UnVeilify: A Path Towards Personalized Face Mask Removal

Jin Hyoung Joo
hyoungjoo.j@gmail.com
Sungkyunkwan University
Republic of South Korea

Minji Kim
kmjj7864@g.skku.edu
Sungkyunkwan University
Republic of South Korea

Seungmin Lee
lsmbest@g.skku.edu
Sungkyunkwan University
Republic of South Korea

Hyeonmin Lee
hyuni7185@g.skku.edu
Sungkyunkwan University
Republic of South Korea

Hohyun Na
skghgus9@g.skku.edu
Sungkyunkwan University
Republic of South Korea

Jaemin You
yjm7455@g.skku.edu
Sungkyunkwan University
Republic of South Korea

ABSTRACT

Our project aims to remove face masks from masked images while retaining the features of the target identity. Previous methods of removing facial masks exist, however do not consider the importance of identity preservation. We introduce a novel neural network architecture that combines the U-Net architecture with the StyleGAN generator to achieve this goal. We experimentally show that our method out-performs the previous methods in both quality and speed. Our code can be found at https://github.com/jinhyoungjoo/UnVeilify

1 INTRODUCTION

As the field of image generation is improving, personalized image generation has drawn attention to many researchers, and as a result, many astounding works have been done on this topic [13, 19]. However, most facial inpainting or facial reconstruction methods that attempt to encode the identity information [9, 12, 23] all seem to generate poor results and unusable quality for production. We belive this is due to the difficulty of successfully embedding the identity information to existing image generation frameworks.

In this paper, we introduce a specific task within the field of image inpainting. We aim to remove facial masks from images, by proposing a novel neural network architecture. Face masks have been, and will be part of many people's lives. We believe that many people would like to remove this masks in images, since they view masks as occlusions rather than accesories.

This problem of removing face masks has also been dealt in other projects. Our project's novelty comes in that these methods do not consider an important aspect of the masked images, the identity features. Also, most of these projects show low quality image generation. Our model is capable of removing masks in a much higher quality within a very short time.

We first introduce other projects of mask removal and brief explanations of their methods. We then explain our approach to this task in much detail. Then we quantitatively and qualitatively analyze our approch, followed by our conclusion to this research.

2 RELATED WORK

Previous methods of mask removal do exist. For example, [2] approached this problem by utilizing two separate networks. By first segmenting the mask regions from the masked image, and inpainting the segmented region. [11] used a ResNet based U-Net architecture. However, all these methods seem to either produce extremely low quality results or disregard identity information and facial attributes entirely. This results in image quality that is unusable in production (Figure 1).

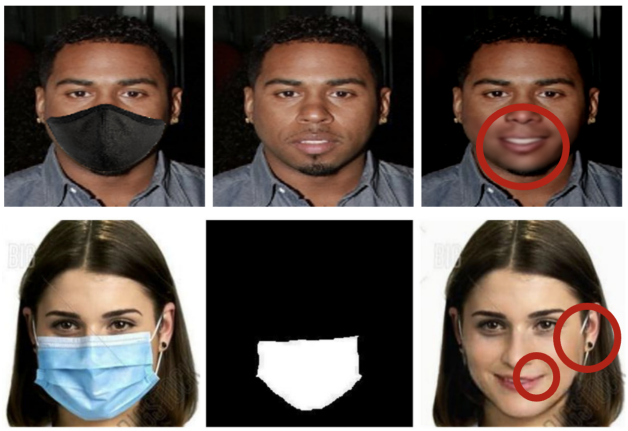

Figure 1: *Reported image results of other works*. The top image is the reported result of [11]. This does not consider the identity information and therefore looks like a different person compared to the ground truth. The bottom image is the reported result of [2]. Since the ground truth is unreported, we could not evaluate the identity preservation of this image. However, it can be seen that artifacts exist that crucially degrades the quality of the image.

3 METHOD

Our goal is to remove artifacts covering the facial area, namely face masks, while preserving the identity of the original person behind the mask. This requires us to generate the facial features by utilizing some sort of identity information extracted from an image that contain no occlusions.

Therefore, we obtain two input images from the user, one as the image containing the face mask to remove (we call this the *masked image*) and the other as the *identity image* that shows the facial features of the identity the user wishes to reconstruct. Here, the identity image isn't constrained to any conditions, *i.e.* the identity image and masked image are expected to be independent in every way except for the identity of the person.

