Facial mask region completion using StyleGAN2 with a substitute face of the same person

Hiroaki Koike and Norihiko Kawai[0000−0002−7859−8407]

Faculty of Information Science and Technology, Osaka Institute of Technology
1-79-1 Kitayama, Hirakata, Osaka 573-0196, Japan
norihiko.kawai@oit.ac.jp

Abstract. In recent years, there has been a worldwide outbreak of coronaviruses, and people are wearing facial masks more and more often. In many cases, people wear masks even when taking photos of themselves, and when photos with the lower half of the face hidden are uploaded to web pages or social networking sites, it is difficult to convey the attractiveness of the photographed persons. In this study, we propose a method to complete the masked region in a face using StyleGAN2, a kind of Generative Adversarial Networks (GAN). In the proposed method, we prepare an image of the same person who is not wearing a mask, and change the orientation and contour of the face of the person in the image to match those of the target image using StyleGAN2. Then, the image with the changed orientation is combined with the target image in which the person is wearing the mask to produce an image in which the mask region is completed.

Keywords: Image Inpainting · Image Completion · GAN · Facial Mask.

1 Introduction

A coronavirus pandemic is sweeping the world, endangering many lives. In order to prevent the spread of the infection, people are asked to refrain from going out unnecessarily and to wear facial masks in public places. As a result, people wear masks even when taking photos of themselves. Therefore, when a photo with the lower half of the face hidden is uploaded to a web page or social networking site, it is difficult to convey the attractiveness of the photographed person. For this reason, if we can generate images with unmasked faces, we can convey the attractiveness of the persons even when taking photos while wearing their masks.

One technique that can be used for this purpose is image inpainting/completion, which has been aimed at plausibly filling in unwanted regions in an image. As image completion approaches, diffusion-based [1, 3], patch-based [2, 4, 6, 9, 10], and machine learning-based [5, 12, 13, 15] methods have been proposed. Since the diffusion-based methods propagate colors to missing regions, they cannot reproduce texture in the missing regions. The patch-based methods search for similar patterns in the image and synthesizes them in missing regions. However, it cannot synthesize textures that do not exist in other regions in the image. For

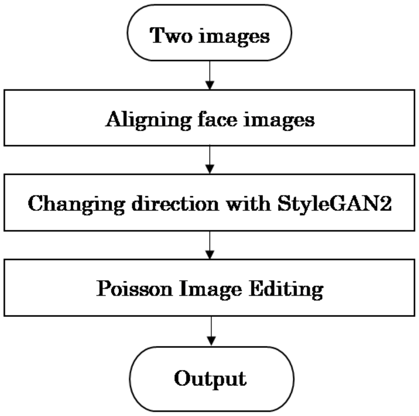


Fig. 1: Flow of the proposed method.

the application of facial mask area completion, these methods cannot be used because the texture of the nose and mouth cannot be properly reproduced.

On the other hand, the machine learning-based image completion methods using Generative Adversarial Networks (GANs), a type of machine learning, understand the structure of face regions and provide good results for face completion. For example, there are studies that introduced edges [13], segmentation [12], and landmarks [15] to improve the consistency of face structure and texture. In addition, the method in [5] automatically identifies and completes facial mask regions from supervised learning. However, since the nose and mouth are completed depending on the face images used for training, the completed face image does not always have the characteristics of the person.

