018 Sfax, Tunisia*

[3]*Department of Electrical and Electronic Engineering Science, Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa*

Keywords: Style Transfer, Pose Control, Segmentation, Garment Interchange.

Abstract: The generation of photorealistic images of human appearances under the guidance of body pose enables a wide range of applications, including virtual fitting and style synthesis. Several advances have been made in this direction using image-based deep learning generation approaches. The issue with these methods is that they produce significant aberrations in the final output, such as blurring of fine details and texture alterations. Our work falls within this objective by proposing a system able to realize the garment transfer between different views of person by overcoming these issues. To realize this objective, fundamental steps were achieved. Firstly, we used a conditioning adversarial network to deal with pose and appearance separately, create a human shape image with precise control over pose, and align target garment with appropriate body parts in the human image. As a second step, we introduced a neural approach for style transfer that can differentiate and merge content and style of editing images. We designed architecture with distinct levels to ensure the style transfer while preserving the quality of original texture in the generated results.

1 INTRODUCTION

The fashion industry is now acting to improve the world of fashion for everyone. As more and more digital technologies become available to fashion enterprises, they become in the age of digital transformation. The need for industrial adaptation is brought on by shifts in consumer demands. Fashion firms must pay attention to their customers' needs and respond with digital solutions. The transformation of the fashion business and the switch from offline to online shopping has been accelerated by the Covid-19 pandemic-related lockdowns when everything was altered for many industries such as fashion industry, which is experiencing difficulties resulted in sales reduction and a change in consumer behavior. When we talk about fashion, one of key offline experiences missed by the on-line consumers is the fitting room where a clothes item can be tried-on.

Virtual try-on solutions have recently been the subject of extensive research in an effort to lower the cost of returns for online shops and provide customers with the same offline experience. Such system could improve the shopping experience by assisting users to make purchase decisions. As an overview of try-on solution, a complete study about virtual fitting system is exposed in our survey (Ghodhbani et al., 2022a), and according to it, we are focused on the most practical solution that is the image-based system allowing garment interchange across images of different persons. The garment image is mapped onto the unique body using an image warping approach. While for now, this kind of system is not mature enough, its results are unrealistic for such proposed systems which cannot produce fine details or allow viewing of textured images from varied angles.

Our solution comes up with the support for personalized clothing transfer to address these

[a] https://orcid.org/0000-0003-1100-0711
[b] https://orcid.org/0000-0003-3178-2116
[c] https://orcid.org/0000-0002-0642-3384

Ghodhbani, H., Neji, M. and Alimi, A.
Neural Style Transfer for Image-Based Garment Interchange Through Multi-Person Human Views.
DOI: 10.5220/0011694200003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 327-335
ISBN: 978-989-758-634-7; ISSN: 2184-4321

problems, and it is developed firstly, in our work called dress-up (Ghodhbani et al., 2022b) that tries to match a person's appearance to another person's image. In this paper, we continued with this challenge of aligning the clothing item with the matching body parts in the person image by exploiting other methods to achieve the system's main goals. The difficulty comes from the fact that the target body and the garment item are typically not spatially aligned.

Our system uses a semantic segmentation-based technique to address this problem, and we suggest a style-based appearance approach to transfer the garment during virtual try-on. Our model is robust to significant misalignments between human and garment photos thanks to this semantic segmentation strategy which makes it more suitable by considering a full-body image in a variety of poses during garment transfer. Thus, our contributions are as follow:

- Interchanging garment across images while preserving the visual quality.
- Analysis of Pose-dependent control with high texture generation.
- Presented the result of stylized images and the semantic maps in correspondent
- Showing the ability of the intermediate segmentation and the style transfer to separate between texture and content in the image.

The structure of this paper is as follows: In Section II, we present the related work to our study. Section III exposed our framework with its different modules. In section IV and section V, the results and comparison are displayed respectively. Finally, section VI presented the conclusion.

