특집논문 (Special Paper)
방송공학회논문지 제27권 제5호, 2022년 9월 (JBE Vol.27, No.5, September 2022)
https://doi.org/10.5909/JBE.2022.27.5.619
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

얼굴 마스크 정보를 활용한 다중 속성 얼굴 편집

Laudwika Ambardi[a)], 박 인 규[a)], 홍 성 은[a)‡]

Multi-attribute Face Editing using Facial Masks

Laudwika Ambardi[a)], In Kyu Park[a)], and Sungeun Hong[a)‡]

요 약

얼굴 인식 및 얼굴 생성이 다양한 분야에서 큰 주목을 받고 있지만, 얼굴 이미지를 모델 학습에 사용하는데 따른 개인 정보 문제는 최근 큰 문제가 되고 있다. 본 논문에서는 소수의 실제 얼굴 이미지와 안면 마스크 정보로부터 다양한 속성을 가진 얼굴 이미지를 생성함으로써 개인 정보 침해 이슈를 줄일 수 있는 얼굴 편집 네트워크를 제안한다. 다수의 실제 얼굴 영상을 이용하여 얼굴 속성을 학습하는 기존의 방법과 달리 제안하는 방법은 얼굴 분할 마스크와 얼굴 부분 텍스처 영상을 스타일 정보로 사용하여 새로운 얼굴 이미지를 생성한다. 이후 해당 이미지는 각 참조 이미지의 스타일과 위치를 학습하기 위한 훈련에 사용된다. 제안하는 네트워크가 학습되면 소수의 실제 얼굴 영상과 얼굴 분할 정보만을 사용하여 다양한 얼굴 이미지를 생성할 수 있다. 실험에서 제안 기법이 실제 얼굴 이미지를 매우 적게 사용함에도 불구하고 새로운 얼굴을 생성할 뿐만 아니라 얼굴 속성 편집을 지역화하여 수행할 수 있음을 보인다.

Abstract

Although face recognition and face generation have been growing in popularity, the privacy issues of using facial images in the wild have been a concurrent topic. In this paper, we propose a face editing network that can reduce privacy issues by generating face images with various properties from a small number of real face images and facial mask information. Unlike the existing methods of learning face attributes using a lot of real face images, the proposed method generates new facial images using a facial segmentation mask and texture images from five parts as styles. The images are then trained with our network to learn the styles and locations of each reference image. Once the proposed framework is trained, we can generate various face images using only a small number of real face images and segmentation information. In our extensive experiments, we show that the proposed method can not only generate new faces, but also localize facial attribute editing, despite using very few real face images.

Keyword : Face editing, Image synthesis, Facial mask, Data privacy

Ⅰ. 서 론

The rise of face recognition and image generation has been increasing over the past decades. With the use of Generative Adversarial Networks (GAN)[1], image generation and editing have seen an increase in quality. More recent GANs having great progress in synthesizing realistic faces[2,3,4], such as face swapping[5,6,7], attribute editing[8,9,10], and face frontalization[11,12]. Although these existing approaches have shown promising results, they still suffer from issues needing large amounts of the real-world dataset and considerable computational power.

However, in recent years, face datasets have been seeing an issue with privacy concerns, with datasets getting recalled due to problems in privacy with face recognition[13] or face-swapping methods. Using whole image faces to classify images or even generate synthetic images requires the image to be in the public domain or curated, which is agreed upon. While images in the public domain are available, they are frequently in-the-wild images; this causes difficulty or tedious labor to perform annotations of labels for tasks such as human parsing and segmentation. Additionally, even though curated data can be gathered, the lack of diversity becomes an issue.

To address these concerns, we propose a method that can use a limited amount of data to create a more significant amount. Our key idea is to create a method that can limit the amount of data used to preserve more privacy of the people being used in the dataset. By utilizing segmentation masks, we can use the masks as a tool to control the geometry of the desired output. And by using segmented textured images, we can maintain the image’s textures. The segmentation mask can also be used more than once, and each combination of the geometry and textured images can be used to create a larger dataset with limited data. Because our method uses pre-existing segmentation masks, we can control the annotations of the segmentations to perform human parsing or segmentation.

