Journal of Theoretical and Applied Information Technology
15th February 2022. Vol.100. No 3

ISSN: 1992-8645 www.jatit.org ISSN: 1817-3195

# ANTI-MASK: AN AUTOENCODER-BASED DEEP NEURAL NETWORK TO REVEAL HIDDEN KOREAN FACES WITH FACE MASKS

**JUNHWI KIM[1], JAE WAN PARK[2]**

[1]Student, Soongsil University, Global School of Media, Seoul, Republic of Korea

[2]Associate Professor, Soongsil University, Global School of Media, Seoul, Republic of Korea

E-mail: [1]kjh046j@soongsil.ac.kr, [2]Corresponding author: jaewan.park@ssu.ac.kr

**ABSTRACT**

The purpose of this study is to draw the face part covered by the mask using deep learning technology to complete the entire image. In this paper, we introduce a method of predicting the lower canal of the face based on the information on the upper part of the face using the autoencoder structure. For this study, we design our anti-mask model based on transfer learning through VGGace and train, evaluate, and experiment with a dataset of 800 Korean frontal faces. Our anti-mask model trained in this way accurately drew a part of the hidden face. Through the evaluation of the drawn face images, we proved that our anti-mask model can sufficiently depict the lower canal from the upper image of the face. Moreover, in this paper, it was demonstrated that drawing by analogy with a part of the face is more accurate than reconstructing by analogy with the entire face. This study is expected to contribute to the development of various applications.

**Keywords:** *Autoencoder, Face Mask, Anti-Mask, Vggface, Image Inpainting*

## 1. INTRODUCTION

Due to the recent COVID-19 pandemic, face masks have become a daily necessity around the world. However, several European countries and several states in the United States have anti-mask laws prohibiting face covering in public places. It is known as a useful law to prevent violence or terrorism in public places [1]. However, these useful laws are currently not functioning properly due to the COVID-19 pandemic.

This research introduces a method of completing the frontal human image using the autoencoder model for the part of the face hidden in the mask. Through this, this study explores to what extent a person's appearance can be inferred from an image of a part of a person.

For this study, we design a neural model that predicts a feature vector representing the entire face by inputting a part of the face. To develop this effectively, transfer learning is performed on VGGFace [2][3], a pre-trained face recognition network that can extract facial features. Using our reconstruction model built through this, we output a frontal image of a human face based on the predicted facial features.

To this end, our deep learning model in this study is designed based on the autoencoder model. This model consists of two sub-models: an encoder and a decoder. That is, the autoencoder has a structure that converts input data into feature vector values that can be understood by a computer and restores it back to original data. It is composed of an encoder that converts input data into feature vector values and a decoder that creates original data from the vector values as it is. Using this model, we create an artificial intelligence with a structure that extracts the feature vector values of the upper part of the face exposed when a mask is worn as an input value, and outputs these feature vector values back to the original entire face.

The encoder we are going to learn extracts 4096-dimensional feature points from the image by learning only the face part that does not use a mask, that is, the face above the nose, and passes through a decoder trained in advance to create a complete face from the extracted feature points. A face image part without. To train our autoencoder model, we use 800 Korean frontal face images. Our method proceeds with unsupervised learning that does not supply additional information.

We have the following research questions in this study. Is it possible to predict the characteristics of other parts of the face from the characteristics of one part of the face? More specifically, can we predict the shape of the mouth and chin from the shape of the eyes and nose? And technically, are facial landmarks effective in implementing this?

We believe that we can support useful applications that can reconstruct the entire face from a part of the face image and retrieve the actual face image based on it. Our paper is the first to explore learning to reconstruct facial images using autoencoder models. In this paper, we will introduce the process of developing an autoencoder model-based deep neural network and investigate how well the represented face image matches the real face image.

## 2. RELATED WORKS

This study can be seen as an image inpainting field that redraws the hidden area in the image. Image inpainting is an area where research has been conducted before the rise of deep learning technology. As in the studies of Bertalmio, Marcelo, et al [5][6], image inpainting was performed by directly making an algorithm without using deep learning in the past. Since then, with the rise of deep learning, various studies have been conducted to apply deep learning to image inpainting.