2 RELATED WORK

2.1 Deep Generative Models

In recent years, Generative Adversarial Networks (GAN) have been successfully applied in image generation. A GAN model architecture involves two components: generator and discriminator. Until the image produced by the generator is convincing enough to fool the discriminator, the two components are trained together. GAN architecture was proposed at the first in 2014 (Makhzani et al., 2014), and as a first proposal, it presented many limits because only it was able to synthesize low-resolution images. Then, other version is appeared (Zhang et al., 2019) to improve the quality of generated result but it cannot differentiate different attributes in images and so it has little control over image synthesis. To overcome this problem, StyleGANs (Karras et al. et al., 2019) method was proposed and the idea is to use an intermediate latent space, which is fed into the generator to control different levels of attributes.

Conditional GAN (cGAN) is a type of GAN, which that gives the generator and discriminator conditional information. The cGAN can be used for different applications such as image-to-image translation (Isola et al, 2017; Wang et al., 2018; Park et al., 2019) which learn how input images can be mapped to output ones and synthesize style in the final output using segmentation masks. This task becomes more complicated when synthesizing the full human appearance with control of body pose and human appearance due to their several changes. Our work aims to resolve this challenge by proposing a method based on cGAN to synthesize realistic images of a full human body with control over poses and appearance.

2.2 Neural Style Transfer (NST)

NST is a technique for combining two images to create a new one by mimicking the style of a different image, often known as style image. Thus, NST refers to the task of images manipulation to adopt the appearance style of another image, and compose images in the style of another one. The foundation of NST is that the style and the content representations can be kept distinct. This technology has emerged in computer vision task by its ability to transfer an artistic style to the content image which can be automatically redrawn with a particular artistic style.

Gatys et al. (Gatys et al., 2016) discovered through the visualization and the analysis of Convolutional Neural Network (CNN) that the content and style representation of images could be easily separated from various layers basing on the specific capacity of CNN to extract features of various scales in different layers. They created a style transfer technique using VGG network and achieved remarkable artistic effects. Since that time, NST has evolved into the primary route for image stylization, garnering interest of many researchers. Thus, for preserving original semantic information, several studies have been developed to improve style transfer (Li et al., 2019; Yao et al., 2019; Wang et al., 2020).

Neural network models are capable to deal with style transfer due to the role of deeper layers to extract more general features. With convolutional neural networks, the lower layers extract very local features, such as edges, corners, and colors which are combined together in deeper layers to depict more global features, such as shapes, faces, etc. Our work focus on integrating this task to realize virtual try-on task and preserving the quality of original texture.

3 MULTI PERSON STYLE TRANSFER

3.1 Overview

A challenging image processing task is to produce a content of an image in various styles. The lack of image representations that directly reflect semantic information and enable the separation of visual content from style may have been a significant limiting factor for earlier methods. Thus, in this work, we used a style neural model to edit content and style of person images, separately, and recombine them as we desire. Our findings show how we can learn deep image representations and how they may be used for sophisticated image generation and manipulation.

The texture transfer issue can be seen on the challenge to separate content from style in images and transferring it from one image to another while preserving semantic information of target one. Recent research have led to the development of neural style transfer systems (Luan et al., 2019; Liu et al., 2021; Cheng et al., 2019) showing remarkable visual quality and artistic picture result by extracting detailed semantic information. However, there is no attention has been paid to multi-fashion image transfer which refers to interchange multiple styles between different images. Our system aims to resolve this issue and attempts to produce realistic images with complete control over pose, shape, and appearance.

3.2 Proposed System

In this section, we present our proposed style transfer system with conditioning pose preservation. Thus the architecture is shown in Figure 1. Given two images of a person, our goal is to synthesize a new image of the person in target body pose wearing the clothes of the second one. Firstly, we extract the pose P and the appearance A from I_B and I_S respectively. Secondly, we encode pose and appearance to obtain target segmentation and reconstruct it. Our method is based on existing methods such as (Ghodhbani et al., 2022b; Gatys et al., 2016; Raj et al., 2018), and it is a continuity of our Dress-up system (Ghodhbani et al., 2022b). In the current work, the main contribution resides on the integration of style transfer network at final phase and the use of intermediate shape to generate final result. Thus, a learning-based method for human image synthesis is proposed.

This approach introduced the application of artistic style transfer model for generation of fashion image. Its fundamental component consisting in allowing control over body pose and appearance from a various views. To separate conditioning of two aspects in several modalities, our strategy disentangles the pose, appearance and body parts. In the following sections, each part of architecture will be described in details.

