CleftLipGAN : Interactive GAN-Inpainting for Post-Operative Cleft Lip Reconstruction

Daniel Anojan Atputharuban[1], Christoph Theopold[2], and Aonghus Lawlor[1]

[1] The Insight Centre for Data Analytics, School of Computer Science, University College Dublin, Dublin, Ireland
[2] Children's Health Ireland at Temple Street, Dublin, Ireland
{daniel.anojan, aonghus.lawlor}@insight-centre.org

Abstract. Synthetic generation of post-surgical outcomes holds significant value in the clinical domain, particularly for Cleft lip and Palate surgery. These synthetic images can be utilized for surgical planning, serve as reference points to evaluate surgical success and assist in educating patients and caretakers about potential outcomes. Image inpainting is effective for selectively generating Cleft-affected regions, making it a promising technique for this task. However, due to the lack of publicly available Cleft-specific datasets, Cleft inpainting models are typically trained on healthy data and applied to Cleft conditions to generate post-surgical lip appearances. Existing Cleft inpainting methods often struggle to capture the complexities of Cleft deformities, leading to implausible outcomes that fail to reflect the unique structural characteristics of Cleft-affected regions. To address this, we propose a Structural Guided Pluralistic Inpainting model, trained on healthy images, which allows for real-time, interactive adjustments to synthesize Cleft-specific images. We demonstrate the model's effectiveness by generating images that closely resemble Cleft conditions and benchmarking it against existing GAN-Inpainting methods. Additionally, we provide a user-friendly interface designed as a tool for post-surgical visualization of Cleft conditions. The source code is available at `https://github.com/danielanojan/CleftLipGAN.git`

Keywords: Synthetic Lips · Pluralistic Inpainting · Cleft Lip and Palate

1 Introduction

Cleft lip and Palate is a prevalent congenital facial anomaly, affecting approximately 1 in 700 childbirths [42]. Cleft condition is primarily caused by genetic and environmental factors, which contribute to varying degrees of craniofacial malformations during fetal development. These malformations result in incomplete fusion of the upper lip and/or palate, leading to the formation of Cleft conditions at birth. [25]. Cleft conditions can significantly impact speech, feeding and dental health, and they also impose socioeconomic and psychological

burdens on families and caregivers [4, 14, 46]. Treatment for Cleft repair typically involves a series of complex surgical procedures and rehabilitation, starting with initial repair surgery performed at 6 to 12 months of age. The primary goal of these surgeries are to restore essential facial functions and enhance both facial appearance and symmetry [56]. Despite advancements in surgical techniques, the inherent complexity of Cleft conditions and the variability of surgical outcomes pose significant challenges in both surgical planning and surgical outcome assessment [38]. Cleft surgical success is measured through functional outcomes such as speech tests, and aesthetic evaluations using facial markers. However, it is noteworthy that there are no universally accepted surgical planning protocols for Cleft lip repair, and the aesthetic assessments remain subjective, leading to inconsistencies and potential bias in clinical outcomes [13,48,49,51]. These challenges underscore the need for standardized methods for surgical planning and surgical evaluation, where machine learning techniques could enhance precision and consistency.

Synthetically generating variations of post-operative lip regions for patients with Cleft conditions plays a pivotal role in surgical planning and evaluation. By visualizing potential outcomes before surgery, surgeons can better tailor their approach, improving preoperative planning accuracy. These synthetic images serve as consistent and objective reference points for assessing surgical success. Additionally, they serve as educational tools, helping surgeons explain outcomes to parents and caregivers, enhancing their understanding of the surgical process. Thus, synthetic generation of post-operative lip images holds substantial value in improving both clinical outcomes and patient education.

Image inpainting models are particularly well-suited for addressing selective image region synthesis, as they focus on reconstructing missing areas of an image while preserving the surrounding context [53]. In particular, facial inpainting models have shown great promise in generating realistic and semantically coherent facial regions, which are used in real world applications such as facial restoration and digital forensics [10,47]. However, a major challenge arises from the lack of publicly available datasets specific to Cleft condition, which hinders the development of facial inpainting models tailored for Cleft condition [1]. To overcome this limitation, Cleft inpainting models are typically trained on images of healthy individuals and later applied to Cleft-affected regions to generate non-Cleft facial areas or simulate ideal post-surgical outcomes resembling those of children without Cleft conditions [1, 6].

However, it has been observed that models trained on healthy facial datasets struggle to capture the complexity and variability of Cleft deformities, making it difficult to produce realistic post-surgical results. These models often generate overly smoothed upper lips, resembling average healthy lips [1], without capturing the unique features of Cleft conditions. The underlying issue lies in the fact that inpainting models pre-trained on healthy faces fail to capture the semantic and structural details specific to Cleft-affected faces.

On the other hand, research in Cleft conditions has demonstrated that Cleft deformities can be effectively modeled using facial landmarks [2,21]. Motivated by this observation, we have developed a structural guided pluralistic inpainting model with real-time editing capabilities. Our inpainting model, trained on images of healthy individuals, effectively reconstructs the facial features of patients with Cleft conditions and can be applied to Cleft-affected areas to generate realistic non-Cleft regions. Additionally, we demonstrate the model's ability to produce plausible results in real time by manipulation of facial keypoints. To enhance usability, we have developed an interactive user interface that facilitates real-time modeling of these adjustments, offering a valuable tool for both surgical planning and outcome assessment.

2 Related Work

In this study, we categorize the relevant prior work into two distinct areas: facial inpainting (Section : 2.1) and the application of machine learning techniques for Cleft-related conditions (Section : 2.2).