In the study of Köhler, Rolf, et al. [7], Multilayer perceptron(MLP) was applied to image inapining and showed better performance than the existing algorithm method, but there was a limitation in that ininting was not performed well for large holes. Later, in a study by Xu, Li, et al. [8], image deblurring was performed using Convolutional Neural Networks(CNN), which is more specialized for image processing than MLP. In the study of Iizuka and Satoshi [9], the performance of image inpainting was improved by using a generative adversarial network(GAN) structure using two discriminators, a global context discriminator and a local context discriminator, but also showed a limitation in not being able to fill a large hole. In the study of Yu, Jiahui, et al. [10], they succeeded in improving the resolution of inpainting by learning to obtain information through a convolution filter from the image around the hole. Liu, Guilin, et al.'s study [11] pointed out the limitation that existing image inpainting studies mainly focus on square holes, and expansive post-processing is required to process them. As a solution to this, a partial convolution network based on UNet [12] was proposed, making it applicable to irregular holes, and a separate preprocessing process was omitted. In the study of Yeh, Raymond, et al [13], GAN was used. After training the generator and discriminator models on the complete image, the information z vector of the image is extracted from the generator using backpropagation, and the hole image is reconstructed by finding the z vector most similar to the hole image.

However, when learning image inpainting using these methods. When random noise is applied to the face image, good results are obtained, but when a part of the face image is cropped, the image is drawn smoothly, but the result is different from the original [13]. This may be inaccurate because most studies related to image inpainting are conducted with reference to the environment around the occluded part. For example, even the same person may have a different face depending on where the person is located. Therefore, more accurate reference information than the surrounding environment is needed for the hidden part. In particular, if you want to inpainting only the face, the performance will be better when intensively learning only the face image information.

## 3. LITERATURE REVIEW

### 2.1 VGGFace

VGGFace proposed by the Visual Geometry Group (VGG) of Oxford University is a network structure. VGGFace was trained with a deep network structure consisting of 5 convolutional layers blocks using the VGG face dataset collected on the Internet, which is a large-scale facial recognition dataset consisting of 2622 identities [2][3]. This model achieves a face recognition accuracy of 98.95% on LFW dataset [19][20]. For effective development, our model was built through transfer learning of the VGGFace model. Figure 1 illustrates the VGGFace Model Layer Architecture.

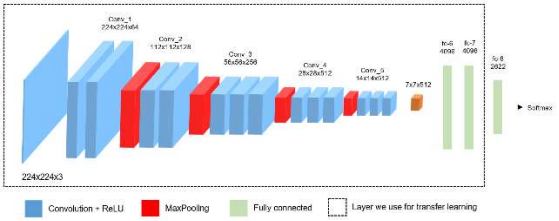


*Figure 1: VGGFace Layer Structure.*

Journal of Theoretical and Applied Information Technology
15th February 2022. Vol.100. No 3

ISSN: 1992-8645 www.jatit.org ISSN: 1817-3195

### 3.2 Transfer Learning

DeepFace [14], VGGface [2][3], and studies by Taigman, Yaniv, et al. [15], etc., used hundreds of thousands of face photos for classification and embedding of face information. Using the model used in these studies to extract the embedding information about the face and use it for reconstructing the face image shows better performance. In the study of Cole, et al. [16], a decoder that redraws a face from the features embedded by VGGFace has been successfully trained. Furthermore, the speech2face model uses a spectrogram visualized from human voice audio as an input and decodes it into an image of a frontal face based on the output of 4096-D facial features [4]. Based on the study of Cole, et al., the speech2face model was trained to feed a human image to a pre-trained face decoder network, extract features from the second to last layer, and output it as a real face image. The model was trained on millions of speech-face embedding pairs using the AVSpeech dataset. This study model and learning method were constructed with reference to the speech2face model and learning method.

Based on the above method, we are going to try to train an autoencoder that draws the lower part of the face that matches the information by extracting information from the upper part of the face using VGGFace. In this case, by referring to the feature information of VGGFace, information about the face can be extracted more accurately than the existing image inpainting. In addition, many existing face-related AI models have a limitation in that most of the Western faces are learned mainly from the master dataset. In order to experiment whether these models are well applied to Korean faces, we will train with the Korean face dataset.

## 4. DATASET

All face images used for training were extracted from the K-FACE dataset [17] and images collected from the Internet, etc. K-FACE, from the Korea Institute of Science and Technology(KIST) and National Information Society Agency(NIA), contains images of 400 Korean faces using various angles, lighting, and props such as glasses. We were able to utilize 400 frontal images as data for the stability of learning. An additional 600 face images were collected through Internet searches and donations from individuals. In this way, a total of 1000 face images of Koreans are secured, 80%(800 images) are used as the training dataset, and 20%(200 images) are used as the test dataset. Table 1 shows the configuration of the dataset.

