Facial Image Unmasking with Imputed Convolution

Seongsu Ha[1] Sungchan Park[1] Hyunbin Kim[2]
Department of Data Science[1] Department of Biological Science[2]
Seoul National University
{sha17, warld2234, khb7840}@snu.ac.kr

Abstract

In this report, we present the application of Imputed Convolution. It could compensate some artifacts of Gated Convolution which is very effective to inpainting free-form mask. But, inpainting of large chunked area is quite a difficult challenge and Gated Convolution has some limitation to this task. Meanwhile, Imputed Convolution could show better performance to this task compared to Gated Convolution. It seems like explicit imputation makes it could unmask which is coherent to overall semantic. We present a method combining Imputed Convolution to GAN along with UNet segmentation network for extracting masked area from original raw masked facial image. Our method performs image inpainting end-to-end in a coarse-to-fine approach that many recent learning-based image inpainting methods employ. At the end of the report, we will compare imputed convolution with other methods including Gated Convolution and show it's effectiveness with the experiment result. Then, we will list some limitations of our method and what kind of future work should be done to develop it more.

1. Introduction

As part of an effort to slow the COVID-19 pandemic, facial masks became ubiquitous in public area. With the expansion of the quarantine policy, computer vision technology based on deep neural technology has begun to be applied to handle images and videos with facial masks.

Face mask detector was put into practice first due to urgent needs. As the pandemic prolonged and became commonplace, studies in different directions, such as face recognition or personal identification, emerged. One of the emerging topics is facial mask removal. Not only are social networks flooded with masked images, but there has also been an increased practical need for mask removal in specific cases, such as group photos of events or weddings.

N. Ud Din *et al.*[2] tackles the problem of removing masks from facial images with the image inpainting method. The paper first applies UNet image segmentation

Figure 1. Results on Training Image: Input, GT, Output

Figure 2. Results on Unseen Image: Input, Output. We need a way to edit only the mask area

over the input image to find the masked area in the face.[15] With this binary segmented output as an additional channel concatenated to the input image, their modified version of UNet GAN performs the image inpainting task. The key approach to note here is that they use two distinct discriminators. This is to follow the coarse-to-fine approach adopted by many current state-of-the-art inpainting methods; they make one discriminator gauge the quality of the output in terms of overall structure, while the other focuses on the mask area where detailed artifacts like lips and noses should be created. Furthermore, to ensure the identity between images before and after inpainting, they train their GAN using L1 pixel-wise loss and Perceptual Loss over feature maps extracted from pretrained VGGNet.[7] Many previous GAN-based methods show limitations where generated image becomes totally different from original image. To address this method, research including mentioned above tries to generate only the masked area's data, which would be quite unnatural due to semantic difference between original image and generated image

area. In an attempt to tackle this task from N. Ud Din *et al*.[2], they present the appliaction of StyleGAN2, Style-based Generator Architecture for Generative Adversarial Network[9], for the generator network a modified version of UNet was used regards to the paper. This approach helps produce more natural face restored within masked area as demonstrated in E. Richardson *et al*.[14]. But, still, contextual naturalness is quite one to accomplish.

Gated Convolution[16] could be very effective alternative. It learns implicitly how to weigh the relative importance of mask area's nearby pixels. Especially, it shows significant performance in inpainting free-form masks. But the larger mask area becomes, the lower it's performance goes.

Compared to the previous methods, the model in the paper produces more natural outputs with mask regions restored with faces. It could generate semantically coherent images using Imputed Convolution which recursively impute masked area's data using it's surrounding data.

The main contributions of this work are as follows:

1. Application of Imputed Convolution for largely masked area's inpainting

2. Comparison between Imputed Convolution and Gated Convolution, and Suggestion some relatively better aspects of Imputed Convolution

2. Related Work

2.1. High-Quality GAN & Latent Space Embedding

Generative Adversarial Network(GANs) has shown dramatic progresses since it was first introduced in the paper by Goodfellow *et al*.[4] in 2014. One of the most active research topic using GANs is to create high-resolution images. This was impossible to achieve until the publication of high-resolution facial image datasets such as FFHQ with Progressive Growing GAN[8]. With the success of generating high-quality images, the need to control over output images is raised for various tasks including image inpainting. Many astounding researches attempted to tackle this problem by controlling image's embedded latent space, the input to the generator. The most popular GAN architecture of this approach is probably StyleGAN. StyleGAN starts from a constant tensor to build images with higher resolutions as layers get deeper. The key contribution of StyleGAN is the control over the output by adding latent space embedding code, $\mathcal{W}$, into each layer of the generator affecting the style of the image as a result. These style vectors help the network avoid being biased towards images in the training dataset, which was the issue of previous GANs, so many flexible image translation tasks were made available afterwards[10].

