World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:5, 2024

Rejuvenate: Face and Body Retouching Using Image Inpainting

H. AbdelRahman, S. Rostom, Y. Lotfy, S. Salah Eldeen, R. Yassein, N. Awny

***Abstract*—**People are growing more concerned with their appearance in today's society. But they are terrified of what they will look like after a plastic surgery. People's mental health suffers when they have accidents, burns, or genetic issues that cause them to cleave certain body parts, which makes them feel uncomfortable and unappreciated. The method provides an innovative deep learning-based technique for image inpainting that analyzes different picture structures and fixes damaged images. This study proposes a model based on the Stable Diffusion Inpainting method for in-painting medical images. One significant advancement made possible by deep neural networks is image inpainting, which is the process of reconstructing damaged and missing portions of an image.#The patient can see the outcome more easily since the system uses the user's input of an image to identify a problem. It then modifies the image and outputs a fixed image.

***Keywords*—**Generative Adversarial Network, GAN, Large Mask Inpainting, LAMA, Stable Diffusion Inpainting.

I. Introduction

PLASTIC surgery has a long history, dating back to ancient times. The ancient Egyptians were known to practice a form of cosmetic surgery, particularly in the area of reconstructive surgery, such as the repair of broken noses. The Edwin Smith Papyrus, which was written around 3000 BC, contains descriptions of surgical procedures for various conditions, including facial injuries [1].

In modern times, plastic surgery has become increasingly popular worldwide, particularly in the areas of cosmetic surgery and reconstructive surgery. Many people seek cosmetic surgery to improve their appearance [1], with common procedures including breast augmentation, liposuction, and rhinoplasty (nose surgery). Reconstructive surgery is also commonly performed, with procedures such as cleft lip and palate repair, burn reconstruction, and facial reconstruction.

The field of plastic surgery has also seen significant advancements in recent years, with the development of new surgical techniques and technologies. Many plastic surgeons are highly trained and experienced, and Egypt has become a destination for medical tourism, with many patients coming from other countries to receive plastic surgery.

However, the popularity of plastic surgery has also led to concerns about the safety and ethical practices of some plastic surgeons. Patients have no idea how they will look after plastic surgery, and not all plastic surgeries are intended to be fun. There are many serious surgeries being done on cases with congenital conditions such as cleft lip problems. Some patients and their parents may panic if they do not know how they or their children will look after a surgery, so the proposed system will help them overcome their apprehension. Furthermore, some doctors are unable to explain to their patients what will happen during the surgery, the outcome, and whether it will leave scars or not. This system is willing to fix the issue by taking a picture of the patient's problem, regardless of whether it is a cleft lip, and presenting the result before having the surgery.

The paper is set out in the following manner. The review of related work is presented in Section II. Data-set and methods description are explained in Section III. The proposed techniques used are in Section IV. Section V describes the experimental results that are found by the above steps. Finally, Section VI draws a conclusion.

H. AbdelRahman, S. Rostom, Y. Lotfy, S. Salah Eldeen, R. Yassein, N. Awny are with Faculty of Computer Science, Misr International University, Cairo, Egypt (e-mail: hossam.csc0225@miuegypt.edu.eg).

II. Related Work

Latent Diffusion Models for High-Resolution Image Synthesis were discussed in [2]. Diffusion models (DMs) are a kind of image synthesis technique that use sequential denoising autoencoders to produce amazing outcomes. Nevertheless, this process can be computationally expensive and time-consuming. To address this issue, DMs in the latent space of pretrained autoencoders have been proposed, which allows for efficient training and improved image quality, (1). By introducing cross-attention layers into the model architecture, a flexible and powerful generator has been created that can be conditioned on various inputs such as text or bounding boxes. This approach results in state-of-the-art image inpainting and class-conditional image synthesis scores, as well as competitive performance in other image-related tasks. Moreover, this method significantly reduces the computational requirements compared to pixel-based DMs. Pairs of images and their corresponding conditioning information have been used to train the conditional Latent Diffusion Model (LDM):

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\right\|_2^2\right] \quad (1)$$

After that, photorealistic Text-to-Image Diffusion Models with Deep Language have been proposed [3]. Understanding Image is a text-to-image diffusion model that offers text-to-image synthesis with an unparalleled degree of photorealism and language understanding by utilising the power of transformer Language Models (LMs) and high-fidelity diffusion models. The key finding of Imagen is that, in

contrast to previous work that just uses image-text data for model training, text embeddings from massive LMs that have been pre-trained on text-only corpora are remarkably effective for text-to-image synthesis [3]. Large frozen language models that have solely been trained on text data have been found to be unexpectedly good text encoders for text-to-image conversion. Additionally, they discovered that increasing the size of the frozen text encoder greatly increases sample quality compared to increasing the size of the image diffusion model.