*Table 1: Configuration of the Dataset.*

| | Total | Train | Test |
|---|---|---|---|
| Dataset(images) | 1000 | 800 | 200 |

## 5. ANTI-MASK

### 5.1 Data Preprocessing

Image preprocessing and data augmentation were done using 'dlib', a machine learning toolkit [18]. Using this, the position of the face was detected in the entire image and only the part with the face was cut out from the image. Since the weights of the VGGface were trained to receive the input image of size (224x224x3), all face images were rescaled to this size.

Then, face data augmentation was performed based on the face data augmentation method of cole et al. [16] to increase the amount of data. The face data augmentation process is shown in Figure 2.

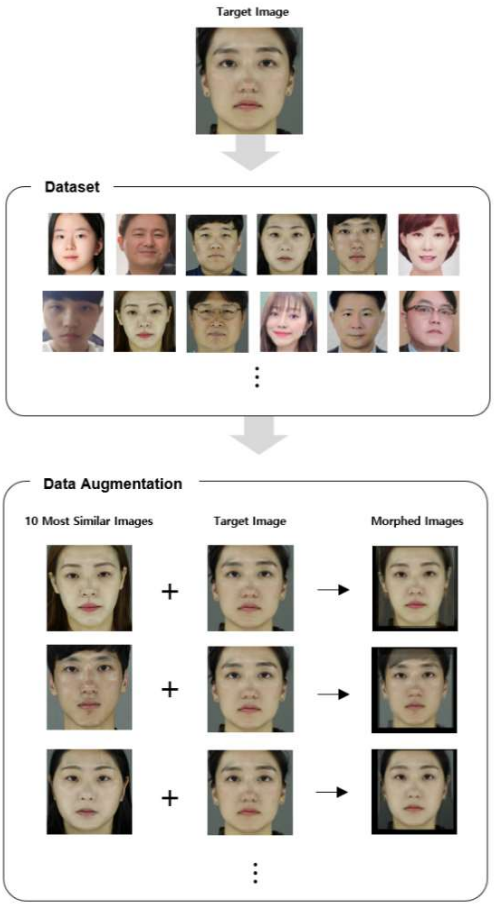


*Figure 2: Face Data Augmentation Process.*

Briefly, it can be said that data amplification was carried out using the face morphing technique, which finds faces that are similar to each other and mixes the images. If you want to morph the target image in Figure 2, first of all, 10 face images most similar to the target are found among the datasets. To measure the degree of similarity, not only image pixel values but also landmarks coordinate information of the face are used. The following formula was used to measure the similarity according to the paper [16]. L is the coordinates of landmarks extracted using 'dlib', and T is the pixel values of the face image. λ=10 was assigned to the equation (1) below:

$$d(A,B) = \lambda||L_A - L_B|| + ||T_A - T_B|| \quad (1)$$

In this way, 10 images most similar to one image was extracted. The selected images was morphed with the target image to create a new face. In this way, we were able to secure an additional data set of about 8,000 images, with 10 images per image. For the stability of learning, the pixel values of all images were normalized to values between 0 and 1. In addition, since landmarks as well as images will be provided as input for stable learning of the decoder, the coordinates x and y of 68 landmarks were extracted from each face using 'dlib' and stored in an array form.

### 5.2 Autoencoder-based Model and Training

Our model is designed based on the autoencoder model. The model of this study looks at the part of the face that is not covered by the mask, reads information such as the shape of the eyes, the position, and the shape of the forehead, and so on and predicts and draws the lower crown of the face according to the information. To construct this autoencoder model, we build an encoder that reads information from the upper part of the face and a decoder that finds and draws a suitable lower face shape from this information. In addition, learning can be performed faster and more efficiently with the help of VGGFace, which is already pre-trained to extract information from hundreds of thousands of faces. So we will train our encoder and decoder sub-models respectively using VGGFace. Figure 3 shows the overall structure of our Anti-mask model.

Our Anti-mask model is completed in three stages of training: (1) face decoder model, (2) Anti-mask's encoder model, and (3) Anti-mask's decoder model.