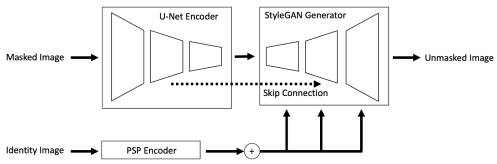

Figure 2: *Overall Architecture*. Our model consist of three main modules. The U-Net encoder extracts the features from the masked image. The PSP encoder extracts the features from the identity image. Finally, the StyleGAN generator uses the extracted features to generate the unmasked image.

We first describe the overall architecture that we proposed. Next, each sub-module of the network is described in detail. Finally, the objective functions used in our research is explained.

3.1 Overall Architecture

Figure 2 shows the overall architecture of the proposed network, named as *UnVeilify*. The input masked image goes through an encoder-decoder network, where the decoder is the StyleGAN generator [6]. Although the architecture of this decoder is different in some ways from the original StyleGAN model, to reduce confusion we leave the name as it is. Further description of this updated generator architecture can be seen in the following sections.

The encoder consists of two main parts, the U-Net encoder that extracts the features from the masked image, and the PSP encoder that extracts the identity features from the identity image.

3.2 Identity Extraction

Our research began with attempting to implement the StyleGAN generator to our case. Since the StyleGAN architecture is capable of producing high quality images while also providing controllable styles via latent style vectors, we assumed that this type of architecture is exactly what we want. However, while the StyleGAN architecture is capable of conditional image generation, it lacks the ability for direct image-to-image translation. This is because finding the exact latent style vectors to extract the features of the original domain is difficult.

The PSP encoder [17] was introduced to address this problem. This encoder uses a ResNet backbone network to extract the features of an image to multiple style vectors that can be inserted in the StyleGAN generator in each layer. By providing style guidance for low-level and high-level features, this encoder network can successfully generate images for image-to-image translation.

Our first approach to our task was to directly use this architecture. The PSP encoder does not change the StyleGAN architecture except for the generation of style vectors. The authors explained that additional conditions can be inserted into the generator by adding or concatenating the vectors onto the style vectors. Our initial approach was to extract the image features from the masked image by this PSP encoder, and then add the identity vectors extracted from the identity image. This identity vector was the output of a pretrained face recognition network [1].

However, we did not realize an important problem. While our task is an image-to-image translation, where the goal is to translate an image where the person is wearing a mask, to another domain where the person is not wearing the mask, in this task the image features should remain exactly the same except for the mask region. Our initial approach does not consider this fact and by experimental results, generates very poor results. Our solution to this problem was to use the PSP encoder as the style extractor from the identity image, and use a separate module that preserves the features of the original masked image. This module is the U-Net encoder and will be explained in the next section.

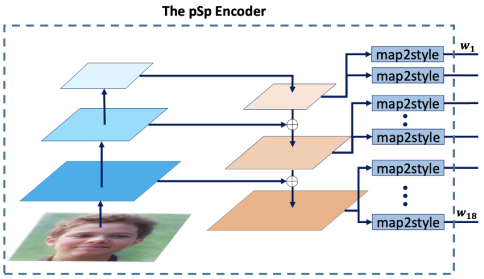

Figure 3: *The PSP encoder*. The PSP encoder extracts information from a given image in multiple layers of information. This figure is taken from [17].

While there are other possibilites of extracting the identity information, we hypothesized that the extraction of identity features is sufficiently done by the hierarchial feature pyramid network that extracts information from multiple different resolutions of the image.

Also, while the authors of [17] utilized a pretrained StyleGAN generator, due to our novel network architecture where the original input image features are intact, there is no need for any pretrained networks. This can help with further expandability to other domains, especially where specialized datasets are difficult to obtain, since our network can train within a end-to-end setting with no pretrained networks.

3.3 U-Net Encoder

The U-Net encoder is based on the U-Net architecture [18]. As mentioned above, the main purpose of this encoder is to extract the features of the original masked image so that the final generated image better preserves the information other than the masked region. The U-Net encoder outputs skip connections and a downsampled image.