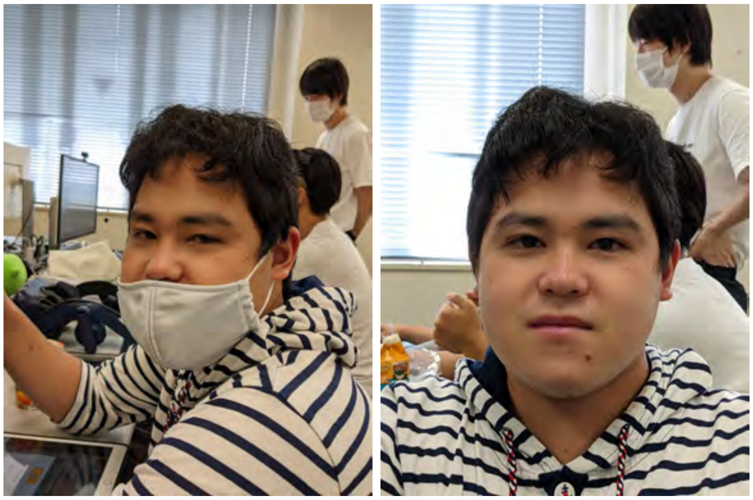
In this study, we prepare an image of the same person who is not wearing a mask, and change the orientation and contour of the face of the person in the image using StyleGAN2[7, 8]. We also change the face expression according to the user's preference. Then, the image with the changed orientation and contour is combined with the target image in which the person is wearing the mask to produce an image in which the mask region is completed. This approach completes the mouth and nose with the characteristics of the person.

2 Proposed method

2.1 Overview

The processing procedure of the proposed method is shown in Fig. 1. The proposed method takes two input images, one with a facial mask and the other without a facial mask. Fig. 2(a) shows example images wearing a facial mask, and Fig. 2(b) shows an example image of the same person without a facial mask.

Next, as a preprocessing step, the faces in the prepared images are aligned so that the parts such as the mouth and nose are in the specified positions. Next, StyleGAN2 is used to estimate the latent variable for the preprocessed image, and generate the image from the estimated latent variable. We then use Style Mixing, one of the features of StyleGAN2, to deform the face of the unmasked image to match the orientation of the face wearing the mask. According to the user's preference, the face expression is also changed by inputting another image with a different expression. Finally, by using Poisson Image Editing[14], the image with the changed face orientation is combined with the image of the face wearing the mask to produce an image in which the masked part is completed. The detailed processing of the proposed method is described below.

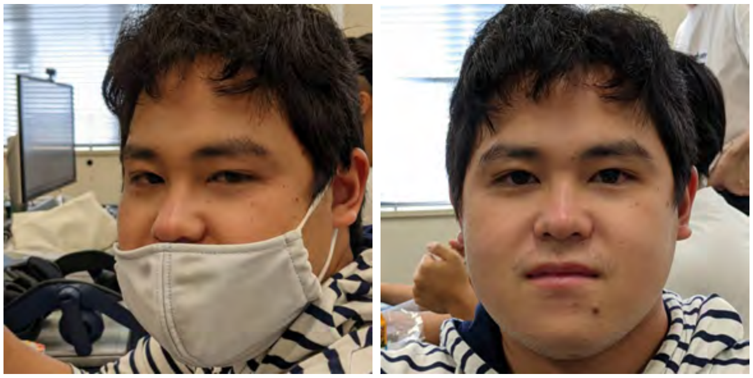


(a) With facial mask (b) Without facial mask

Fig. 2: Example input images.

2.2 Alignment of face images

The StyleGAN2 model used in this study was trained and generated on the FFHQ dataset, which contains a large number of face images, and the landmarks of the faces were aligned to the same position for training. Therefore, if the images are not aligned in the same way before inputting them into StyleGAN2, the quality of image generation will be low. We use Dlib[11], a machine learning library, to detect landmarks on the face and align them. The image is rotated so that the eyes are at the same height, and the center coordinates are obtained from the positions of the eyes and the mouth, and the image is cropped to form a square. Fig. 3 shows the aligned images of Fig. 2.



(a) With facial mask (b) Without facial mask

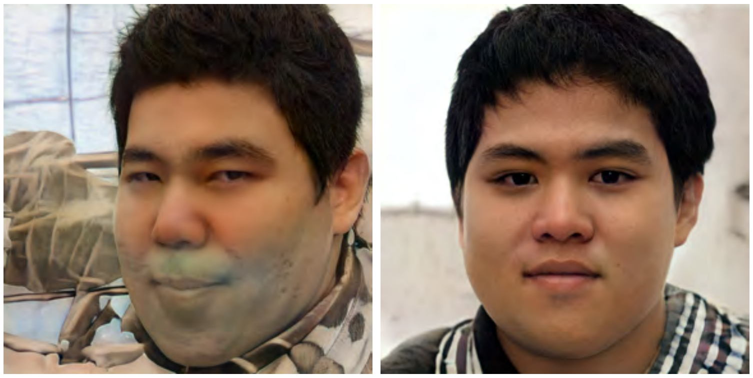
Fig. 3: Aligned example input images.