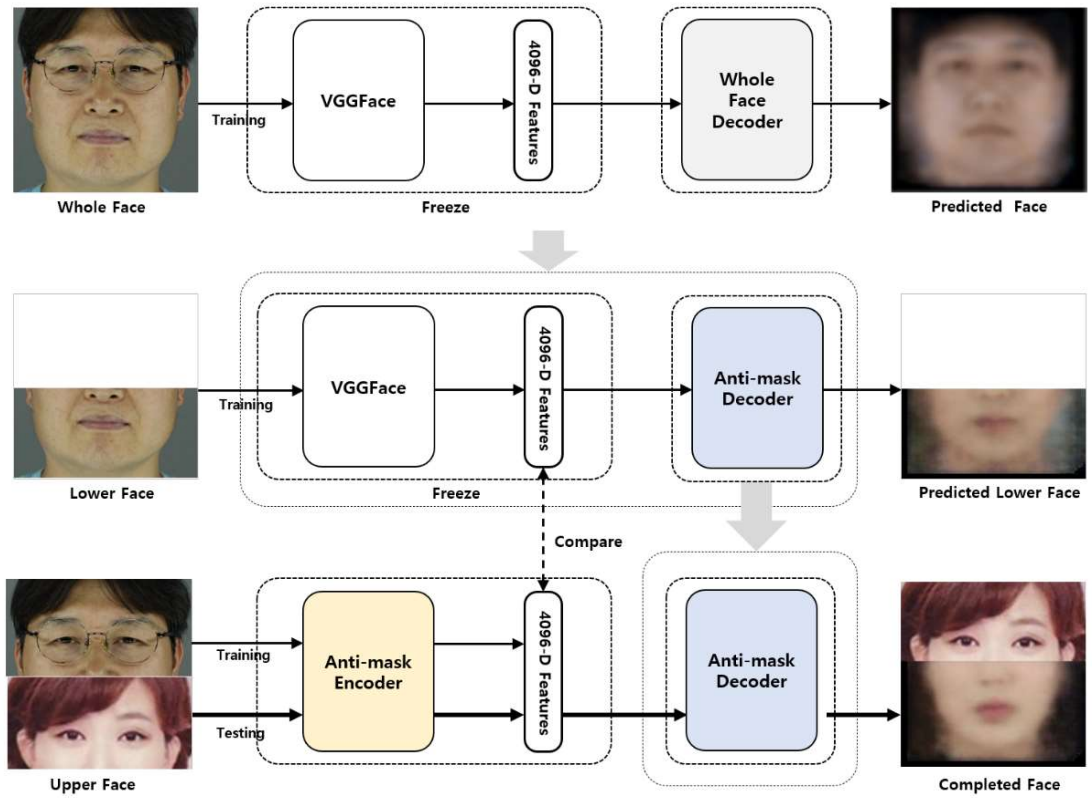


*Figure 3: Structure of Anti-mask Model.*

Journal of Theoretical and Applied Information Technology
15th February 2022. Vol.100. No 3

ISSN: 1992-8645 www.jatit.org ISSN: 1817-3195

*Figure 4: Face Image Reconstruction of the Test Dataset.*

### 5.2.1 Face decoder

Since VGGFace was mostly trained based on western face images, transfer learning was performed based on Korean face frontal images. To this end, after designing the decoder, we connected it after the 4096-d fully connected layer of the VGGFace and frozen the VGGFace before proceeding with training [4]. Through this, the VGGFace model can learn the characteristics of Korean faces. Figure 4 shows the results of face image reconstruction of the test dataset.

### 5.2.2 Decoder

Our goal is to extract the information from the upper part of the face and draw the rest of the lower crown. However, VGGFace is currently learning to extract information from the entire face. Therefore, our decoder extracts only the upper part of the face information from the entire face information provided by VGGFace and learns to draw the lower part of the face based on this. At this time, you can draw a more accurate face by learning the landmark together [16].

For this, the first 1000-d fully connected layer of our decoder model is connected after the 4096-d fully connected layer of the VGGFace. This layer is shared by the landmarks learning layer and the face image reconstruction layer so that landmarks and image reconstruction can be learned together. Image reconstruction outputs an image array of 128x224 size by concatenating the transposed convolution layers behind it. In the case of landmarks, only the landmark of the lower part of the face is extracted by building a fully connected layer on a 1000-dimensional dense layer. When the landmarks was trained together, the shape of the face was drawn more clearly. Figure 5 shows the detail structure of our decoder.