Pretrained text encoders are the first stage in this paper's five-step process for displaying a stable diffusion outcome. It compares different text encoders for text-to- image generation, such as image-text models, large language models, and pre-trained models like BERT, T5, and CLIP. It argues that freezing the weights of pre-trained models has advantages and scaling the text encoder size improves the image quality. It also reports that human evaluators prefer T5-XXL over CLIP on a challenging benchmark called Draw Bench. The second step is Diffusion models and classifier-free guidance, it explains diffusion models, which are generative models that transform noise into data samples by denoising them iteratively. It also introduces two techniques to improve the quality of conditional samples: classifier guidance, which uses a pre-trained model to guide the sampling process, and classifier-free guidance, which trains a single model on both conditional and unconditional objectives and adjusts the predictions based on a guidance weight. It states that Imagen, a text-to-image model, relies on classifier-free guidance for text conditioning (2):

$$\mathbb{E}\mathbf{x}, \mathbf{c}, \boldsymbol{\epsilon}, t\left[w_t \|\hat{\mathbf{x}}\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2\right] \quad (2)$$

Also, it describes two methods to enhance the quality of conditional samples from diffusion models: classifier guidance, which uses a pre-trained model to influence the sampling, and classifier-free guidance, which trains a single model on both conditional and unconditional tasks by randomly dropping the condition. It gives the formula for the adjusted prediction in classifier-free guidance (3):

$$\tilde{\boldsymbol{\epsilon}}\theta(\mathbf{z}_t, \mathbf{c}) = w\boldsymbol{\epsilon}\theta(\mathbf{z}t, \mathbf{c}) + (1 - w)\boldsymbol{\epsilon}\theta(\mathbf{z}_t) \quad (3)$$

The third and fourth step are to discusses two techniques to improve the quality and alignment of text-guided diffusion models: static thresholding and dynamic thresholding. Static thresholding clips the predictions to a fixed range, while dynamic thresholding adjusts the range based on the pixel values. It also describes how Imagen uses a pipeline of cascaded diffusion models with noise conditioning augmentation to generate high-fidelity images at different resolutions. The paper not only presents a deep learning technique, but also explains how the noise level is chosen randomly during training and swept over during inference to achieve optimal results. The last step is Neural Network architecture that has 2 models' Base model and a Super-resolution model. The base model uses a U-Net architecture to generate 64x64 images from text. The model uses text embeddings and cross-attention to condition the text at different levels. It also mentions that layer normalization helps improve the performance of the model. The Super-resolution model uses a modified U-Net architecture to generate 256x256 and 1024x1024 images from 64x64 images. It explains how the models are trained on the crops of the images and how they use text cross attention to condition the text. Abd highlights the advantages of the modified U-Net over the original one.

Meanwhile, Resolution-robust Large Mask Inpainting with Fourier Convolutions has been released [4]. The issue of accurately filling in missing portions of a picture is known as image inpainting. Understanding the large-scale structure of natural pictures and executing image synthesis are both necessary solutions to this challenge. Prior to deep learning, the subject was explored, but with the advent of adversarial learning and deep and large neural networks, the field has advanced quickly. Inpainting algorithms are often trained using a sizable autonomously generated dataset produced by randomly masking real-world photos. Complex two-stage models with intermediate predictions are frequently used, along with segmentation maps, edges, and smoothed pictures. In the mentioned study, they use a straightforward single-stage network to get cutting-edge outcomes. For the purpose of comprehending an image's overall structure and resolving the inpainting problem, a broad effective receptive field is essential.

The information required to produce high-quality inpainting may also be difficult to obtain when the mask is big, despite having a limited yet broad receptive field. Common convolutional designs may not have an effective receptive field that is suitably big. To solve the issue and unleash the power of the one-stage solution, careful intervention has been taken in each system component. Particularly: An inpainting network based on recently discovered Fast Fourier Convolutions (FFCs) has been suggested. Even at the network's foundational levels, FFCs provide a receptive field that completely encloses a picture. This FFC characteristic enhances the network's parameter efficiency while also enhancing perceptual quality. The network can interestingly generalize to high resolutions that are never observed during training because of the inductive bias of the FFC. Since fewer training data and computations are required as a result of this discovery, it has considerable practical benefits of using the perceptual defect.