Our method is also able to perform visual editing tasks. One such task is facial editing; we can do localized edits in the task to change certain aspects of the face. Either by changing the geometry and reconstructing the images, or even changing the texture of the face.

Overall, our contributions are as follows:

1. We propose a generic method that utilizes a combination of masks and textured images from a limited dataset to create a more extensive dataset.
2. Our framework can be used to perform multiple tasks, such as facial reconstruction, face synthesis, and texture swapping.

Ⅱ. 관련 연구

1. Face synthesis

Face synthesis is a topic that is still popular to this day. Works such as StyleGAN[4] have results that are high in resolution and highly detailed. Even though StyleGAN has excellent results, editing the latent space to get the desired attribute is still being researched. Causes editing of minor attributes or the control of attributes in the latent space to

a) 인하대학교 전기컴퓨터공학과(Department of Electrical and Computer Engineering)
‡ Corresponding Author : 홍성은(Sungeun Hong)
E-mail: csehong@inha.ac.kr
Tel: +82-32-860-7427
ORCID: https://orcid.org/0000-0003-1774-9168

※이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2021R1F1A1054569, No. 2022R1A4A1033549). 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2022-00155915, 인공지능융합혁신인재양성(인하대학교)).
※This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1054569, No. 2022R1A4A1033549). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).
· Manuscript June 30, 2022; Revised July 25, 2022; Accepted July 25, 2022.

be complex[10].

While noise-to-image face synthesis can work well, it is difficult to control the geometry and predefined textures using noise-to-image methods. Conditional GANs[14], however, help solve the issue by introducing labels to allow the generator to learn. But learning information from labels generates images that are not easily edited to get the desired attribute.

Similar to conditional GANs, which generate images from masks, we propose a generic framework that utilizes masks and segmented attributes for image generation and editing. Though similar, our method takes in the geometry and styles rather than labels. This allows us to create face images from the segmented masks, reconstruct images from the latent space, and even edit the segmentation mask to perform image editing for more minor features.

2. Image-to-Image translation

The image-to-image translation is the method of translating a specific image from the source domain to the target domain[15,16] and using the source domain to control the geometry of the output image and using the target domain to control the texture. With this, we can preserve the contents required from the source domain.

We use this method to get our input images from the domain of multiple-segmented attributes to combine them into one singular image. Conventional image-to-image translations use a source domain image and a target domain image. Our approach differs by using the style transfer[17] method of the Adaptive Instance Normalization to translate the given geometry from a segmentation mask to the desired attributes of our segmented attributes.

3. Image editing

Image editing has seen a lot of progress with the use of neural networks. Face editing has seen its fair share of progress[8,9,10]. These works are similar in that they use a mask to control the regions in which they intend to edit the facial parts−using style transfer to extract and inpaint the textures according to the intended input.

Our approach is similar in that we use the mask to control the geometry of the image. However, our work takes in segmented textures to maintain the desired textures of the image, in the sense that the segmentation mask should control the geometry. The textured images should learn the location and textures to be put into the geometry.

Ⅲ. 제안하는 기법

The key idea of our work is to limit the amount of dataset needed, thus preserving most of the privacy of the people used in the dataset. To do this, we use the segmentations from the CelebA-HQ[19] dataset to extract each of the five attributes such as skin, hair, left eye, right eye, and mouth. We use five individual VGG encoders for each to learn the textures and one single encoder for the segmentation mask to learn the geometry. Using Inputting both geometry and texture into an Adaptive Instance Normalization (AdaIn)[18], we take the normalized feature into a decoder to get our synthetic image.

