# A Feasibility Study on Image Inpainting for Non-cleft Lip Generation from Patients with Cleft Lip

Shuang Chen[1], Amir Atapour-Abarghouei[1], Jane Kerby[2], Edmond S. L. Ho[3],
David C. G. Sainsbury[2], Sophie Butterworth[2], Hubert P. H. Shum[1,†]

[1] Department of Computer Sciences, Durham University, Durham, United Kingdom
[2] Department of Plastic and Reconstructive Surgery, The Newcastle Upon Tyne Hospitals NHS Foundation Trust
[3] School of Computing Science, University of Glasgow, Glasgow, United Kingdom

*Abstract*—A Cleft lip is a congenital abnormality requiring surgical repair by a specialist. The surgeon must have extensive experience and theoretical knowledge to perform surgery, and Artificial Intelligence (AI) method has been proposed to guide surgeons in improving surgical outcomes. If AI can be used to predict what a repaired cleft lip would look like, surgeons could use it as an adjunct to adjust their surgical technique and improve results. To explore the feasibility of this idea while protecting patient privacy, we propose a deep learning-based image inpainting method that is capable of covering a cleft lip and generating a lip and nose without a celft. Our experiments are conducted on two real-world cleft lip datasets and are assessed by expert cleft lip surgeons to demonstrate the feasibility of the proposed method.

*Index Terms*—Cleft Lip, Image Inpainting, Deep Learning, Multi-task

## I. INTRODUCTION

A cleft lip is a congenital condition which arises during pregnancy in the early stages of development where the upper lip does not fuse together. One in every 700 births in the UK has a cleft lip and palate [1]. Patients with an orofacial cleft require surgical treatment by a cleft lip and palate surgeon at an average age of three months to correct cleft lip [2].

Achieving symmetry and improving nasolabial appearance is a fundamental goal of cleft lip surgery [3]. There are various surgical approaches to repairing a cleft lip. The most commonly used worldwide at present are a Millard repair or a Fisher repair [4]. Repairing cleft lips is a specialist skill and training in the UK requires an extended period of subspeciality training. Evaluating the outcome of cleft lip and palate surgery is an essential part of being able to improve surgical technique. The current gold standard for assessing outcomes is the Asher-McDade rating scale by using a 5-point standard scale to assess nasolabial profile, nasal symmetry, nasal form, and vermilion border [5].

Recently, with the rapid advancement of AI, technologies based on deep learning have emerged to locate cleft lip surgical annotation and incisions to facilitate surgery [6]. This may help junior surgeons in the early stages of their career and also surgeons who may not be as familiar with repairing cleft lips. Other clinical applications include being able to predict the outcome of a cleft lip repair which would enable surgeons to adjust their surgical procedure to provide the best outcome possible.

Our work aims to generate a non-cleft lip from an image of a baby with a cleft lip while protecting patient privacy. StyleGAN [7] would be one candidate, given its impressive performance in style transfer tasks. However, due to their confidential nature, images of an individual with a cleft lip are not readily available, making it challenging to feed the data-hungry StyleGAN. Even with sufficient data for training, once approaches such as StyleGAN fall victim to a model inversion attack, malicious users will be able to steal photos of real patients with a cleft lip [8], posing a grave privacy risk to patients. In this case, image inpainting is preferable, since we do not use images of individuals with cleft lip for training, the patient photographs leave no traces in the model.

In the facial inpainting task, the surgical evaluation criteria correspond to the semantic plausibility of the face shape and the image quality. Existing advances in facial inpainting typically ensure the accuracy of generated facial attributes by supplying supplementary facial geometry information. EdgeConnect [9] first generates a structure map, then combines the corrupted image to perform image inpainting in the second stage. However, the correlation between structural information and texture is frequently redundant and unreliable [10]. Human face can be modelled using landmarks and their geometrical features [11]. Lafin [12] uses landmarks as indicators to more precisely described facial attributes. Nonetheless, both of EdgeConnect and Lafin have a multi-stage limitation: the final image quality is highly dependent on how well the indicator generation in the first stage works.

In this paper, we propose a single-stage end-to-end multi-task image inpainting framework to generate non-cleft lip from patients with cleft lip. We use adaptive feature fusion and landmark indicator to boost parameter sharing and utilize the second task more efficiently. The landmark prediction is guided by both masked image and partial inpainted information, resulting in a more precise geometry indicator for repairing facial attributes. Moreover, since we do not use images of individuals with cleft lips to train the model, the patient is protected against privacy violations.