- *High Receptive Field (HRF) Perceptual Loss:* It is responsible for the supervised signal and consistency of the global structure.

$$\mathcal{L}_{HRFPL}(x, \hat{x}) = \mathcal{M}([\phi_{HRF}(x) - \phi_{HRF}(\hat{x})]^2) \quad (4)$$

- *Adversarial Loss:* It is used to make sure that details in the generated image looks natural as much as possible. A discriminator has been defined to differentiate between real and fake areas, fake areas are defined when they are found in the intersection with the masked area. HRF perceptual loss helps in copying the original parts of the

image [4]. Equation (5) illustrates the adversarial loss:

$$\begin{gathered}\mathcal{L}_D = -\mathbb{E}x[\log D\xi(x)] - \mathbb{E}x, m[\log D\xi(\hat{x}) \odot m] \\ -\mathbb{E}x, m\big[\log\big(1 - D\xi(\hat{x})\big) \odot (1 - m)\big] \\ \mathcal{L}G = -\mathbb{E}x, m[\log D\xi(\hat{x})] \\ LAdv = \mathrm{sg}_\theta(\mathcal{L}D) + \mathrm{sg}\xi(\mathcal{L}G) \to \min\theta, \xi\end{gathered} \tag{5}$$

- *The Final Loss Function:* Gradient penalty, and feature matching loss are both used in the final loss function, it improves the performance and keep the training more stable. Also, it is defined as the sum of the previous discussed losses (6):

$$\mathcal{L}_{\text{final}} = \kappa L_{Adv} + \alpha\mathcal{L}_{HRFPL} + \beta\mathcal{L}_{\text{DiscPL}} + \gamma R_1 \tag{6}$$

Jo and Park developed a system that offers image completion, and the system receives as inputs free-form mask, sketch, and color [5]. The masked region is being restored using an encode-decode generator. The discriminator is used to make sure the output result was real or fake. Some of the produced images have some awkward edges, many algorithms are used to solve this problem but every time there was another problem, such as FaceShop that produces unclear images when there is a large region is wiped out, also GuidedInpating plays on restoring the missing region from another image but this did not work well due to the difficulty to restore some details. So, SN-patchGAN is used to deal with those limitations. This system provides realistic images to the user. The same masking operations are used just as Deepfillv2 for facial images and adding hair masks using GFC. To complete the missing parts of the image this paper is using a generator of type encoder-decoder similar to the U-Net and gated convolution. The encoder downsamples the input and applies dilated convolutions, while the decoder uses transposed convolutions for upsampling. While the discriminator is based on SN-patchGAN. ReLu function is not applied on GAN loss. Convolution kernel of size 3 x 3 is used. The generated images are high quality images of size 512 x 512. Equations (7)-(10) describe the loss functions:

$$L_{G\_SN} = -\mathbb{E}\big[D\big(I_{\text{comp}}\big)\big] \tag{7}$$

$$\mathcal{L}_D = -\mathbb{E}_x\big[\log D_\xi(x)\big] - \mathbb{E}_{x,m}\big[\log D_\xi(\hat{x}) \odot m\big] - \mathbb{E}_{x,m}\Big[\log\Big(1 - D_\xi(\hat{x})\Big) \odot (1 - m)\Big] \tag{8}$$

$$\mathcal{L}_G = -\mathbb{E}_{x,m}\big[\log D_\xi(\hat{x})\big] \tag{9}$$

$$L_{Adv} = \mathrm{sg}_\theta(\mathcal{L}_D) + \mathrm{sg}_\xi(\mathcal{L}_G) \to \min_{\theta,\xi} \tag{10}$$

According to the final results, SC-FEGAN generates blurry hair. It gives results better than the two other discussed approaches in the same paper. Armanious et al. focus on repairing medical images such as MRI and CT using Generative Adversarial Networks (GANs) [5]. GAN has many types, the used type in this study is MedGAN. It is a combination of cascaded U-net generator (CasNet) with non-adversarial losses. The output image must be realistic and must fit the given information. In order to improve the inpainting performance, a discriminator is added. Ip-MedGAN approach has been discussed, based on conditional GAN. cGAN consists of two convolution networks, generator, and a discriminator. A picture input is given to a generator. The resulting image is sent to a discriminator, which determines whether or not it is a genuine equation (11):

$$\min_G \max_D \mathcal{L}_{\text{adv}} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{\hat{x},y}\big[\log\big(1 - D(\hat{x},y)\big)\big] \tag{11}$$

