electronics

MDPI

Article

Development of a Hairstyle Conversion System Based on Mask R-CNN

Seong-Geun Jang, Qiaoyue Man and Young-Im Cho *

AI&SC Lab, Department of Computer Engineering, Gachon University, Seongnamsi 461-701, Gyeonggido, Korea; zzangga97@gachon.ac.kr (S.-G.J.); manqiaoyue@gmail.com (Q.M.)
* Correspondence: yicho@gachon.ac.kr; Tel.: +82-10-3267-4727

Abstract: Interest in hairstyling, which is a means of expressing oneself, has increased, as has the number of people who are attempting to change their hairstyles. A considerable amount of time is required for women to change their hair back from a style that does not suit them, or for women to regrow their long hair after changing their hair to a short hairstyle that they do not like. In this paper, we propose a model combining Mask R-CNN and a generative adversarial network as a method of overlaying a new hairstyle on one's face. Through Mask R-CNN, hairstyles and faces are more accurately separated, and new hairstyles and faces are synthesized naturally through the use of a generative adversarial network. Training was performed over a dataset that we constructed, following which the hairstyle conversion results were extracted. Thus, it is possible to determine in advance whether the hairstyle matches the face and image combined with the desired hairstyle. Experiments and evaluations using multiple metrics demonstrated that the proposed method exhibits superiority, with high-quality results, compared to other hairstyle synthesis models.

Keywords: face and hair segmentation; data analysis; convolutional neural network; generative adversarial network

Citation: Jang, S.-G.; Man, Q.; Cho, Y.-I. Development of a Hairstyle Conversion System Based on Mask R-CNN. *Electronics* **2022**, *11*, 1887. https://doi.org/10.3390/electronics11121887

Academic Editor: Amir Mosavi

Received: 28 May 2022
Accepted: 15 June 2022
Published: 15 June 2022

1. Introduction

As the development of convolutional neural networks (CNN) and generative adversarial networks (GAN) has rapidly accelerated in the past several years, artificial intelligence (AI)-based deep learning has been used in various fields, including beauty, which was the focus of this study. Image recognition and separation are areas of considerable interest in fields such as automation and smart cities. Owing to the recent COVID-19 epidemic, the manner in which people express their personalities has been diversifying. An individual's personality may be revealed by applying makeup and wearing various accessories. Hairstyling is one area in which people take an interest, and through which they can best express themselves to others. Thus, hairstyling is one of the most important aspects of modern means of self-expression. Hairstyling-related content has increased considerably on various SNSs such as Instagram and YouTube, and the public expresses strong interest in this content. Moreover, as content that focuses on topics such as on hairstyling has increased, interest in hair loss treatment has also increased. Thus, the importance of hairstyling to modern people has increased considerably. When changing one's hairstyle, the new style can be chosen based on the usual hairstyle that is recommended subjectively by a beauty expert, or on one that is trending or has been popularized by celebrities. However, when a new hairstyle is adopted, the hair may be changed to a trendy style, which may be something that does not suit the person; alternatively, a style that is recommended by a beauty expert may not be liked. In particular, if a woman with long hair changes her hairstyle to short hair, a long time is required to return to the long hairstyle if she does not like the new style, or if it does not suit her; thus, she should find a hairstyle that suits her prior to styling [1]. It is almost impossible to attempt all hairstyle options, such as perms and dyeing, and to find a suitable hairstyle, because substantial time and money

are required. It is also quite difficult for a person to see the style they wish to adopt in advance and to decide if it suits them. To ease this difficulty, in recent years, apps have emerged that display different hairstyles on heads using filters, which makes it possible to experience different styles. However, these are not detailed or accurate compared to the proposed method. Applications recognize the head part of selfies and display other prepared hairstyles; therefore, the head may stick out or feel unnatural. In order to overcome both the disadvantages of new hairstyles that require a lot of time and money, and the existing methods of showing protruding or unnatural hairstyles, in this paper, we propose a network that uses deep learning image classification technology, based on AI algorithms, to recognize the hair and face, and execute hairstyle conversion through classification. The contributions in this paper are as follows:

- We created a new dataset that divided the hair and face based on facial image data.
- We extracted hair and facial features more effectively using the pyramidal attention network (PAN), which effectively extracts context from the pyramidal scale for more accurate segmentation.
- We propose the hairstyle conversion network (HSC-Net), which is a network model that combines the Mask R-CNN with PAN and GAN, to provide hairstyle conversion.

2. Related Work

Mask-RCNN was designed to deal with object segmentation by adjusting the network parameters, and it outperformed existing models in all parts of the 2016 COCO challenge (object recognition and segmentation) [2,3]. This network, which can be used to separate the hair and face, is useful in beauty as well as in other fields. In the biomedical and medical fields, it is applied to separation algorithms that are used to segment automatic nuclear instances in the predicted boundary boxes by determining the boundary boxes for objects, thereby allowing researchers to manually replace separation with automation systems. From a mechanical perspective, this method is sometimes used to deal quickly with traffic accident compensation problems through a car damage detection network, by extracting a dataset label from a damaged car photograph. Moreover, Mask R-CNN models are used to obtain the location and size of photovoltaic generators based on satellite datasets [4–6]. In this study, the training and testing data were constructed using a cut image, in which the hair and face were separated using Mask R-CNN.