To evaluate our model and inform the feasibility of our proposed method, we collected two datasets, CleftLip10 and CleftLip24, from real patients for testing. Specifically, we use segmentation masks to cover the cleft lip and medical equipment. Then our model automatically synthesizes a lip and nose without a cleft. We compare our results with EdgeConnect [9], Lafin [12], and CTSDG [13], which also
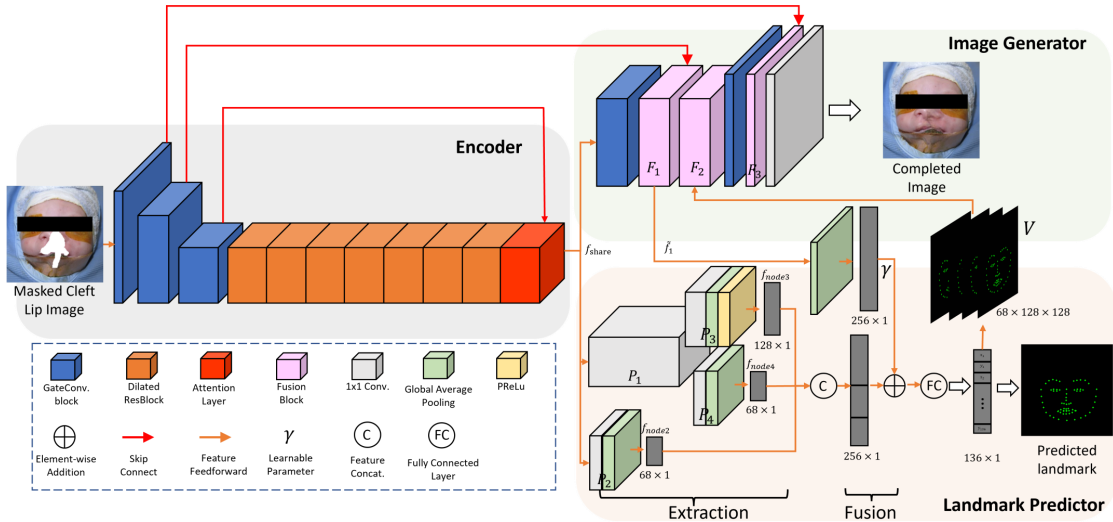
Fig. 1. Overview of the proposed method.

challenged facial inpainting. We invited three professional cleft lip surgeons to rank the results. We also include quantitative results on CelebA [14] to show the advantages of our design.

Our main contributions are summarized as follows:

- We propose an image inpainting approach to produce an non-cleft lip image from patients with cleft lip. The code is publicly available to enable better reproducibility⋆.
- We propose a multi-task network in which branches cooperate with each other through parameter sharing between tasks, which can achieve both landmark prediction and image inpainting at the same time.

## II. METHODOLOGY

We propose an end-to-end multi-task model, the framework is shown in Fig. 1. We first train our model on CelebA, then perform inference step on real patient images with cleft lip to generate non-cleft lip with semantic plausible facial attributes. Our model can simultaneously perform image inpainting and facial landmark prediction. The paramters in two tasks are shared through image-to-landmark and landmark-to-image feature fusion operations. Formally, the whole pipeline could be denoted as:

$$(\hat{I}, \hat{L}) = G(I \odot (1 - M)), \quad (1)$$

where $G$ is our multi-task model, $I$ is the real image and $M$ denotes the segmentation mask that occludes the cleft lip and medical equipment. $\hat{I}$ and $\hat{L}$ are completed image and predicted landmarks respectively.

### A. The Dataset

To verify that our model can work on real clinical patient pictures, we collected CleftLip10 and CleftLip24 datasets. The datasets composed of frontal face images of patients who underwent a cleft lip repair at the Royal Victoria Infirmary (RVI) in Newcastle upon Tyne during the outpatient clinic. CleftLip10 contains images taken from 10 patients pre-operatively and immediately post-operatively. The images were taken using a Canon PowerShot G1 X Mark II Camera

⋆https://github.com/ChrisChen1023/NCLG-MT

with a resolution of $3072 \times 2048$. CleftLip24 is comprised solely of pre-surgery images of twenty-four patients. The images were taken using a Canon EOS 20D or 5D Mark II DSLR Camera (with a 105mm lens) with a resolution of $2574 \times 3861$. All images were used in our experiments.

### B. Encoder and Image Generator

The encoder and the image generator jointly perform the inpainting task. The masked cleft lip image is downsampled three times and fed into the dilated convolutional residual blocks used to improve the receptive field, followed by a short-long attention layer to match feature more efficiently. We use gated convolutions instead of vanilla convolutions only in the image downsampling and downsampling stages. This is because 1) using gated convolution is more efficient for irregular masks [15], 2) its sensitivity to valid and missing pixels seems to be significant only for encoder and decoder [16] and 3) extensive use of gated convolution lead to a significant increase in parameter count. The shared feature is extracted at the end of the encoder:

$$f_{\text{share}} = E(I \odot (1 - M)), \quad (2)$$

where $E$ is the encoder and $f_{\text{share}}$ is the deep feature from the attention layer (See Fig.1 (Encoder)).