3.2.1 Pose and Style Encoding

We used existing methods to detect the human pose P from the image I_B (Liang et al., 2018; Gong et al., 2019), and the style S from the image I_s (Omran et al., 2018; Lassner et al., 2017). Then, these representations are fed on the pose encoder and style encoder, respectively. We obtain two representations from this process, one correspondent to the pose features pose and the other for the style features. These representations are then concatenated to create a clothing segmentation of I_s that rigorously adheres to the body shape and pose from I_B. We frame this issue as a conditioned generating process, where the clothing being conditioned on clothing segmentation and the body is conditioned on body segmentation.

To solve the dual conditioning issue, we use a dual path network (Ronneberger et al., 2015) composed of two encoders, one for the body and the other for the clothes, and a decoder combining the two encoded representations to produce the output image. The clothes encoder generates a feature map, where each channel holds the probability map of one clothing type. The body encoder creates a feature map to represent the target body from a color-coded channel body segmentation. The residual blocks are combined and then applied to these encoded feature maps. The desired garment segmentation is created by upsampling the obtained feature map.

The body segmentation has a large influence over the resulting image, while the style segmentation has a weaker influence. The generated representation aiming to capture style information while restricting the generated segmentation to be close to target pose and consisting with the desired pose. In this level, we can generate a clothing segmentation in target body to perform desired shape change. The advantage of this intermediate representation is that the body and style segments do not need to be extremely clean for our framework to perform. Existing human parsing and body parsing models were used to create our segmentations, however their predictions are frequently inaccurate and noisy. But the noise in these

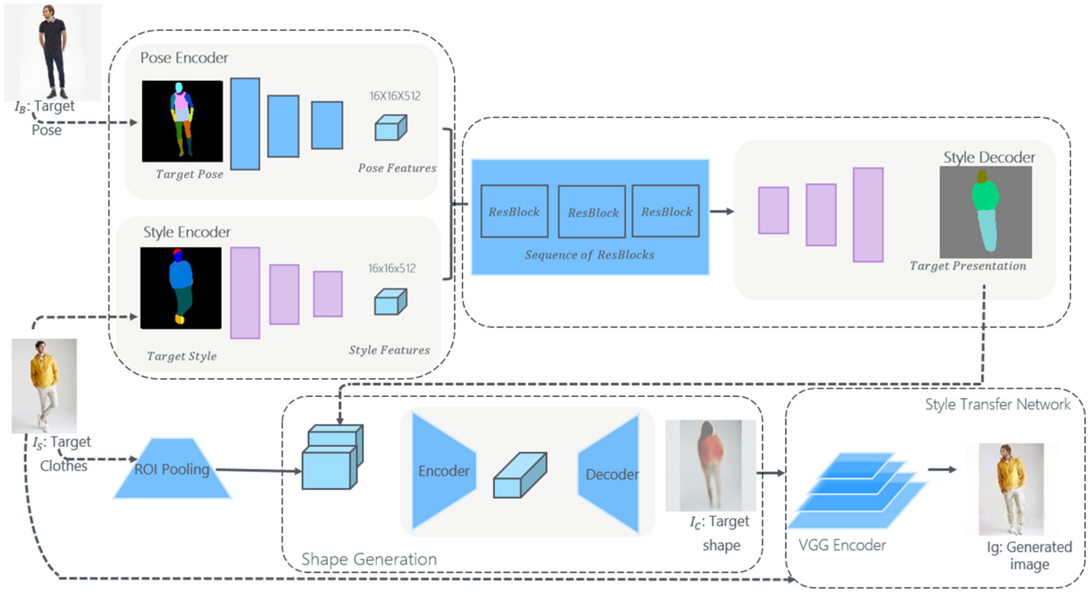

Figure 1: Our pipeline: from the pose and style encoding to the shape generation and finally style transfer.

intermediary representations can be made up for by our network. Some results from this phase are presented Figure 2.

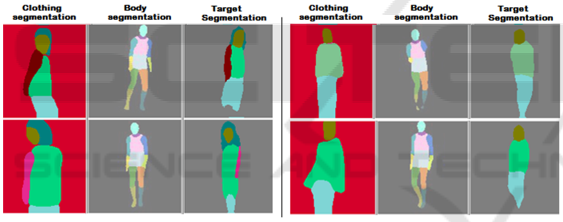

Figure 2: Results of clothing segmentation generation.