With the recent development of AI technology, GAN [7] has been used in various fields. Since the advent of GAN, high-quality image generation studies have increased considerably, and the quality of the generated images has improved dramatically. Owing to the fairly high quality of the generated image, it is difficult to distinguish it from the actual image. Many fake faces without copyright have been developed using image creation. Although writing these images without copyright represents a fairly strong advantage, GAN is also used in neural network-based fake face detectors to prevent them from being used maliciously, without being detected through image detection algorithms [8]. GAN plays a significant role in image generation and restoration. The reconstructed image is extracted by reconstructing the details and background of the image using the GAN model, in instances where the resolution is low or the image is damaged owing to noise, blur, etc. [9]. In this study, the image of a new hairstyle was extracted by training the GAN model based on a dataset that was composed using Mask R-CNN.

3. Method

3.1. HSC-Net

GAN models include latent optimization of hairstyles via orthogonalization (LOHO) [10], segmentation, using networks such as the Graphonomy segmentation network [11], and 2D-FAN [12]. The model proposed in this study segmented hair and face more accurately using Mask R-CNN with PAN, which extracts context from pyramid scale in order to extract context more effectively. Typical Mask R-CNNs use ResNet and a feature pyramid network as the backbone networks. However, HSC-Net uses PAN as the backbone network of Mask

R-CNN for stronger feature expression capabilities. Hairstyle conversion is performed using GAN with the mask extracted from Mask R-CNN (Figure 1).

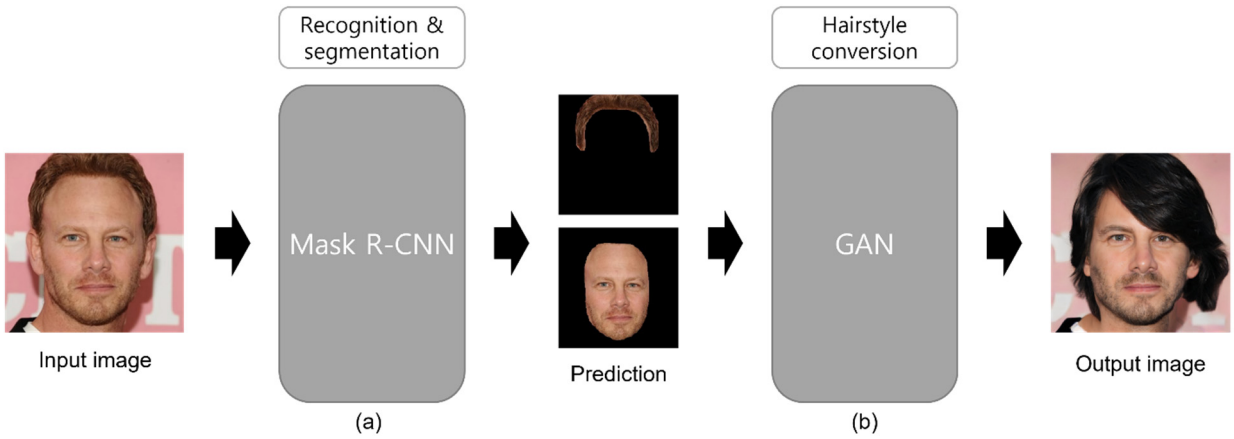

Figure 1. Overview of the structure of our proposed model, HSC-Net: (**a**) recognition and separation of hair and face in input image through Mask R-CNN, and (**b**) hairstyle conversion using GAN.

3.2. Mask R-CNN

The Mask R-CNN consists of a structure in which Faster R-CNN and the mask branch are incorporated. The region of interest (RoI) pooling that is used in Faster R-CNN undermines the information on the adjacent pixel spaces because it forcibly ignores decimal places or lower values by rounding operations if the size includes decimal points, while performing a pooling operation to generate the RoI bin. Therefore, in this study, image classification was performed using RoIAlign rather than Faster R-CNN, in order to perform image segmentation more effectively. The Mask R-CNN framework is depicted in Figure 2.

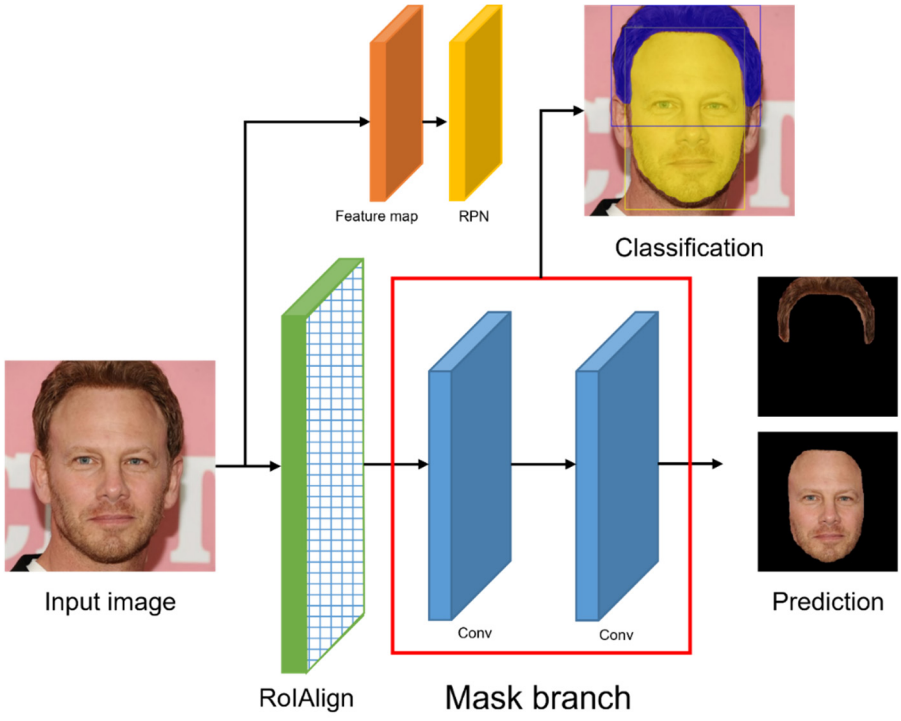

Figure 2. Basic framework of Mask R-CNN.

