Plastic Surgery Image Classification and Generation

Yiwei Han
University of Rochester
Rochester
yhan32@u.rochester.edu

Kaiyi Qi
University of Rochester
Rochester
kqi2@ur.rochester.edu

ABSTRACT

Plastic surgery is a very important surgical procedure and has received a lot of attention by public. It is often difficult to distinguish whether a person has had plastic surgery or not, and people who want to undergo plastic surgery often need to have a basic idea of what they might look like afterwards. In our paper, we use classification models and generative models from deep learning to solve the above problems, and we perform a cross-sectional comparison among many models and select the one with the best performance for further modification. Result shows that the classification model based on Res50 shows a strong classification ability, while the generation model based on AttentionGAN shows a relatively good ability to generate images of plastic surgery.

KEYWORDS

Classification, Aspect Ratio, GAN, Plastic Surgery, Pytorch

ACM Reference Format:
Yiwei Han and Kaiyi Qi. 2023. Plastic Surgery Image Classification and Generation. In *Proceedings of ACM Conference.* ACM, New York, NY, USA, 6 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Plastic surgery has become increasingly popular in many Asian countries in recent years, and it is important to understand the impacts of plastic surgery on individuals and society. According to statistics, 5.4 millions surgical plastic surgeries, and 8.4 million non-surgical plastic surgeries are done in Asia in 2019. There is a 33.3% increase of plastic surgeries over the past 4 years. [2] Moreover, it is estimated that 33.3% of Korean women aged between 19 to 29 has done plastic surgery [8] , and the top 3 popular plastic surgeries done in Asia are double eyelid surgery, rhinoplasty and facial contouring surgery. [5]

In this project, we collected a dataset of Asian before and after plastic surgery images to investigate the effectiveness of different plastic surgeries and their impact on facial appearance. We used various state-of-the-art deep learning models, including VGG16, ResNet50, AlexNet, and GoogleNet, to classify whether a person has undergone plastic surgery, and we applied several methods like 5-fold cross-validation and data augmentation, relugarization techniques to improve the classification accuracy.

We also focused on identifying which plastic surgery procedures were performed on an individual's face. Specifically, we developed algorithms that utilized the aspect ratio of eyes to determine double eyelid surgery, the aspect ratio of nose compared to the face, straightness of the nose bridge, and the color of the side of the nose and nose bridge to determine rhinoplasty surgery, and the aspect ratio of the chin and face to determine whether a person underwent face contouring surgery.

Finally, we explored the use of Generative Adversarial Networks to generate before and after photos of a person's plastic surgery. We trained various GANs, including UNIT, cycleGAN, and attentionGAN, to generate realistic images of an individual's face before and after plastic surgery. The generated images could be used to help individuals make informed decisions about whether or not to undergo plastic surgery and to better understand the potential outcomes of plastic surgery.

Overall, our project aims to provide a comprehensive understanding of the impact of plastic surgery on facial appearance and explore new ways to analyze and visualize the changes that occur after plastic surgery.

2 RELATED WORK

Plastic surgery is not a hot topic in the filed of computer vision and the generation of plastic surgery has no previous researches. Saksham et al. in 2018 [6] propose a model LARS to distinguish the plastic surgery images from normal people and get a good result. What they focused on is to avoid the deception by plastic surgery images on face recognition but not the plastic surgery itself. A machine learning framework for diagnose plastic surgery published by Paul et al. 2019 [3] shows another great work which could help in diagnostics, risk stratifcation, and treatment simulation. What it is focusing on is to get the clinical 3D morphable model instead of plastic surgery images. Different from above two methods, our goal in classification is on the 2D computer vision level and only focused on the plastic surgery itself. Together with the plastic surgery and non-plastic surgery image inter-conversion we will provide a useful tool for both patients and surgeons in plastic surgery.