2.2. Image Segmentation & Inpainting

Facial mask removal, or unmasking, is as a task which needs both image segmentation and inpainting. Image segmentation has already reached technically sufficient accuracy even with smaller models, but image inpainting has not yet been as mature as detection. Before the dominance of deep learning, the researches on the image inpainting tasks have been largely conducted in two flows; patch-based and diffusion based approaches[3].

Patch-based image inpainting is an approach of filling the masked region by iterative search for best-matching patches for replacement, while diffusion-based approach fills the region inward from the boundary. With statistical and algorithmic improvements, state-of-the-art non-deep learning methods produce decent results.

However, these non-deep learning based algorithms have a limitation that visual semantics are not reflected in the process. This leads to two problems, one is that these methods do not perform well with large missing regions or images with complex textures, and the other is that errors occur when the algorithm tries to copy from patch of wrong objects.

By learning high-level structures through hidden layers, deep learning-based methods enabled image inpainting that reflects visual semantic information. Inpainting methods based on CNNs succeeded in reducing semantic errors by end-to-end learning of large scale training data, but the early CNN-based inpainting methods had mask-value-dependent visual artifacts occurring in the process of replacing the values with convolutional filters of fixed values. Liu *et al*.[11] proposed partial convolution operation as a solution for artifacts from hole placedholder values, which mask and re-normalize convolutions so that the convolution results should be dependent only on the valid pixels.

Another major artifact in CNN-based approaches is blurry results which tend to have smaller average pixel level difference. EdgeConnect, proposed by Nazeri *et al*.[12], reduced this type of artifacts with two GANs, one for predicting edge maps from the masked grayscale image and the other which uses edge map from the previous GAN and the raw masked images as inputs and generates inpainted images.

2.3. Gated Convolution & Imputed Convolution

Many recent learning-based image inpainting methods employs coarse-to-fine approach where two distinct discriminators are used to gauge the quality of overall structure and details independently. One of the recent models, DeepFill v2 [17] further developed the ideas of previous studies, using gated convolutions, which improved partial convolution layer to be trainable, and user-guided sketch channel that borrowed the idea of EdgeConnect. Gated convolution is a method of learning soft mask from the data by adding

Figure 3. First Row: original CelebA-HQ Facial images; Second Row: results applying MaskTheFace

an additional convolutional layer with a sigmoid function as an activation function. The output gating values between zeros and ones are element-wisely multiplied by the output of the feature convolutional layers to obtain the final output value. With gated convolution, the model can dynamically select features in each channel and location.

However, additional network modules doubles the number of parameters when applying gated convolutions, which leads to the need to reduce the size of the model. DeepPrivacy [5] proposed imputed convolution which can reduce the increase of number of parameters from doubling to 2%. Instead of adding a neuron dedicated to learn certainty, imputed convolution takes a similar architecture to the partial convolution while maintaining the certainty map learnable and imputing uncertain locations with weighted average of spatially close values.

3. Method

3.1. CelebA-HQ Facimal Image Dataset

In order to prepare pairs of facial images with and without masks, we applied MaskTheFace [1] module directly onto the CelebA HQ Facial iamge dataset from Kaggle.[13] Examples of images can be found in Figure 3. We synthesized a variety types of mask so that the model wouldn't be overfitted.

3.2. Binary Mask Segmentation

To input the mask information to GAN, we have to extract binary mask first from the original input image. This binary mask later will be inputted to generator part. Imputed convolution, or gated convolution require this to predict only the masked area.

Segmentation has been implemented using UNet, which is very effective for segmentation. The structure is depicted in Figure 4. We gave 256x256 RGB images as a input image, and the model outputs the same size of binary mask representing target area.

