electronics

MDPI

Article

Readily Design and Try-On Garments by Manipulating Segmentation Images

Yoojin Jeong and Chae-Bong Sohn *

Department of Electronics and Communications Engineering, Kwangwoon University, Seoul 01897, Korea; yoojin2115@kw.ac.kr

* Correspondence: cbsohn@kw.ac.kr

Received: 28 August 2020; Accepted: 17 September 2020; Published: 22 September 2020

Abstract: Recently, fashion industries have introduced artificial intelligence to provide new services, and research to combine fashion design and artificial intelligence has been continuously conducted. Among them, generative adversarial networks that synthesize realistic-looking images have been widely applied in the fashion industry. In this paper, a new apparel image is created using a generative model that can apply a new style to a desired area in a segmented image. It also creates a new fashion image by manipulating the segmentation image. Thus, interactive fashion image manipulation, which enables users to edit images by controlling segmentation images, is possible. This allows people to try new styles without the pain of inconvenient travel or changing clothes. Furthermore, they can easily determine which color and pattern suits the clothes they wear more, or whether the clothes other people wear match their clothes. Therefore, user-centered fashion design is possible. It is useful for virtually trying on or recommending clothes.

Keywords: deep learning; generative adversarial networks (GAN); cloth image generation; fashion design

1. Introduction

Fashion research using artificial intelligence can be divided into four main categories: Detection, recommendation, analysis, and synthesis. Detection is the most basic study, recognizing where the clothing region is. It is possible to search for a picture of the clothes you are looking for and find the clothes or clothes of similar styles. When garments are difficult to express in words, an image-based search is useful for finding similar or identical items. Fashion recommendation is the problem of understanding clothing and learning suitability between other fashion items. It includes recommending outfits that reflect the user's taste and suggesting goods that fit the current style. Fashion analysis is a study that analyzes the characteristics of outfits, the latest trends, and people's styles. It has potential in the fashion industry, primarily in marketing fields. Lastly, fashion synthesis involves creating an image that reflects style changes and pose changes. Especially in the field of fashion design, research based on a generative model such as generative adversarial networks (GAN) [1] is actively being conducted. There are various studies applying such generation models, such as a model for making clothes when given the desired text explanation [2], a model for making clothes when people and clothes are given [3], a model for applying clothes worn by others [4], and a model for making clothes with a given pattern [5].

In this work, we introduce a novel approach that can transform clothes into the desired style and shape. In our approach, a person (P) can try-on clothes worn by others (A, B). There are some differences from other virtual try-on networks. First, we can manipulate the cloth pattern or shape as we want. For example, P can wear not only A's clothing but also A's t-shirt and B's pants. P can change

the shape of the clothing such as sleeve length and neckline. Further, it is possible to change different kinds of clothes such as pants to skirt. Second, we can generate a model image in a different pose.

The main research contributions of this work are: (a) Collect additional images of people wearing colorful or patterned clothing for better results. (b) Semantic region-adaptive normalization (SEAN) [6] among a style transfer GAN was modified and applied in order to transform clothes into the desired styles. Style transfer is a change from the current style to another style, and in the paper, it converts the desired area from the semantic segmentation map to other styles. (c) Add an image correction step to make the generated image look more realistic.

2. Related Works

2.1. Fashion Parsing

Image analysis can be largely divided into three categories: Classification, object detection, and segmentation. Classification finds classes such as 'shirt', 'pants', and 'dress' in the image. Object recognition indicates the position of an object within the image. Unlike classification or recognition, segmentation is the division of all objects in the image into semantic units (Figure 1). The goal is to create a segmentation map by classifying all pixels into a specified class. For example, the pixels related to the pants are colored green, and the pixels related to the dress are colored blue. It is a high-level task that requires a complete understanding of the image because the class of each pixel must be predicted. A typical model is fully convolutional networks (FCNs) [7]. It uses convolutional layers instead of fully connected layers in a general classification convolutional neural network (CNN). Therefore, after the pixel's class is predicted while retaining the pixel's information, the reduced image size can be restored through up-sampling.

Figure 1. Image analysis applied to apparel. (**a**) Classification, (**b**) object recognition, and (**c**) segmentation.