Among the networks that recognize and classify images, ResNet [13,14] learns in the direction of minimizing $H_{(x)} - x$, such that the output value of the network becomes x, where $H_{(x)} - x$ is referred to as a residual, and it indicates the network where this residual has been learned. Regardless of the depth of the layer, the residual has a value of 1 or more, thereby solving the gradient vanishing problem. In ResNext [15], the bottleneck in ResNet was changed. The input value of 256 channels is 128 channels through 1×1 conv, and the channels are divided into 32 groups to form four channels per group. By connecting four feature maps that are formed in 32 groups, 129 channels are created, and 256 channels are again formed through 1×1 conv. Increasing the cardinality is more effective and provides excellent performance compared to increasing the depth, such as in ResNet; therefore, it is applied in various computer vision fields. Figure 3 presents the frameworks of ResNet and ResNext.

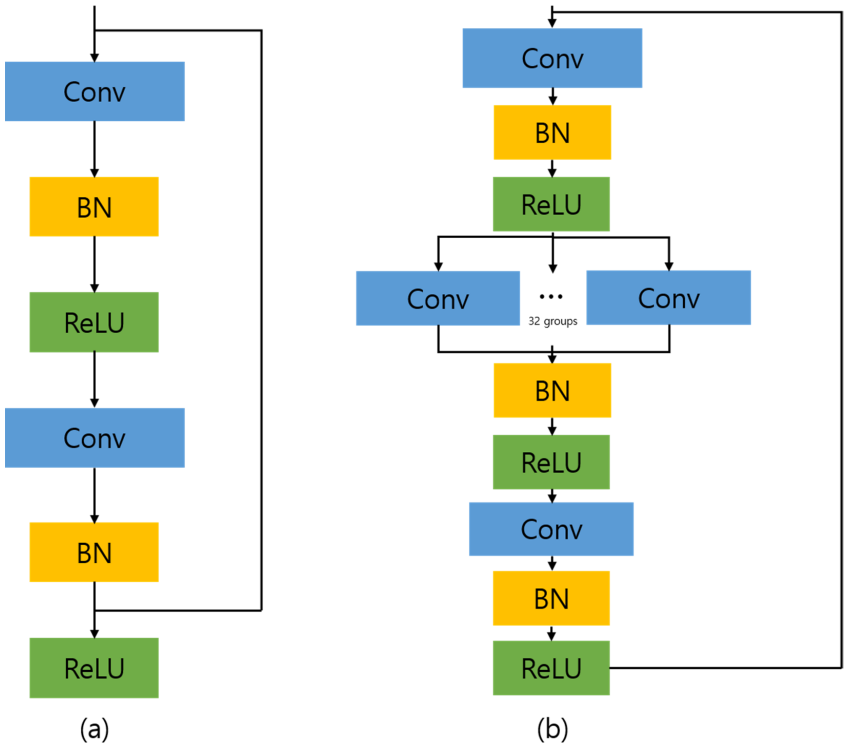

Figure 3. (**a**) ResNet framework and (**b**) ResNext framework.

For the loss of Mask R-CNN (L_t), L_c is the classification loss, L_b is the bounding box loss, and L_m is the binary cross-entropy loss for the mask loss. The Mask R-CNN L_t is expressed as follows:

$$L_t = L_c + L_b + L_m. \quad (1)$$

The mask branch generates an output value of N × M × M for each RoI. N is the number of classes, and M is the size of the mask, which is the mask branch that calculates the output value for each N class; only the mask with the class output from the class branch calculates the loss.

$$L_m = -\sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)], \quad (2)$$

where y_i represents the prediction probability that is calculated from the sigmoid function per pixel, and $\hat{y}_i$ refers to the true class label.

PAN

PAN consists of two modules: feature pyramid attention (FPA) and global attention upsample (GAU) [16–18]. The FPA (Figure 4) uses 3×3, 5×5, and 7×7 convolutions

to extract the context from other pyramid scales effectively. The pyramid structure can integrate the adjacent scale of the context function more accurately by integrating different information sizes, step by step. The three feature maps are reduced by a 1×1 convolution layer after being connected and multiplied by pixels through 1×1 convolution again. The extracted features add a global pooling branch to obtain an improved FPA. GAU (Figure 5) performs 3×3 convolution on the low-level features to reduce the feature map channels of the CNN. The global context that is created by the high level of functionality is multiplied by the low level of functionality that is generated through the 1×1 convolution, including batch normalization and ReLU nonlinearity. A high-level feature is added to the weighted low-level feature to be upsampled to generate a GAU feature.