And we adopted Dice Loss, and it's calculated like below. p represents for predicted output, and t represents for

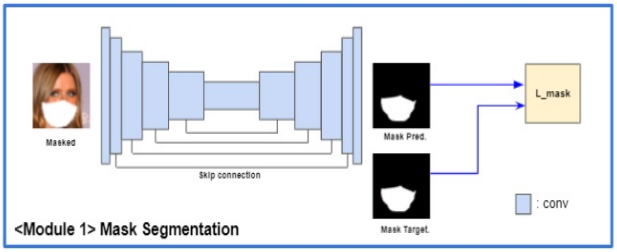

Figure 4. Model Structure - Module 1

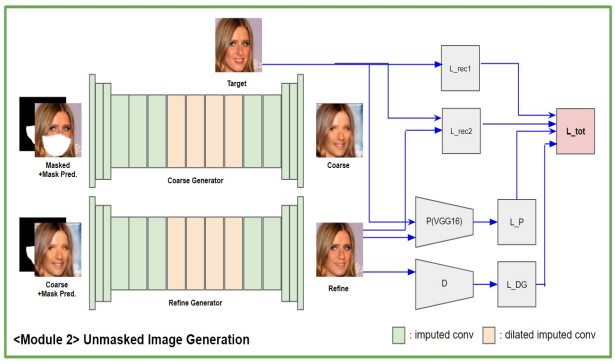

Figure 5. Model Structure - Module 2

target output.[6] It has it's lowest value of 0 when two outputs are exactly same, otherwise, it has value larger than 0.

$$D(Dice\ Loss) = 1 - \frac{2 * \sum_i^N p_i t_i}{\sum_i^N p_i^2 + \sum_i^N t_i^2} \quad (1)$$
$$p : predictedoutput, t : targetoutput$$

3.3. Imputed Convolution

Imputed convolution uses certainty map to represent the valid area. Each layer's input feature map is computed by weighted summing previous layer's input feature map and imputed value of invalid area(masked area) estimated by it's surrounding pixels' values. Certainty map determines the weight between these two.

In the paper about imputed convolution[5], authors insist that it could reduce the number of parameters comparing to gated convolution. While gated convolution has to learn weights about both gating and input feature, imputed convolution only needs to learn about certainty map, so the insist is reasonable. At the experiment part of this paper, they mentioned that imputed convolution even generated more semantically coherent results compared to previous solutions including gated convolution. We judged that it's because imputed convolution explicitly handles certainty map from end-to-end, which means it could distinguish whether values are valid or not, and it also explicitly implements

imputation while weighted summing input feature map and it's surrounding, which is meaningful for relatively large masked area to be coherent semantically with the whole image. Meanwhile, we couldn't specify gating that is determined within it's gated convolution operation. And we could say that it implicitly implements imputation while operating convolution for input feature map come from previous layer, which is also done in imputed convolution after explicit imputation.

With the reasons so far, we'd tried to implement network using imputed convolution. For that, based on the preceded gated convolution research[16], we replaced refinement network to the same one with coarse network to reduce the number of parameters and reduced the number of layers in discriminator. And we had checked it's overall performance didn't change significantly. After then, we replaced it's loss function following the previous research on unmasking task[2], which computes perceptual loss both with coarsely generated image and elaborately generated image, since it's important to weigh both on coarse result and refinement result as we impute masked area corresponding to the image's overall context.

We'd tried various activation functions(Tanh, LeakyReLU), normalization methods(layer normalization, pixel normalization) and whether use normalization to each encoder and decoder. Also tried for kernel sizes of explicit imputation used for weighted averaging input feature map and it's surroundings. Then, compared it to result from gated convolution.

3.4. Model Architecture

Overall model architecture is depicted in Figure 4 and Figure 5. It is composed with 2 modules. One is for mask segmentation. As mentioned in section 3.2, it uses UNet to do the task. Second is for unmasked image generation. It's structure is like Figure 5. After extracting binary mask from the first module, it is inputted to coarse network along with original masked image. Coarse network generates coarse image. And then, it is inputted to refine network with binary mask. In this stage, inputted coarse image is a bit manipulated. Non-masked area is replaced to original masked image so that refine network could do it's task without missing important semantic information. Final output is generated from refine network and it is also manipulated with original masked image in the same way.

Coarse network and refine network have exactly same structure. Each is consisted of imputed convolution blocks and dilated imputed convolution blocks. Dilated imputed convolution blocks' main purpose is to consider overall context of the whole compressed feature map so that the output could be semantically coherent.