As seen in Figure 1, our method works by taking a real image from the CelebA-HQ dataset and the segmentation mask from the CelebA-HQ Annotated dataset. Using the information from the annotated dataset we decompose the real images into five separate textures. This method involves transferring the style between a segmentation map and multiple style attributes. By utilizing the segmentation mask, we can control the geometry of the intended output as seen in Figure 3, while the styles control the textures as seen in figure 4. We input each image into a VGG network and concatenate the style inputs into one feature vector. We then utilize the AdaIn to perform style transfer

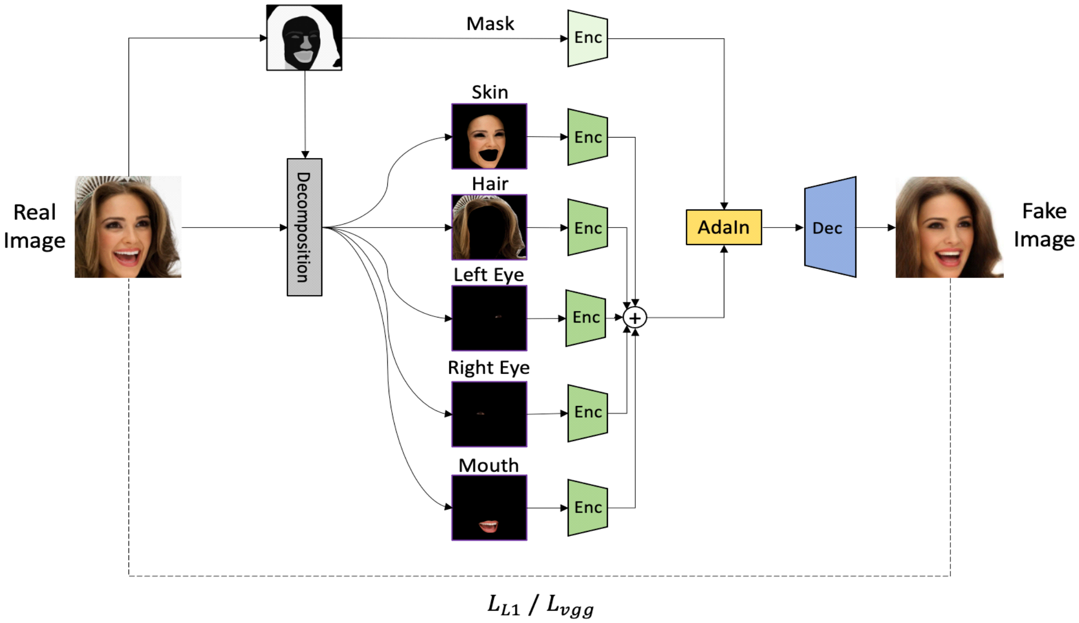


그림 1. 제안하는 프레임워크의 아키텍처 개요
Fig. 1. Overall architecture of the proposed framework

between the segmentation mask and the style images. While the first VGG encoder learns the geometry of the image, meanwhile the other five learn the style textures of each image as follows:

$$AdaIn(x,y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right)+\mu(y), \tag{1}$$

Where $\sigma(\cdot)$ is the standard deviation, and $\mu(\cdot)$ is the mean and all operations are computed along the spatial dimensions of the norm. We then take the normalized feature and input it into a decoder to get our synthetic image. When using a decoder with upscaling features, the training objective involves two losses. First, we utilize the L1 loss as a reconstruction objective function:

$$L_{L1}(x,\hat{x}) = |x-\hat{x}|, \tag{2}$$

where x denotes the real image and $\hat{x}$ is the generated image by the decoder. We also apply perceptual loss by summing all the squared errors between the feature-level values between real and generated images as follows:

$$L_{vgg}(x,\hat{x}) = E_{vgg}(x,\hat{x}), \tag{3}$$

Finally, we can define overall objective loss as follows:

$$L = L_{l1}(x,\hat{x}) + L_{vgg}(x,\hat{x}) \tag{4}$$

Our overall training schema is to split the dataset into multiple attributes. Using the CelebA-HQ dataset, we take the wj of the attributes. The textures are then decomposed from the geometry of the annotated dataset to extract the textures. This gives us both the geometry and texture needed in our framework. We then have the real image, the full segmentation mask, and the five different attributes. In our framework, we use the segmentation mask as the original style and the five attributes as the styles to feed into the AdaIn. When we get the fake image, we compare it with the original real image using our losses.