2.1 Facial Inpainting

Image inpainting refers to the process of synthetically generating missing parts of an image, ensuring that the result is both visually realistic and contextually meaningful. GAN-based inpainting models have demonstrated state-of-the-art performance by generating plausible content for large corrupted regions. These models typically follow an encoder-decoder architecture. Different methods have explored to improve the representation capability in inpainting process such as coarse-to-fine training [15,27,57], structural guidance [15,37,54,57,58], local and global discriminators [18], and semantic segmentation-based losses to preserve facial structure [28]. Specialized network modules, such as partial [31], Fourier [47], and gated [57] convolutions, are employed to effectively handle corrupted image regions. Additionally, techniques like pixel shuffle upsampling [7] and attention-based upsampling [60] help mitigate feature degradation during image propagation. With the success of transformers in machine translation tasks, vision transformer-based GAN-Inpainting models have been proposed to enhance long range dependencies present in images, surpassing the performance of traditional convolution-based GAN-Inpainting frameworks [11, 23, 27, 61, 62]. However, the quadratic complexity of self-attention poses a bottleneck for implementing end-to-end vision transformers in inpainting. To overcome this limitation, variations of transformer-based inpainting models have been developed with linear computational complexity, integrating the strengths of both convolution and attention mechanisms [7, 9, 60].

Structure-guided inpainting methods incorporate prior guidance to generate semantically consistent inpainting results. Traditional GAN-Inpainting models employ various forms of structural guidance, such as facial keypoints [54], canny edges [15, 37, 60], segmentation maps [5], and gradient maps [12]. While edge,

segmentation, and gradient information generally offer more robust guidance compared to landmark-based methods, the accuracy of the structural predictions is critical for achieving high-quality inpainting results. Inaccurate or redundant structural information can significantly degrade the performance of inpainting models, leading to artifacts or inconsistencies in the generated regions [54]. In parallel, pluralistic inpainting methods aim to generate multiple plausible solutions for a missing region, enhancing the diversity of inpainting outputs [3, 66]. Guided pluralistic inpainting models further refine these outputs by incorporating user input [30, 65] or prior information [24, 34].

Building on these approaches, in this work we incorporate edge priors to guide the inpainting process, ensuring semantically consistent results while enabling the generation of pluralistic outputs that can represent Cleft conditions. During the training phase, we use edge contours derived from a face parsing module for structural guidance. In the inference phase, we leverage facial keypoints, connecting them to form edge contours which guide the inpainting process. Facial keypoints are employed in inference phase for their ease of prediction and flexibility in modifying landmark positions, which facilitates the generation of diverse and realistic outcomes.

In GAN-based inpainting models, discriminators are crucial for enhancing the realism of generated textures by guiding the generator to produce visually coherent and perceptually pleasing results. Traditionally, GAN inpainting models employ DCGAN based [41] image level discriminators. However, image level discriminators fails to generate inpainting regions that maintain local consistency with surrounding areas. To address this limitation, global and local discriminators have been introduced to ensure overall image consistency and local coherence [18].

Building on this concept, patch-wise discriminators were proposed as a more generalized approach, focusing on small patches rather than the entire image. This design enhances the model's ability to distinguish between real and synthesized regions locally, contributing to more precise inpainting results [32, 63, 64]. Further advancements in patch-wise discriminators include the incorporation of spectral normalization, which constrains the spectral norm of the discriminator's weight matrix, leading to smoother gradient updates and more stable training processes [**?**, 37, 57]. Yang et al. [54] improved this approach by integrating attention layers into the discriminator to adaptively manage features, thus enhancing image consistency. Additionally, WGAN [55, 59] and LSGAN [43] based discriminators have been explored for faster and more stable training. Despite these innovations, the design of discriminators in GAN-inpainting models remains comparatively under explored, especially when contrasted with the extensive modifications applied to generators.

Inpainting models are often trained with free-form masks, where irregularly shaped masks are randomly placed on images during training [8, 31, 37, 54]. In contrast, our method employs fixed-form masks [15, 50, 58], specifically conditioning the inpainting of Cleft-prone upper lip and philtrum regions based on

lower lip and surrounding regions which remains unaffected in Cleft conditions. This targeted approach ensures a more focused reconstruction of the affected areas. Our proposed inpainting mask covers between 10% and 30% of the total image area.

2.2 Machine Learning Applications for Cleft Condition

Machine learning applications for Cleft Lip and Palate analysis can be broadly divided into three subdomains: Cleft surgical marker localization [29,44,45], Cleft severity prediction [35,39] and synthetic generation of Cleft-specific facial regions [1, 6, 16]. Facial landmark models are instrumental in predicting Cleft-specific landmarks for tasks such as prenatal diagnosis, Cleft severity assessment, surgical planning, and Cleft surgical success assessment. Petcas et al. [39] used pretrained Facial Beauty Prediction (FBP) models to predict Cleft severity, benchmarking their results against human ratings. However, their study shows that FBP models trained on healthy individuals fail to capture the key visual indicators specific for Cleft severity.

Due to the lack of publicly available Cleft-specific facial datasets, models for generating Cleft-specific synthetic faces are trained using publicly accessible facial datasets of healthy individuals. These synthetic faces are designed to resemble post-operative Cleft lips, which appear similar to normal lips, using GAN-based inpainting or inversion techniques. Hayajneh et al. [16] have deployed GAN inversion technique on Stylegan2 [20] latent space to generate natural looking normal lips for Cleft patients. They further extend the work to generate a synthetic Cleft dataset using GAN inversion techniques. While this work has explored the development of a synthetic Cleft dataset to address the lack of publicly available data, the authors report an inherent limitation in the model's ability to accurately categorize Cleft faces within StyleGAN's latent space. This limitation makes it challenging to modify the severity of the Cleft condition without altering surrounding facial features [16]. A GAN-Inpainting approach has been proposed by Chen et al. to generate post-operative Cleft faces [6]. This model is trained on the CelebA dataset [33] using free-form masks and processes the entire facial image at a resolution of 256x256 to generate post-operative Cleft faces. Atputharuban et al. [1] trained a GAN-inpainting model to generate upper lip and philtrum regions that resemble post-operative Cleft conditions and developed a surgical success score using the GAN-inpainted images as reference points to assess Cleft surgery outcomes. However, It can be observed that the inpainting models generate flat, asymmetric lips that do not accurately represent Cleft conditions. These models struggle to capture the unique anatomical features of Cleft-affected lips. We build upon this work by generating realistic Cleft lips with real-time editable capabilities to model varying Cleft severities.