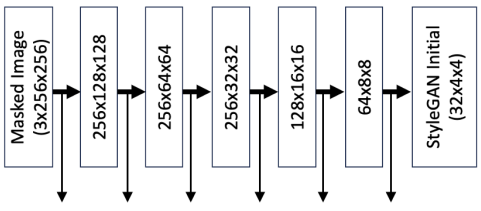

Figure 4: *Detailed architecture of the U-Net encoder.*

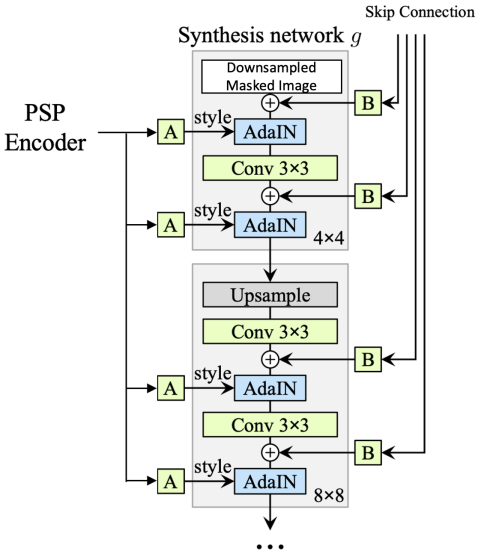

Figure 5: *The modified StyleGAN architecture.* Our modified StyleGAN generator utilizes the two input from the U-Net encoder and PSP encoder.

First, the skip connections are inserted into the StyleGAN generator. This is different to the original StyleGAN architecture, where additional vector inputs are not expected other than the style vectors. We modified this by removing the single-channel noise inputs and instead use the multi-channel skip connections from the U-Net encoder. The single-channel noise inputs in the original StyleGAN is used for generating stochastic detail, however we believed that this is unneccesary, since our task is not focused on diversity of the generated image, but rather has a ground truth image that it should match.

The downsampled image outputted from the U-Net encoder is used as the initial starting point of the StyleGAN generator. This is different in that the original StyleGAN generator uses a constant learned parameter as the starting point. Also, the style vectors and skip connections are added to the image channel-wise, where the weight of each vector is learned by the network. Our final model adds these two vectors at almost the same weight, showing that for quality image generation, both the style vector and skip connection are neccesary. The final modified StyleGAN generator can be seen in Figure 5.

3.4 Objective Function

Our models uses several loss functions to achieve the best quality possible. The final objective function follows the form similar to the objective functions of other image-to-image translation models. We first use the $L1$ content loss in order to generate less blurry images compared to that of using the $L2$ loss [5, 7]. This loss is used as a guide for generation by penalizing the distance between the ground truth unmasked image and generated unmasked image.

$$L_{content}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_1$$

We also use a perceptual loss in order to generate images with high visual qualities. We utilize the LPIPS loss [21] in the same way as [17] has done, since it has been shown that the LPIPS loss preserves image quality better than other perceptual losses [3].

$$L_{perceptual}(\mathbf{y}, \hat{\mathbf{y}}) = \|F(\mathbf{y}) - F(\hat{\mathbf{y}})\|_2$$

Here, F is the feature extractor used for the perceptual loss calculation.

This task focuses highly on identity preservation, where a large portion of the network architecture design process is shifted towards this specific goal. Hence, in the objective function we needed a method to also enforce identity preservation. This is achieved by introducing an identity loss. By taking the cosine similiarity between the features of a pretrained face recognition network (here we use ArcFace [1]), we can assume that if the network trains to reduce this loss then the identity information loss would be minimized.

$$L_{identity}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - C(R(\mathbf{y}) - R(\hat{\mathbf{y}}))$$

Here, R is the face recognition backbone model and C is the cosine similiarity function.

This model follows the form of a generative adversarial network, and incorporates a discriminator and GAN loss. We use the objective function of the original GAN without any modifications. However, since the identity image is given as a form of a conditional image, we give the identity image paired with the unmasked image to the discriminator. The discriminator is a simple PatchGAN [8] discriminator.

$$L_{GAN}(G, D) = \mathbb{E}_{y,c}\left[\log D(y, c)\right] + \mathbb{E}_{\hat{y},c}\left[\log(1 - D(\hat{y}, c))\right]$$

The total objective function is the combination of these 4 separate objective functions.