Fashion parsing uses segmentation to classify apparel such as tops, pants, and dresses to obtain the desired information. Graphonomy [8] showed good fashion parsing performance by introducing a hierarchical graph. For example, the head area can be divided into hat, face, and hair. The head contains the face, and the face is next to the hair. Using a unique hierarchical relationship, all human parsing can be done from different domains or labels of different levels in one model. In this paper, Graphonomy was used in the preprocessing process to obtain more accurate segmentation maps than other segmentation networks.

2.2. Segmentation Image to Real Image

GAN cannot generate data in the desired direction. To improve this, a conditional GAN was developed to generate data based on additional information y. In other words, condition y is added to G and D of GAN so that y can generate data in the desired way [9].

$$min_G max_D V(D,G) = E_{x\sim P_{data}(x)}\left[logD(x|y)\right] + E_{z\sim P_z(z)}\left[\log\left(1 - D\left(G(z|y)\right)\right)\right] \quad (1)$$

The condition y can be a label, text, or image. If the label is a condition, images of the class are created. For example, it creates a pants image when the label is pants or a T-shirt image when it is a

T-shirt. If the text is a condition, an image corresponding to the text is created. For example, given 'blue pants' as text, an image of blue pants is created. When an image is given as a condition, it is also called image-to-image translation. Image-to-image translation is learning the relationship between the input image and the result image. One of the image-to-image translation methods, Pix2Pix [10], converts an image to another style image. Using this, the segmentation images can be synthesized into actual images. The Pix2PixHD [11] is capable of high-definition image synthesis following the Pix2Pix. GauGAN [12] proposed by Nvidia produces a realistic image within seconds by modifying the segmentation image with a specified label. It can also be synthesized into any style. Figure 2 shows image-to-image translation's results.

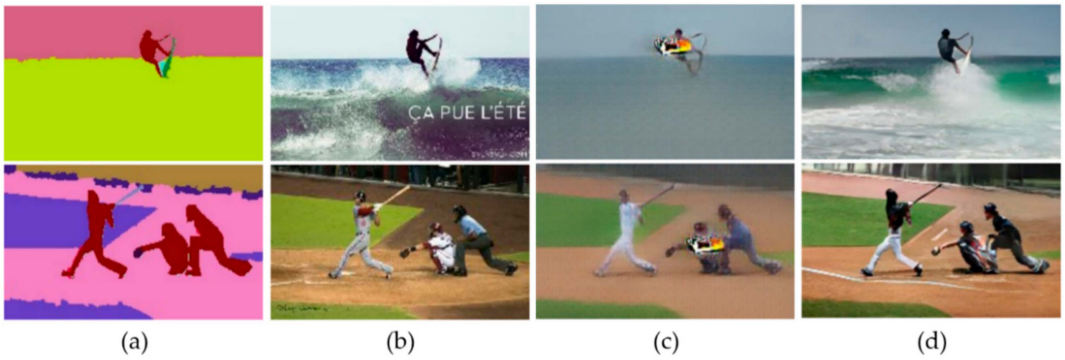

Figure 2. (**a**) Segmentation image, (**b**) original image. When (**a**) is given as input, (**c**) is the result of Pix2PixHD, and (**d**) is the result of GauGAN.

3. Materials and Methods

Given a person image P and a model image M, we propose a costume design system that generates a person image P' wearing M-style outfits. It is possible to partially change the outfits, and new forms of clothing can be created through segmentation modification. This system is divided into three stages: Generating style code, applying modified SEAN, and image correction.

3.1. Dataset

Images were collected through publicly available DeepFashion [13] and web crawling. DeepFashion is a large-scale garments database that provides images and annotation data for clothing-related tasks such as clothing detection, landmark prediction, clothing segmentation, and search. In the experiment, an In-shop Clothes Retrieval Benchmark was used among the datasets provided by DeepFashion. The In-shop Clothes Retrieval Benchmark is largely divided into men's and women's, and each is categorized by the type of clothing (T-shirt, pants, jacket, etc.). In each item, there are images of people wearing that kind of apparel and posing in various poses from different angles. In this paper, only full-body and upper-body images of people dressed were used for better learning results. However, the DeepFashion dataset lacks diversity in patterns and colors of clothes. These are important factors in fashion design. Therefore, we collected additional images to complement this through web crawling. We collected images of people wearing patterned clothing such as stripes, dots, checks, flowers, tie-died, leopards, and camouflage, and people wearing clothing of various colors. Moreover, for the consistency of the dataset, only the full body and upper body image were added to the dataset. The images used in the experiment totaled 53,000 with 45,000 DeepFashion images and 8000 additional images, divided into 48,000 learning images and 5000 test images. The size of the images was 256 × 256.