2.3 Processing by StyleGAN2

In StyleGAN2, by inputting a latent variable, which is a point in the latent space where various elements of the training data are distributed, the corresponding image is generated. In order to edit a face image using StyleGAN2, the face image to be edited must be generated by StyleGAN2. To do this, we first estimate the latent variable of an image that is close to the prepared image from the latent space, and then generate an image using the estimated latent variable. Then, we perform Style Mixing on the generated image to change the orientation and contour of the face.

Estimation of latent variable Estimation of the latent variable is done by inputting random values and iterating the process so that there is no difference from the image to be estimated. Fig. 4 shows the images generated from the latent variable estimated by searching for images close to the images in Fig. 3 300 times each. It should be noted that StyleGAN2 is trained on face images without facial masks. Therefore, even if we input images with facial masks such as Fig. 3(a), face images with masks removed are outputted as shown in Fig. 4(a).

Orientation and contour change by Style Mixing StyleGAN2 is a neural network that generates images by gradually increasing the resolution from 4×4 images to 1024×1024 images. In this process, latent variables are input as style information to generate images for different resolutions, and each resolution has a different effect. StyleGAN2 allows different latent variables to be input at different resolutions, and by inputting two latent variables at different resolutions, it is possible to generate an image that mixes the features of the two images.



(a) From Fig. 3(a) (b) From Fig. 3(b)

Fig. 4: Images generated from the estimated latent variable.

Specifically, we mainly control the orientation and contour of the face in the low resolutions, and the hair and skin color in the high resolutions.

In this study, since we want to change the orientation of the face in Fig. 4(b) to match Fig. 4(a), we input the latent variable that generates Fig. 4(a) when generating images with low resolutions such as 4x4 and 8x8 resolutions, and the latent variable that generates Fig. 4(b) when generating images with the other resolutions. This reflects the features in Fig. 4(b) in the generation of the face orientation and contour. The output result from the two latent variables is shown in Fig. 5. The masked region in Fig. 3(a) are completed using this result.

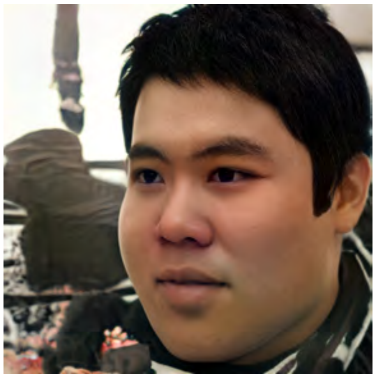


Fig. 5: Result of style mixing using Fig. 4.

Facial expression change by Style Mixing In addition to the orientation and contour, we can also change the facial expression. Specifically, we input the latent variable for generating a face image with some expression such as Fig. 6 with smile when generating a medium 16x16 resolution image, and the latent variable that generates Fig. 5 for the other resolutions. By doing this, the feature of the mouth is reflected, and the face shown in Fig. 5 becomes a smile as shown in Fig. 7. It should be noted that images with different facial expressions do not necessarily have to be images of the same person.

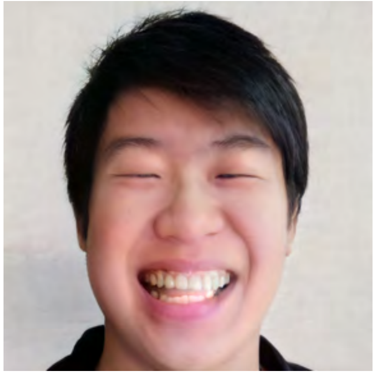


Fig. 6: Smile face.