Ⅳ. 실험 결과

To demonstrate the effectiveness of the proposed method, we use CelebA-HQ as it is annotated with the same colors without requiring other methods to generate the segmentation masks. In training, we use 15,000 images of the first half of the CelebA-HQ dataset with a dimension of 256256. We then use a subset of the CelebA-HQ to validate our proposed method to evaluate our networks on the same dimensions. The subset of the dataset consists of 1,000 segmentation masks and 5,000 segmented images for a total of 6,000 images. The training and validation set are randomly separated from Celeba-HQ, so there is no overlapping subject ID, but statistics (age, gender) are similar.

1. Experimental settings

We first use the segmented data from CelebA-HQ to extract the desired attributes. We then train our model using a single RTX A6000 on a batch size of 16 with a dimension of 256256 on attributes of the same person. We call the data of a list attributes of the same person as paired data. While attributes of a list of different people as unpaired. We use a learning rate of 0.0002 using the Adam optimizer for 50 epochs.

2. Reconstruction

To evaluate our model, we perform a paired reconstruction, meaning the data belong to the same identity of our geometry and attributes. We can see in Figure 2 that our model can reconstruct paired images synthesizing images to the ground truth image. Except for accessories, our model can recreate images using the latents learned by each encoder.

We can also edit the segmentation mask using any image editing tools to reconstruct the image using a newly edited segmentation image. This allows image editing to more minor details. As seen in Figure 3, we edit the segmentation mask of the hair, allowing it to be shorter, the mouth to be more pursed, and the eyes to be smaller.

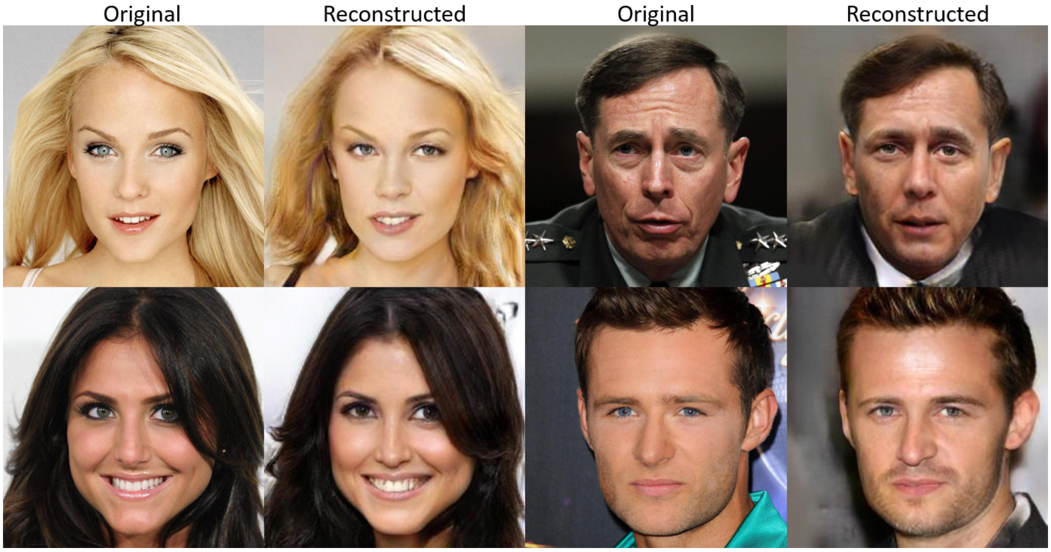


그림 2. 제안 기법을 이용한 페어 기반 이미지 합성 결과
Fig. 2. Results of our work on paired image synthesis