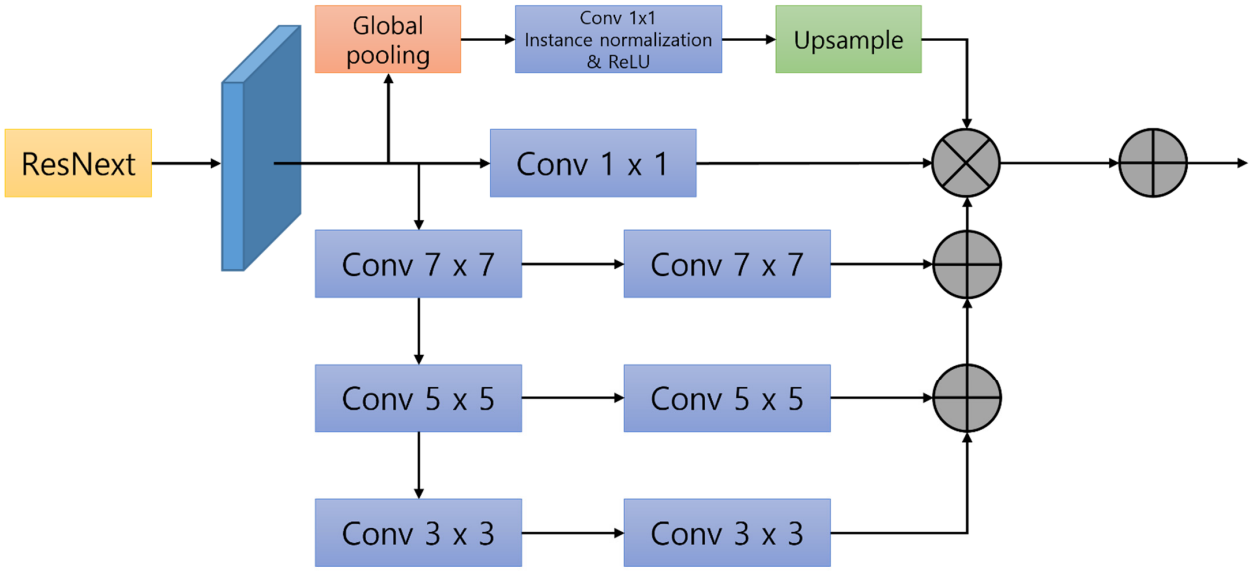

Figure 4. Structure of FPA.

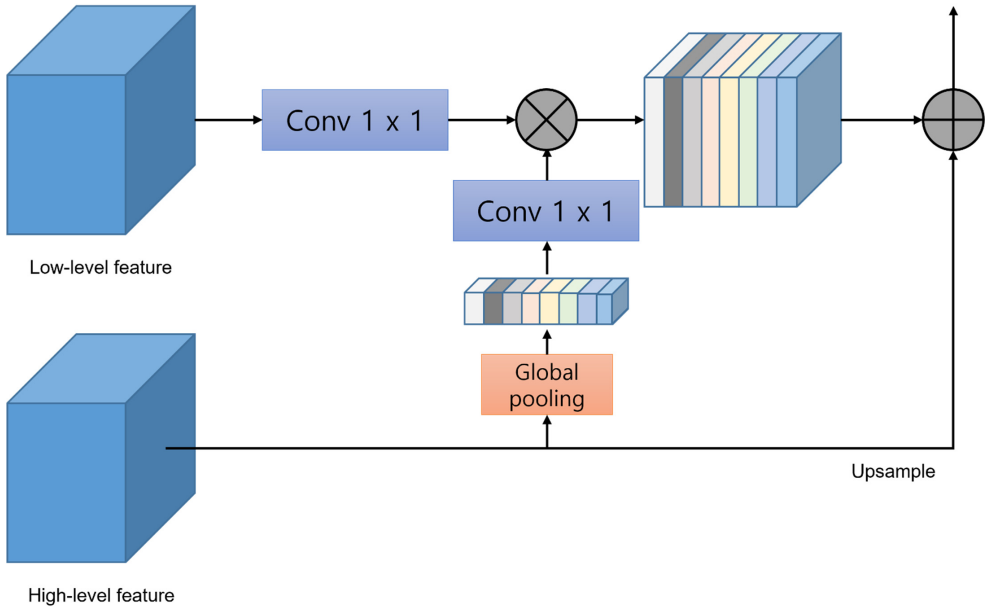

Figure 5. Structure of GAU in which high-level and low-level features are added together and upsampled to create features.

3.3. GAN

GAN has a structure in which two networks compete to produce better outcomes, and various GAN models are available. For example, CycleGAN [19] overcomes the need for pairs of training data, which is a limitation of pix2pix [20] that succeeded in improving the performance using an objective function combined with the GAN loss. StarGAN [21], which is a new and scalable approach, has difficulty in adjusting the attributes of images that are synthesized through generators, because different models can overcome the limitations that are created independently for each image domain pair, and one model can be used to perform image-to-image translation. Multi-input conditioned hair image GAN (MichiGAN) [22] consists of a conditional hair image transfer and generation method based on StyleGAN [23], and LOHO is an approach for transferring hairstyles by optimizing the noise space to the faces of other people. In this study, we proposed a model that connects Mask R-CNN and GAN. The architecture is depicted in Figure 6.

