Lip Abnormality Detection for Patients with Repaired Cleft Lip and Palate: A Lip Normalization Approach

Karen Rosero
karen.rosero@utdallas.edu
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

Ali N. Salman
ali.salman@utdallas.edu
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

Rami R. Hallac
rami.hallac@childrens.com
Department of Plastic Surgery
University of Texas Southwestern Medical Center
Dallas, Texas, USA

Carlos Busso
busso@utdallas.edu
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

ABSTRACT

The cleft lip condition arises from the incomplete fusion of oral and labial structures during fetal development, impacting vital functions. After surgical closure, patients commonly present with abnormal lip shape, which may require secondary revision surgery for both aesthetic and functional improvement. However, a lack of standardized evaluation methods complicates decision-making for secondary surgery. To address this limitation, we propose a transformer-based lip normalization approach that filters out abnormalities and achieves a standardized appearance while preserving individual anatomy. An innovation of our approach is a lip transformation method using available face datasets to mimic repaired cleft lip shapes, enabling the training of deep learning models without using patients' data. We employ a Siamese convolutional neural network that processes pre- and post-normalization images to detect lip abnormalities with an accuracy of 89%. We compare our approach with a single-branch model without lip normalization, which reached an accuracy of 60%. Our approach has the potential to provide an impartial view to determine the need for revision surgery while also assisting in the selection of healthcare tools specialized for patients with repaired cleft lip. The code for this work is available in our official repository [1].

CCS CONCEPTS

• **Applied computing** → **Imaging**.

KEYWORDS

Medical Imaging, Image Restoration, Image Abnormality Detection, Cleft Lip, AI Interfaces for Healthcare.

[1]Official repository: https://icmi2024-1257.github.io/

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685726

ACM Reference Format:
Karen Rosero, Ali N. Salman, Rami R. Hallac, and Carlos Busso. 2024. Lip Abnormality Detection for Patients with Repaired Cleft Lip and Palate: A Lip Normalization Approach. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, San Jose, Costa Rica, 9 pages. https://doi.org/10.1145/3678957.3685726

1 INTRODUCTION

Advances in computer vision have provided clear benefits in healthcare, leveraging various medical imaging modalities such as x-rays, magnetic resonance imaging, ultrasound, and 2D and 3D photographs [22]. These images, when analyzed using computer vision techniques, offer valuable insights to physicians, assisting in both diagnosis and treatment planning for patients. By detecting subtle abnormalities or patterns not easily discernible to the human eye, computer vision aids in enabling early disease detection, including cancers [5], and facilitating timely interventions, improving patient outcomes. An important area in computer vision that is relevant to healthcare is facial expression analysis, which has been applied for the assessment of several conditions involving the face, such as facial paralysis [3], epileptic seizures [10], and cleft lip [1].

The focus of this paper is on the *cleft lip* (CL) condition, with or without a fissure in the palate. This congenital condition arises from the incomplete fusion of oral and labial structures during fetal development. The clinical manifestations of CL compromise vital functions such as feeding, swallowing, breathing, and speaking [7]. Hence, pediatric patients typically undergo the surgical closure of the cleft in the first months of life. Unfortunately, complications such as lip scarring or abnormal lip shape commonly persist [21]. As a result, individuals with repaired CL often endure negative social experiences and suffer from negative self-perception, especially during childhood and adolescence [2]. The severity of the initial cleft influences the necessity for secondary revision surgery, aimed at improving lip appearance or correcting velopharyngeal incompetence [27]. However, the decision for lip revision is subjectively based on clinical criteria, leading to disagreements among clinicians regarding the necessity of such surgeries [28]. Since there is no agreed gold standard to evaluate lip abnormality in patients

with repaired CL [26], there is a need for a computational tool capable of detecting aesthetic outcomes that highly deviate from the appearance of control subjects [24].