3.2. Generate Style Code

Style code generation is a process of generating style codes for each part after filtering styles for each field in order to synthesize images based on the segmentation area (Figure 4). A pre-learned Graphonomy network was used to obtain segmentation images. Entering a human image P into the network, it outputs the class label of each pixel. By assigning colors for each class, visualization is possible as shown in Figure 3. In the experiment, it produces segmentation images P_seg with 20 classes (background, hat, hair, gloves, sunglasses, top clothes, dress, coat, socks, pants, torso, scarf, skirt, face, left arm, right arm, left leg, right leg, left shoe, right Shoes). The style encoder takes P and P_seg as inputs and creates a (512 × 20)-dimension style matrix ST. As there are 20 classes, it creates 20 columns. Each column of ST corresponds to the style code of the area. The column of the class area that does not exist in the image becomes 0. The style encoder is trained to filter out region-specific style codes from the input image according to the corresponding segmentation mask. This style code creates a style map by performing convolution for each style and broadcasting to the corresponding area according to the segmentation mask (Figure 4).

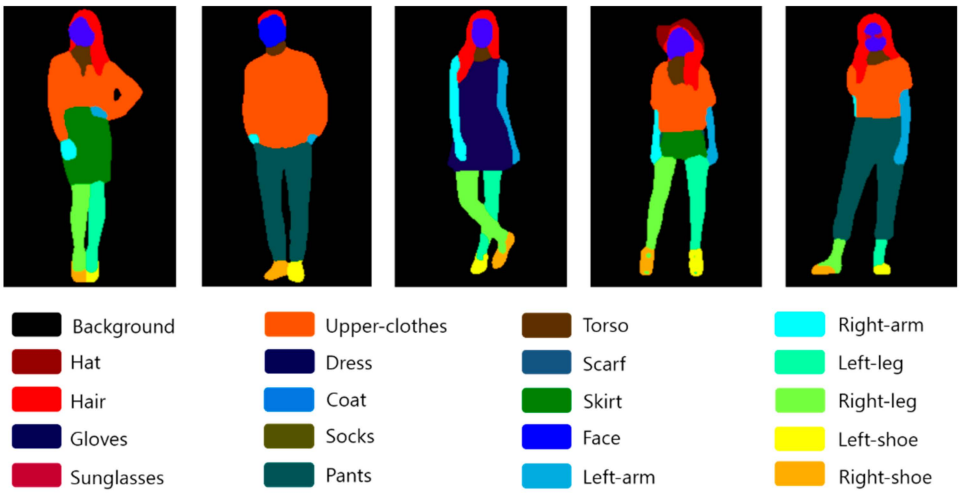

Figure 3. Segmentation images and labels by color used in the experiment.

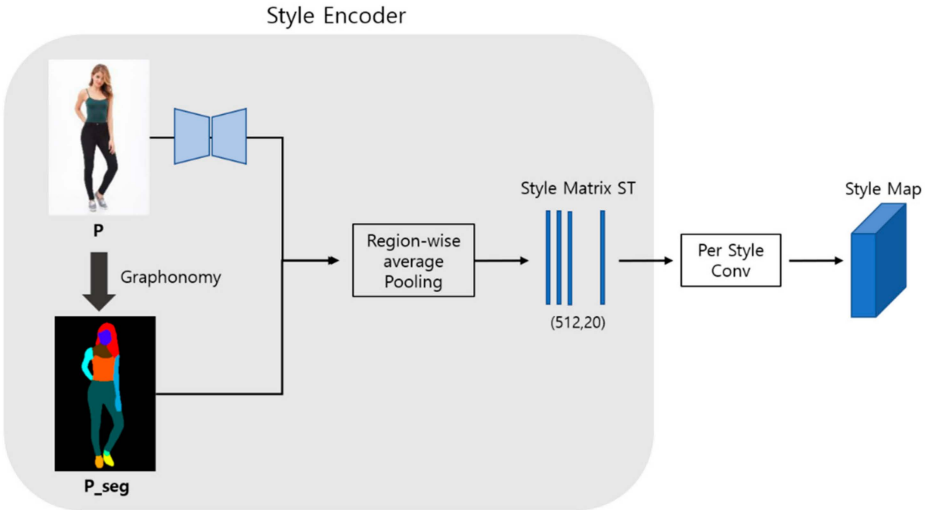

Figure 4. Style code generation process.