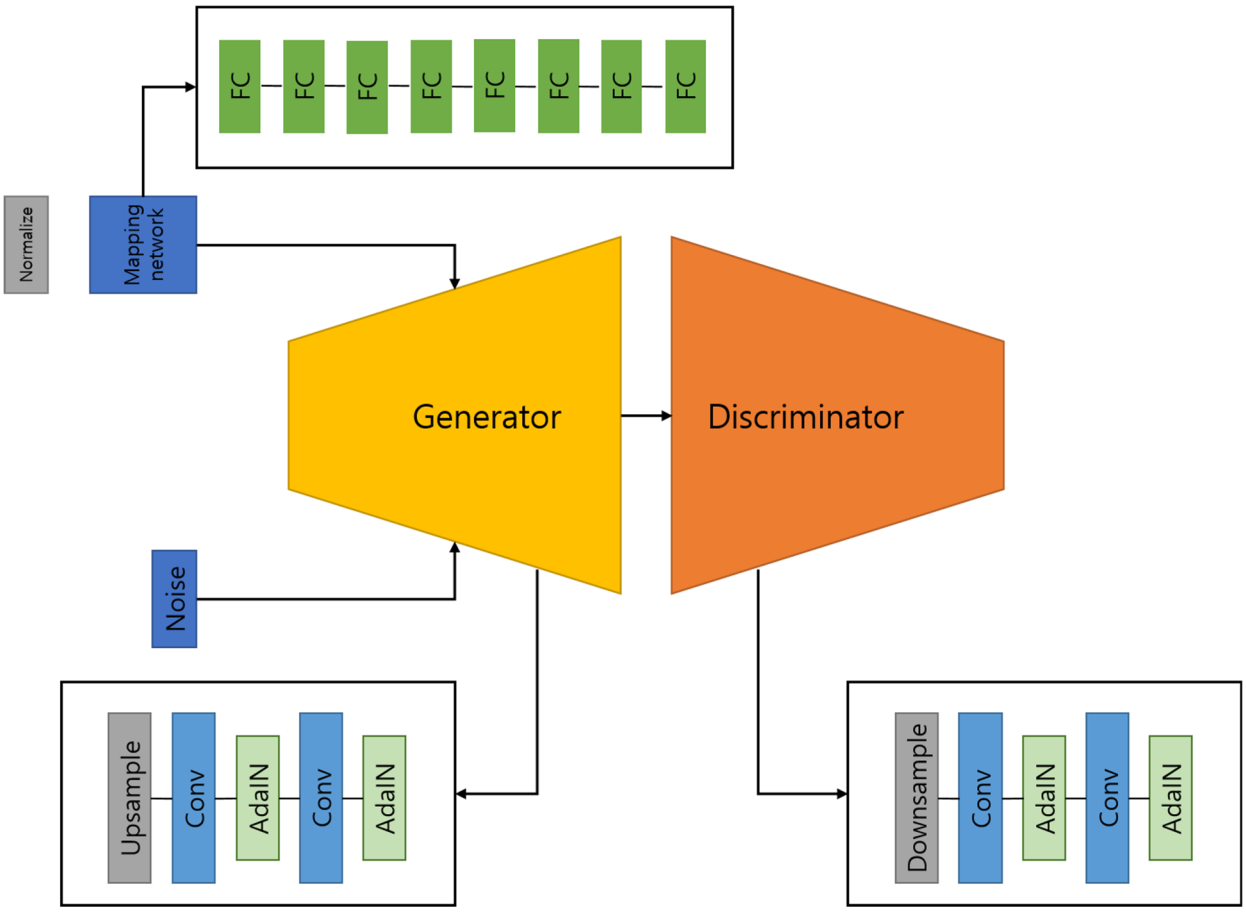

Figure 6. Structure of GAN module that performs hairstyle conversion.

Adaptive instance normalization (AdaIN) [24] was performed by transferring the style conversion from the feature space to the feature statistics. We used the following loss:

$$L = L_{content} + \lambda L_{style}, \tag{3}$$

$$L_{content} = ||f(G(m)) - m||_2, \tag{4}$$

where $L_{content}$ represents the content loss, λ represents the style loss weight, and L_{style} represents the style loss. Instead of using the feature response of the commonly used content image, we set the AdaIN output m, which is the target feature map, as the content target, where G represents the randomized initialized decoder.

4. Results

4.1. Implementation Details

Datasets

We trained the model using 30,000 high-resolution (1024 × 1024) images from open-source CelebA-HQ [25] (Large-scale CelebFaces attributes) and performed a comparison with other models. Because our proposed model executes the GAN through Mask R-CNN, we generated 512 × 512 label datasets, as shown in Figure 7, using Labelme [26] to generate a dataset to train the Mask R-CNN. The generated dataset formed a 512 × 512 mask dataset through Mask R-CNN, as illustrated in Figure 8, and, on this basis, hairstyle transformation was performed using the GAN network. The configuration parameters of the proposed model are presented in Table 1.

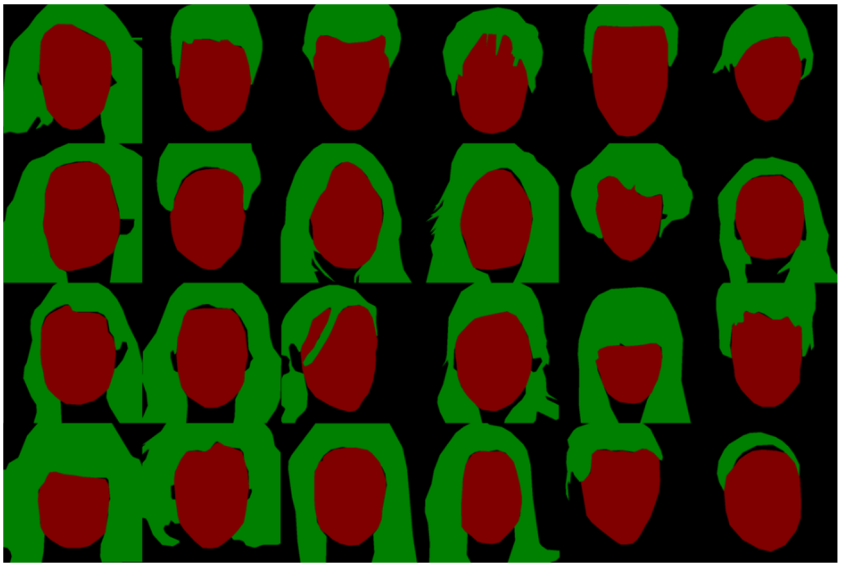

Figure 7. Label datasets created after separating hair and face using Labelme.