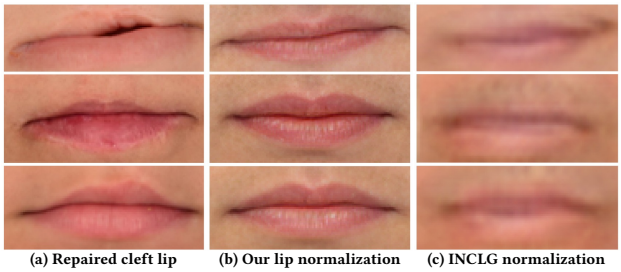
We present a novel system for lip abnormality detection specifically designed for patients with repaired CL. We propose a Siamese neural network that operates by comparing a lip image displaying an anomaly with a normalized version of the same image, which filters out the existing abnormalities. By leveraging both versions of the image, our model improves its ability to detect lip abnormalities by explicitly contrasting their differences. The normalized version of the image is obtained using a novel *lip normalization* strategy. We employ a transformer-based prediction network designed for face reconstruction. This network is trained on extensive open datasets of facial images with no reported abnormalities, allowing it to capture typical facial appearances found within the control population. Through this approach, the system can adapt the unique anatomical features of each patient, particularly focusing on the region affected by the sequelae of the repair surgery. Our deep learning model requires facial images displaying lip abnormality to effectively learn from them. However, the availability of such images from patients with repaired CL is limited. To address this challenge, we propose a *lip transformation* technique based on landmarks. This method modifies the lip shape of control subjects to closely resemble the characteristics of the target population while preserving the inherent features of the control subject lips. By employing this approach, we overcome the problem of insufficient training data for our model to generalize to unseen subjects. We source control images from openly available datasets of facial images, thereby ensuring a representative sample set. Furthermore, by avoiding the use of patient images to train our models, we protect their privacy, which is an additional benefit of our methodology.

Our experimental results demonstrate the effectiveness of our lip transformation approach along with the lip normalization stage for accurate lip abnormality detection. We compare our approach with a single-branch baseline model that does not incorporate the lip normalization stage. Notably, our approach exhibits a remarkable 22.3% (absolute) point improvement in accuracy for detecting lip abnormalities compared to the baseline. These findings are evaluated on 395 images from patients with repaired CL.

The contributions of our paper are as follows:

- A lip transformation approach for images of open available datasets that imitates the lip shape of patients with repaired CL.
- A lip normalization technique able to standardize images of patients with CL displaying lip abnormalities.
- A Siamese neural network that compares the representations of the images before and after normalization for improved lip abnormality detection.

This paper is organized as follows: Section 2 presents previous work related to this study. Section 3 describes the methodology for our lip transformation and normalization approaches, and the baseline that will be used for comparison. Section 4 describes the experimental setup. Section 5 describes our experiments and results. Finally, Section 6 summarizes the main findings and insights of our work.



(a) Repaired cleft lip (b) Our lip normalization (c) INCLG normalization

Figure 1: Comparison of our proposed lip normalization approach with the INCLG normalization method [6], illustrated through images of three patients with repaired cleft lip.

2 RELATED WORKS

Previous technological tools developed for patients with CL have focused on diagnosing the condition before surgery. Agarwal et al. [1] explored the detection and classification of unrepaired pediatric cleft. The authors collected a training set of facial images containing 136 bilateral cases, 412 unilateral cases, and 670 control images. Later, they implemented data augmentation techniques that do not distort the face. The pre-trained AlexNet architecture [15] was used to extract features from the images after cropping the orofacial area. Subsequently, a *support vector machine* (SVM) classifier was trained for each of the 256 feature maps extracted from earlier layers of the model. They achieved an accuracy of 84.12% on a testing set of 58 bilateral, 78 unilateral, and 97 control images. Similarly, McCullough et al. [19] trained a *convolutional neural network* (CNN) with 800 preoperative images of patients with unilateral cleft lip. All images were manually annotated with cleft-specific landmarks to rate the cleft severity. The model achieved a correlation of 0.89 with the severity assessment performed by expert reviewers. Both previously presented works used patients' images to train the model, which is not ideal since limits the size of labeled samples in the training set, and raises privacy concerns by using images from the patients. In contrast, our approach does not rely on images of patients. Instead, we develop a transformation approach that uses images of control subjects for lip abnormality detection on repaired CL.