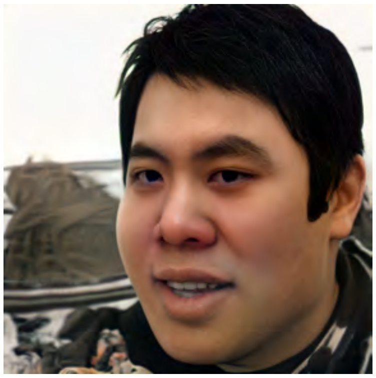


Fig. 7: Result of style mixing using Figs. 4 and 6.

2.4 Completion with Poisson image editing

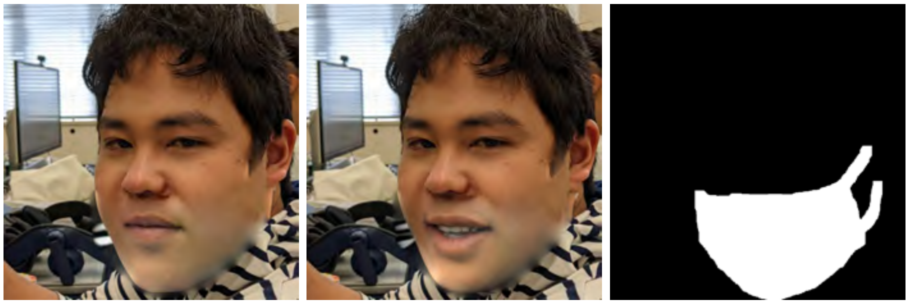
When Fig. 5 is simply merged with the mask in Fig. 3(a), the different color tones of the images make the boundary of the merged region clear and unnatural. Therefore, we use Poisson image editing to create a seamless image composition.

In Poisson image editing, the pixel values after composition are calculated to minimize the squared error of the gradient between the destination image and the source image. Concretely, f_p is obtained such that the cost function E shown in Equation (1) is minimized.

$$E = \sum_{(p,q)\in N} ((f_p - f_q) - (g_p - g_q))^2, \tag{1}$$

where f is the pixel value of the destination image, g is the pixel value of the source image, p is the pixel of interest, and q is its neighboring pixel. By doing this, the color tone of the composite region is made to resemble the destination image, and the texture is made to resemble the source image.

Fig. 8 shows the results of combining the images of Figs. 5 and 7 into Fig. 3(a) by Poisson image editing for the specified mask area shown in Fig. 8(c). As shown in the figure, the color tones are corrected to match the destination image.



(a) From Fig. 5 (b) From Fig. 7 (c) Mask region

Fig. 8: Results of Poisson image editing.

3 Experiments

To demonstrate the effectiveness of the proposed method, we prepared face images of persons wearing a mask and frontal images of the same persons as shown in Fig. 9. In this study, we conducted two types of experiments. In the first experiment, we checked the results of completing the mask regions and compared them with the actual face images with the mask removed. In the second experiment, we estimated latent variables in the images showing various expressions and verified the effect of facial expression changes using them.

3.1 Experiment 1

Fig. 10 shows the completion results of the mask regions in Fig. 9 and the actual face image at that angle. The experimental results show that the orientations and contours of the faces are properly modified in all the images. In the case of Persons 1, 2, and 3, the color tones of the faces are properly adjusted by Poisson image editing, and the boundary between the completed area of the facial mask region and the other region in the face is not noticeable. On the other hand, If the quality of the prepared image is poor, as in the image of Person 4, the boundary between the area where the mask is completed and the original image is noticeable.