Figure 8. Mask datasets: test results of Mask R-CNN trained with label datasets.

Table 1. Configuration parameters of HSC-Net.

Stage	Output Size	Attention
Conv 1	112 × 112	7 × 7, 64, stride 2
Conv 2	56 × 56	3 × 3 max pool, stride 2
Conv 3	28 × 28	$\begin{bmatrix} 1\times1,\ 128 \\ 3\times3,\ 128 \quad C=32 \\ 1\times3,\ 256 \end{bmatrix} \times 3$
Conv 4	14 × 14	$\begin{bmatrix} 1\times1,\ 256 \\ 3\times3,\ 256 \quad C=32 \\ 1\times3,\ 512 \end{bmatrix} \times 4$
Conv5	7 × 7	$\begin{bmatrix} 1\times1,\ 512 \\ 3\times3,\ 512 \quad C=32 \\ 1\times3,\ 1024 \end{bmatrix} \times 6$
Global average pool	1 × 1	$\begin{bmatrix} 1\times1,\ 1024 \\ 3\times3,\ 1024 \quad C=32 \\ 1\times3,\ 2048 \end{bmatrix} \times 3$
		softmax

4.2. Comparison

We differentiated between two GAN algorithms, MichiGAN and LOHO, resulting in a more natural composite image, because they segmented the hair and face through Mask R-CNN and synthesized the separated hair and face mask images through GAN. We evaluated the model by comparing it with MichiGAN and LOHO. For a fair comparison, the same dataset was used for training, and all hyperparameters and configuration options were set to the default values. The results for the sampled 512 × 512 resolution images are displayed in Table 1. As indicated in the table, the results when using various metrics, including the learned perceptual image patch similarity (LPIPS) [27], Fréchet inception distance (FID) [28], the peak signal-to-noise ratio (PSNR) [29], and the structural similarity index map (SSIM) [30], were far superior to those of the other models. Table 2 presents the LPIPS and FID results, for which lower values indicate better performance, and PSNR and SSIM, for which higher values indicate better performance. It can be observed that the metrics outperformed both the baseline and the other models, MichiGAN and LOHO.

Table 2. Our method performed the best in all metrics, including LPIPS, FID, PSNR, and SSIM.

Model	LPIPS↓	FID↓	PSNR↑	SSIM↑
Baseline	0.20	44.10	24.12	0.85
MichiGAN	0.13	28.23	28.15	0.89
LOHO	0.19	42.45	24.49	0.85
HSC-Net (ours)	**0.09**	**23.55**	**29.79**	**0.91**

Furthermore, Table 3 presents the results of the dataset that was trained and tested using human facial image data that differed from Flicker-Faces-HQ (FFHQ), which also differed from CelebA-HQ, which we used. It can be observed that the proposed model exhibited superior results in all GAN evaluation metrics compared to MichiGAN and LOHO, and the results were consistent, without significant differences in performance, even when the data were changed. Assuming that the FID value, which is the most-used metric for evaluating GAN, was 100%, and the baseline value was 44.10, our proposed HSC-Net yielded a difference rate of 0.68%, and the next smallest model, MichiGAN, produced a difference rate of 1.36%. On this basis, our proposed HSC-Net exhibited approximately twice the performance difference. Figure 9 depicts the components that are required by our model when generating composite images. It displays the results of combining different styles and colors in terms of identity. Figure 10 presents a graph comparing the FID values. Because a lower FID indicates a better model, it can be observed that the FID was lower when using CelebA-HQ and FFHQ compared to the other models. As it was approximately

two times lower than the baseline, which was the basis of comparison, our proposed model exhibited significantly better performance. Figure 11 presents a comparison of LOHO and the composite images that were output using our model. The testing of our model using hair mask images that were separated using Mask R-CNN demonstrated that the images were more natural and similar to the hair color of the referenced image than the results of LOHO. Although the proposed model did not exhibit good performance on certain images, it produced images of the same natural and high quality in various hairstyles and different hair colors, as illustrated in Figure 12.

Table 3. Comparison of CelebA-HQ and FFHQ data with other models using GAN performance evaluation metrics, including LPIPS, FID, PSNR, and SSIM.

Model	CelebA-HQ				FFHQ			
	LPIPS↓	**FID↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**FID↓**	**PSNR↑**	**SSIM↑**
MichiGAN	0.13	28.23	28.15	0.89	0.12	28.17	28.30	0.90
LOHO	0.19	42.45	24.49	0.85	0.17	40.97	25.02	0.88
HSC-Net (ours)	**0.09**	**23.55**	**29.79**	**0.91**	**0.09**	**23.52**	**29.94**	**0.93**

Figure 9. Resulting images of our model using different styles and colors for identity.