The loss function of the decoder is as follows (2). L stands for landmark loss and T stands for image texture loss. The weight of each loss λ is best learned when it is assigned as $\lambda_1$=10 and $\lambda_2$=100.

$$Loss = \lambda_1(L_A - L_B)^2 + \lambda_2||T_A - T_B|| \quad (2)$$

After several experiments, Adam was used as the optimizer and the learning rate was set to 0.0001. The best results were obtained without overfitting when training was conducted for 100 epochs. Figure 6 is a graph of the loss per epoch during the decoder training.

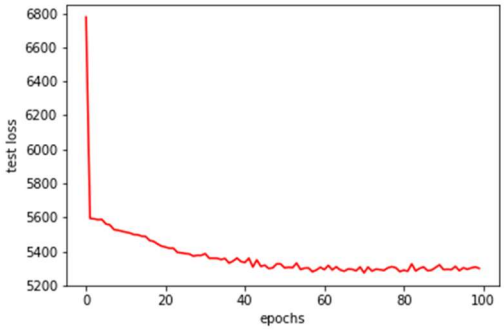


*Figure 6: Loss per Epoch During the Decoder Training.*

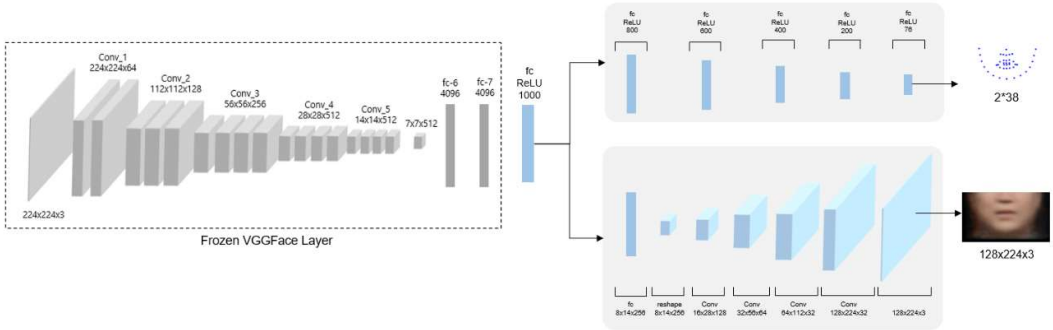


*Figure 5: Detail Structure of our Decoder.*

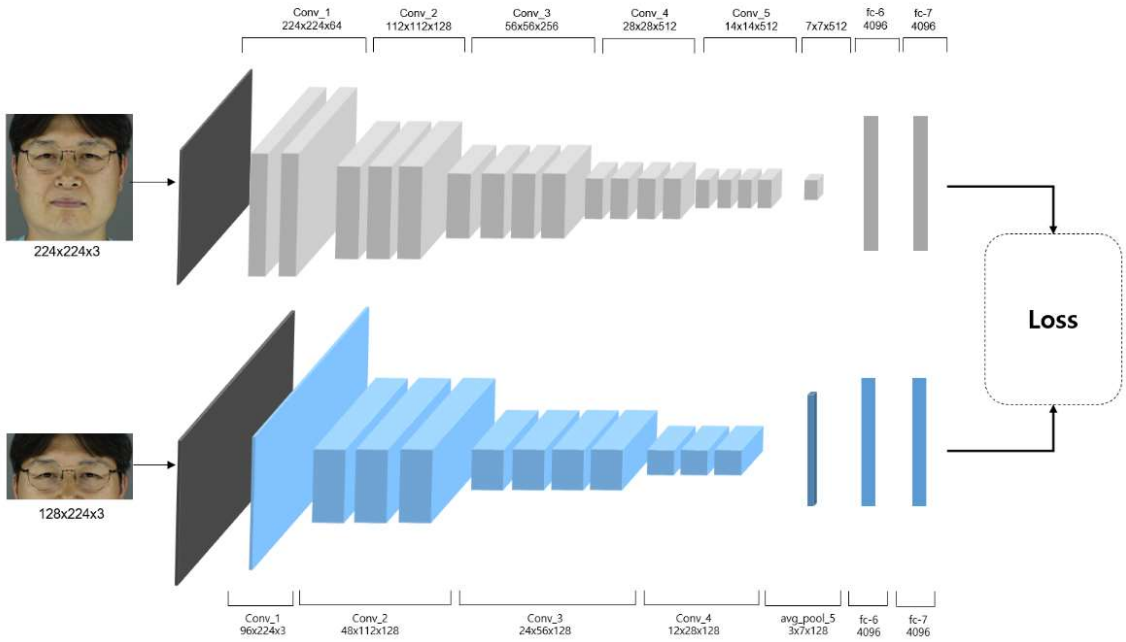


*Figure 7: Detail Structure of our Encoder.*

### 5.2.3 Encoder

Currently, the decoder has learned to extract and redraw information about the lower face image from the VGGFace. However, since the VGGFace is a layer that extracts information from the entire image of the face, for our goal, we need a new encoder that will extract enough information about the face from only the