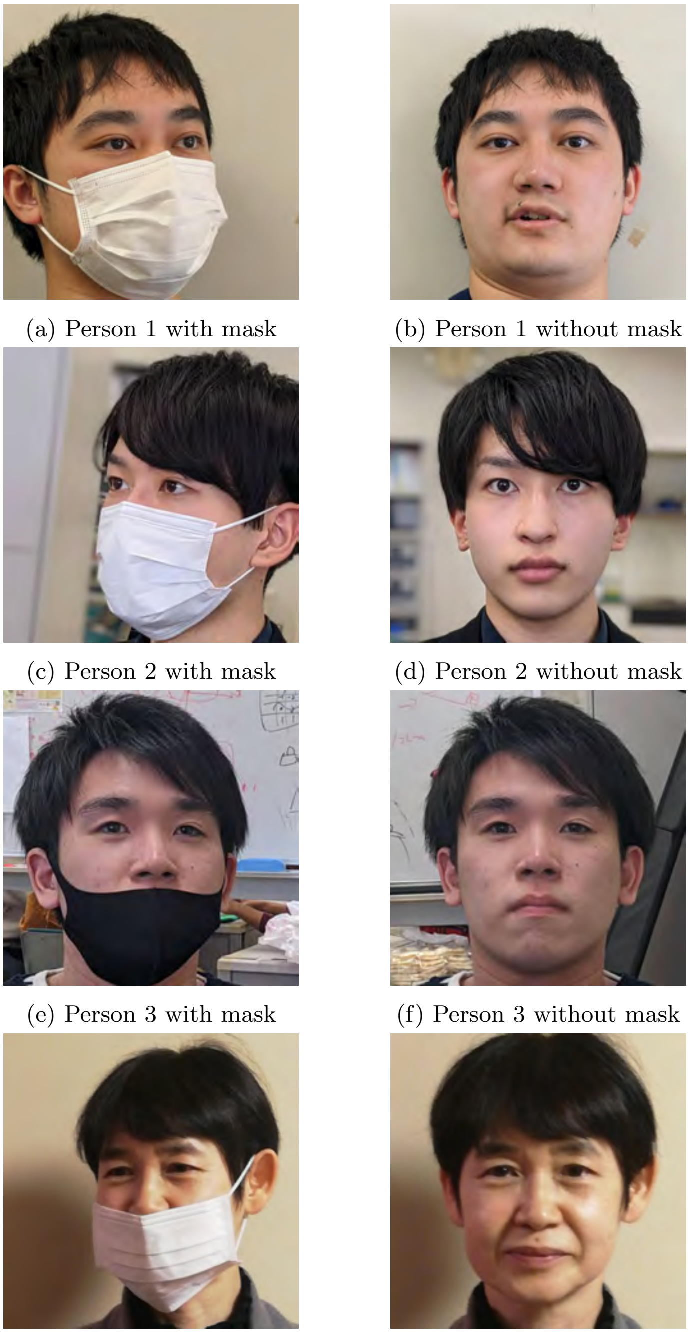
Next, we compare the completion result of mask regions with the images in which the persons actually took off the facial masks. In the case of Person 1, the shape of the mouth, cheek bulge, and beard as shown in Fig. 9(b) are reflected in Fig. 10(a). However, features of moles were not represented when projecting the real image into the latent space and the moles disappeared. In the case of Person 3, the texture of the skin is reflected, and there are no major differences when compared to the actual person.

In the case of Person 2, the noses and mouths are misaligned in Figs 10(c) compared to Fig. 10(d), indicating that the parts are synthesized on the front side of the face, rather than the positions of the parts corresponding to the orientation of the actual face. This is because most of the nose in Fig. 9(c) is hidden by a facial mask, making it difficult for StyleGAN2 to estimate the orientation of the face. In fact, in the training of StyleGAN2, there are not many images in which the faces are largely turned sideways. Therefore, when the latent variables are estimated and the face images are created, the orientation of the nose is different from the actual one as shown in Fig. 11.

Another difference is that the completion images in Fig. 10 appears to have slightly longer chins than the actual images. The reason for this is that when the latent variable corresponding to the image with the facial mask is worn is found in the latent space, the image generated from the latent variable tends to have a wider and longer chin due to the puffiness of the facial mask. In style mixing, not only the orientation of the face but also the contour of the face is affected, resulting in the generation of a completion result with a long chin. An improvement would be to use some kind of image inpainting to complete the mask area in advance, and then use that image for style mixing.