Then loss is calculated. It is consisted of 4 parts. The firse one is discriminator loss(L_DG). Generated refine image goes through discriminator network and the loss is calculated using the output feature map of this network. Second one is perceptual loss(L_P). Generated refine image and target image is processed together within perceptual network whose structure is same with VGG-19 network. Third and Fourth one are reconstruction loss. They focus on relative similarity between target image and coarse image, and refine image, respectively. Finally, the total sum is calculated by weighted summing this four loss elements.

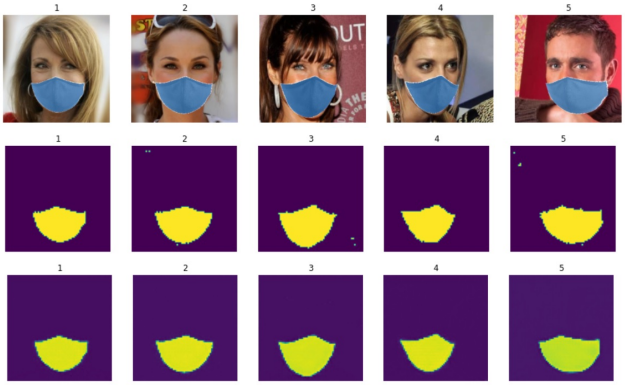

Figure 6. First Row: Original Image; Second Row: Target Image; Third Row: Predicted Image

4. Experiment

4.1. Binary Mask Segmentation

We traine UNet used for mask segmentation for 5 epochs with 30,000 of synthesized celeb-A facial masked dataset, and used learning rate of 0.001. Due to there are a variety types of mask in real world, we synthesized this dataset using different types of colors, textiles, and shapes.

The pixel-wise accuracy was about 99%. As depicted in Figure 6, the model segments the mask area very well.

4.2. Image Generation

We trained image generator part of the main model for 7 epochs with the same dataset above, and used learning rate of 0.001. We used LeakyReLU for the activation, and Layer Normalization. Using Tanh or using pixel-wise normalization make the model malfunction. The performance gets very bad with this choice. Also, eliminating normalization in the decoder part was very important. Just deleting it makes the performance much better. We concluded it's because we have to make detailed distribution we target in the decoder part, but with normalization, decoder couldn't build the detailed one properly.

Figure 7 9 shows the procedure of training. Each columns represent the target image, masked image with mask segmentation, coarse generator output, refine generator output, and the final output, which is combination of

refine generator output and the masked image. Later row is the one of later in training comparing to the former one.

At the beginning, it just output all blacked. At epoch 1, it became to recognize a vague shape of face. At epoch 2, it imputes the missing part very well, and refine image became much more detail than before. As Figure shown, coarse generator output predicts the overall image shape, and refine generator output concentrate on the detail, especially the one about masked area. As the train gone by, the model generates mostly similar output. But, as some samples shown, if some facial image have irregular facial expression or some special accessories, model cannot predict that kind of characteristic, but it rather generate the average facial expression. At the end, epoch 7, the model seems like be overfitted. So we stopped training at that stage and implement validation to check whether the model trained well or not.

The validation result is shown in Figure 10. The result was very good, and we could conclude the model was well-trained with this dataset.

4.3. Ablation Study

4.3.1 Comparison with Gated Convolution

We compared generated images of gated convolution and imputed convolution. In the referred paper of imputed convolution, Author insist that their method could generate more semantically coherent results compared to previous methods including gated convolution[5]. We experimented if their statement was right. The result is shown below. We carefully adjusted the kernel sizes of compositions. Especially, kernel size of filters for convolution of input feature map of two cases were controlled as the same, due to they would role as a implicit imputation.

As it depicts, imputed convolution shows better performance to understanding the whole context as we could see in the coarse generator output. The one of gated convolution seems like it just refer to the area nearby masked area, because the further pixels are blurred or colored very differently. But, the final result of two are both in good shape in this case. So it was a bit hard to say that imputed convolution shows vividly better performance than gated convolution. We might need some more experiments.

4.3.2 Comprison with StyleGAN2 using pSp Encoder

StyleGAN2 is very recent GAN model and is generating image of higher resolution gradually from lower resolution with it's latent space embedding code. And using this latent code is it's key contribution.