3 Methodology

The proposed approach consists of three key components. First, the dataset acquisition pipeline, which involves obtaining Cleft-prone orofacial regions of the face, the inpainting mask, and the corresponding edge contour (Section. 3.1). Next, CleftLipGAN inpainting model is described in detail (Section. 3.2). Finally, the custom landmark detector module, for localizing keypoints in the orofacial region, is presented (Section. 3.3).

3.1 Dataset Acquisition

Experiments in this study are focused on Cleft-prone orofacial regions, as this is the area primarily affected by Cleft conditions. Clinicians use these specific cropped regions to assess surgical outcomes, making it highly relevant for our study. However, due to the unavailability of Cleft-specific datasets capturing orofacial regions, we created a dataset from the high resolution FFHQ facial dataset [19] to support our experiments.

FFHQ dataset [19] comprises 70,000 in-the-wild images of healthy individuals, representing a wide variation in age and racial demographics with the resolution of 1024 x 1024. We cropped the orofacial regions and generated custom fixed-form masks following the approach outlined in [1]. We utilized the facial keypoint detector and the face parsing model proposed by FaRL [67] for generating masks cropping the region of interest (ROI) and generating edge contour maps. Resulting images are resized to the resolution of 256 x 256. The dataset was then manually curated to remove images with occlusions, extreme poses, and blurred facial regions. The final dataset consists of 51,000 images, of which 46,000 are used for training inpainting models and remaining 5000 images are used for testing.

We use a post-operative Cleft dataset to evaluate the performance of inpainting models on Cleft condition. This fully anonymized dataset, consisting of 164 post-operative Cleft repair surgery images, was collected at Children's Health Ireland at Temple Street between 2009 and 2012 under a research agreement between Children's Health Ireland at Temple Street and University College Dublin. We refer to this post-operative Cleft dataset as Cleft164 dataset. Cleft surgeries were performed on children aged between 6 months to 1 year, with post-operative images captured during follow-up consultations when the children were 5 years old. Cleft164 dataset was used as a test set to assess the performance of CleftLipGAN model on abnormal post-operative Cleft images. Inpainting masks for the Cleft164 dataset was manually generated, encompassing the upper lip and philtrum region.

3.2 CleftLipGAN Model for Post-Operative Cleft Lip Reconstruction

We have proposed CleftLipGAN, an interactive inpainting model designed for generating post-operative Cleft images. Formally, the inpainting pipeline is for-

mulated as follows: the input for the inpainting module, I_{input}, is obtained by concatenating the masked image, $I_{\text{masked}} = I \odot M$, the mask image M and the corresponding edge contour I_{edge}. The input image, I_{input}, is then processed by the proposed inpainting model, CleftLipGAN, to produce a semantically appealing inpainted image, I_{out}. The overall formulation of Cleft lip reconstruction is denoted as $I_{\text{out}} = \text{CleftLipGAN}(I_{\text{input}})$. The following subsections describe sub modules of CleftLipGAN: namely generator, discriminator and the structural guidance pipeline. CleftLipGAN inpainting module pipeline is illustrated in Fig. 1.

Generator We adapt the generator module proposed by Xiankang et al. [60] for our inpainting model. The module is built in encoder-decoder fashion with U-shaped skip connections to pass shallow features from the encoder to decoder. The generator architecture is composed of Spatial Attention Based Gated Convolution(SAGC) module and Channel Attention based Gated Convolution(CAGC) module. The SAGC module is employed in both the encoder and decoder blocks of the network, aiding in the extraction of structural information. In the SACG module, spatial attention is used instead of self-attention to achieve linear complexity when handling high-resolution features in both the encoder and decoder. The bottleneck layers of the network are built using the CAGC block, which employs Squeeze-and-Excitation attention for semantic feature extraction. SAGC and CAGC blocks, along with the gated convolution layer, produce semantically consistent images without blurring or watermark artifacts. Gated convolution layers are employed for their ability to dynamically extract features using learned gating mechanisms, which help guide the weights of each pixel based on prior information. We choose the number of the bottleneck layers to be 4 for this task. The model inputs for the generator are masked RGB image, grayscale inpainting mask and grayscale edge contour.

Discriminator We adapt the discriminator proposed by [26] for image super-resolution task. This discriminator module incorporates semantic guidance by fusing features extracted by a pretrained feature extractor, allowing the discriminator to learn fine-grained distributions. The feature extraction branch built using a pretrained CLIP 'RN50' module [40], is selected for its robust representation capabilities. As illustrated in Fig. 2, the ground truth image is passed through the CLIP model to obtain semantic details, which are then fused with the patch-wise discriminator via the semantic-aware fusion block to guide the assessment of realism. Unlike vanilla discriminators that focus on coarse-grained image distributions, this enhanced module leverages self-attention and cross-attention mechanisms to discriminate fine-grained details, such as textures, aiding the inpainting process. The inpainted image and mask are input into the discriminator to condition the evaluation specifically on the inpainted regions.