The added patch-based local discriminator focuses more on the details of the generated image.

$$\min_G \max_{D_{\text{L}}} \mathcal{L}_{\text{local}} = \mathbb{E}_{x_{\text{L}}}[\log D_{\text{L}}(x_{\text{L}})] + \mathbb{E}_{\hat{x}_{\text{L}}}\big[\log\big(1 - D_{\text{L}}(\hat{x}_{\text{L}})\big)\big] \tag{12}$$

To enhance the inpainting performance, there are two adversarial losses used in this approach in order to train the generator:

Using intermediate feature maps from a trained network, the generator is guided by the style reconstruction loss to match the style and textures of the target pictures.

$$\mathcal{L}_{\text{style}} = \sum_{n=1}^{N} \lambda_{sn} \frac{1}{4d_n^2} \|G_n(\hat{x}) - G_n(x)\|_F^2 \tag{13}$$

A non-adversarial loss called perceptual loss reduces pixel variances and inconsistencies to produce pictures that are more uniform. The global discriminator network's mean absolute error (MAE) between inputs and intermediate feature maps is used to assess it.

$$\mathcal{L}_{\text{percep}} = \sum_{n=0}^{B} \lambda_{pn} \|D_n(\hat{x},y) - D_n(x,y)\|_1 \tag{14}$$

To guarantee homogeneity, improve in-painted output, and reduce perceptual and style reconstruction losses, ip-MedGAN employs a CasNet generator, global discriminator, local discriminator, and local discriminator. Equation (15) illustrates the loss function:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{adv}} + \lambda_2\mathcal{L}_{\text{local}} + \lambda_3\mathcal{L}_{\text{style}} + \lambda_4\mathcal{L}_{\text{percep}} \tag{15}$$

For the final results, in-painted regions lacked details and had artificial tilting artifacts, whereas MedGAN enhanced sharpness and global uniformity. The patch-based local discriminator, proposed by the ip-MedGAN framework, improved textural quality, and eliminated tilting artifacts, yielding the top results across measures.

III. Dataset Description

Medical datasets for cleft lip are vital for gaining insights into the causes, treatments, and prevention of this congenital anomaly. However, due to the sensitive nature of patient information, these datasets are not published anywhere on the internet. Instead, researchers collect the data from different hospitals and clinics while ensuring that patient privacy is maintained at all times. There are strict protocols in place to

de-identify patient data, and access to the dataset is typically restricted to authorized researchers who have undergone rigorous training in data security and confidentiality. By collecting and analyzing these datasets, medical professionals can better understand the prevalence and incidence of cleft lip, identify risk factors, and develop new treatments and therapies to improve patient outcomes. It is essential that patient privacy is protected at all times, and medical professionals and researchers must continue to prioritize this aspect of data collection and analysis. The used dataset was a medical one. It contains images that are converted into RGB and resized to be 512*512. For the following dataset, data augmentation techniques have been applied to enrich the training set by adding new copies of the dataset in the same data as shown in Fig. 1.