3.2 Experiment 2

In this experiment, we examine the effect of facial expression change. Fig. 12 shows images representing various facial expressions used for expression change. Fig. 13 shows the completion results using the latent variables estimated from



(a) Person 1 with mask

(b) Person 1 without mask

(c) Person 2 with mask

(d) Person 2 without mask

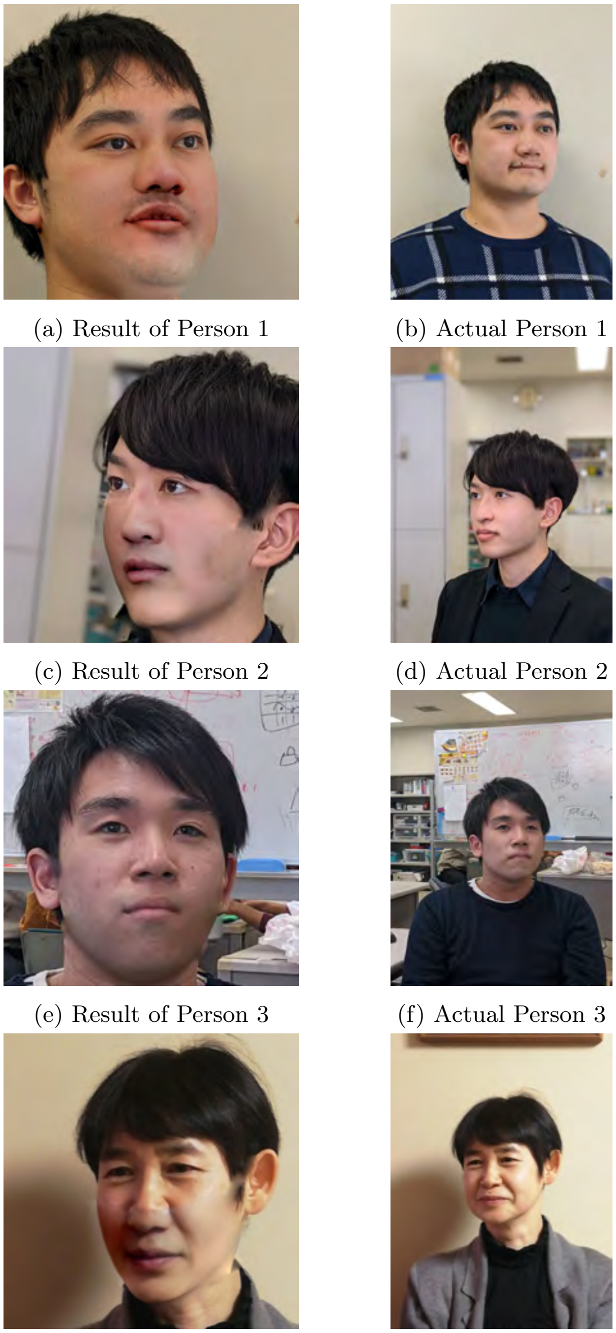
(e) Person 3 with mask

(f) Person 3 without mask

(g) Person 4 with mask

(h) Person 4 without mask

Fig. 9: Input images for experiments.



(a) Result of Person 1

(b) Actual Person 1

(c) Result of Person 2

(d) Actual Person 2

(e) Result of Person 3

(f) Actual Person 3

(g) Result of Person 4

(h) Actual Person 4

Fig. 10: Completion results and comparison with the actual faces.

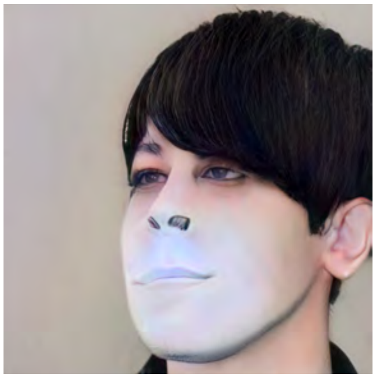


Fig. 11: Image generated from Fig. 9(c) by estimating its latent variable.