Structural Guidance We train the inpainting network using edge contours corresponding to the lip region. We used the FaRL face parsing model [67] to

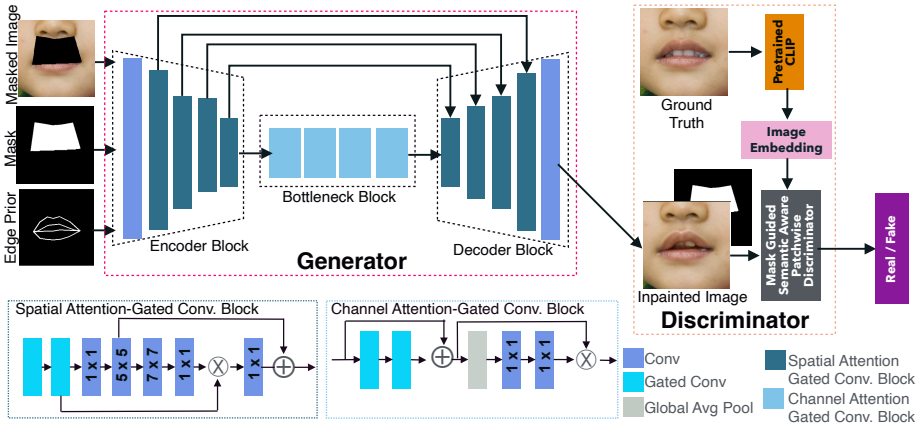

Fig. 1: Proposed CleftLipGAN inpainting model for reconstructing post-operative Cleft images: The model takes a masked image, inpainting mask, and edge prior as inputs to synthetically generate Cleft-prone regions.

obtain segmentation maps, from which we extract lip contours by outlining the segmented regions. During inference, we use facial keypoints generated by a custom keypoint detection model to construct lip contours by connecting the keypoints, following the approach proposed in [57]. This approach facilitates more flexible modifications, and by utilizing contours, richer structural information can be provided for inpainting.

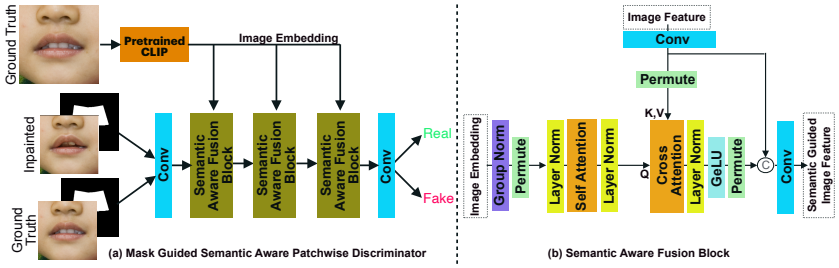

Fig. 2: Mask-guided semantic-aware patch-wise discriminator for the CleftLipGAN inpainting model: Image embedding features are extracted using a pretrained CLIP model. These features are fused with image features from the discriminator network through a Semantic-Aware Fusion block. This approach enhances the semantic extraction of patch-wise features, helping the discriminator better assess the realism of inpainted regions.

Loss Functions To produce high quality inpainting results we have incorporated a combination of loss functions in line with GAN-Inpainting literature [60].

We train the inpainting network with L_1 loss to ensure pixel-wise consistency and ensure realness in generated pixel values. Perceptual loss L_{perc} is used to enforce high level structural and semantic features, while adversarial loss is employed L_{adv} to improve overall quality of the output. The Loss function is denoted as

$$L_{\text{total}}(I_{\text{gen}}, I_{\text{gt}}) = \lambda_1 L_1 + \lambda_2 L_{\text{perc}} + \lambda_3 L_{\text{adv}}$$

where $\lambda_1 = 5$, $\lambda_2 = 0.4$ and $\lambda_3 = 0.05$ are chosen based on experiments.

3.3 Landmark Detection for Orofacial Region

We also developed a custom lips keypoint detector module specifically designed to predict keypoints in cropped orofacial region. In our approach, the predicted keypoints act as guidance during inference, enabling users to adjust and refine them to create edge contours, facilitating the generation of pluralistic results.

Our facial keypoint detector model is trained to predict 20 keypoints specifically from the cropped facial region. For this purpose, we have used a subset of 6000 images randomly obtained from CelebA-HQ [33] dataset, in which 5000 images were used for training and 1000 images were used for validation. We have pre-processed the images as outlined in Section. 3.1 and resize the images to 256 x 256 resolution. The ground truth keypoints for the lip region were obtained using the FaRL facial landmark model [67]. The keypoint detection model is built with a MobileNet [17] backbone and trained using L1 Loss. We measure the performance of keypoint detector model against the test dataset with Normalized Mean Squared Error(NSME), with the lip width used as the normalization factor. NMSE on test set is reported to be 0.23.

4 Experiments & Discussion

We have performed the experiments on a single NVIDIA GForce RTX 4090 GPU with the batch size of 4. We train the CleftLipGAN inpainting model with Adam optimizer [22] with $\beta_1 = 0.99$ and $\beta_2 = 0.5$. Model is trained with a learning rate of 5×10^{-4} and trained for 300000 iterations. It utilizes a warm-up phase consisting of 2 epochs, where only the generator is trained. Following this, generator and discriminator are concurrently trained to fine tune the adversarial loss.

We benchmark the performance of the CleftLipGAN inpainting model against state-of-the-art inpainting models using both quantitative and qualitative evaluations. For quantitative evaluation, we employ PSNR, SSIM [52], and Brisque [36] to assess the performance of the inpainting models. PSNR measures pixel-wise reconstruction accuracy, indicating how closely the inpainted result matches the original image in terms of visual fidelity. SSIM [52] evaluates structural similarity, ensuring that the inpainted regions maintain continuity with the surrounding pixels and remain contextually aligned. Brisque [36] evaluates the perceptual quality of an image by identifying distortions and visual artifacts that affect overall image quality.

Table 1: Quantitative comparison on the FFHQ test set shows that the proposed CleftLipGAN model outperforms in all three metrics employed.