Fig. 1 Data Augmentation

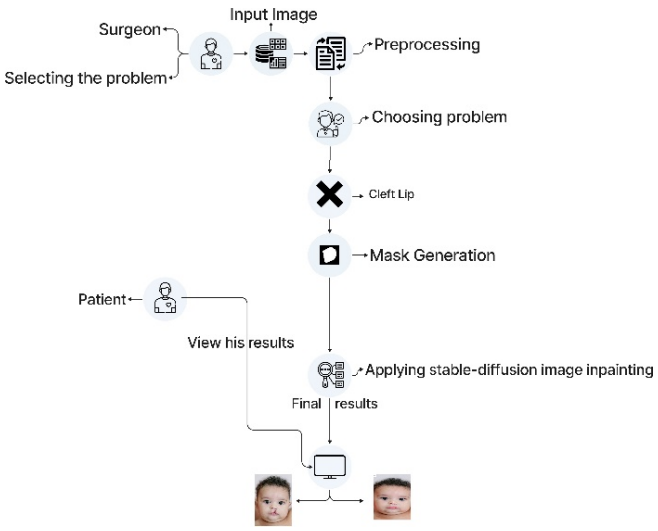

Fig. 2 System Overview

IV. Proposed Methodology

Fig. 2 illustrates stages within system and how they are related to each other. In the first place, a preprocessing phase had been begun in which input image is converted into RGB and resized to be 512*512. At that point, a new stage starts to make the provided image be clear as much as possible to be suitable for the model and apply the image inpainting approach. Then a mask is generated for the input image. Moving to the image inpainting stage, we experiment with three different approaches: Generative Adversarial Networks (GANs), Learned Masked Autoencoders (LaMa), and Stable Diffusion, to determine the most effective method for this task. In the proposed system, Stable Diffusion image inpainting showed the best result which makes the output image as real as possible. This shall show the patient the final result post operation as an output.

A. GAN

In GAN, both a Generator and a Discriminator are present. The Generator's primary role involves producing fabricated instances of data, spanning various domains such as images or audio, with the aim of deceiving the Discriminator. Conversely, the Discriminator strives to differentiate between genuine and counterfeit samples. Both the Generator and Discriminator function as Neural Networks and engage in a competitive relationship during the training stage.

B. LaMa

In Large Mask Inpainting, ResNet-like architecture has been used. The aim of this approach is to inpaint colored image *x*, using the mask *m*. The training is done on a dataset containing the original image and its mask. A feed-forward inpainting network which is known as generator has been used as well. There may be many fillings for the missing areas, so there are three loss functions are used.

C. Stable Diffusion

A pre-trained model is used for inpainting, using diffusers. This model takes two input images one for the original image and the other for the mask generated which is related to the provided image [7]. There are some operations done on the input image as a preprocessing stage, which are resizing the image to fit the model and converting the image into RGB, then it starts to process them as shown. They provide a short overview of diffusion models in this section. Diffusion models are a type of generative model that transforms Gaussian noise into samples that follow a learned data distribution through a repeated process of removing noise. These models can be conditional on various factors, such as class labels, text, or low-resolution images, Fig. 3.

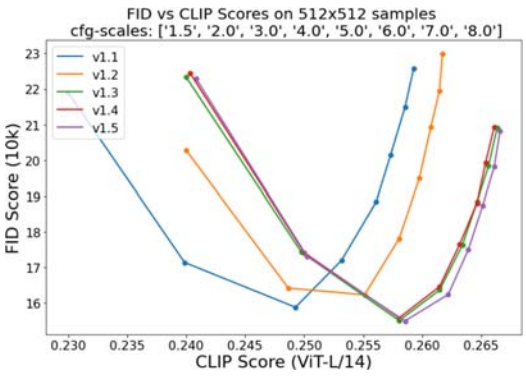

Fig. 3 Stable-Diffusion Result

D. Mask Generation

First, each image in the dataset is taken to a labeling tool to select the cleft lip part from each image individually. Second, each approach (GAN, LaMa, and Stable Diffusion) will be evaluated by generating an output JSON file containing the in-painted image data. To analyze the results, a separate method will process the generated JSON files from each approach (GAN, LaMa, and Stable Diffusion). This method will extract relevant image information and create a dictionary for each image, including its associated annotations. Third, a mask is generated by training these labeled images of cleft lip dataset on a Mask R-CNN model using Detectron2 library. After training the model, it extracts the masks from the output. Then "Visualizer" class from Detectron2 visualizes the predicted masks on the input image and saves the generated mask of each image separately. But due to the shortage of dataset the output mask was not selecting that part exactly but it was so close to it [6]. This means that by collecting more images for the dataset it would get more accurate.

$$PSNR(x,y) = \frac{10\log_{10}(\max(\max(x),\max(y))^2}{|x-y|^2} \quad (16)$$

- *SSIM:* It is a metric that measures the deterioration of image quality brought on by processing (like compression). Equation (17) shows how SSIM is being calculated:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad (17)$$

E. Measurement Units Results

Table I stated that the best result in both measurements is Stable-Diffusion. It is clear that a higher PSNR and SSIM values provide a higher image quality.

TABLE I
MEASUREMENTS

Functions	PSNR	SSIM
GAN	28.1	0.54
LAMA	28.5	0.48
Diffusion	30.1	0.63

F. SVM

Support Vector Machine is a machine learning algorithm for supervised learning that can be applied to regression and classification tasks. In the proposed system, it was used to detect whether the image imported contains a cleft lip or not.