3 DATASET

In light of the absence of any existing dataset containing before and after plastic surgery photos, particularly focused on Asian individuals, our team took the initiative to gather this data ourselves. We aimed to create a comprehensive and diverse dataset featuring Asian individuals who had undergone various plastic surgery procedures, with the intent to generate realistic images of Asian

Authors addresses: Yiwei Han, University of Rochester, RochesterNY; Kaiyi Qi, University of Rochester, RochesterNY.

ACM Conference, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/0000001.0000001

faces post-surgery and accurately classify whether a person has done plasic surgery.

This dataset, consisting exclusively of Asian before and after plastic surgery photos, was curated from a range of sources including Asian plastic surgery hospitals website, social media apps and other relevant platforms. Leveraging this unique dataset, our goal is to develop a robust classification model that can accurately determine whether a person has undergone plastic surgery, which plastic surgery the person has done and the generation of before after plastic surgery of an individual

The dataset contains two classes, first is non-plastic surgery class and the second is plastic surgery class. Each class has 1000 images with Asian faces. Each class is independent on whether the faces have makeup, since both classes has make up and non make up images.

By focusing on this specialized and inclusive dataset, we hope to contribute significantly to the fields of facial recognition, aesthetic research, and plastic surgery classification, ultimately enhancing the understanding and analysis of plastic surgery outcomes in the Asian population.

4 CLASSIFICATION

In our project, we tackled the task of plastic image classification using deep learning models. We utilized several well-known architectures including VGG16, ResNet50, AlexNet, and GoogleNet, and applied various techniques such as 5-fold cross-validation, data augmentation, and learning rate decay to improve the models' accuracy. The plastic image classification task aims to categorize plastic items into different classes based on their visual appearance.

4.1 Improvements

Before utilizing any improvement methods, our best model accuracy was achived by Resnet 50, which has an accuracy of around 87%. The result was not ideal, so we decided to utilize several improvement methods, and hopefully can increase our accuracy. After implementing the methods below, the highest accuracy raised to 92.25% by Resnet50.

4.1.1 Choosing Number of Epoch. By running a 5-fold cross-validation loss, we are able to plot the loss plot (Figure 1). We first trained the model for 50 epochs, and from the plot, it is pretty clear that the model converges at 10 epochs. As a result, we could train 10 epochs for each model furthermore in order to save time.

4.1.2 Data Augmentation. We did image augmentation method by various steps, which could potential increase the amout of our data. First, we resized to a standard size of 224 x 224 pixels to ensure uniformity in image dimensions. Next, the image is converted to grayscale with 3 output channels to reduce the computation required and apply random brightness, contrast, saturation, and hue adjustments to each image in order to augment the dataset and improve the generalizability of the model. Then, we standardize the pixel values of the image, with a mean and standard deviation of 0.5. Finally, random horizontal and vertical flips, as well as rotations up to 30 degrees, are applied to the image to further augment the dataset.

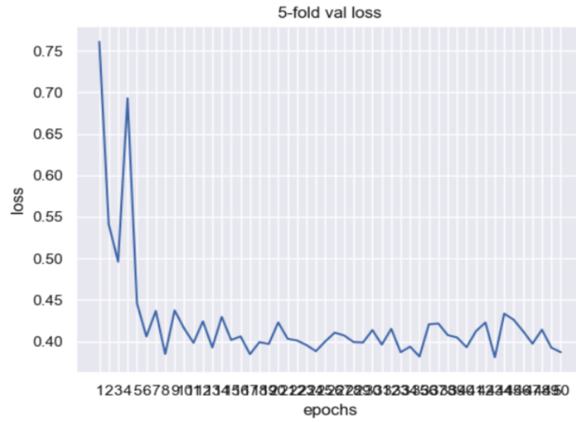


Figure 1: Loss Plot

4.1.3 Regularization. Weight decay is a regularization technique that penalizes large weights during training by adding a regularization term to the loss function. It helps prevent overfitting and can improve the generalization performance of a model. In the given code, a weight decay parameter of 10^-4 is applied, which means that the loss function will be modified to include a term that penalizes large weight values by adding 0.0001 times the square of each weight to the overall loss. Before the gradient descent step, the current weight values are multiplied by a factor slightly less than 1 to implement a form of L2 regularization. This process slightly decreases the weight values before performing the gradient descent step, thereby preventing overfitting and improving the model's generalization performance.