4 EXPERIMENTS

Our dataset requires a dataset of a masked image and an unmasked image that has everything except for the masked area to be the same. Since no dataset fits this criteria, we use a synthesized dataset made from [4], where the masked image is generated by morphing face masks onto the detected facial landmarks. The final dataset consists of 8,746 masked and unmasked image pairs, with 31,604 identity images. We use 80% of the dataset as the training set and 10% of the dataset as the test set and no identities overlap between the training set and test set.

Also, due to the lack of properly working open-source code on mask removal, we set an image inpainting model [14] as our baseline model. We train this baseline model on our dataset, with the goal of generating the covered lower half of the masked image. We use the following metrics as evaluation metrics. We also use metrics computed by [2]. These metrics are not evaluated on our test set, however can be used as one comparison criteria against other mask removal projects.

In order to evaluate the pixel-wise difference between the ground truth and generated image, we incorporate the PSNR metric and SSIM metric. To compare the perceptual differences, we use the LPIPS metric [21] that uses a pretrained VGG network.

We trained our model on a single NVIDIA V100 GPU for 87 epochs. We used early stopping with the validation set. Adam optimizers are used for both the generator and discriminator with the learning rate set as 0.0002 for both models.

Our experimental results can be seen in Table 1. Our model performs better in terms of PSNR, and although slightly performs lower in SSIM and LPIPS, we believe this is because of the blurry results generated by the baseline model [15, 16]. This can be seen in a qualitative analysis (Figure 6), our model generates less blurry results compared to the baseline model and also generates images with identity information closer to the ground truth.

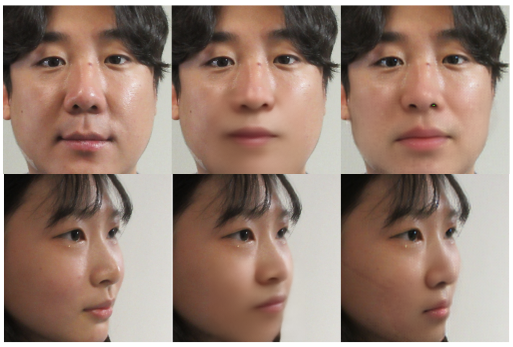

Figure 6: *Image results of our model compared to the baseline model.* Generated image results from the test set. Starting from the left are from: the ground truth, the baseline model (DMFN), and our model (UnVeilify).

Our model also out-performs the GAN-based mask removal method, showing that our method can generate more quality results compared to other mask removal methods. Our model can also generate images with highly less blurry results, which seem to be a common theme among all other mask removal methods.

Figure 7: *Image results of our model compared to the baseline model.* Generated image results from the test set. The baseline model fails to generate images where the person is not facing forward. Starting from the left are from: the ground truth, the baseline model (DMFN), and our model (UnVeilify).

Our method can also process masked images where the pose conditions are extreme. The baseline model cannot generate results correctly when the person is not facing forward (7). Rather, the baseline model, and other mask removal methods generate the mouth region in a deformed way. Our model generates images with quality even in these different pose conditions.

5 DISCUSSION

5.1 Limitations and Further Improvements

Our model's current biggest problem is that it fails to generalize to real-world data (Figure 8). We suspect that this is due to the synthetic dataset having a different data distribution compared to actual real-world data. Real masked images contain images where the mask itself is deformed by the person wearing it. Our synthetic dataset does not contain any information on this sort of information, hence impossible to generate quality data. Also, the lack of diversity in our dataset may lead to this problem as well.

Figure 8: *Failure cases of our model.* Generated images from actual real-life masked images. Starting from the left are: the identity image, masked image, and the generated output.

To alleviate this problem, we believe that in future works, by implementing a process of generating masked images within the model framework (similar to CycleGAN [24]), then the masked images may follow the real-world data distribution. Also, using a more diverse dataset like CelebA [10] or FFHQ [6] can help with this problem as well.

Another problem is that noise-like artifacts appeared in the generated images. We have mostly solved this problem by normalizing the output to the PSP encoder. However certain results still show these artifacts (Figure 9). We currently suspect this

Finally, while our generated images show higher quality than most other projects, our results still present mild blurriness and misrepresented facial attributes. We believe this is because of the high L1 loss [22], however our experiment results show that low L1 loss does not lead to optimal results. This may be correlated to our changed architecture, however further insight is required to know the exact reasons and method for improvement.

Other improvements can be done to our model, such as encoding additional information such as emotion as a condition. There are also other methods of identity extraction [20] that could give higher performance.