3.3. Application of The Modified SEAN

3.3.1. SEAN

In deep learning, normalization is used to normalize the output of the intermediate layer. Stable learning is possible by forcing the distribution of the output values of the activation function, but a lot of information is lost. GauGAN devised a spatially adaptive normalization (SPADE) that can preserve spatial information and produce better results. However, because SPADE has one style code, it is impossible to modify the image in detail because it can only be changed into one style. Therefore, semantic region-adaptive normalization (SEAN) was proposed in [6] so that each region can be individually controlled. As the image is created under the condition of the segmentation area to be changed, only that part can be controlled in a desired style. Using a simple convolutional network, γ^s(scale) and β^s (shift) for each pixel are obtained from the previous style. At the same time, normalization is performed in the same way as SPADE, and γ^o and β^o are created. They are combined by multiplying the weights. γ, β values and weights are learned in SEAN. As γ and β are not scalar values but tensors that are dimensions of space, they have the advantage of not losing spatial information. This can be expressed as follows.

$$\gamma_{c,y,x}(ST,\ SM) = \alpha_\gamma \gamma^s_{c,y,x}(ST) + \left(1-\alpha_\gamma\right)\gamma^o_{c,y,x}(SM), \beta_{c,y,x}(ST,\ SM) = \alpha_\beta \beta^s_{c,y,x}(ST) + \left(1-\alpha_\beta\right)\beta^o_{c,y,x}(SM) \quad (2)$$

$$\gamma_{c,y,x}(ST,\ SM)\frac{h_{n,c,x,y}-\mu_c}{\sigma_c} + \beta_{c,y,x}(ST,\ SM) \quad (3)$$

ST is the style matrix, SM is the segmentation mask, N is the batch size, C is the number of channels, and H and W are the height and width of the activation map, respectively ($n \in N$, $c \in C$, $y \in H$, $x \in W$). μ_c, σ_c are the mean and variance of the activation on the c channel, respectively. $\gamma_{c,y,x}$ and $\beta_{c,y,x}$ are the weighted sum of $\gamma^s_{c,y,x}$ and $\gamma^o_{c,y,x}$, $\beta^s_{c,y,x}$ and $\beta^o_{c,y,x}$, respectively (Equation (2)). Normalize $h_{n,c,x,y}$ to the mean μ_c and distributed σ_c of the activation on the c channel. Conduct the denormalization under the condition of $\gamma_{c,y,x}$ and $\beta_{c,y,x}$ obtained by adding segmentation and style matrix ST, respectively. Figure 5 shows the structure of semantic region-adaptive normalization.

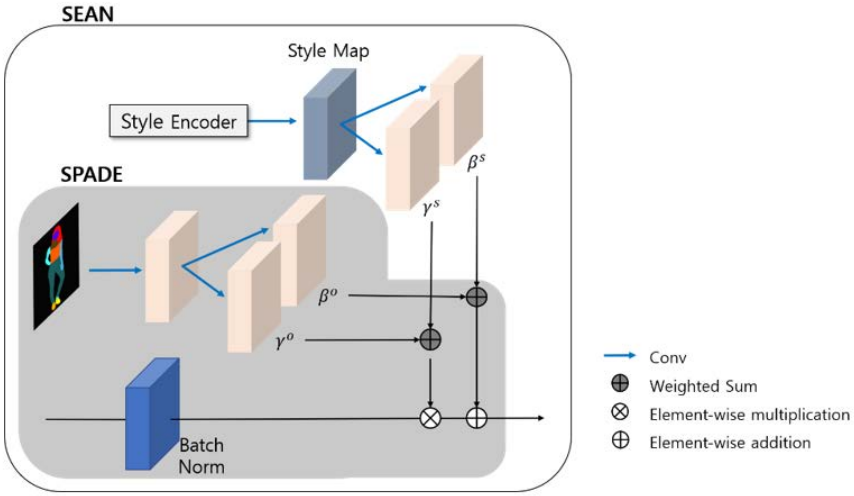

Figure 5. Structure of semantic region-adaptive normalization (SEAN).