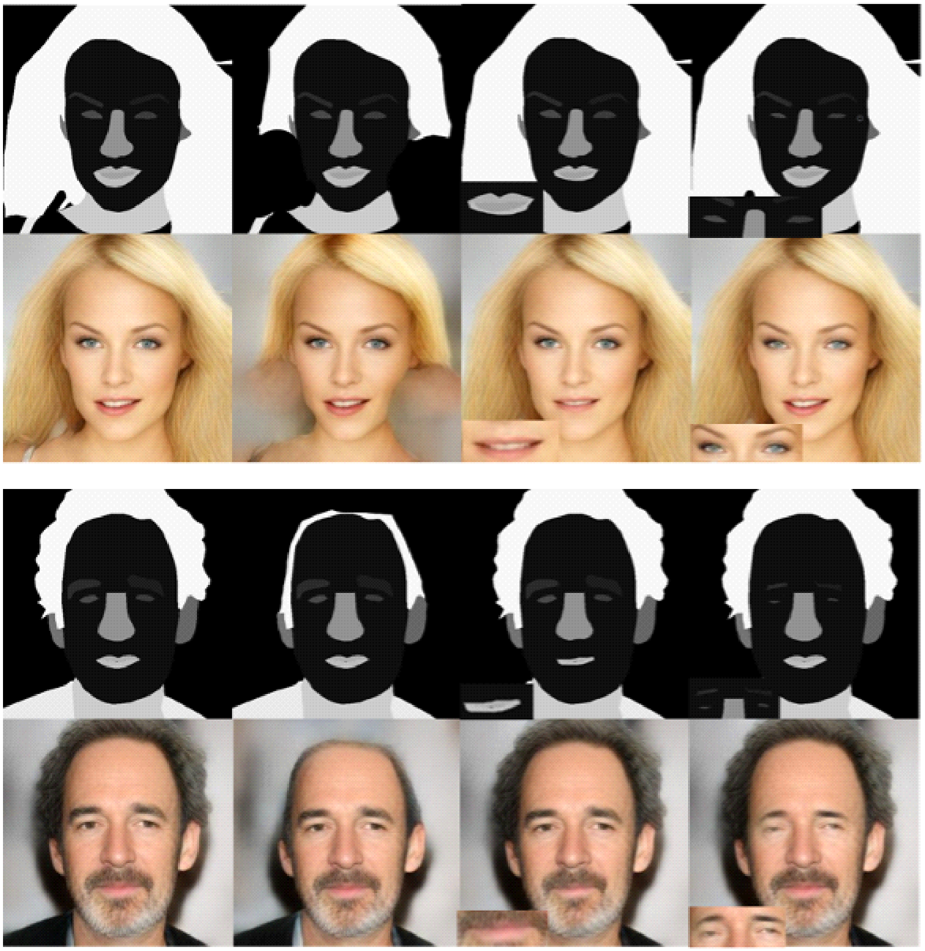


그림 3. 제안 기법을 활용한 얼굴 분할 마스크 변경에 따른 이미지 편집 결과
Fig. 3. Image editing result according to the change of the facial segmentation masks

Table 1 shows our quantitative results, using Frechet Inception Distance (FID) to calculate the distance score between the real dataset and our generated dataset. LPIPS is used to calculate the similarity score between the reals and reconstructed images, while MS-SSIM is used to calculate the structural similarity of the images. As our evaluation metrics. As our work is a conditional GAN that is able to recreate images, we use LPIPS and SSIM to evaluate the similarities of the image. We can see that although we have high FID scores compared to other methods[8], we do not consider background while training our images; we have low results on our LPIPS and high marks on our MS-SSIM. QSNGAN[20] works with unconditional GANS using the Hamilton Product, ours use a more classic approach with the AdaIn, resulting in a lower FID score. Unfortunately as QSNGAN is unconditional, we are unable to get LPIPS and MS-SIM scores.