3.2.2 Shape Encoding

Once the segmentation representation was obtained, it was fed into a U-Net architecture that had been trained to provide form details given the style segmentation at the desired body shape and pose, and an embedding of the intended clothes depicted in image I_B . Then, a feature maps are created and upsampled to the original picture size, by ROI pooling on each body parts of I_B. Before submitting these feature maps to the U-Net, we stack them with the generated style segmentation. In this part of architecture, we aim to exploit the style information from desired style image I_S to synthesize the target shape I_C as shown in Figure 1. The obtained results from this phase are presented in the Figure 3.

3.2.3 Image Generation with a Style-Based Generator

Our style transfer network takes a content image I_C and a style image I_S as inputs, and synthesizes an output image I_g that recombines the content of the image and the style latter. We adopt a simple pre-trained model called VGG-19 (Simonyan et al., 2014) which is employed as a feed-forward encoder to extract features of the input pairs. Thus, we proceed to the texture transfer task using this network after presenting the desired shape I_c. VGG19 network was frequently employed in style transfer implementation to get the results as near as feasible to realistic image.

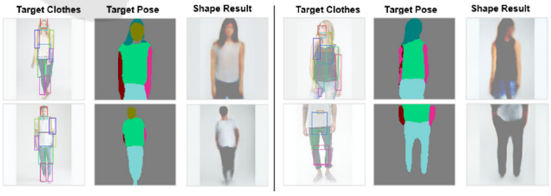

Figure 3: Result of shape generation using ROI Pooling.

To transfer the style of one image to another, we produce a new image that is consistent to both the content representation and the style representation. As a result, we obtained a new image of a person from a two single view images by realizing the transfer of garment textures from an image (style image) to another (content image), these results are exposed in the following Figure 4 (Further results in Figure 6).

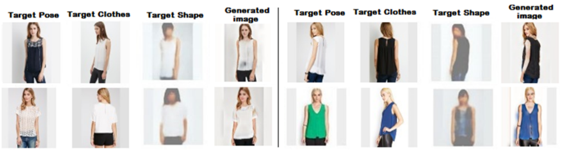

Figure 4: Demonstration of style transfer between multiple views using of shape generation as intermediate phase.

The adopted approach in this style encoding phase relies on a slow optimization procedure that updates the image iteratively in order to reduce both the content loss and the style loss computed by a loss network. They made use of the feature space offered by a normalized version of the 16 convolutional and 5 pooling layers of the 19-layer VGG network. The model is normalized by adjusting the weights in order that the average activation of convolutional filter is equal to one for. Since the VGG network only has rectifying linear activation functions and neither normalization nor pooling over feature maps, this rescaling can be performed on it without affecting the output. The images displayed were created using average pooling since we discovered that doing so for image synthesis produces slightly more pleasing results than doing for maximum pooling.

To transfer the style of one image to another, we produce a new image that is consistent to both the content representation and the style representation (Figure1). As a result, we jointly reduce the distance between the feature representations of a white noise image and the Convolutional Neural Network layers defining the style representation of the style and the content representation of the photo, respectively.

The loss function we try to reduce is:

$$L_{total}\,(\vec{p}, \vec{a}, \vec{x}) = \alpha\, L_{content}(\vec{p}\,, \vec{x}) + \beta\, L_{style}(\vec{a}, \vec{x}) \qquad (1)$$

where α and β are the weighting factors for content and style reconstruction, respectively.

As mentioned in our based work (Gatys et al., 2016), at each stage of processing, the CNN represents a particular input image as a collection of filtered images (Figure 5). The overall number of unit's per-layer of the network decreases as the number of distinct filters grows along the processing hierarchy and the size of filtered images is decreased by downsampling method (e.g. max-pooling). This representation shows two kinds of reconstructions:

Content Reconstructions. The information can be seen at multiple CNN processing stages by reconstructing input image using only the network's responses in a certain layer. The layers "conv1 2" (a), "conv2 2" (b), "conv3 2" (c), "conv4 2" (d), and "conv5 2" (e) of the original VGG network are used to rebuild the input image. The rebuilding using the lower layers is practically perfect (a-c). Higher levels of the network lose detailed pixel information while maintaining the images high-level content (d, e).