3.3.2. Modified SEAN

The SEAN Block has regional style codes, segmentation, and noise as input. In this paper, an additional network is added after the SEAN block to improve performance. ResNet [14] improves performance by solving the gradient vanishing problem. The unit of SEAN with Resnet structure is called SEAN Resnet block (ResBlK). The structure of SEAN ResBlK is shown in Figure 6. SENet [15] consists of a squeeze step and an excitation step. When it is added to the model, there is not much increase in hyperparameters, so model complexity and computations do not increase significantly. However, the performance improvement of models is quite high. By adding a network like SENet before the up-sampling step, features can be emphasized in consideration of SEAN ResBlK's importance per channel. It also improves performance by making the model converge faster and the synthesized image look smoother. We called this step SENet block (SEBlK). Structure of SEBlK is shown in Figure 7.

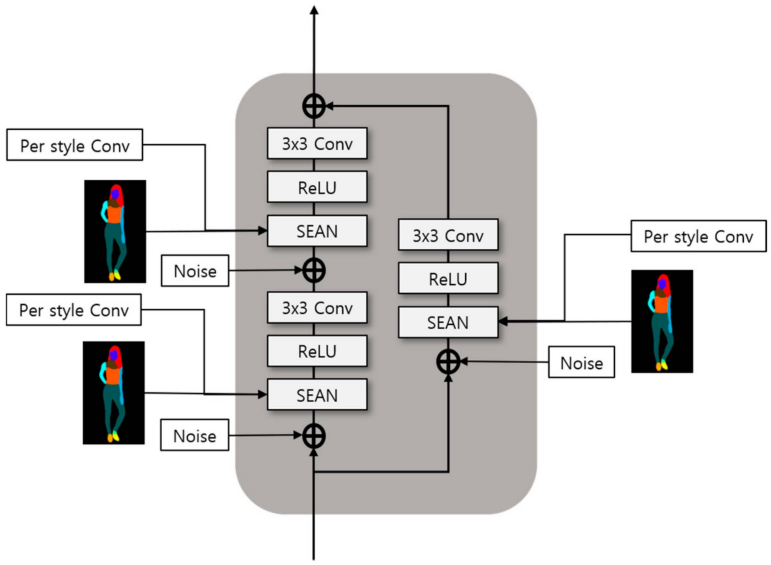

Figure 6. Structure of SEAN ResBlK.

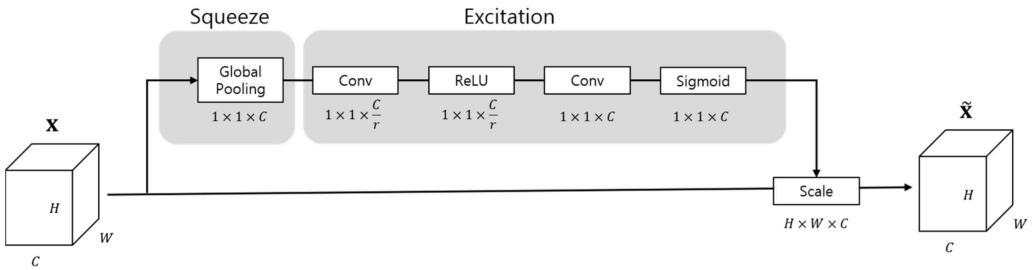

Figure 7. Structure of SEBlK. The squeeze step converts the feature maps of H × W of C channels into a feature map of 1 × 1 size. One value is obtained by averaging the two-dimensional feature map. That is, the H × W × C feature map is converted into a 1 × 1 × c feature map. The excitation step is performed to find out the relative importance of each channel through the feature map Conv and the activation function. The relative importance of each channel can be expressed as a value between 0 and 1, which is multiplied by x to obtain $\widetilde{x}$ in the scale process.

Therefore, it is trained to input the mask, style code, and noise into SEAN and reconstruct the input image through SEBlK and up-sampling. Image generation process is shown in Figure 8. As a result of learning, it is possible to modify the image by changing the style image and the segmentation mask.