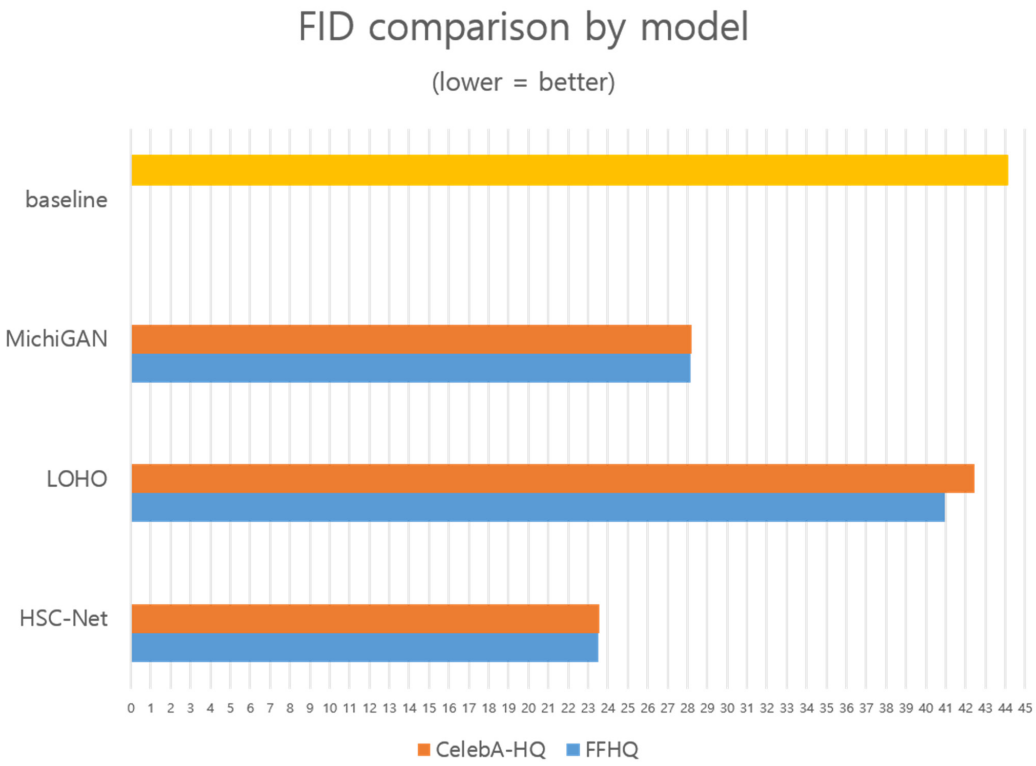

Figure 10. Comparison of FID, which is the most-used metric for evaluating GANs, by model.

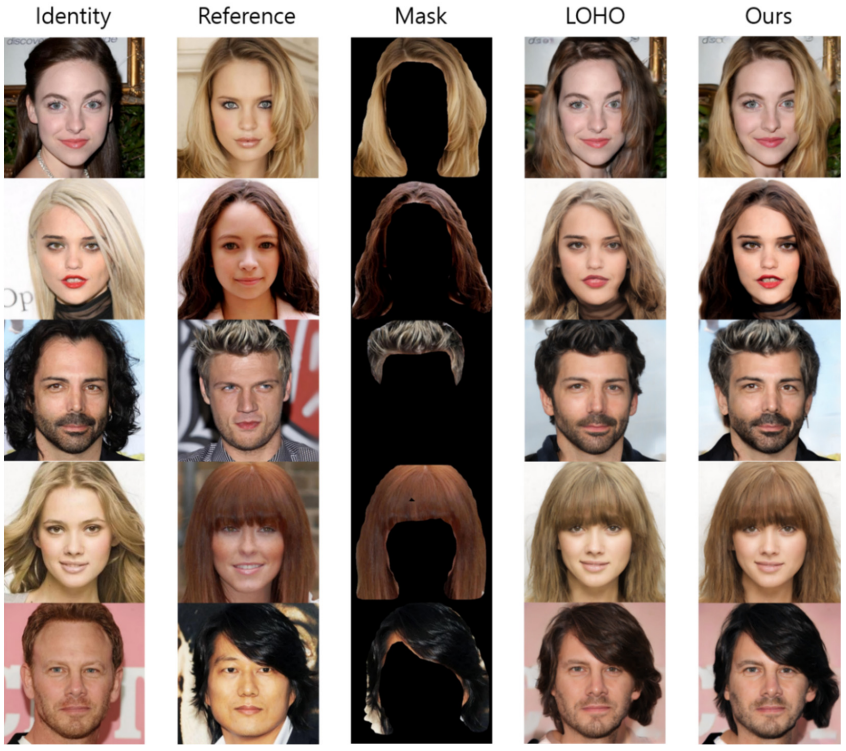

Figure 11. Comparison of LOHO and our model. Compared to the other models, our results were quite similar to the referenced image of hair color, and the images were synthesized more naturally.

Figure 12. Left image: style; top image: color. High-quality composite images are displayed in various styles and images by synthesizing various styles and colors based on the first facial figure.

5. Discussion

We focused on how people tend to convert their hairstyles, regardless of gender or age, because men may want to convert to long hair styles, or women may want to convert to short hair styles. We have created an artificial intelligence convolutional neural network system that detects and segments hairstyles and faces of all ages. It is helpful for the beauty industry and can be utilized in the fields of image classification and segmentation and image generation. Using low-quality images to create datasets resulted in an inaccurate hairstyle conversion, unnatural and inaccurate results, and a complete hairstyle conversion was not achieved. Therefore, the dataset was created using high-quality images of men and women of all ages. Through future research, it can be used not only for hairstyle conversion but also for hairstyle recommendation algorithms.