4.1.4 Learning Rate Decay. We set the learning rate of the optimizer to the initial learning rate decayed by a factor of 10 every 2 epochs. It does this by calculating a new learning rate value using the formula modellrnew = modellr * (0.1 ** (epoch // 2)) and then updating the learning rate of each parameter group in the optimizer.

4.1.5 Fine-Tune Strategy. We also applied fine-tuning strategy, which we freeze earlier layers to preserve low-level features and prevent overfitting, while fine-tuning later layers responsible for higher-level feature extraction and task-specific prediction. This strategy allows us to adapt pre-trained models to new tasks or domains, reducing the amount of training data required to achieve good performance.

4.1.6 Cross-Validation. In order to evaluate the performance of our plastic surgery classification model, we have implemented a 5-fold cross-validation technique on our dataset. By utilizing 5-fold cross-validation, we ensure that our model's performance assessment is both reliable and robust. This technique involves dividing the dataset into five equal folds, and training the model on four of the folds while testing it on the remaining one. This process is repeated five times, with each fold serving as the test set exactly once.

Through this methodology, we are able to ascertain the model's generalizability and account for any potential overfitting. By averaging the performance metrics across the five iterations, we obtain a comprehensive understanding of the model's performance on unseen data. Overall, it also help us find the best model during our training process

4.2 AlexNet

The AlexNet model consists of multiple convolutional, pooling, and fully connected layers, which enable the extraction of complex features and patterns within images. By training the model on our curated dataset of before and after plastic surgery photos, we expect the network to learn relevant and discriminative features that can accurately distinguish between individuals who have undergone plastic surgery and those who have not. In order to achieve this, we will fine-tune the AlexNet architecture, optimizing the hyperparameters and adapting the final layers to suit our binary classification task.

4.3 GoogLeNet

The GoogLeNet model is known for its Inception modules, which enable the extraction of complex features and patterns within images using multiple convolutional and pooling layers with varying kernel sizes. These Inception modules, combined with the overall depth of the architecture, allow for efficient and effective feature extraction.

4.4 VGG16

The VGG16 model consists of a series of 16 layers, including 13 convolutional layers, interleaved with max-pooling layers, followed by 3 fully connected layers at the end. Its architecture features small 3x3 convolutional filters, which enable the network to learn more complex and expressive features from the input images. Despite its relative simplicity compared to other deep learning architectures, VGG16 has proven to be highly effective in extracting hierarchical features from images, making it ideal plastic surgery image classification tasks.

4.5 Resnet50

The ResNet50 model consists of 50 layers, including convolutional and fully connected layers, arranged in a series of residual blocks. The key innovation in ResNet50 is the introduction of skip connections, that enable the network to bypass certain layers. These connections allow gradients to flow more easily through the network during the back propagation process, mitigating the vanishing gradient problem and enabling the training of much deeper models without sacrificing performance.

4.6 Comparison

We used several evaluation metrics, include accuracy, F-1 score to evaluate the overall performance of the models, while using the time elapsed parameter to evaluated the computational cost of our models. Furthermore, we provide the confusion matrix to make a visualization of the four classes of the results.

4.6.1 Evaluation Metrics. From the metrics table, we can see that our Resnet 50 achived the highest accuracy and F-1 score, while AlexNet has the lowest accuracy and GoogLeNet has the lowset F-1 score. The main reason ResNet50 achieved the highest accuracy in image classification is due to its deeper architecture compared to AlexNet. ResNet50 contains 50 layers while AlexNet only has 8 layers. The deeper architecture of ResNet50 enables it to model more complex relationships between features and achieve better accuracy. This makes AlexNet less effective at capturing fine-grained features and modeling complex relationships between features.