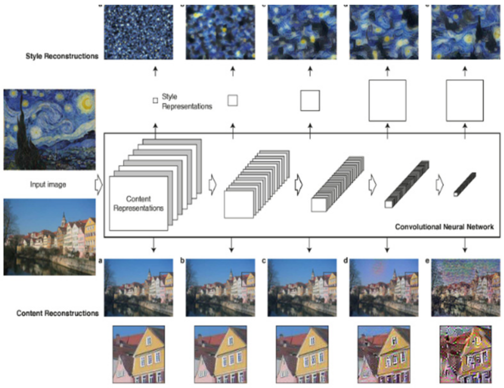

Figure 5: Image representations in CNN (Gatys, 2016).

Style Reconstructions. A feature space is employed to capture the texture data of an input image. The style representation determines correlations between the various features in the various CNN layers. We reconstruct the style of the input picture using a style representation created using various CNN layer subsets, including "conv1 1" (a), "conv1 1" and "conv2 1" (b), "conv1 1" and "conv3 1" (c), "conv1 1" and "conv2 1" and "conv3 1" and "conv4 1" (d), and "conv1 1" and "conv2 1" and "conv3 (e). Thus, images that are increasingly similar to a particular image's style are produced while the details of the scene's overall composition are discarded.

4 RESULTS

In this section, we present the generated results on DeepFashion datasets (Liu et al., 2016), from all the components of our architecture. Different results are presented in Figure 6 where we are showing the link between each level of the whole process. The key finding of this work is that the representations of the content (body of the target person) and the style (target clothes) are well separable. That is, we can manipulate both representations independently to produce new meaningful images. To demonstrate this finding, we generate images that mix the content and style representation from two different source images. The texture information is obtained using ROI pooling for different body parts (upper clothes, left arm, right arm, left leg, right leg, face, hair, etc.).