6. Conclusions

In this paper, we create a hairstyle conversion system based on a generative adversarial network, focusing on how to segment hair and faces more accurately and reliably. It can be used in various situations, such as style change in daily life and style preview in the beauty industry. The dataset consists of men and women of various ages, which helps to train models without age- or gender-based exceptions. The network was trained and verified using various hairstyles, such as long hairstyles, short hairstyles, perm hairstyles, and straight hairstyles, and various facial types, such as heart-shaped, long, oval, round, and square. We primarily generated our own label dataset from a large amount of data, using Labelme to train Mask R-CNN. We focused on how natural-looking the results of the hairstyle conversion were. In addition, we proposed a new network that adds a mask R-CNN to the GAN with a pyramid attrition network; experiments show that our network model is effective in generating more accurate images through better segmentation. In this paper, only hairstyle conversion was discussed, but in future studies, we will analyze the matter of hairstyle recommendations based on face shape and the location of facial features.

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

Funding: This paper is funded by Gyeonggido "Development and Construction of Virtual Hair Styling Service Using AI Technology" in 2021, and supported by the Gachon University research fund of 2021(GCU-202110250001) and partly supported by Gachon University Research Fund (2019-0389).

Data Availability Statement: The data are not publicly available because of privacy concerns. The data presented in this study are available upon request from the corresponding author.

Acknowledgments: This work was supported by Gyeonggido and Gachon University.

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

Abbreviations

Abbreviation	Meaning
CNN	convolutional neural networks
GAN	generative adversarial networks
PAN	pyramidal attention networks
HSC-Net	hairstyle conversion network
LOHO	latent optimization of hairstyles via orthogonalization
FAN	feature attention network
RoI	region of interest
FPA	feature pyramid attention
GAU	global attention upsample
ReLU	rectified linear unit
MichiGAN	multi input conditioned hair image GAN
AdaIN	adaptive instance normalization
Conv	convolution layer
CelebA-HQ	CelebFaces attributes-HQ
LPIPS	learned perceptual image patch similarity
FID	Fréchet inception distance
PSNR	peak signal-to-noise ratio
SSIM	structural similarity index map
FFHQ	Flicker-Faces-HQ

References

1. Sunhem, W.; Pasupa, K.; Jansiripitikul, P. Hairstyle recommendation system for women. In Proceedings of the 2016 Fifth ICT International Student Project Conference (ICT-ISPC), Nakhon Pathom, Thailand, 27–28 May 2016; IEEE: New York, NY, USA, 2016; pp. 166–169.
2. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
3. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European conference on computer vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
4. Vuola, A.O.; Akram, S.U.; Kannala, J. Mask-RCNN and U-net ensembled for nuclei segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; IEEE: New York, NY, USA, 2019; pp. 208–212.
5. Zhang, Q.; Chang, X.; Bian, S.B. Vehicle-damage-detection segmentation algorithm based on improved mask RCNN. *IEEE Access* **2020**, *8*, 6997–7004. [CrossRef]
6. Liang, S.; Qi, F.; Ding, Y.; Cao, R.; Yang, Q.; Yan, W. Mask R-CNN based segmentation method for satellite imagery of photovoltaics generation systems. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; IEEE: New York, NY, USA, 2020; pp. 5343–5348.
7. Liu, M.-Y.; Tuzel, O. Coupled generative adversarial networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5 December 2016; pp. 469–477.

8. Tariq, S.; Lee, S.; Kim, H.; Shin, Y.; Woo, S.S. Detecting both machine and human created fake face images in the wild. In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, Toronto, ON, Canada, 15 October 2018; pp. 81–87.
9. Yang, T.; Ren, P.; Xie, X.; Zhang, L. GAN prior embedded network for blind face restoration in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 672–681.
10. Saha, R.; Duke, B.; Shkurti, F.; Taylor, G.W.; Aarabi, P. Loho: Latent optimization of hairstyles via orthogonalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1984–1993.
11. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 7450–7459.
12. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1021–1030.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 630–645.
14. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
15. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
16. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
17. Huang, Z.; Zhong, Z.; Sun, L.; Huo, Q. Mask R-CNN with pyramid attention network for scene text detection. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; IEEE: New York, NY, USA, 2019; pp. 764–772.
18. Zhang, X.; An, G.; Liu, Y. Mask R-CNN with feature pyramid attention for instance segmentation. In Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 12–16 August 2018; IEEE: New York, NY, USA, 2018; pp. 1194–1197.
19. Almahairi, A.; Rajeshwar, S.; Sordoni, A.; Bachman, P.; Courville, A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 195–204.
20. Qu, Y.; Chen, Y.; Huang, J.; Xie, Y. Enhanced pix2pix dehazing network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8160–8168.
21. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
22. Tan, Z.; Chai, M.; Chen, D.; Liao, J.; Chu, Q.; Yuan, L.; Tulyakov, S.; Yu, N. Michigan: Multi-input-conditioned hair image generation for portrait editing. *arXiv* **2020**, arXiv:2010.16417. [CrossRef]
23. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
24. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
25. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
26. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
27. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
28. 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 Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4 December 2017. pp. 6629–6640.
29. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; IEEE: New York, NY, USA, 2010; pp. 2366–2369.
30. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; IEEE: New York, NY, USA, 2003; pp. 1398–1402.