upper part of the face. Since the current decoder draws a 128x224 image from a 224x224 image, the encoder receives the 96x224 size input of the upper part of the face that the decoder will refer to. The input image goes through convolution layers, extracts important information on the upper part of the face, and finally outputs 4096-D features. Figure 7 shows the detail structure of our encoder.

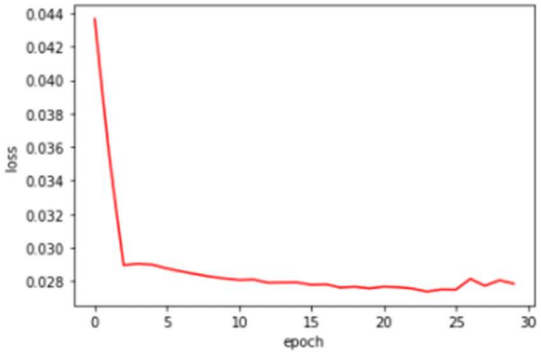


*Figure 8: Loss per Epoch During the Encoder Training.*

In order to efficiently extract information from the upper part of the face, the encoder also learns using the VGGFace. Learning is carried out so that the 4096-D features output by our encoder is similar to the 4096-D features output by the VGGFace. By doing this, it can harmonize well with the decoder familiar with the features of the VGGFace, and at the same time, it is possible to improve the performance by learning the information extraction ability of the VGG face. After different tests, Adam was utilized as the optimizer and the learning rate was set to 0.0001. The best results were obtained without overfitting when training was conducted for 30 epochs.