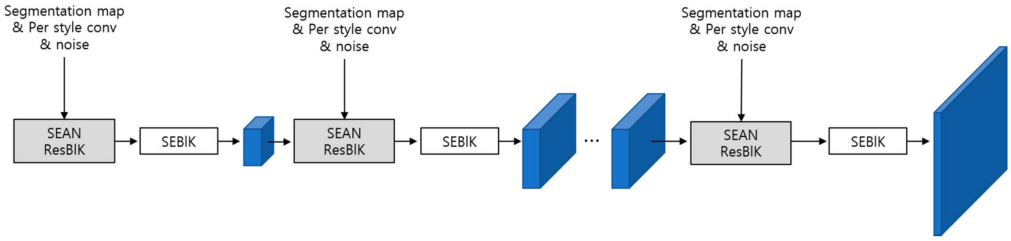

Figure 8. Image generation process. SEAN ResBlK, SEBlK, and up-sampling are repeated to generate an image.

3.4. Generated Image Correction

The face is unclear because the face area is also created when the image is created. In addition, as the style of model M is reflected, the face area has the face of model M, not the face of person P. Therefore, correction is needed for better results and preserving P's face in the generated image. Figure 9 shows the editing process. In this paper, after extracting the human face (P_face) through the operation of the human image (P) and the segmentation image (P_seg), the generated image (P') is corrected by combining P_face and P'.

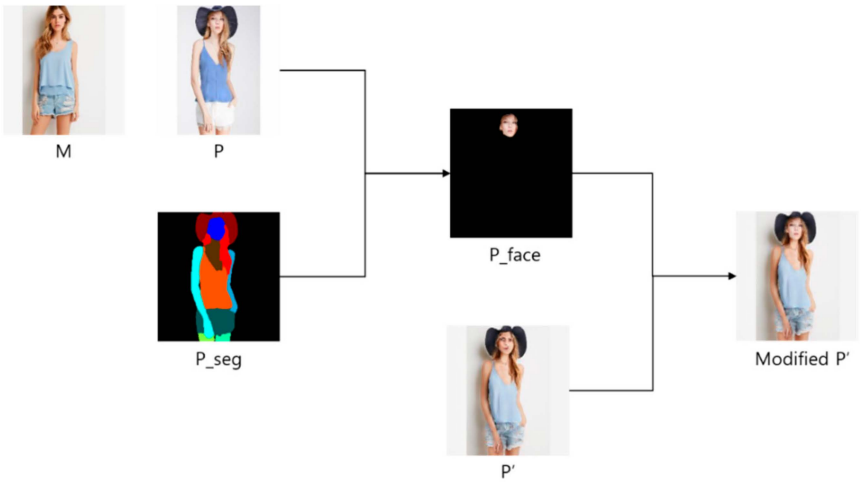

Figure 9. Correction process of the generated image.

4. Results

4.1. Qualitative Results

Figure 10 shows the results of partial style changes. (a) is the input image of a person who wants to change outfits. There are three different style model images to try-on in (b). Convert the input image to the model image's style. (c) is the attempt to change the colored parts (full, upper-clothes, pants) of the segmentation images. (d) is the result images of changing area (c) of (a) to the (b) style. The resulting images show that only the colored region has been changed while retaining the input image.

Figure 11 shows that various and unique fashion items can be made by modifying the segmentation image. As an image is generated based on the semantic map, if the semantic map changes, a new result image is obtained. By using this, clothes of various shapes and lengths can be designed. Length or shape of outfits can be easily revised from long sleeves to short sleeves, long pants to short pants, or an A-line skirt as an H-line skirt. A skirt can be trousers by shifting the shape. Furthermore, if switching the area label from pants to top, you can apply the top style to the pants area.