GoogLeNet's relatively lower F1-score compared to its accuracy could be due to the fact that the model is designed to detect multiple objects within an image, rather than classify a single object into one of several categories. As such, its predictions may be less precise on a per-class basis, but more comprehensive in terms of overall object detection.

As analyzed above, the time complexity of the AlexNet is the least, showing the shortest training time, while other more complex models has larger time complexity.

Table 1: Metrics of 4 Models

Name	Accuracy	F-1 Score	Time
Resnet50	0.9225	0.9182	1469.65s
VGG16	0.9000	0.8980	1811.55s
GoogLeNet	0.8825	0.8550	1037.90s
AlexNet	0.8525	0.8862	984.05s

4.6.2 Confusion Matrix. From the confusion matrix comparison below (Figure 2), we can see that the true positive cases by AlexNet, VGG16 and AlexNet are relative the same, while GoogLeNet has more predicted as true positive.Our highest accuracy model, which is Resnet50, has significantly less false positive cases and more true negative cases comparing to the rest of the models. While AlexNet, which is our least accurate model, has the most number of false negative cases, with other 3 models has relative same number.

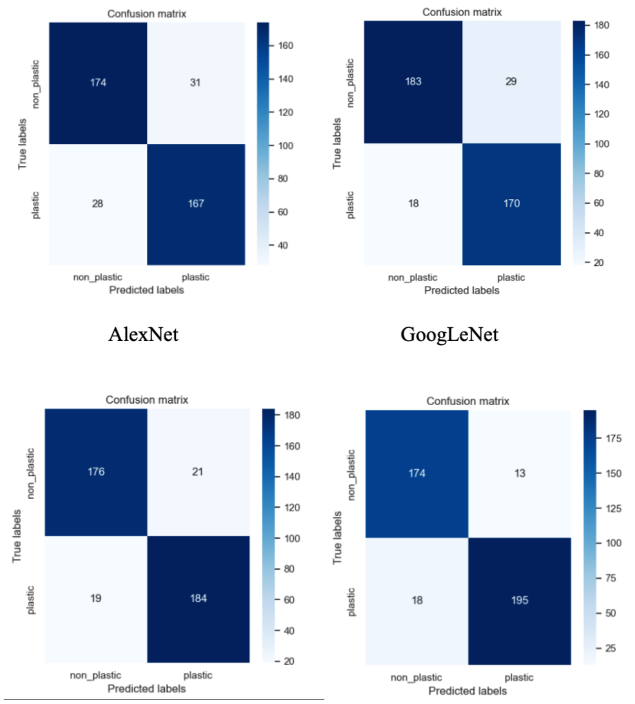
4.7 Wrong Results

Though our model achieved a relatively high accuracy, there are still some wrong results provided by our model. Below are some common mistakes made by our classifiers.

4.7.1 False Negative. This wrong result occurs when a person did plastic surgery, but our classifier classify it as non-plastic surgery. (Figure 3)

We can see that common false negative cases include that person has done plastic surgery, but are subtle. For example, the left image still has single eyelid, and the right image has round face. These features are classified as non-plastic surgery images, leading to the wrong result, while the plastic surgery they have done are not as obvious as the features that are classified as non-plastic surgery images.

4.7.2 False Positive. This wrong result occurs when a person did not do plastic surgery, but our classifier classify it as plastic surgery. (Figure 4) The images typically have features that are like plastic surgery faces. For example, sharper chin, bigger eyes and straighter nose bridge. In reality, this could happen with celebrities, since most of them already has features that like plastic surgery faces.