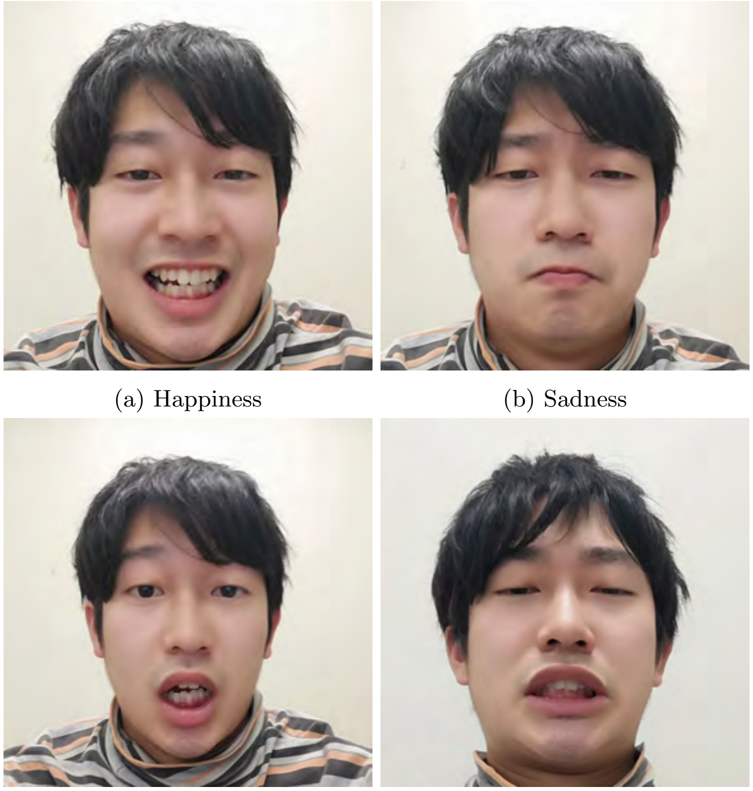
these images with various expression for Figs. 10(a) and (e), which were relatively successfully completed in Experiment 1.

The results of the experiment show that the mouths of both Persons 1 and 3 changed significantly in response to the expressions of happiness, surprise, and disgust, and that the mouth of Person 1, whose mouth was originally slightly open, closed in response to sadness. The vertical and horizontal openings for happiness, sadness, surprise, and disgust were reflected, but subtle changes such as the lowering of the corners of the mouth in Fig. 12(b), which was used for the sadness image, were not. In general, not only changes in the mouth but also changes in the eyes can be combined to recognize the entire facial expression. However, since this study focuses on complementing the facial mask region, we believe that the two types of facial expressions, happiness and surprise, which can be easily recognized by changes in the opening and closing of the mouth alone, are practical based on the above results.

4 Conclusion

In this study, we proposed a mask region completion method using StyleGAN2 and another substitute image of the same person without a facial mask. In the proposed method, successful completion was achieved by correcting the orientation and contour of the face without a facial mask using StyleGAN2, correcting its color tones and synthesizing it using Poisson image editing. In addition, by estimating the latent variable of another image with a different expression and introducing it in the middle of style mixing, we changed the face expression.

In experiments, we showed that the completion with the features of the person can be achieved, but we also confirmed that small features such as moles are lost due to the low accuracy of latent variable estimation. As for the changes in facial expressions, we confirmed that the changes were reflected in the vertical and horizontal changes of the mouth, but not in the detailed features.