We wanted to compare our model to this popular network whether our one could outperform it. For StyleGAN2, pSp Encoder was used to make latent code from the original input image. pSp encoder is being widely used with it's capacity for encoding in various domains.

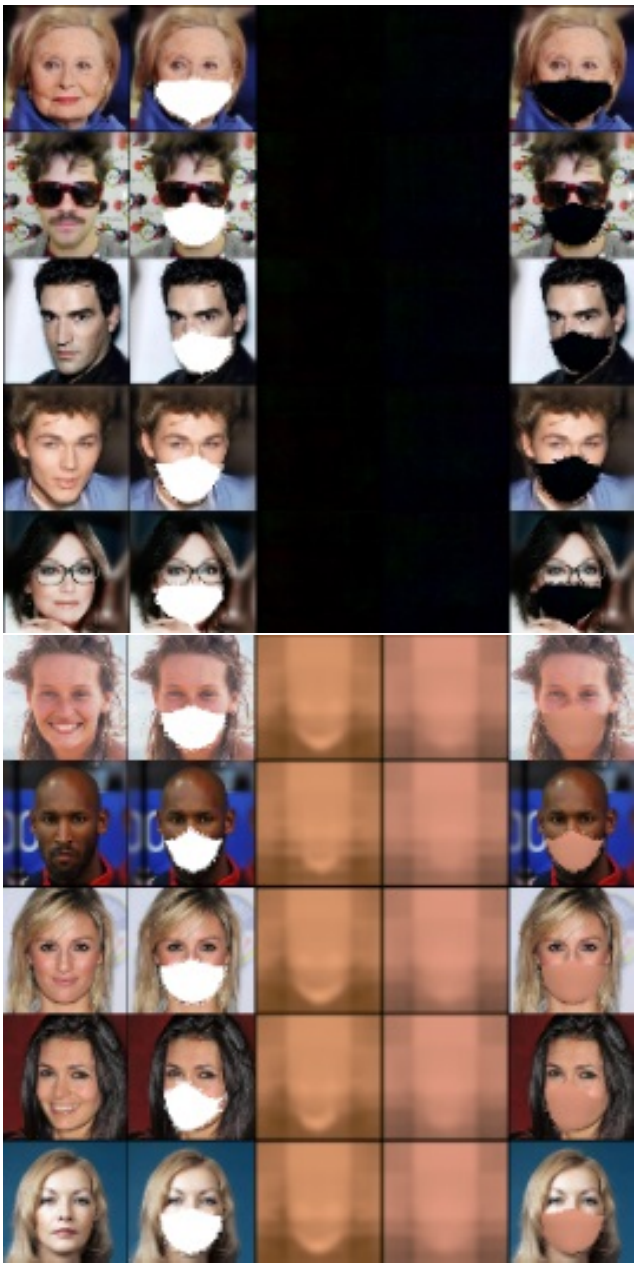

Figure 7. Epoch 1 2 training result

During training, StyleGAN2 discriminator was fixed, and only the encoder and generator was fine-tuned using the pairs of facial images with and without masks. As explained in the pSp paper, the encoded latent space shows robustness to the defective images. So we fine-tuned pSp Encoder initialized with pretrained model on the FFHQ dataset. The architecture of the network is described in Figure 4.

In the experiment section, we provide results applying loss functions given in the pSp paper, and also perceptual