V. EXPERIMENTAL RESULTS

After applying different image inpainting approaches, final result of each approach should be compared with each other to find the best result and the most real one.

A. Measurement Units

The comparison has been made based on the following two measurement units between two images the original input image and the resulting output image. For each of both measurements Mean Square Error (MSE) should be calculated.

- PSNR: Peak Signal-To-Noise Ratio: The ratio of an image's maximum potential power to corrupting noise's power determines how well the image is represented [9]. If the calculated MSE equals zero then both images are identical, otherwise that ratio will be presented. Equation (16) shows how PSNR is being calculated:

$$PSNR(x,y) = \frac{10\log_{10}(\max(\max(x),\max(y))^2}{|x-y|^2} \quad (16)$$

- *SSIM:* It is a metric that measures the deterioration of image quality brought on by processing (like compression). Equation (17) shows how SSIM is being calculated:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad (17)$$

B. SVM Results

Equations (18)-(21) are used to calculate the accuracy of the algorithm and how many right and wrong images got detected [8].

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{18}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{19}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{20}$$

$$\text{F}_1\text{-Score} = 2X\frac{\text{Precision } X \text{ Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{21}$$

Table II shows the results after the training of the algorithm.

TABLE II
SVM RESULTS

	precision	recall	f1-score	support
0	0.88	1.00	0.93	7
1	0.75	0.60	0.67	5
2	0.89	0.89	0.89	9
accuracy			0.86	21
macro avg	0.84	0.83	0.83	21
weighted	0.85	0.86	0.85	21

C. Final Results

As shown in Fig. 4, GANs model has some limitations: The output was not clear, it did not fix the cleft lip problem, also it reduced the resolution of the image. LaMa model did not give us the best result; it did not work on all images that holds cleft lip problem, and it cannot fill up all the missing or damaged areas in the input image. However, Stable-Diffusion was giving the best result so far; it filled up the image's missing and damaged areas.

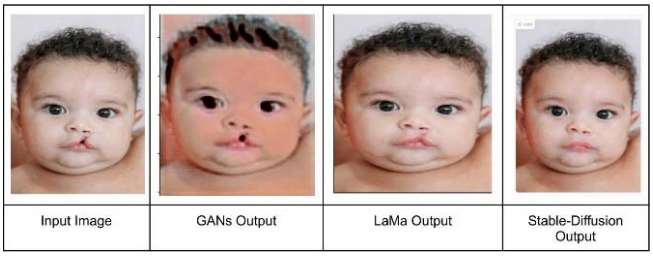

Fig. 4 Input image and final results of each approach

VI. CONCLUSION AND FUTURE WORK

This system gives the ability to the patient to see his appearance pre- and post-surgery before taking his decision to be more comfortable with it. An input image should be provided to the system to apply the model on it, then presenting what the patient will look like after the surgery as the output image. The model uses an image inpainting approach to generate a new image from the provided image by the user after selecting the part that holds the issue in the input image. Different algorithms were tried such as GAN, LaMa, and Stable Diffusion. After comparing the results using PSNR and SSIM, Stable Diffusion was giving the best results so far, while others were giving much lower results. In the near future, this system aims to be more generalized on different levels of problems people/patients are suffering from including burning scars and wrinkles.

REFERENCES

[1] Shuang Chen, Amir Atapour-Abarghouei, Jane Kerby, et al. "A Feasibility Study on Image Inpainting for Non- cleft Lip Generation from Patients with Cleft Lip". In: *arXiv preprint arXiv:2208.01149* (2022).

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. "High-resolution image synthesis with latent dif- fusion models. 2022 IEEE". In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10674–10685.

[3] Ilkin Sevgi Isler, Chase Walker, Dominic Simon, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding".

[4] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, et al. "Resolution-robust large mask inpainting with fourier convolutions". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022, pp. 2149–2159.

[5] Youngjoo Jo and Jongyoul Park. "Sc-fegan: Face editing generative adversarial network with user's sketch and color". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1745–1753.

[6] Karim Armanious, Youssef Mecky, Sergios Gatidis, et al. "Adversarial inpainting of medical image modalities". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3267–3271.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10684–10695.

[8] SVM Vishwanathan and M Narasimha Murty. "SSVM: a simple SVM algorithm". In: *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*. Vol. 3. IEEE. 2002, pp. 2393–2398.

[9] Alain Hore and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM". In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 2366–2369.