Model	PSNR ↑	SSIM [52] ↑	BRISQUE [36] ↓
Ours	**31.46**	**0.9382**	**21.89**
AGG-Net [60]	30.33	0.9246	23.37
HINT [7]	26.25	0.9048	25.49
E2F-GAN [15]	27.89	0.9087	26.67
HourglassAttention [8]	29.12	0.9178	25.89
DeepfillV2 [57]	25.39	0.8979	26.17

For comprehensive evaluation, we compare our model against three structural guided inpainting models, E2F-Net [15], DeepfillV2 [57] and AGGNet [60]. We retrained the models using the hyperparameters specified by the authors, but incorporated our edge contour-based guidance for the inpainting process. We also benchmark against HINT [7] inpainting model, a vision transformer based model which has demonstrated superior performance compared to contemporary state-of-the-art inpainting models. Additionally, we evaluate our model's performance against the Cleft inpainting model proposed by Atputharuban et al. [1] for synthesizing Cleft-prone region, which is built on the HourglassAttention [8] architecture. We benchmark the performance of CleftLipGAN against this model to assess its effectiveness in Cleft inpainting.

Quantitative studies are performed on the test set proposed in Section. 3.1 consisting 5000 images and the results are presented in Table 1. It can be observed that our model outperforms the other state-of-the-art inpainting models on all evaluation metrics. This demonstrates that incorporating semantic-aware discriminators enhances inpainting results when combined with carefully selected generator architecture.

This observation can be further validated with qualitative evaluation. We conduct qualitative evaluation on Cleft164 dataset specified in Section. 3.1. Figure. 5 illustrates the post operative Cleft lip reconstruction results. It can be observed that non structure based inpainting models produce flat upper lips that lack symmetry and are not semantically accurate. In contrast, by adjusting facial keypoints in inference stage, structure based inpainting models which were trained on healthy individuals can generate lips which resemble that of Cleft condition. When compared to other structure guided inpainting models, CleftLipGAN model surpasses them by generating structurally coherent and visually appealing lips for Cleft conditions, with fine textural details preserved. Although E2F-GAN [15] can produce accurate structural results, it consistently fails to capture fine details and struggles to produce smooth textures. Additionally, the DeepFillV2 [57] model generates blurry results with visible artifacts.

We also evaluate the possibility of generating images resembling Cleft conditions, with the results presented in Fig. 3. It can be observed that CleftLipGAN produces images closely resembling Cleft conditions with fewer artifacts compared

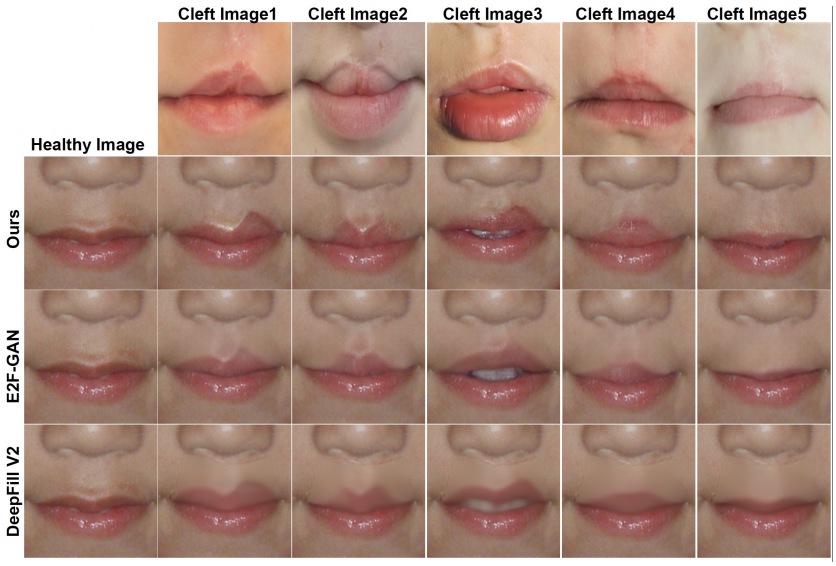

Fig. 3: Synthetic generation of Cleft lips from healthy lips. We generate lips that resemble the Cleft condition from healthy lip images by fine tuning facial landmark locations. All three structure guided inpainting models successfully produce Cleft-like lips; however, our model captures finer details and achieves a more accurate structural representation of Cleft lips, enhancing the realism and anatomical correctness of the generated results.

to other structure-guided inpainting models. We believe that the CleftLipGAN model can be employed to generate a synthetic dataset resembling Cleft conditions, particularly in this domain where no publicly available Cleft datasets exist.

Additionally, we have developed an interactive user interface, as illustrated in Figure. 4. The user interface integrates lip landmark prediction module (Section. 3.3) for identifying keypoints in the orofacial region and the inpainting model (Section. 3.2) for reconstruction of Cleft prone areas. The interface receives the image of the orofacial region and the corresponding mask, where the inpainting will be applied, as inputs. Baseline lip keypoints are predicted using the keypoint detector module, which can be interactively adjusted by the user to generate pluralistic results. Once the adjustments are finalized, an edge contour map is generated by connecting the keypoints. The masked image, inpainting mask, and edge contour map are then fed into the inpainting model to generate the inpainted image, which is displayed in the interface for further fine-tuning.

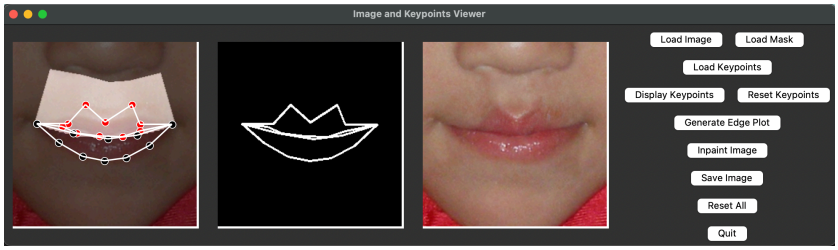

Fig. 4: User interface for selective inpainting for Cleft region. Lip keypoints can be manually adjusted to obtain a edge contour map, which in turn guides CleftLipGAN model to produce pluralistic inpainting results.