(a) Happiness (b) Sadness

(c) Surprise (d) Disgust

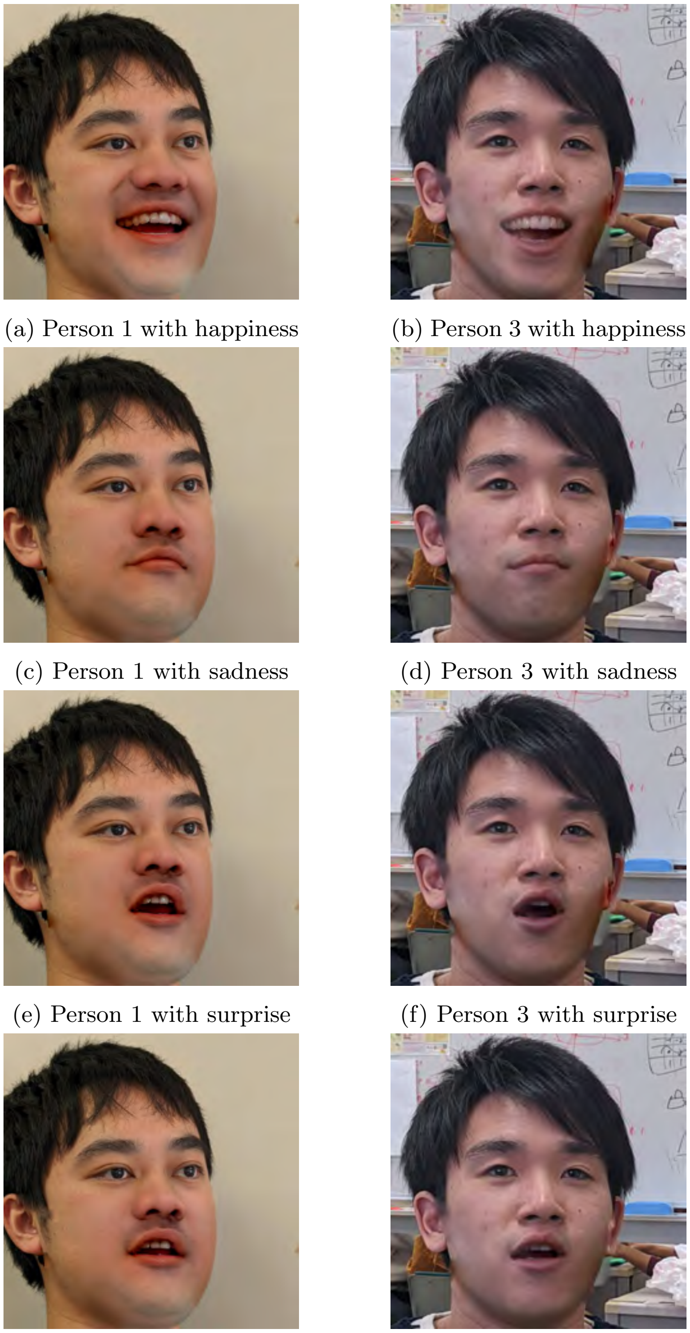
Fig. 12: Various expressions.

Future work includes the introduction of image inpainting to further improve orientations and contours.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers JP18H03273, JP18H04116, JP21H03483.

References

1. Ballester, C., Bertalmío, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing **10**(8), 1200–1211 (2001)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics **28**, 24:1–24:11 (2009)



(a) Person 1 with happiness

(b) Person 3 with happiness

(c) Person 1 with sadness

(d) Person 3 with sadness

(e) Person 1 with surprise

(f) Person 3 with surprise

(g) Person 1 with disgust

(h) Person 3 with disgust

Fig. 13: Input images for experiments.

3. Bertalmío, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of SIGGRAPH. pp. 417–424 (2000)
4. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing **13**(9), 1200–1212 (2004)
5. Din, N.U., Javed, K., Bae, S., Yi, J.: A novel gan-based network for unmasking of masked face. IEEE Access **8**, 44276–44287 (2020)
6. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of IEEE International Conference on Computer Vision. vol. 2, pp. 1033–1038 (1999)
7. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
8. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
9. Kawai, N., Sato, T., Yokoya, N.: Image inpainting considering brightness change and spatial locality of textures and its evaluation. In: Proceedings of Pacific-Rim Symposium on Image and Video Technology. pp. 271–282 (2009)
10. Kawai, N., Yokoya, N.: Image inpainting considering symmetric patterns. In: Proceedings of IAPR International Conference on Pattern Recognition. pp. 2744–2747 (2012)
11. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)
12. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. pp. 5892–5900 (2017)
13. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. pp. 3265–3274 (2019)
14. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on Graphics **22**(3), 313–318 (2003)
15. Yang, Y., Guo, X.: Generative landmark guided face inpainting. In: Proceedings of Chinese Conference on Pattern Recognition and Computer Vision. pp. 14–26 (2020)