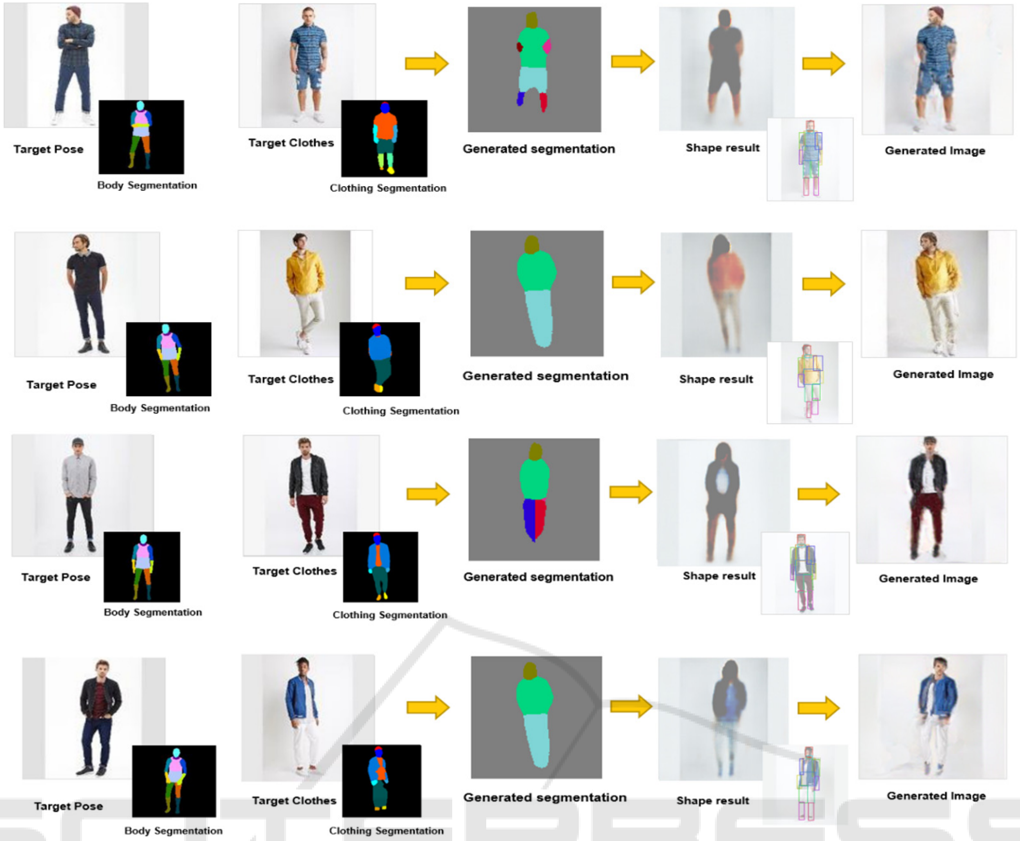

Figure 6: Obtained results from the whole process of our pipeline.

All the steps described previously are presented in Figure 6 such as pose and style encoding, shape encoding and texture generation. In each task, we demonstrate a high quality results that conducted to realize the correspondence of the texture to the body shape successfully. Thus, our process that using intermediate segmentation to generate target content, improves the visual quality and matching robustness.

5 COMPARISON

To evaluate the effectiveness of our model, we conduct a comparison with existing methods for the two main tasks achieved by our work: pose transfer and garment transfer, and with both qualitative and quantitative comparisons.

5.1 Qualitative Comparison

The visual comparisons of style transfer methods with pose control are shown in Figure 7. We compared the results of our work with four state-of-the-art pose transfer: PG2 (Ma et al., 2017), DPIG (Ma et al., 2018), Def-GAN and (Siarohin et al., 2018) PATN (Zhu et al., 2019).

As we can see, our method produced more realistic results. The facial identity is better preserved and even detailed textures of clothes and body are successfully synthesized. Another comparison of garment transfer task with DiOr method (Cui et al., 2021) is presented in Figure 8 which shows the performance of our approach to transfer all the clothes items from a person to another while preserving original texture.

5.2 Quantitative Comparison

In Table 1, we show the quantitative comparison with various metrics such as Inception Score (IS) (Salimanis et al., 2016), Structural Similarity (SSIM) (Wang et al., 2004) and Detection Score (Siarohin et al., 2018). IS and SSIM are two most commonly-used evaluation metrics in the person image synthesis task, which were firstly used in PG2 (Ma et al, 2017). Later, Siarohin et al. (Siarohin et al, 2018) introduced Detection Score (DS) to measure whether the person can be detected in the image. The results show that

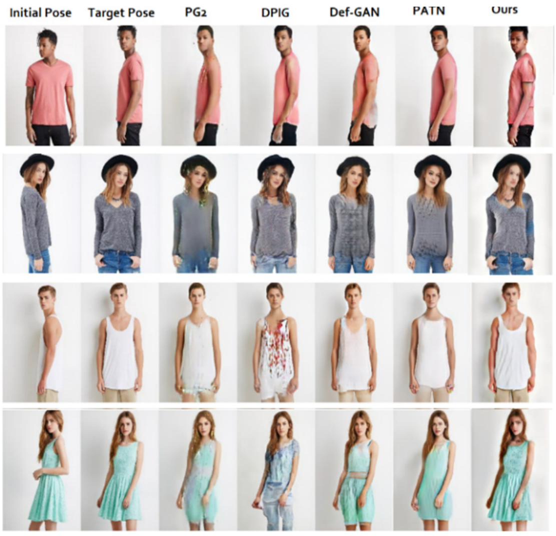

Figure 7: Qualitative comparison of Pose transfer task.

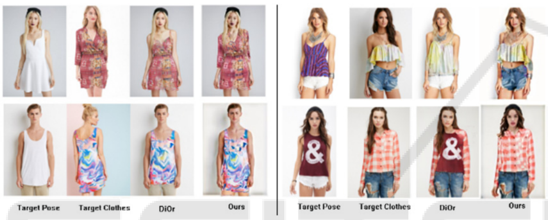

Figure 8: Qualitative comparison of Garment transfer task.

our method generates more realistic details with the highest IS value, and more detailed textures consisting to the source image and target image.

Table 1: Quantitative comparison with state-of-the-art methods on DeepFashion dataset.