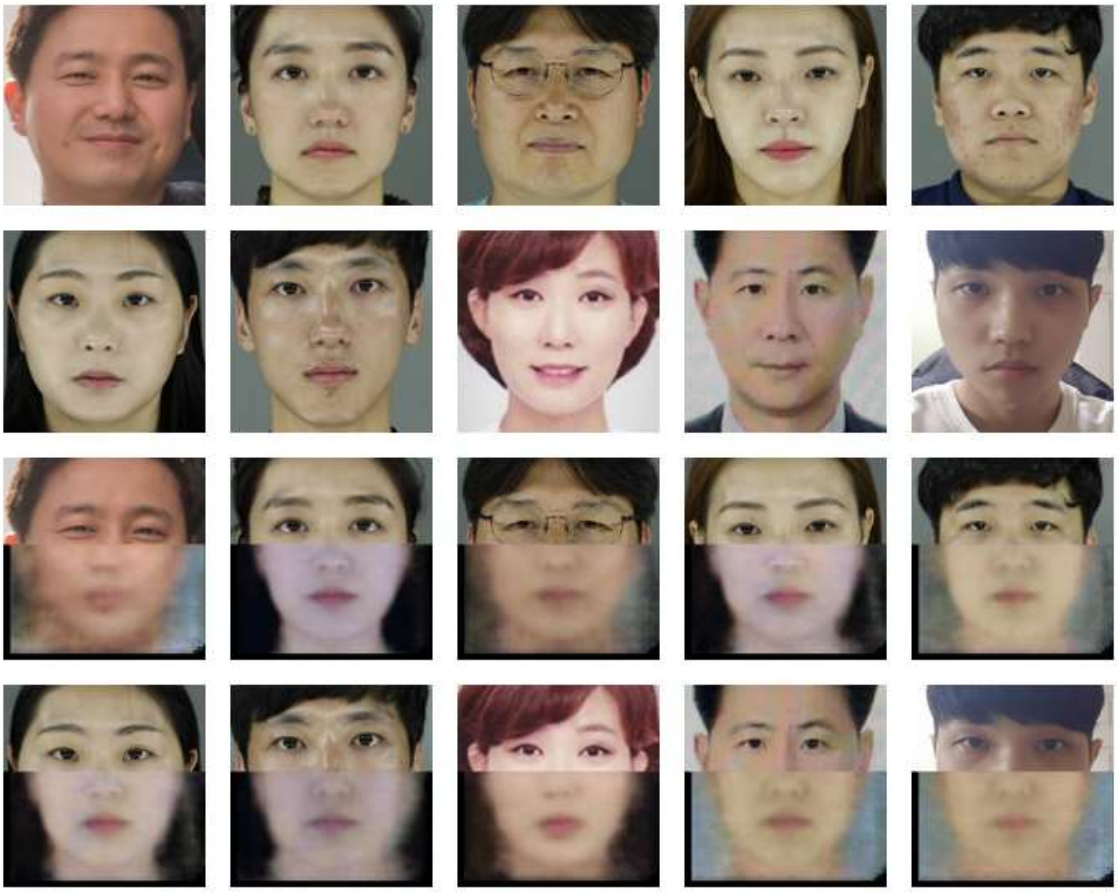

*Figure 9: Results by our Anti-mask.*

### 5.3 Evaluation

The results drawn by our anti-mask for the test datasets are shown in Figure 9. When viewed with the naked eye, the faces drawn by the anti-mask appear to have drawn the shape of the chin and nose accurately compared to the original face. In particular, like the ID photo, the light intensity is uniform and the full frontal face image is drawn more accurately.

How similar is the anti-mask predicted face to the original face when judged by a computer rather than a human? To evaluate this, we compared the original face image with the face predicted by Anti-mask using the VGG Face pretrained model. Figure 10 is a scatter plot drawn by extracting the values of 4096-d features from two images using VGGFace. The red points represent the feature values of the face drawn by the anti-mask, and the blue pointss represent the feature values of the original face.

We extracted 4096-d features from two images using VGGFace and compared their values. As a result of comparing the two features using cosine similarity, the average similarity was 0.89. Judging from this, the computer seems to be judging that the two images are similar.

## 6. DISCUSSION

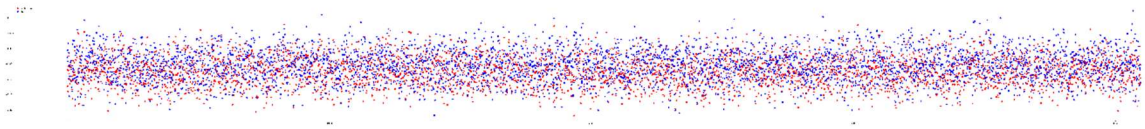

*Figure 10: Comparison of Feature Values of the Original Image and the Image Drawn by our Anti-Mask.*

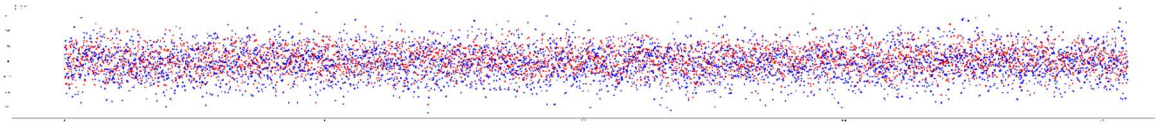

*Figure 11: Comparison of feature values of the original image and the image drawn in full.*

In this study, we conducted various experiments to derive optimal results. Here we would like to discuss some questions: (1) Is it effective to draw a partial image or to redraw the whole image? (2) What is the effect of anti-mask model on landmarks? And (3) why do we get inaccurate results for some images?

In order to answer the first question of whether it is effective to draw a partial image or to redraw the whole image, we designed a decoder in VGGFace and then trained a Korean face image. At this time, the cosine similarity extracted from 4096-d is 0.86. (Figure 11) It is lower than the cosine similarity of 0.89 when the part is drawn. In other words, you can see that it is more accurate than drawing the entire face, which is drawing part of the face.

For the second question, we printed the results of learning without landmarks to investigate the effect of landmarks on anti-mask learning. As shown in Figure 12, when learning without landmarks, it can be seen that the image is incomplete and the outline of the face cannot be clearly captured. Through this, it can be confirmed that the landmark information is influencing the anti-mask to draw the face clearly.