표 1. 다양한 평가 지표에 대한 정량적 결과
Table 1. Quantitative results across various evaluation metrics

Evaluation metric	FID↓	LPIPS↓	MS-SSIM↑
QSNGAN[20]	29.41	-	-
Ours	25.98	0.285	0.745

3. Texture Transfer

We then test our model on an unpaired segmentation mask to transfer the texture between different attributes belonging to varying identities from the dataset. As seen in Figure 4, our framework can successfully transfer the texture to edit the face. Transferring texture only requires the segmented attributed of the required texture to be switched with the original segmented texture. This allows our method to only change the location needed, interfering with the rest of the attributes.

Our model can reconstruct images using paired attributes. But it is also able to generate images with different attributes. As seen in Figure 4, while the attributes of the paired reconstruction work well, we can change specific attributes. Figure 4 shows the results of changing the attribute to darker skin color and when we change the attributes. We can see that we can edit specific attributes while keeping the rest intact.

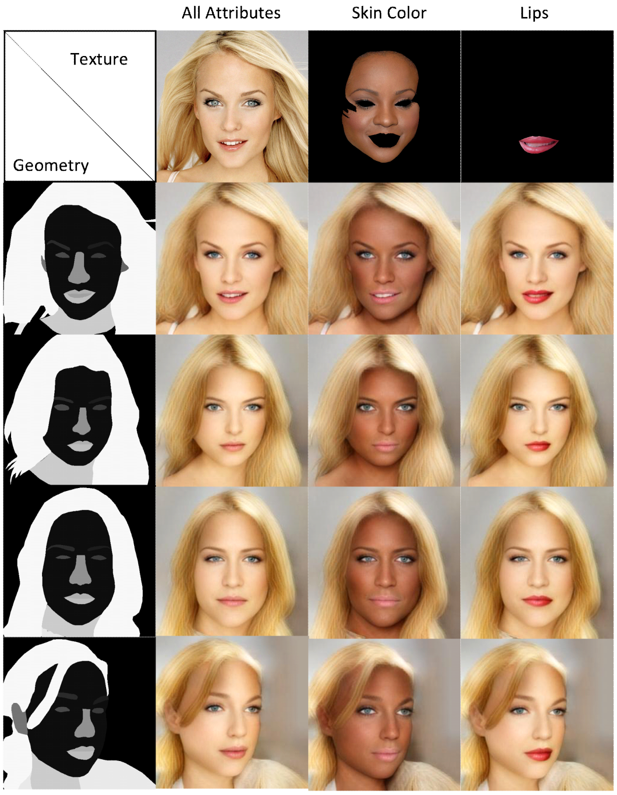


그림 4. 원하는 속성만 변경할 수 있는 텍스처 전이 결과
Fig. 4. Result of texture transfer where only desired target attributes can be changed

4. Unpaired reconstruction

We test our model on an unpaired segmentation mask to transfer the texture between different attributes belonging to varying identities from the dataset. This can be used to create a more extensive dataset using unpaired attributes. Figure 5 shows that while our method was trained on paired data, our approach can achieve the unpaired image reconstruction showing results of similar quality between paired and unpaired data.