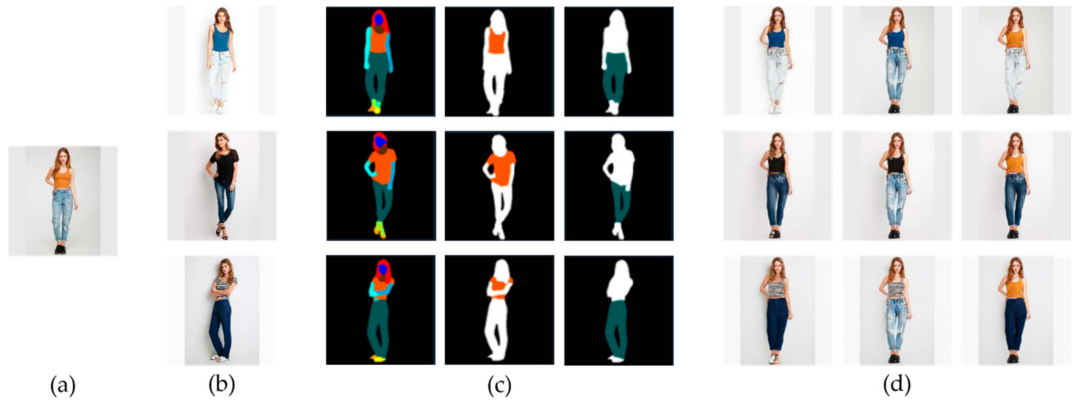

Figure 10. Change the clothes that (**a**) is wearing to (**b**)'s. (**c**) is the parts of the style to modify. (**d**) shows the results.

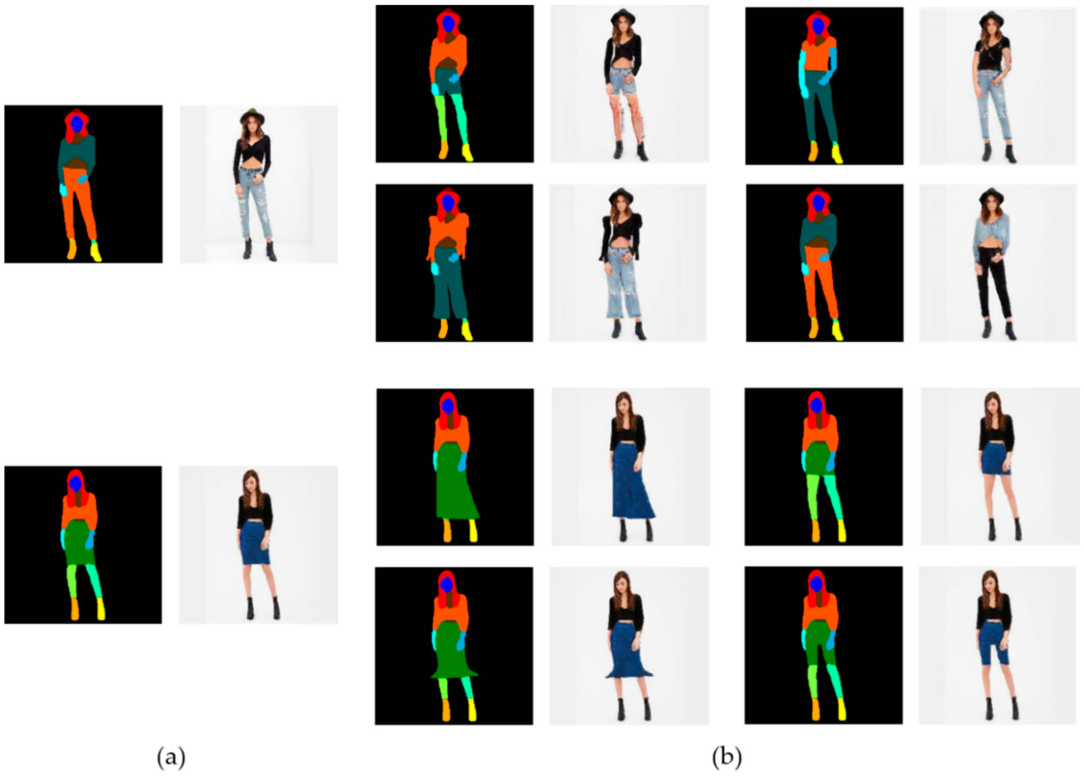

Figure 11. Segmentation images and corresponding generated images. (**a**) Original segmentation and style images. (**b**) Manipulated segmentation images and synthesized results based on them.

SEAN can be applied to other fashion synthesis-related problems. For example, it can create an image of a model that poses differently with one image (Figure 12). Or it can be converted to an image from another angle. Additionally, if only the image of the clothes is given, try-on wearing the clothes is possible.

Figure 12. Creating images of a model in various poses.

4.2. Quantitive Results

The method of judgment by a person is ambiguous and involves human subjectivity. In addition, the greater the number of images to evaluate, the more time it takes. The following metrics are used to quantify the performance of our model and compare our results to SEAN. Four indicators were used: Root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [16], Frechet Inception distance (FID) [17], and Naturalness Image Quality Evaluator (NIQE) [18]. PSNR is the most common criterion for evaluating images and SSIM is an indicator of the similarity between the two images. FID measures the difference between the two normal distributions. These methods measure the quality difference between two images. However, NIQE is a method that allows quality evaluation without referring to the original image. This method has been used in some applications [19,20]. For SSIM and PSNR, the higher the better. For RMSE, FID, and NIQE, the lower the better.