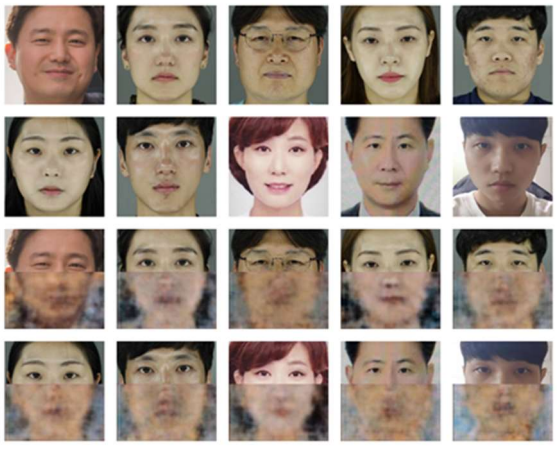

*Figure 12: Results without landmarks.*

Finally, as shown in Figure 13, some predicted images gave inaccurate results. It was found that this occurs when there is not a complete frontal image. A problem was also found that the predicted image quality of the face was not yet sufficiently clear. Further research on this is needed in the future.

*Figure 13: Prediction on Non-frontal Face.*

## 7. CONCLUSION

The purpose of this study is to complete the frontal image of the face by drawing the face part covered by the mask using the autoencoder model. In this paper, we introduce the development process for this and explore that the whole face can be inferred from a part of a human face.

For this study, we designed and trained our anti-mask model using 800 Korean frontal face images based on transfer learning using VGGFace through various experiments. For the evaluation of our model, a value of 0.89 was derived as a result of measuring the cosine similarity of the resulting image drawn by the original image and the anti-mask model. That is, we were able to prove that we can infer the whole face from a part of a human face. In this paper, we have demonstrated for the first time that we have succeeded in drawing the whole from a part of a face using our autoencoder model. Therefore, this study is expected to contribute to the development of various applications.

Future research for this study is to calculate a more precise frontal face image. Further enhancements for our Anti-maks model will include:

(1) developing more sophisticated landmarks, (2) collecting more Korean face data sets, and (3) experimenting with transfer learning based on various models.

## REFERENCES:

[1] Lawrence, Caroline V, et al. "Masking Up: A COVID-19 Face-off between Anti-Mask Laws and Mandatory Mask Orders for Black Americans." Calif. L. Rev. Online 11, 2020, p.479.

[2] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition.", 2015.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman. "Deep face recognition", In British Machine Vision Conference (BMVC), 2015.

[4] OH, Tae-Hyun, et al, "Speech2face: Learning the face behind a voice", In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7539-7548.

[5] BERTALMIO, Marcelo, et al, "Image inpainting", In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 417-424.

[6] Bertalmio, Marcelo, et al, "Simultaneous structure and texture image inpainting." IEEE transactions on image processing, 12.8, 2003, pp.882-889.

[7] KÖHLER, Rolf, et al. "Mask-specific inpainting with deep neural networks", *German conference on pattern recognition*. Springer, Cham, 2014, pp. 523-534.

[8] Xu, Li, et al, "Deep convolutional neural network for image deconvolution.", Advances in neural information processing systems, 27, 2014, pp.1790-1798.

[9] Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa. "Globally and locally consistent image completion." ACM Transactions on Graphics (ToG) 36.4, 2017, pp. 1-14.

[10] Yu, Jiahui, et al. "Generative image inpainting with contextual attention.", Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 5505-5514.

[11] Liu, Guilin, et al. "Image inpainting for irregular holes using partial convolutions." Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 85-100.

[12] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015, pp. 234-241.

[13] Yeh, Raymond A., et al. "Semantic image inpainting with deep generative models." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 5485-5493.

[14] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp.1701-1708.

[15] Taigman, Yaniv, et al. "Web-scale training for face identification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 2746-2754.

[16] Cole, Forrester, et al. "Synthesizing normalized faces from facial identity features." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 3703-3712.

[17] Choi, Yeji, et al. "K-FACE: A Large-Scale KIST Face Database in Consideration with Unconstrained Environments." arXiv preprint arXiv:2103.02211, 2021.

[18] King, Davis E. "Dlib-ml: A machine learning toolkit." *The Journal of Machine Learning Research* 10, 2009, pp. 1755-1758.

[19] Huang, Gary B., et al. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments." *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 2008.

[20] Huang, Gary B., and Erik Learned-Miller. "Labeled faces in the wild: Updates and new reporting procedures." *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep* 14.003 (2014).