AlexNet GoogLeNet

VGG16 Resnet50

Figure 2: Confusion Matrix of 4 Models

Figure 3: False Negative Cases

Figure 4: False Positive Cases

5 IDENTIFY TYPE OF PLASTIC SURGERY

This part analyzes facial features and determine whether a person has undergone certain plastic surgeries. As stated before, the top three plastic surgery that done in Asia are double eyelid surgery, rhinoplasty surgery and face contouring surgery. In particular, we would like to detect whether a person has undergone these three types of plastic surgeries. The first step is to read in an image and convert it to grayscale using the OpenCV library. The grayscale image is then passed to the dlib library, which provides a facial landmark detection algorithm to detect faces in the image and obtain a set of 68 facial landmarks.

5.1 Double EyeLid Surgery

By looking from eyes, it is pretty clear that plastic surgery eyes has two features: first is the ratio of height to width is greated than non-plastic surgery eyes; second is the area proportion to the whole face is larger.

In order to determine whether a person has done eye plastic surgery, we extract the landmarks for the left and right eyes and calculate the ratio of the distance between specific pairs of landmarks on the eye to the distance between the top and bottom eyelids. It takes a six-element array of facial landmarks that correspond to the points around one eye and computes the eye aspect ratio using the Euclidean distance between the landmarks. Specifically, it calculates the distance between the vertical landmarks (the top and bottom of the eye) and the horizontal landmarks (the inner and outer corners of the eye) and divides the sum of the distances by twice the distance between the left and right vertical landmarks. This yields a normalized ratio that represents the height of the eye-opening relative to the width. Moreover, we also calculate the ratio of the area of eyes comparing to the area of the face. Combining these two criteria together, we are able to detect the surgery relative accurately.

5.2 Rhinoplasty Surgery

By looking from eyes, it is pretty clear that the noses undergone rhinoplasty surgeries have three features in general: first is the ratio of the length of the nose to the face length is greater and the width of the nose to the face width is smaller; second is the nose bridge tend to be more straight; third is the color of the two sides of the nose bridge tend to be darker comparing to the nose bridge, since the nose are more 3-D shaped.

Firstly, we calculated the ratios of the nose width to the face width, and the ratio of the nose height to the face height. These ratios can be used to compare the size of a person's nose relative to the size of their face.

Secondly, we calculated the average color of the skin on either side of the nose, and the average color of the skin on the bridge of the nose. If the difference of color is large, then it is more likely the person has done rhinoplasty surgery.

Thirdly, we gave a measure of the straightness of the bridge of the nose. This function uses the Sobel operator to calculate the vertical edges in the image, and then calculates the average edge intensity along a line that spans the length of the nose bridge. A straighter nose bridge will have a higher average edge intensity.

5.3 Face Contouring Surgery

By looking from eyes, we can see that people that have done face contouring plastic surgery generally has V-shaped smaller face, moreover, the face line are smoother comparing to who has not done the surgery.

The face width is calculated by taking the distance between the first landmark point (left-most point on the face) and the 17th landmark point (right-most point on the face). The chin width is calculated by taking the distance between the 5th and 11th landmark points, which are the points on the left and right sides of the chin, respectively. The function then returns the ratio of the chin width to the face width, which indicates how wide the chin is relative to the overall size of the face.

6 INTER-CONVERSION OF PLASTIC SURGERY AND NON-PLASTIC SURGERY IMAGES

To achieve the inter-conversion of plastic surgery and non-plastic surgery images, we decide to use GANs [1] for the generation of new images. GANs has a remarkable success in generating high-quality images and once the model is well-trained, a large number of images can be transformed in seconds, which is difficult to achieved by other methods. We adopt three GANs models and compared the performance based on their coherence, plastic surgery image generation and non-plastic surgery image generation.