5 Conclusion and Future Work

In this study, we demonstrate the ability of image inpainting models trained on images of healthy individuals to capture the semantics of Cleft conditions and generate anatomically accurate lips for Cleft patients using structural guidance. To achieve this, we propose the CleftLipGAN model, which features a novel mask-guided, semantic-aware, patch-wise discriminator. Our results indicate that the proposed model outperforms existing state-of-the-art inpainting methods in producing semantically coherent Cleft lips for both normal and post-operative condition. These findings are supported by comprehensive quantitative and qualitative analyses. Our pipeline, which includes facial landmark localization and reconstruction, is applicable to privacy-sensitive Cleft images, focusing on the cropped orofacial region to generate post-operative outcomes. Additionally, the interactive user interface enables the generation of synthetic Cleft faces, addressing the challenge of limited Cleft-specific datasets by providing synthetic Cleft data, which can be used, for example, in training facial landmark models tailored to Cleft conditions.

Future work will focus on clinically validating the results and expanding the evaluation to a broader distribution of Cleft faces. Additionally, we aim to further explore the semantics of Cleft facial features and automate the generation of Cleft lips based on specific Cleft conditions. With thorough validation, this approach could serve as a tool for generating post-operative Cleft facial reconstructions, aiding in patient and caretaker education about potential outcomes, as well as providing an objective measure to assess surgical success. Furthermore, we plan to explore alternative conditional inpainting methods to produce high-resolution, plausible lip reconstructions for Cleft conditions.

6 Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number [12/RC/2289_P2]. For

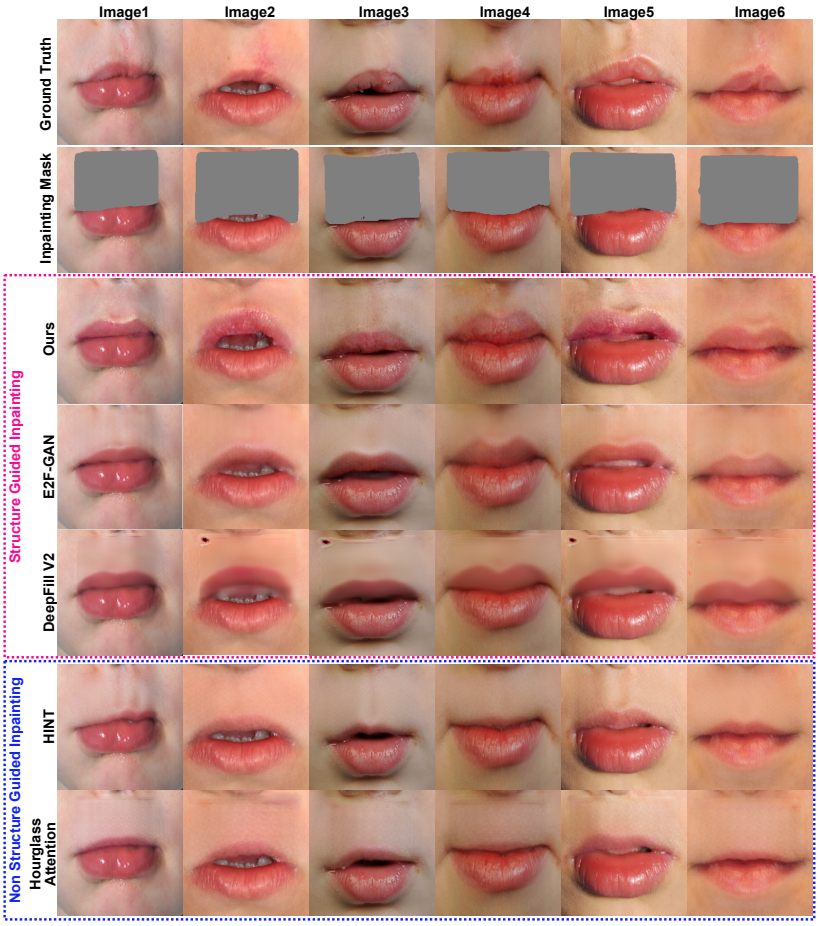

Fig. 5: Synthetic Non-Cleft Lip generation for patients with Cleft condition: Notably, we can observe that non structure based inpainting models, HourglassAttention [8] and HINT [7] fail to capture the specific semantics of Cleft conditions. In contrast, structure guided inpainting models can generate semantically appealing results specific for Cleft condition with user guidance. Additionally, DeepFillV2 [57] model produces visible artifacts making it unsuitable for Cleft reconstruction. In comparison, E2F-GAN [15] and our model generates semantically appealing lips with minimal artifacts. But E2F-GAN [15] fails to capture fine details in the inpainted region. On comparison to both structure-guided and non-structure-guided inpainting models, our model generates lips that are both semantically and structurally accurate for Cleft conditions.