The image generator is designed to up-sample $f_{\text{share}}$ and reconstruct a non-cleft lip and nose. We employ three feature fusion blocks to facilitate parameter sharing, which are denoted by $F_1$, $F_2$, $F_3$ respectively. $F_1$ and $F_3$ aim to fuse the uncompleted image features from encoder by skip connections to generate more exquisite results by combining low-level and high-level feature.

$$\widetilde{f_l} = \begin{cases} F_i\left(\text{Concat}\left(f_{ei}, f_{di}\right)\right), & \text{if } (i = 1, 3) \\ F_i\left(\text{Concat}\left(\widetilde{f_1}, V\right)\right), & \text{if } (i = 2) \end{cases}, \quad (3)$$

where $\widetilde{f_i}$ is the result from fusion block $F_i$ ($i = 1, 2, 3$). $f_{ei}$ and $f_{di}$ is the feature map from corresponding encoder and decoder layer. After $F_3$ followed by a vanilla convolution layer, completed image is genereated. $F_2$ is designed to fuse the Landmark map $V$ from the landmark predictor, which will be detailed in the next subsection.

## C. Landmark Predictor

The landmark predictor involves extraction, fusion block and a fully-connected layer, aims to predict facial landmarks and inform the generator for assisting image inpainting.

The extraction step is designed to collect the landmark information from the encoded image feature. Specifically, $f_{\text{share}}$ is fed to a $1 \times 1$ convolutional layer $P_1$ to increase dimensionality, then we conduct dimensionality reduction followed by global average pooling to extract the feature into two vectors with different lengths ($P_4$ and $P_3$). Particularly, there is a $PReLu$ layer at the end of $P_3$ for non-linear projection. Simultaneously, $P_2$ also returns a vector after dimensionality reduction and global pooling directly acting on $f_{\text{share}}$, then we concatenate them:

$$f_{lmk} = \text{Concat}\left(f_{\text{node2}}, f_{\text{node3}}, f_{\text{node4}}\right), \quad (4)$$

where $f_{\text{node}i}$ is the corresponding vector from $P_i$. In existing multi-stage networks [9], [12], generated indicators are assumed as perfect and used in final inpainting stage directly. A faulty indicator may mislead image inpainting. To involve both corrupted and regenerated information in landmark predictor, and strengthen the parameter sharing between two tasks, we adaptively borrow $f_1$ from inpainting task, followed by a global average pooling, we merge it with the concatenated landmark feature vector:

$$f'_{lmk} = \gamma * \widetilde{f_1} \oplus f_{lmk}, \quad (5)$$

where $\gamma$ is a trainable weight with zero initialization and $\oplus$ is element-wise addition. Finally, we apply a fully-connected layer to predict facial landmark points.

To strengthen the parameter interaction between the two tasks and improve the completed image quality, we further map the landmark points into a binary feature map $V$, which is integrated with texture information in $F_2$. Formally, let $v_{pq}$ be the value in $V$ at position $(p, q)$:

$$v_{pq} = \begin{cases} 1, & \text{if } (p = [\alpha x_i], q = [\alpha y_i]) \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

where $\alpha$ is a scale factor corresponding the size of the feature map in $F_2$, [.] means integer operation. We create a $68 \times 128 \times 128$ tensor with landmark annotations, and transfer it to $F_2$ to provide facial geometry indicators.

## D. Loss Function

We follow Yang et al. (2019) to design our loss function, since our work applies a multi-task architecture, the landmark loss is involved in a joint loss function:

$$\mathcal{L}_{lmk} = \left\| \hat{L} - L_{gt} \right\|_2^2. \quad (7)$$

We consider $L_1$ loss, adversarial loss, style loss, perceptual loss, total variation loss and landmark loss. Given a masked image $I$, the ground truth image $I_{gt}$ and corresponding landmark ground truth $L_{gt}$. The overall loss function is:

$$\begin{aligned} \mathcal{L}_{\text{total}}(I, I_{gt}, L_{gt}) = &\mathcal{L}_{\text{pixel}} + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}} + \lambda_{sty}\mathcal{L}_{\text{style}} \\ &+ \lambda_{tv}\mathcal{L}_{tv} + \lambda_{adv}\mathcal{L}_{adv_G} \\ &+ \lambda_{lmk}\mathcal{L}_{lmk}, \end{aligned} \quad (8)$$

where $\lambda_{\text{perc}} = \lambda_{sty} = \lambda_{tv} = 0.1$, $\lambda_{adv} = 0.01$, $\lambda_{lmk} = 0.00046$.