Figure 8. Epoch 3 4 training result

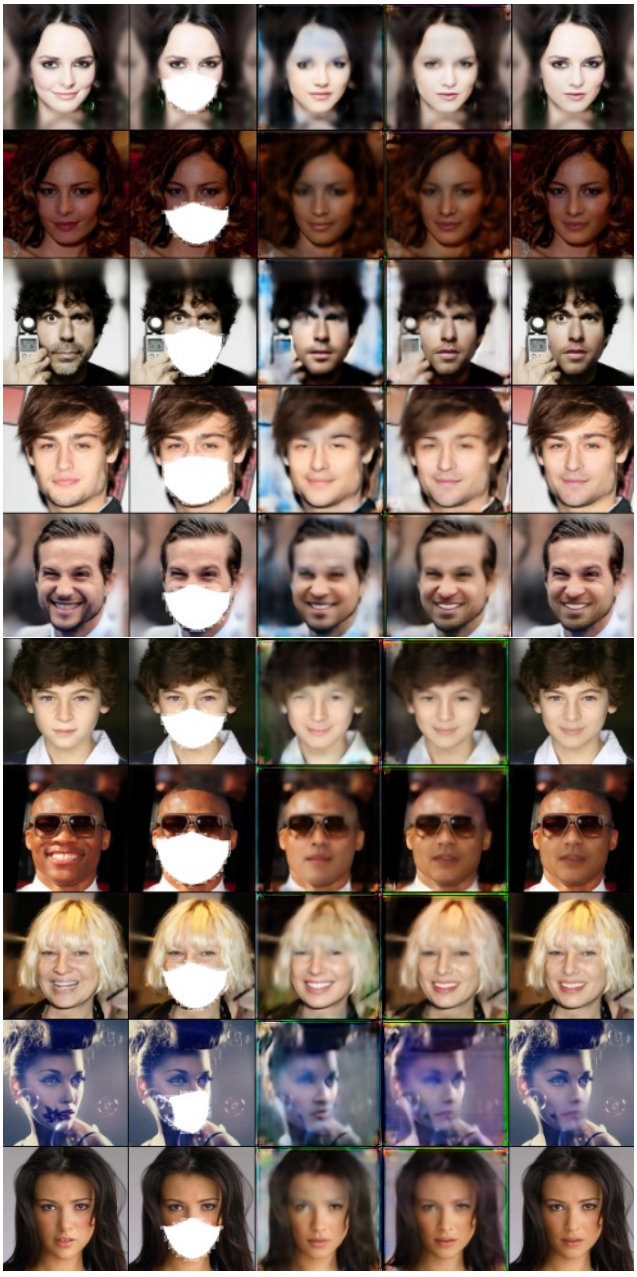

Figure 9. Epoch 5 7 training result

loss suggested in N. Ud Din *et al*. as well.

The loss function is the conjunction of three different losses as follows.

$$L(x) = \lambda_1 L_2(x) + \lambda_2 L_{LPIPS}(x) + \lambda_3 L_{ID}(x), \quad (2)$$

$$L_2(x) = ||x - pSp(x)||_2, \quad (3)$$

$$L_{LPIPS}(x) = ||F(x) - F(pSp(x))||_2, \quad (4)$$

$$L_{ID}(x) = 1 - < R(x), R(pSp(x)) > \quad (5)$$

L_2 loss measures pixel-wise distance. L_{LPIPS} measures perceptual loss by comparing the distance between feature maps from pretrained ArcFace Network. The pSp framework paper shows the better result with ArcFace network for Perceptual loss instead of VGGNet used in N. Ud Din *et al*.. Lastly, L_{ID} measures the identity loss between the input and output image.