Model	IS↑	SSIM↑	DS↑
PG2	3.202	0.773	0.943
DPIG	3.323	0.745	0.969
Def-GAN	3.265	0.770	0.973
PATN	3.209	0.773	0.976
Ours	**3.367**	**0.773**	**0.986**

Our method has the highest confidence for person detection with the best DS value. For SSIM that relies on global covariance and means of the images to assess the structure similarity, we see that the scores of all methods are clustered around similar values. This metric predicts that the generations are very close to ground truth. We adopted the pose transfer task to be able to compare on a subset of data for which we have paired information.

6 CONCLUSION

The development of neural style transfer technology has allowed people to significantly speed up the process of various fields such as fashion design. In this work, we proposed a framework, which successfully interchange garment across fashion images by using the tasks of image-to-image translation and style transfer. The generated images preserved the content of the input image while altering its style according to a reference appearance.

In this work, we demonstrated how to use intermediate feature representations generated from cGAN to transfer image style between arbitrary views. The objective from the implementation of this phase is to obtain a human shape in the target pose, then we proceed to apply the style transfer task by using simple style transfer network. The effectiveness of our method is presented in the previous section, and we presented further results in Figure 9.

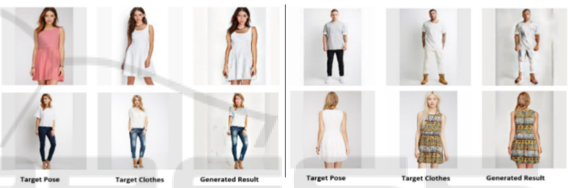

Figure 9: Further results for style transfer.

In our work, we realized garment transfer system by merging two essential tasks: the conditioning segmentation and the style transfer. After generating image-based fashion representations from fashion data, we used the image segmentation and style transfer method to finish the fashion style transfer process by transferring the appropriate texture. The use of adjustable aspects for style transfer can be taken into consideration in the future using style transfer strategies. In practical, a stylized works need to be adjusted according to various criteria, explore more techniques, and incorporate them into the style system to generate images with fine details and develop an effective image stylization system.

REFERENCES

Ghodhbani, H., Neji, M., Razzak, I., & Alimi, A. M. (2022a). You can try without visiting: a comprehensive survey on virtually try-on outfits. *Multimedia Tools and Applications*, 1-32.

Ghodhbani, H., Neji, M., Qahtani, A. M., Almutiry, O., Dhahri, H., & Alimi, A. M. (2022b). Dress-up: deep neural framework for image-based human appearance transfer. *Multimedia Tools and Applications*, 1-28.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644. (2015)

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354-7363). PMLR.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).

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 *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8798-8807).

Wang, Z., Zhao, L., Lin, S., Mo, Q., Zhang, H., Xing, W., & Lu, D. (2020). GLStyleNet: exquisite style transfer combining global and local pyramid features. *IET Computer Vision*, *14*(8), 575-586.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2337-2346).

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414-2423).

Li, X., Liu, S., Kautz, J., & Yang, M. H. (2019). Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3809-3817).

Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y. J., & Wang, J. (2019). Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1467-1475).

Luan, F., Paris, S., Shechtman, E., & Bala, K. (2017). Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4990-4998).

Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., & Ding, E. (2021). Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6649-6658).

Cheng, M. M., Liu, X. C., Wang, J., Lu, S. P., Lai, Y. K., & Rosin, P. L. (2019). Structure-preserving neural style transfer. *IEEE Transactions on Image Processing*, *29*, 909-920.

Raj, A., Sangkloy, P., Chang, H., Lu, J., Ceylan, D., & Hays, J. (2018). Swapnet: Garment transfer in single view images. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 666-682).

Liang, X., Gong, K., Shen, X., & Lin, L. (2018). Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, *41*(4), 871-885.

Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., & Lin, L. (2019). Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7450-7459).

Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., & Schiele, B. (2018, September). Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)* (pp. 484-494). IEEE.

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M. J., & Gehler, P. V. (2017). Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6050-6059).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, *29*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600-612.

Siarohin, A., Sangineto, E., Lathuiliere, S., & Sebe, N. (2018). Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3408-3416).

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose guided person image generation. *Advances in neural information processing systems*, *30*.

Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., & Fritz, M. (2018). Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 99-108).

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2347-2356).

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1096-1104).

Cui, A., McKee, D., & Lazebnik, S. (2021). Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14638-14647).