## III. EXPERIMENTS

### A. Training Details

We train our model with CelebA [14], which is a popular human face dataset containing over 160 thousands training face images. For CelebA, we remove a few images which can not be obtained landmark ground truth. During training, the images are resized to $256 \times 256$ and we use irregular masks as in [17]. We use Adam optimizer and follow [9] to set $\beta_1 = 0$ and $\beta_2 = 0.9$. The learning rate $= 2.92 \times 10^{-4}$ and $2.92 \times 10^{-5}$ for discriminator, with a learning rate decay ratio of $0.78$. Batch size $= 4$.

TABLE I
VALID POSSIBILITY ON CLEFT LIP DATASET.

| Method | EC [9] | Lafin [12] | CSTDG [13] | Ours |
|---|---|---|---|---|
| CleftLip10 | 0.233 | 0.233 | 0.233 | **0.5** |
| CleftLip24 | 0.319 | 0.222 | 0.264 | **0.333** |

TABLE II
AVERAGE RANKING ON THE CLEFT LIP DATASET.

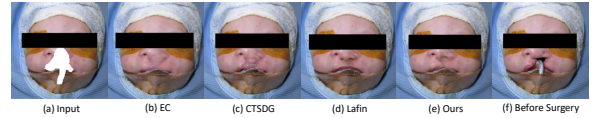| Method | EC [9] | Lafin [12] | CSTDG [13] | Ours |
|---|---|---|---|---|
| CleftLip10 | 1.857 | 1.714 | 2.429 | **1.267** |
| CleftLip24 | 1.696 | 1.813 | 1.947 | **1.208** |



Fig. 2. Visual comparison of different facial inpainting methods on real Cleft Lip dataset: (a) input masked image, (b) EdgeConnect [9], (c) CTSDG [13], (d) Lafin [12], (e) Ours, and (f) Before Surgery

### B. Experimental Validation

*1) Cleft Lips Repair:* We use the CleftLip10 and Cleft-Lip24 as test sets to compare our model with the current state-of-the-art facial inpainting methods [9], [12], [13]. The visualization results are shown in Fig. 2. We crop and resize them to $256 \times 256$, then design a mask to cover the cleft lip, as well as the medical equipment used during surgery, according to the type (unilateral and bilateral) and the severity of cleft lip for each patient. To better evaluate the feasibility of the proposed method, we invited NHS specialist cleft lip surgeons to assess the results based on the quality, consistency and validity. For each patient, results from four models are presented together. To avoid bias, the results are mixed and unlabelled. Images are deemed invalid if it is excessively blurry or illogical, e.g. flying lip or three nostril (see Fig. 2(d)). The valid probability represents the success rate of models in repairing cleft lips images (see Table I), and the average ranking represents the performance of each models in the valid repaired results (see Table II).

From our observation, each of the four models is capable for repairing small cleft lip areas. However, our method performs best for relatively complex situation, such as Fig. 2(f) with severe cleft lips and large medical equipment. The result from EC [9] is too blurry and CTSDG [13] leads obvious artifacts in regenerated region. Lafin [12] seems to be suffering from model collapse and was seriously misled by the input indicator, generating a full nose at the right nostril. From the surgeons assessment, our model generates more natural and semantically plausible images with a higher

## TABLE III
### QUANTITATIVE COMPARISON ON CELEBA.

| Mask Ratio | Model | PSNR | SSIM | FID |
|---|---|---|---|---|
| 0-20% | EC [9] | 36.1340 | 0.9880 | 0.4706 |
| | Lafin [12] | 35.9544 | 0.9870 | 0.5845 |
| | CTSDG [13] | 37.9275 | 0.9908 | **0.3420** |
| | Ours | **38.1083** | **0.9911** | 0.3421 |
| 20-40% | EC [9] | 28.3684 | 0.9486 | 3.1275 |
| | Lafin [12] | 28.2797 | 0.9476 | 3.3880 |
| | CTSDG [13] | 29.3860 | 0.9570 | 2.8436 |
| | Ours | **29.6678** | **0.9595** | **2.8327** |
| 40-60% | EC [9] | 23.4513 | 0.8561 | 6.1253 |
| | Lafin [12] | 23.5109 | 0.8614 | 6.5367 |
| | CTSDG [13] | **24.3130** | **0.8762** | 8.7051 |
| | Ours | 24.2076 | 0.8726 | **4.3419** |

valid possibility. Additionally, our model is able to generate textures similar to post-surgical scars while we leave certain intimation to the model (see Fig. 2(e)).