	PSNR	SSIM	LPIPS	Execution Time
Baseline Model (DMFN) [14]	26.92	*0.90*	*0.03*	1404 ms
Baseline Model (GAN-based Mask Removal) [2]	26.19	0.86	Unreported	Unreported
UnVeilify (No U-Net Encoder)	3.42	0.00	0.91	34 ms
UnVeilify (No Identity Image Given)	6.57	0.01	0.89	55 ms
UnVeilify (Unnormalized Identity Features)	19.16	0.75	0.14	
UnVeilify (Final Model)	*27.04*	0.84	0.06	

Table 1: *Experimental results of our model.*

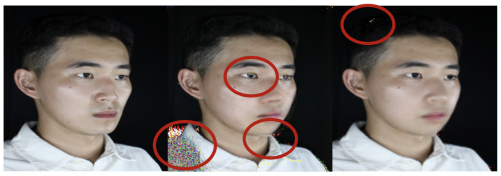

Figure 9: *Noise artifacts in the images.* Although the problem of large noise-like artifacts are greatly reduced, generated images still contain these artifacts. Starting from the left are: the ground truth, results with the unnormalized identity features, and our final model.

5.2 Other Tasks

We believe our proposed network can be expanded to tasks other than mask removal. Our model has the advantage of doing image-to-image translation while adding additional information as a form of an image. This has many use cases, especially in the case of personalized image generation, where a user might attempt to translate an image while having the identity information intact. Facial reconstruction, image inpainting could all be one of many examples where our proposed architecture can be used.

6 CONCLUSION

We introduced a novel neural network architecture for personalized image-to-image translation by focusing on removing face masks. Our network has achieved considerable quality generation compared to other baseline models and even generating under extreme conditions, all while retaining the source identity with high accuracy. This network is very fast and can be expanded to other tasks and is also applicable to real-world use cases.

REFERENCES

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[2] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. 2020. A novel GAN-based network for unmasking of masked face. *IEEE Access* 8 (2020), 44276–44287.

[3] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. 2020. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758* (2020).

[4] Human ICT. 2021. https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=528

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[6] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[7] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*. PMLR, 1558–1566.

[8] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam , The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 702–716.

[9] Xiaoming Li, Guosheng Hu, Jieru Zhu, Wangmeng Zuo, Meng Wang, and Lei Zhang. 2020. Learning symmetry consistent deep cnns for face completion. *IEEE Transactions on Image Processing* 29 (2020), 7641–7655.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August* 15, 2018 (2018), 11.

[11] Koucky Lukas and Maly Jan. 2021. Mask2Face: How we built AI that shows the face beneath the mask. https://www.strv.com/blog/mask2face-how-we-built-ai-that-shows-face-beneath-mask-engineering#ideas-for-additional-improvements

[12] Wuyang Luo, Su Yang, and Weishan Zhang. 2023. Reference-Guided Large-Scale Face Inpainting with Identity and Texture Control. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[13] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2023. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410* (2023).

[14] Evangelos Ntavelis, Andrés Romero, Siavash Bigdeli, Radu Timofte, et al. 2020. AIM 2020 Challenge on Image Extreme Inpainting. In *European Conference on Computer Vision Workshops*.

[15] Amy R Reibman, Robert M Bell, and Sharon Gray. 2006. Quality assessment for super-resolution image enhancement. In *2006 International Conference on Image Processing*. IEEE, 2017–2020.

[16] Amy R Reibman and Thilo Schaper. 2006. Subjective performance evaluation of super-resolution image enhancement. *Second Int. Wkshp on Video Proc. and Qual. Metrics (VPQM'06)* (2006).

[17] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2287–2296.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich , Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[19] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411* (2023).

[20] Yu-Chuan Su, Kelvin CK Chan, Yandong Li, Yang Zhao, Han Zhang, Boqing Gong, Huisheng Wang, and Xuhui Jia. 2023. Identity Encoder for Personalized Diffusion. *arXiv preprint arXiv:2304.07429* (2023).

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[22] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging* 3, 1 (2016), 47–57.

[23] Yajie Zhao, Weikai Chen, Jun Xing, Xiaoming Li, Zach Bessinger, Fuchang Liu, Wangmeng Zuo, and Ruigang Yang. 2018. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772* (2018).

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.