Some virtual try-on networks have similar goals as us. References [4,21] can try-on what others wear. When a product clothes image is given, References [3,22] can generate the person wearing it. However, they differ from ours in that they cannot manipulate outfits. To compare with a network that performs the same function, we compared the modified SEAN proposed in the paper with SEAN. Each network was trained for 20 epochs. RMSE, PSNR, SSIM, and FID evaluate the reconstruction performance. Images created based on the segmentation map and original style were compared to the original images. NIQE was used to evaluate images changed to various styles through each network. Table 1 shows that our model has better performance on all metrics.

Table 1. Quantitative comparison.

Methods	SEAN	Ours
RMSE	5.574	5.463
PSNR	24.376	24.898
SSIM	0.855	0.865
FID	80.353	74.983
NIQE	36.946	34.503

5. Conclusions

In this paper, modified SEAN was applied to apparel design. In the experiment, it was shown that only certain clothing parts of the image can be modified. In addition, through the change in the semantic map, it was possible to change to a new shape, style, or type of clothes. User-centered design is possible because users can quickly apply the style they want, so it is useful for choosing or recommending garments and fashion items. In addition, it showed that it can be applied to various fashion issues.

Unlike ordinary objects, it is difficult to apply artificial intelligence to clothes because they have many changes in design and style. As real fashion is much more complex and diverse, there are some problems to be solved to apply this paper's methods directly to the fashion industry. Furthermore, compared to the designer's ability, the completeness of the result is low. However, as an auxiliary means, experts can easily prototype ideas and stimulate imagination, which will help in design

development. In addition, anyone can design because it can be easily expressed with a simple touch. There is a possibility of development through continuous research.

Author Contributions: Conceptualization, Y.J.; methodology, Y.J.; software, Y.J.; formal analysis, Y.J.; investigation, Y.J.; writing—original draft preparation, Y.J.; writing—review and editing, Y.J. and C.-B.S.; visualization, Y.J.; supervision, Y.J.; project administration, Y.J.; funding acquisition, C.-B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2016-0-00288) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

Acknowledgments: The present research has been conducted by the Research Grant of Kwangwoon University in 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 2672–2680.
2. Zhu, S.; Urtasun, R.; Fidler, S.; Lin, D.; Change Loy, C. Be your own prada: Fashion synthesis with structural coherence. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1680–1688.
3. Pandey, N.; Savakis, A. Poly-GAN: Multi-Conditioned GAN for Fashion Synthesis. *arXiv* **2019**, arXiv:1909.02165.
4. Wu, Z.; Lin, G.; Tao, Q.; Cai, J. M2e-try on net: Fashion from model to everyone. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 293–301.
5. Xian, W.; Sangkloy, P.; Agrawal, V.; Raj, A.; Lu, J.; Fang, C.; Yu, F.; Hays, J. Texturegan: Controlling deep image synthesis with texture patches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8456–8465.
6. Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5104–5113.
7. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
8. Gong, K.; Gao, Y.; Liang, X.; Shen, X.; Wang, M.; Lin, L. Graphonomy: Universal human parsing via graph transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7450–7459.
9. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
10. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
11. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
12. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2337–2346.
13. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1096–1104.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

15. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
16. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
17. 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 *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6629–6640.
18. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
19. Kwan, C.; Budavari, B.; Bovik, A.C.; Marchisio, G. Blind quality assessment of fused worldview-3 images by using the combinations of pansharpening and hypersharpening paradigms. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1835–1839. [CrossRef]
20. Kwan, C. Demosaicing Mastcam Images using A New Color Filter Array. *Signal Image Process. Int. J. (SIPIJ)* **2020**, 11. [CrossRef]
21. Liu, Y.; Chen, W.; Liu, L.; Lew, M.S. Swapgan: A multistage generative approach for person-to-person fashion style transfer. *IEEE Trans. Multimedia* **2019**, *21*, 2209–2222. [CrossRef]
22. Han, X.; Wu, Z.; Wu, Z.; Yu, R.; Davis, L.S. Viton: An image-based virtual try-on network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7543–7552.