References

1. Atputharuban, D., Theopold, C., Lawlor, A.: Enhancing Surgical Visualization: Feasibility Study on GAN-Based Image Generation for Post Operative Cleft Palate Images. Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods pp. 939–945 (2024). https://doi.org/10.5220/0012576900003654 2, 5, 6, 10
2. Brief, D.J., Behle, D.J.H., Stellzig-Eisenhauer, D.A., Hassfeld, D.S.: Precision of landmark positioning on digitized models from patients with cleft lip and palate. The Cleft Palate Craniofacial Journal **43**(2), 168–173 (2006). https://doi.org/10.1597/04-106.1, https://doi.org/10.1597/04-106.1, pMID: 16526922 3
3. Cai, W., Wei, Z.: Piigan: Generative adversarial networks for pluralistic image inpainting. IEEE Access **8**, 48451–48463 (2020). https://doi.org/10.1109/ACCESS.2020.2979348 4
4. Chaudhari, P.K., Kharbanda, O.P., Chaudhry, R., Pandey, R.M., Chauhan, S., Bansal, K., Sokh, R.K.: Factors affecting high caries risk in children with and without cleft lip and/or palate: A cross-sectional study. The Cleft Palate Craniofacial Journal **58**(9), 1150–1159 (2021). https://doi.org/10.1177/1055665620980206, https://doi.org/10.1177/105566562098020, pMID: 33349037 2
5. Chen, Q., Qiang, Z., Zhao, Y., Lin, H., He, L., Dai, F.: Rdfinet: reference-guided directional diverse face inpainting network. Complex & Intelligent Systems pp. 1–12 (2024) 3
6. Chen, S., Atapour-Abarghouei, A., Ho, E.S., Shum, H.P.: INCLG: Inpainting for non-cleft lip generation with a multi-task image processing network. Software Impacts **17**, 100517 (2023). https://doi.org/10.1016/j.simpa.2023.100517 2, 5
7. Chen, S., Atapour-Abarghouei, A., Shum, H.P.H.: Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention (2024), https://arxiv.org/abs/2402.14185 3, 10, 13
8. Deng, Y., Hui, S., Meng, R., Zhou, S., Wang, J.: Hourglass Attention Network for Image Inpainting, pp. 483–501 (11 2022). https://doi.org/10.1007/978-3-031-19797-0_28 4, 10, 13
9. Deng, Y., Hui, S., Zhou, S., Meng, D., Wang, J.: T-former: An efficient transformer for image inpainting. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, ACM (Oct 2022). https://doi.org/10.1145/3503161.3548446, http://dx.doi.org/10.1145/3503161.3548446 3
10. Ding, F., Zhu, G., Li, Y., Zhang, X., Atrey, P.K., Lyu, S.: Anti-forensics for face swapping videos via adversarial training. IEEE Transactions on Multimedia **24**, 3429–3441 (2022). https://doi.org/10.1109/TMM.2021.3098422 2
11. Dong, Q., Cao, C., Fu, Y.: Incremental transformer structure enhanced image inpainting with masking positional encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11358–11368 (June 2022) 3
12. Dong, Q., Cao, C., Fu, Y.: Incremental transformer structure enhanced image inpainting with masking positional encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 3
13. Duggal, I., Talwar, A., Duggal, R., Chaudhari, P., Samrit, V.: Comparative evaluation of nasolabial appearance of unilateral cleft lip and palate patients by professional, patient and layperson using 2 aesthetic scoring systems: A cross sectional study. Orthodontics craniofacial research **26** (04 2023). https://doi.org/10.1111/ocr.12663 2

14. Grollemund, B., Dissaux, C., Gavelle, P., Martínez, C.P., Mullaert, J., Alfaiate, T., Guedeney, A.: The impact of having a baby with cleft lip and palate on parents and on parent-baby relationship: the first french prospective multicentre study. BMC pediatrics **20**, 1–11 (2020) 2
15. Hassanpour, A., Daryani, A.E., Mirmahdi, M., Raja, K., Yang, B., Busch, C., Fierrez, J.: E2f-gan: Eyes-to-face inpainting via edge-aware coarse-to-fine gans. IEEE Access **10**, 32406–32417 (2022) 3, 4, 10, 13
16. Hayajneh, A., Serpedin, E., Shaqfeh, M., Glass, G., Stotland, M.A.: CleftGAN: Adapting A Style-Based Generative Adversarial Network To Create Images Depicting Cleft Lip Deformity. arXiv (2023). https://doi.org/10.48550/arxiv.2310.07969 5
17. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017), https://arxiv.org/abs/1704.04861 9
18. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and Locally Consistent Image Completion. ACM Transactions on Graphics (Proc. of SIGGRAPH) **36**(4), 107 (2017) 3, 4
19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018), http://arxiv.org/abs/1812.04948 6
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019), https://arxiv.org/abs/1812.04948 5
21. Kau, C.H., Medina, L., English, J.D., Xia, J., Gateno, J., Teichgraber, J.: A comparison between landmark and surface shape measurements in a sample of cleft lip and palate patients after secondary alveolar bone grafting. Orthodontics: the art and practice of dentofacial enhancement **12**(3), 188 (2011) 3
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), https://arxiv.org/abs/1412.6980 9
23. Ko, K., Kim, C.S.: Continuously masked transformer for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13169–13178 (October 2023) 3
24. Lahiri, A., Jain, A.K., Agrawal, S., Mitra, P., Biswas, P.K.: Prior guided gan based semantic inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
25. Leslie, E.J., Marazita, M.L.: Genetics of cleft lip and cleft palate. American Journal of Medical Genetics Part C: Seminars in Medical Genetics **163**(4), 246–258 (2013). https://doi.org/10.1002/ajmg.c.31381 1
26. Li, B., Li, X., Zhu, H., Jin, Y., Feng, R., Zhang, Z., Chen, Z.: Sed: Semantic-aware discriminator for image super-resolution (2024), https://arxiv.org/abs/2402.19387 7
27. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 3
28. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion (2017), https://arxiv.org/abs/1704.05838 3
29. Li, Y., Cheng, J., Mei, H., Ma, H., Chen, Z., Li, Y.: CLPNet: Cleft Lip and Palate Surgery Support With Deep Learning. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) **00**, 3666–3672 (2019). https://doi.org/10.1109/embc.2019.8857799 5