6.1 CycleGAN

CycleGAN [9] is an unpaired image-to-image translation model from the source domain X to the target domain Y. The key idea is the mapping function G: X → Y and F: Y → X. An input I of class A will first get converted into I(B) in class B with mapping function G, then I(B) will be converted back to class A and get I(A) with mapping function F. I and I(A) with a cycle consistency loss to ensure that I(A) should be similar to I. A discriminator D will check I(B) to evaluate the simulation of the generated image. We trained the CycleGAN for 250 epochs, 1000 iterations in each epoch.

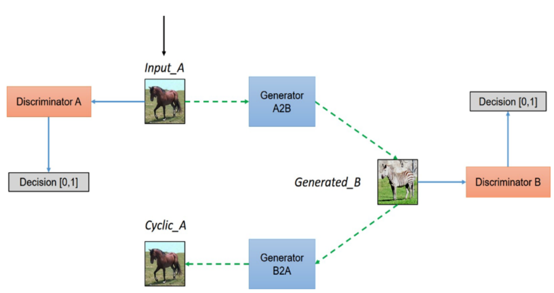


Figure 5: Framework of CycleGAN

6.2 AttentionGAN

AttentionGAN [7] is a new model based on CycleGAN with a generator consist of a Parameter-Sharing Encoder, Content Mask Generator and an Attention Mask Generator. The Attention Mask Generator will generate a foreground attention mask and a background attention mask. The foreground attention mask will select content together with the content mask which is generated by the content mask generator, and the background attention mask will select content together with the original input. Combine these two parts will be the new generated image. We trained AttentionGAN with the same number of epochs and iterations in each epoch to make an impartial comparison between the two models.

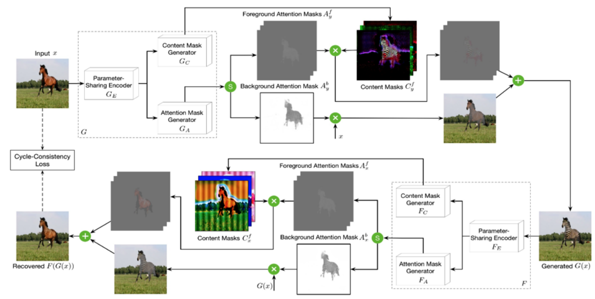


Figure 6: Framework of AttentionGAN

6.3 UNIT

UNIT [4] introduces an idea of latent space. Images x1 and x2 in domain X1 and X2 can be mapped into a same latent code in the latent space Z. Encoding function E1 and E2 will map from image to latent code and generation function G1 and G2 is used to map from latent code to images. E1, E2, G1 and G2 are implemented with the structure of CNNs and the shared latent space is represented as a weight sharing constraint, ties the connection weights of the last few layers in E1, E2, and first few layers in G1 and G2. For UNIT, we have trained two models. The first one is 260 epochs with 1000 iterations each just similar to the previous two model. We find out the UNIT model shows a good transformation from plastic surgery image to non-plastic surgery image but the reverse is not that good. Thus, we decide to keep training this model unto 420 epochs, trying to also reach a good performance on the reverse case.

6.4 Comparison

The following shows a sample case of the result of the inter-conversion between plastic surgery images and non-plastic surgery images.

After comparing for 30 pair of images, we find out all of the three models did pretty well on eye opening but none of them shows a good modification on the nose. Moreover, AttentionGAN is the only one that can change the shape of the face properly. For UNIT model, the one with 260 epochs and the one with 420 epochs doesn't show a huge difference. We also notice that the translation from non-plastic surgery image to the plastic surgery image is likely to make the skin color lighter while from plastic surgery image to the non-plastic surgery one will generally make it becomes darker.