*2) Facial Inpainting:* We compare our model with current state-of-the-art facial inpainting models on CelebA. The evaluation metrics involve peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [18] and Frechet Inception Distance (FID) [19], which is shown in Table III. Higher PSNR, SSIM, and lower FID, indicate better generated image quality. We observe that our model overall suppresses state-of-the-art inpainting models in terms of small and medium masked ratio. The latest CTSDG outperforms ours by a small margin in large missing regions case in terms of PSNR and SSIM, but it is much lower than ours in FID.

## TABLE IV
### ABLATION STUDY ON CELEBA.

| | Mask | Irregular Mask | | | Regular |
|---|---|---|---|---|---|
| | Mask Ratio | 0-20% | 20-40% | 40-60% | Mask |
| PSNR | Baseline | 37.1742 | 29.1189 | 23.7541 | 25.9412 |
| | Base+Lmk | 37.3932 | 29.2115 | 23.7948 | 26.074 |
| | Ours | **38.1083** | **29.6678** | **24.2076** | **26.685** |
| SSIM | Baseline | 0.9895 | 0.9545 | 0.8605 | 0.9113 |
| | Base+Lmk | 0.9897 | 0.9554 | 0.8626 | 0.9144 |
| | Ours | **0.9911** | **0.9595** | **0.8726** | **0.9231** |
| FID | Baseline | 0.5330 | 3.3655 | 5.7635 | 3.6037 |
| | Base+Lmk | 0.4614 | 2.9762 | 5.1124 | 3.410 |
| | Ours | **0.3421** | **2.8327** | **4.3419** | **3.274** |

To validate the effectiveness of our multi-task architecture, we remove the parameter sharing between two tasks and take encoder followed by image generator as the baseline. Then, we implement the landmark predictor (Base+Lmk) and gated convolution (Ours) progressively. As shown in Table IV, the integration of both the multi-task model and gated convolutions improve the performance on both irregular and regular masks.

## IV. CONCLUSION

In this paper, we propose a novel approach to provide an guidance image for cleft lip surgery by masking the cleft lip part and generating lip and nose without cleft. To achieve this task, we design a multi-task image inpainting model that can better protect patient privacy. We collected two real-world patient datasets to demonstrate the feasibility of proposed approach. Three expert cleft lip surgeons assessed that our

design outperforms state-of-the-art methods in both valid possibility and image quality, while the performance of our model on CelebA also suppresses the state-of-the-art facial inpainting counterparts.

## V. ACKNOWLEDGMENT

## REFERENCES

[1] "Cleft lip and palate," 2019. [Online]. Available: https://www.nhs.uk/conditions/cleft-lip-and-palate/

[2] W. Wellens and V. Poorten, "Keys to a successful cleft lip and palate team," *B ENT*, p. 3, 2006.

[3] D. G. Mosmuller, L. M. Mennes, C. Prahl, G. J. Kramer, M. A. Disse, G. M. Van Couwelaar, B. N. Frank, and J. Don Griot, "The development of the cleft aesthetic rating scale: a new rating scale for the assessment of nasolabial appearance in complete unilateral cleft lip and palate patients," *The Cleft Palate-Craniofacial Journal*, vol. 54, no. 5, pp. 555–561, 2017.

[4] T. A. Patel and K. G. Patel, "Comparison of the fisher anatomical subunit and modified millard rotation-advancement cleft lip repairs," *Plastic and reconstructive surgery*, vol. 144, no. 2, pp. 238e–245e, 2019.

[5] C. Asher-Mcdade, C. Roberts, W. C. Shaw, and C. Gallager, "Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate," *The Cleft Palate-Craniofacial Journal*, vol. 28, no. 4, pp. 385–391, 1991.

[6] Y. Li, J. Cheng, H. Mei, H. Ma, Z. Chen, and Y. Li, "Clpnet: cleft lip and palate surgery support with deep learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3666–3672.

[7] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[8] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[10] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Image inpainting guided by coherence priors of semantics and textures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6539–6548.

[11] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?" *Computer vision and image understanding*, vol. 102, no. 2, pp. 117–133, 2006.

[12] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "Lafin: Generative landmark guided face inpainting," *arXiv preprint arXiv:1911.11394*, 2019.

[13] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 134–14 143.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[15] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Freeform image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[16] C. Cao and Y. Fu, "Learning a sketch tensor space for image inpainting of man-made scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 509–14 518.

[17] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.