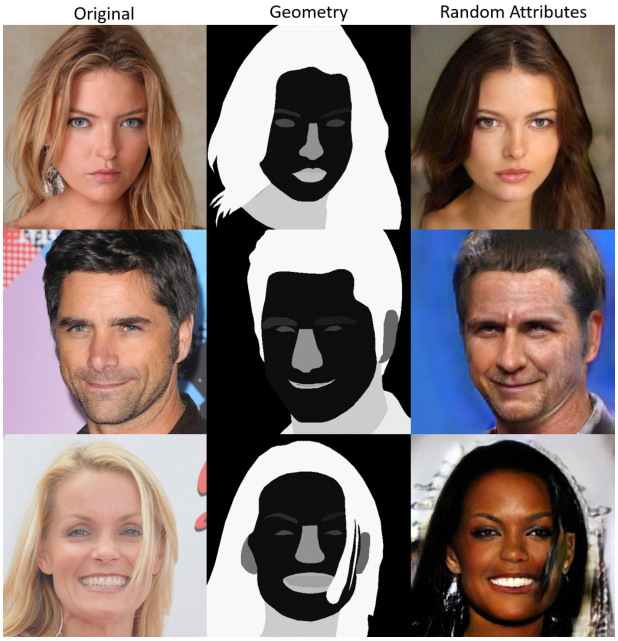


그림 5. 입력 이미지의 속성을 임의의 다른 속성으로 변경할 수 있는 페어 이미지가 필요 없는 이미지 합성 결과 (왼쪽에서부터 순서대로 실제 이미지, 분할 마스크, 합성 결과 이미지)
Fig. 5. Unpaired image synthesis result (left: real image, center: segmentation mask, right: fake image) where the specific attributes of the input image can be changed to any arbitrary attribute

Using unpaired data, we can see in Figure 5 that unpaired image synthesis allows us to generate a larger dataset. We can sample the attributes using the same geometry and create new facial images. Since we can fill in the geometry using our unpaired method, we can also use the geometry to learn segmentation tasks. By reusing geometry with different textures, we can use the data to learn tasks that require facial geometry such as face segmentation and face parsing. As a result, we are able to generate face images with varied attributes using a small number of real face images and facial masks in an unpaired image-to-image translation manner.

V. 결 론

This paper proposes a generic framework using segmentation masks and segmented textures. With interchangeable inputs and facial masks, our framework can edit specific attributes of the image. Concretely, we leverage AdaIn to transfer the style between geometry and texture. Our experiments on face editing show that our method can learn the location of the attributes well enough only to edit the desired attribute. Additionally, although we trained using paired images, our unpaired image synthesis can perform well enough to generate new facial images by using only a few numbers of training images. By only needing limited amounts of data, the method learns where to map each texture to the corresponding geometry. And these results being able to reuse geometry and randomly applied attributes to generate a larger amount of segmentation data. Although our method can work to generate new images based on segmentation masks, it is also possible for our method to edit images while keeping the textures intact, resulting in applications such as face editing to be possible.

참 고 문 헌 (References)

[1] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets," Proceedings of Advances in Neural Information Processing Systems, 2014.
doi: https://doi.org/10.48550/arXiv.1406.2661

[2] Wang, X. and Gupta, A., “Generative image modeling using style and structure adversarial networks,“ Proceedings of European Conference on Computer Vision, pp. 318-335, 2016.
doi: https://doi.org/10.1007/978-3-319-10578-9_24

[3] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434, 2015.
doi: https://doi.org/10.48550/arXiv.1511.06434

[4] Choi, Yunjey, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8789-8797, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00916

[5] Zhu, Yuhao, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. "One shot face swapping on megapixels," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4834-4844, 2021.
doi: https://doi.org/10.1109/CVPR46437.2021.00480

[6] Korshunova, Iryna, Wenzhe Shi, Joni Dambre, and Lucas Theis. "Fast face-swap using convolutional neural networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 3677-3685, 2017.
doi: https://doi.org/10.1109/ICCV.2017.397

[7] Xu, Yangyang, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. "High-resolution face swapping via latent semantics disentanglement," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7642-7651, 2022.
doi: https://doi.org/10.48550/arXiv.2203.15958

[8] Gu, Shuyang, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. "Mask-guided portrait editing with conditional gans," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3436-3445, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00355

[9] Zhu, Peihao, Rameen Abdal, Yipeng Qin, and Peter Wonka. "Sean: Image synthesis with semantic region-adaptive normalization," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5104-5113, 2020.
doi: https://doi.org/10.48550/arXiv.1911.12861

[10] Kim, Hyunsu, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. "Exploiting spatial dimensions of latent in gan for real-time image editing," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 852-861, 2021.
doi: https://doi.org/10.48550/arXiv.2104.14754

[11] Zhou, Hang, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang.

"Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5911-5920, 2020.
doi: https://doi.org/10.48550/arXiv.2003.08124

[12] Wang, Ting-Chun, Arun Mallya, and Ming-Yu Liu. "One-shot free-view neural talking-head synthesis for video conferencing," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10039-10049, 2021.
doi: https://doi.org/10.1109/CVPR46437.2021.00991

[13] Vitadhani, Agastya, Kalamullah Ramli, and Prima Dewi Purnamasari. "Detection of Clickbait Thumbnails on YouTube Using Tesseract-OCR, Face Recognition, and Text Alteration," Proceedings of International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), pp. 56-61, 2021.
doi: https://doi.org/10.1109/ICAICST53116.2021.9497811

[14] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784, 2014.
doi: https://doi.org/10.48550/arXiv.1411.1784

[15] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125-1134, 2017.
doi: https://doi.org/10.1109/CVPR.2017.632

[16] Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "High-resolution image synthesis and semantic manipulation with conditional gans," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8798-8807, 2018.
doi:https://doi.org/10.1109/CVPR.2018.00917

[17] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2414-2423, 2016.
doi:https://doi.org/10.1109/CVPR.2016.265

[18] Huang, Xun, and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization," Proceedings of the IEEE International Conference on Computer Vision, pp. 1501-1510, 2017.
doi: https://doi.org/10.48550/arXiv.1703.06868

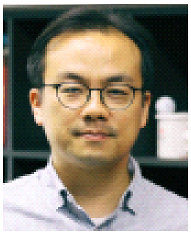
[19] Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation," Proceedings of International Conference on Learning Representations, 2018.
doi: https://doi.org/10.48550/arXiv.1710.10196

[20] Grassucci, Eleonora, Edoardo Cicero, and Danilo Comminiello. "Quaternion generative adversarial networks," In Generative Adversarial Learning: Architectures and Applications, pp. 57-86. Springer, 2022.
doi: https://doi.org/10.1007/978-3-030-91390-8_4

저자소개

Laudwika Ambardi

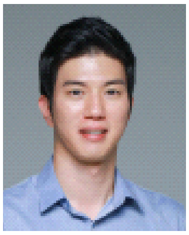
- 2020년 5월 : Universitas Bina Nusantara, Computer Science
- 2020년 9월 ~ 현재 : 인하대학교 전기컴퓨터공학과 석박통합과정
- ORCID : https://orcid.org/0000-0002-7892-8066
- 주관심분야 : Face Generation, Face 3D Modeling, Deep Learning



박인규

- 1995년 2월 : 서울대학교 제어계측공학과 학사
- 1997년 2월 : 서울대학교 제어계측공학과 석사
- 2001년 8월 : 서울대학교 전기컴퓨터공학부 박사
- 2001년 9월 ~ 2004년 2월 : 삼성종합기술원 멀티미디어랩 전문연구원
- 2007년 1월 ~ 2008년 2월 : Mitsubishi Electric Research Laboratories (MERL) 방문연구원
- 2014년 9월 ~ 2015년 8월 : MIT Media Lab 방문부교수
- 2018년 7월 ~ 2019년 6월 : University of California, San Diego (UCSD) 방문학자
- 2004년 3월 ~ 현재 : 인하대학교 정보통신공학과 교수
- ORCID : http://orcid.org/0000-0003-4774-7841
- 주관심분야 : 컴퓨터비전 및 그래픽스 (영상기반 3차원 형상 복원, 증강현실, computational photography), GPGPU

저자소개



홍성은

- 2010년 2월 : 한양대학교 컴퓨터공학과 학사
- 2012년 8월 : 카이스트 컴퓨터공학과 석사
- 2018년 2월 : 카이스트 컴퓨터공학과 박사
- 2018년 1월 ~ 2020년 8월 : SK telecom T-Brain, AI Center 연구원
- 2020년 9월 ~ 현재 : 인하대학교 정보통신공학과 교수
- ORCID : https://orcid.org/0000-0003-1774-9168
- 주관심분야 : Domain Adaptation, Face Recognition, Face Manipulation, Video Object Segmentation