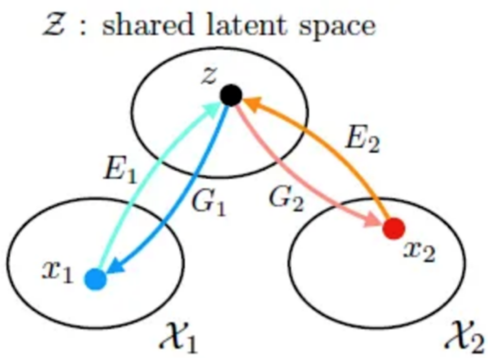


Figure 7: Shared Latent Space Assumption

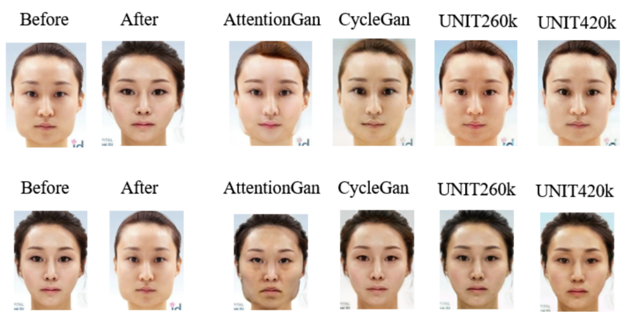


Figure 8: Result of the inter-conversion with different model. The top line is from non-plastic surgery image to plastic surgery image, the bottom line is the opposite.

7 CONCLUSION

For our current work in classification and generation, we compared multiple models and test their performance on this task. In classification, Resnet50 shows the highest accuracy with 92%. However, there might be some other feature other than plastic surgery in the dataset that leads the result. Except the color of skin and brightness of the images, we also find out that the plastic surgery images usually has brighter skin

For future work, we plan to apply some GANs model on the dataset to get all of them makeup and see if the classification result will get affected. For the generation part, all three models are working but they all have some defects (like no modifications on the nose). We are planning to test some larger GANs model like BigGAN to seem if we can get some more properly generated images. If we can get a satisfiable result eventually, then our work can be used in multiple aspects such as the help patients to get some insights of how they might looks like after plastic surgery, or help the surgeons to get some idea on how to do the plastic surgery.

8 LINK TO CODE

AttentionGAN: https://github.com/Ha0Tang/AttentionGAN

CycleGAN: https://github.com/junyanz/pytorch-CycleGAN-and -pix2pix

UNIT: https://github.com/mingyuliutw/UNIT

REFERENCES

[1] I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, and Y Bengio. 2014. Generative adversarial nets. *NIPS* (2014).
[2] ISAPS. 2023. The Latest Global Survey from ISAPS Reports a Significant Rise in Aesthetic Surgery Worldwide. https://www.prnewswire.com/in/news-releases/the-latest-global-survey-from-isaps-reports-a-significant-rise-in-aesthetic-surgery-worldwide-301715433.html
[3] P G Knoops, A Papaioannou, A Borghi, R W Breakey, A T Wilson, O Jeelani, S Zafeiriou, D Steinbacher, B L Padwa, D J Dunaway, and S Schievano. 2019. A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. *Scientific Reports* 9, 1 (2019).
[4] Y Liu, T Breuel, and J Kautz. 2018. , 4-4 pages. https://arxiv.org/abs/1703.00848
[5] Asian Cosmetic Surgery. 2023. BLOG: TOP 3 FACIAL ASIAN PLASTIC SURGERY PROCEDURES. https://www.asiancosmeticsurgery.com/blog/top-3-facial-asian-plastic-surgery-procedures/
[6] S Suri, A Sankaran, M Vatsa, and R Singh. 2018. On matching faces with alterations due to plastic surgery and disguise. *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (2018).
[7] H Tang, D Xu, N Sebe, and Y Yan. 2019. , 4-4 pages. https://arxiv.org/abs/1903.12296
[8] Erin Yi. 2022. True Beauty: The Economics of Plastic Surgery in South Korea. https://uweb.berkeley.edu/2022/01/23/true-beauty-the-economics-of-plastic-surgery-in-south-korea/
[9] J. Y Zhu, T Park, P Isola, and A A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV* (2017).