30. Liu, C., Xu, S., Peng, J., Zhang, K., Liu, D.: Toward interactive image inpainting via robust sketch refinement. IEEE Transactions on Multimedia **26**, 9973–9987 (2024). https://doi.org/10.1109/TMM.2024.3402620 4
31. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions (2018), https://arxiv.org/abs/1804.07723 3, 4
32. Liu, H., Jiang, B., Song, Y., Huang, W., Yang, C.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations (2020), https://arxiv.org/abs/2007.06929 4
33. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. CoRR **abs/1411.7766** (2014), http://arxiv.org/abs/1411.7766 5, 9
34. Lu, W., Zhao, H., Jiang, X., Jin, X., Yang, Y., Wang, M., Lyu, J., Shi, K.: Do inpainting yourself: Generative facial inpainting guided by exemplars. arXiv preprint arXiv:2202.06358 (2022) 4
35. McCullough, M., Ly, S., Auslander, A., Yao, C., Campbell, A., Scherer, S., Magee, W.P.: Convolutional Neural Network Models for Automatic Preoperative Severity Assessment in Unilateral Cleft Lip. Plastic and Reconstructive Surgery **148**(1), 162–169 (2021). https://doi.org/10.1097/prs.0000000000008063 5
36. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR). pp. 723–727. IEEE (2011) 9, 10
37. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019) 3, 4
38. Paradowska-Stolarz, A., Mikulewicz, M., Duś-Ilnicka, I.: Current Concepts and Challenges in the Treatment of Cleft Lip and Palate Patients—A Comprehensive Review. Journal of Personalized Medicine **12**(12), 2089 (2022). https://doi.org/10.3390/jpm12122089 2
39. Patcas, R., Timofte, R., Volokitin, A., Agustsson, E., Eliades, T., Eichenberger, M., Bornstein, M.M.: Facial attractiveness of cleft patients: a direct comparison between artificial-intelligence-based scoring and conventional rater groups. European Journal of Orthodontics (2019). https://doi.org/10.1093/ejo/cjz007 5
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020 7
41. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016), https://arxiv.org/abs/1511.06434 4
42. Rahimov, F., Jugessur, A., Murray, J.C.: Genetics of nonsyndromic orofacial clefts. The Cleft palate-craniofacial journal **49**(1), 73–91 (2012) 1
43. Ren, K., Meng, L., Fan, C., Wang, P.: Least squares dcgan based semantic image inpainting. In: 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). pp. 890–894. IEEE (2018) 4
44. Rosero, K., Salman, A., Sisman, B., Hallac, R., Busso, C.: Enhanced facial landmarks detection for patients with repaired cleft lip and palate. pp. 1–10 (05 2024). https://doi.org/10.1109/FG59268.2024.10582022 5
45. Sayadi, L.R., Hamdan, U.S., Zhangli, Q., Hu, J., Vyas, R.M.: Harnessing the Power of Artificial Intelligence to Teach Cleft Lip Surgery. Plastic and Reconstructive Surgery - Global Open (2022). https://doi.org/10.1097/gox.0000000000004451 5

46. Stiernman, M., Österlind, K., Rumsey, N., Becker, M., Persson, M.: Parental and health care professional views on psychosocial and educational outcomes in patients with cleft lip and/or cleft palate. European Journal of Plastic Surgery **42**, 325–336 (2019) 2
47. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3172–3182 (2022). https://doi.org/10.1109/WACV51458.2022.00323 2, 3
48. Trotman, C.A.: Faces in 4 dimensions: Why do we care, and why the fourth dimension? American Journal of Orthodontics and Dentofacial Orthopedics **140**(6), 895–899 (Dec 2011) 2
49. Trotman, C.A., Phillips, C., Essick, G., Faraway, J., Barlow, S., Losken, H., van Aalst, J., Rogers, L.: Functional outcomes of cleft lip surgery. part i: Study design and surgeon ratings of lip disability and need for lip revision. Cleft Palate-Craniofacial Journal **44**(6), 598–606 (Nov 2007) 2
50. Ud Din, N., Javed, K., Bae, S., Yi, J.: A novel gan-based network for unmasking of masked face. IEEE Access **8**, 44276–44287 (2020). https://doi.org/10.1109/ACCESS.2020.2977386 4
51. Wadde, K., Chowdhar, A., Venkatakrishnan, L., Ghodake, M., Sachdev, S.S., Chhapane, A.: Protocols in the management of cleft lip and palate: A systematic review. Journal of Stomatology, Oral and Maxillofacial Surgery **124**(2), 101338 (2023) 2
52. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861 9, 10
53. Xiang, H., Zou, Q., Nawaz, M.A., Huang, X., Zhang, F., Yu, H.: Deep learning for image inpainting: A survey. Pattern Recognition **134**, 109046 (2023) 2
54. Yang, Y., Guo, X.: Generative landmark guided face inpainting. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 14–26. Springer (2020) 3, 4
55. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting (2020), https://arxiv.org/abs/2005.09704 4
56. Yilmaz, H.N., Ozbilen, E.O., and, T.U.: The prevalence of cleft lip and palate patients: A single-center experience for 17 years. Turkish Journal of Orthodontics **32**(3), 139–144 (2019) 2
57. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.: Free-form image inpainting with gated convolution (2019), https://arxiv.org/abs/1806.03589 3, 4, 8, 10, 13
58. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention (2018), https://arxiv.org/abs/1801.07892 3, 4
59. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5505–5514 (2018). https://doi.org/10.1109/CVPR.2018.00577 4
60. Yu, X., Dai, L., Chen, Z., Sheng, B.: Agg: attention-based gated convolutional gan with prior guidance for image inpainting. Neural Computing and Applications **36**(20), 12589–12604 (2024) 3, 7, 8, 10
61. Yu, Y., Zhan, F., WU, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 69–78. MM '21,

Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475436, https://doi.org/10.1145/3474085.3475436 3
62. Yu, Y., Zhan, F., WU, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475436, https://doi.org/10.1145/3474085.3475436 3
63. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting (2019), https://arxiv.org/abs/1904.07475 4
64. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling (2020), https://arxiv.org/abs/2005.11742 4
65. Zhang, X., Ma, W., Varinlioglu, G., Rauh, N., He, L., Aliaga, D.: Guided pluralistic building contour completion. The Visual Computer **38**(9), 3205–3216 (2022) 4
66. Zheng, C., Cham, T.J., Cai, J.: Pluralistic free-from image completion. International Journal of Computer Vision pp. 1–20 (2021) 4
67. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. arXiv preprint arXiv:2112.03109 (2021) 6, 7, 9