Figure 10. Validation result

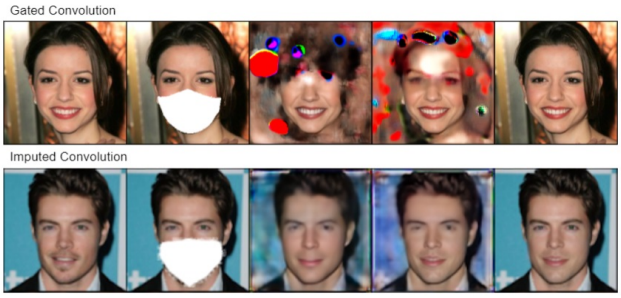

Figure 11. Comparison between Gated Convolution and Imputed Convolution

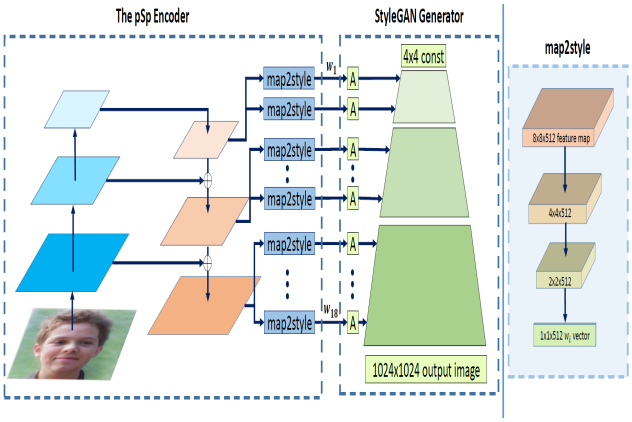

Figure 12. StyleGAN2 using pSp Encoder

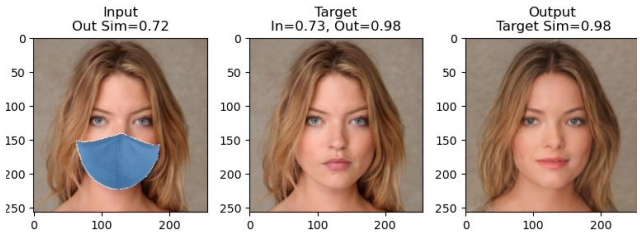

Figure 13. Output of StyleGAN2 using pSp Encode

As a result, the generator produced a realistic image with its mask removed. Results can be found in Figure 8. It's natural but generated image has been changed even in the non-mask area. Remaining the non-mask area same with the original image could be a solution, but it makes the image becomes like heterogeneous.

It is a kind of limitation of StyleGAN. And it is hard to combinate generated image and original image because they have quite different spatial distribution. The output is rather seems like newly image based on the style of the original image.

So for inpainting task, for which preserving original image's detailed characteristics is very important, we could judge that GAN with selective convolution like imputed convolution suits better than StyleGAN.

5. Conclusion

5.1. Summary

We adopted imputed convolution to largely masked area's inpainting task, unmasking, which is very practical and in needs of our society to solve these days.

We showed the training processes so that other researcher could understand how our model refining the image generating process toward the target. UNet was very capable for segmenting mask area, but it wouldn't be suitable for real world images, becuase we'd just artificially synthesized mask to the raw images. And the main model, it started from very vague shape, and develop it from coarse to fine. Coarse Generator catches the overall context of image so that the target area could be coherent to it. And the receptive field of coarse generator was much larger and clearer comparing to the one of GAN using gated convolution. We think it is caused by the essential difference of the two, explicit imputation. Then, Refine Generator elaborates to make the masked area in detailed appearance. After 7 epochs of training, the model generates very well-shaped of images and it works very well with the validation dataset.

Performance or image generating wasn't in significant gap between imputed convolution and gated convolution. We need more experiments.

But, our model shows significantly better performance comparing to StyleGAN, which actually generate newly image based on the style of original image rather preserving important detailed characteristics.

5.2. Limitations and Future Plans

The dataset was artificially generated. The types of masks are limited, and there was a distinguishable boundaries between the boundary of artificial mask and original facial image. We think the great performance of our UNet is based on this limitation. Also, the races of people were limited. Especially, the number of Asian's face were significantly less than the others. We might have to fine tune our model to adopt it in the real world images. And regarding to the data minority problem for some race, it could be serious problem while adopting our model to real world unless it address properly.

There also remains a future work about proof of excellence of imputed convolution. We'd succeeded this about the output of coarse generator, but felt hard to prove it regarding to the overall performance. We have to diversify the type of task and figure out what should we choose to perform better in these various cases, for each.

For the last, we could tackle other large area inpainting problems with the model trained for this task. And it will be much more meaningful if we could build generalized model

for the variety types of large area inpainting problems.

References

[1] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication, 2020. 3

[2] N. Ud Din, K. Javed, S. Bae, and J. Yi. A novel gan-based network for unmasking of masked face. *IEEE Access*, 8(10.1109/ACCESS.2020.2977386):44276–44287, 2020. 1, 2, 4

[3] Omar Elharrouss, Noor Almaadeed, Somaya Al-Máadeed, and Younes Akbari. Image inpainting: A review. *CoRR*, abs/1909.06399, 2019. 2

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[5] Håkon Hukkelås, Frank Lindseth, and Rudolf Mester. Image inpainting with learnable feature imputation, 2020. 3, 5

[6] Shruti Jadon. A survey of loss functions for semantic segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct 2020. 3

[7] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 1

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 2

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 2

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. 2

[11] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *CoRR*, abs/1804.07723, 2018. 2

[12] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *CoRR*, abs/1901.00212, 2019. 2

[13] Moses Odhiambo. Celeba-hq resized (256x256). https://www.kaggle.com/badasstechie/celebahq-resized-256x256?select=celeba_hq_256. 3

[14] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation, 2021. 2

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1

[16] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution, 2019. 2, 4

[17] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 2