Studies have proposed normalization approaches for patients with repaired or unrepaired CL. Boyaci et al. [4] adapted a generative model to produce normalized versions of facial images containing a congenital or acquired deformity. The generated images filter out the face anomalies while preserving the identity and the normal features of the patient. The StyleGAN [15] generator was optimized to find a latent vector that results in a similarity loss that preserves the identity, and an average loss that corrects the structural anomalies. However, this model presented an instability issue across different predictions of the same image. Hayajneh et al. [9] improved this problem by employing StyleGAN2 [12] to provide a more realistic facial normalization and a more stable performance.

Chen et al. [6] proposed an *inpainting technique for non-cleft lip generation* (INCLG) that predicts normalized facial images for patients with CL. They employed a landmark guided face inpainting framework [29] that consists of a convolutional encoder followed by dilated ResBlocks and an attention layer that outputs a feature

representation of an input image. Control images were masked to define and condition the area to be synthesized. Since the model was trained with only control images, the predicted facial landmarks and the masked regions follow a distribution that aligns with the typical patterns learned during training from control individuals. The authors of this study made available the weights of this model and the code needed to reproduce the findings so that we can mask the lip area of any local image of the target population and predict the normalized area. Figure 1 exemplifies the INCLG normalization for three patients with repaired CL. The generated images display low fine details and decreased correlation with the original lip format compared with our lip normalization approach.

3 METHODOLOGY

Our work relies on three stages: 1) a lip transformation approach that imitates the lip shape of patients with repaired CL using images of openly available face datasets, 2) a lip normalization approach applied to both control and modified images, and 3) a Siamese CNN-based model that takes as input the control and modified images before and after normalization. We elaborate on these stages in the following sections.

3.1 Lip transformation

The first block in our approach is the lip transformation. The goal is to modify the shape of control lip images to closely resemble the lip characteristics of patients with repaired CL while preserving the inherent features of the original subject. The approach takes a repaired CL image and a random image from a control subject, which can be obtained from openly available face datasets. The lip transformation transfers the shape abnormally to the controlled image. After CL repair surgery, typical outcomes often include: lip scarring tissue, unilateral lip asymmetry with a raised peak of the upper lip, or misalignment of the upper lip's cupid's bow with the vertical plane of facial symmetry. These outcomes can be identified by the human eye as abnormalities. Given the role of the upper lip, our lip transformation approach is guided by the upper lip landmarks extracted from images of patients with repaired CL (Figure 2a).

After preprocessing the images following the indications of Section 4.2, we extract the upper lip landmarks of the patient image (blue dots in Figure 2a). Then, we scale these CL landmarks based on the oral commissure points and displace them to the cupid's bow original midpoint to match the lip size of the control sample. As the original CL landmarks undergo resizing and displacement, their partial dependency on each control subject introduces variability to the transformations. The scaled CL landmarks are later combined with the remaining face landmarks of the control subject. Next, we triangulate the control facial image based on the control landmarks and apply affine transformations to manipulate the geometric properties of the lip area to satisfy the CL landmarks on every small triangulated region (Figure 2b). This process affects just the image region with modified landmarks (i.e., the upper lip). The modification applied to control images is randomly selected from the lip shapes extracted from images of patients with repaired CL. Our lip transformation approach enables the generation of large amounts of modified images needed for training accurate deep learning models.

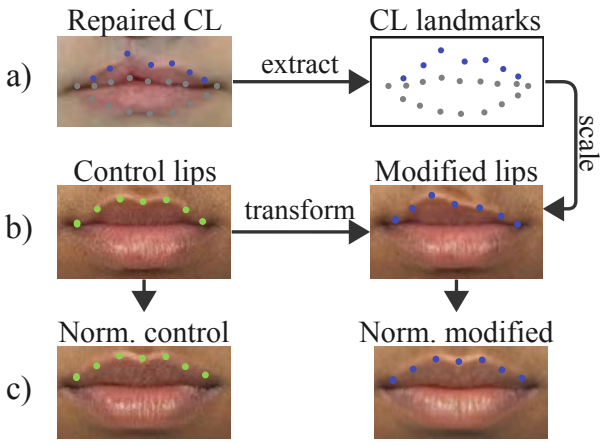


Figure 2: Illustration of a) lip landmarks extraction from an image of a patient with repaired CL, b) lip transformation process, and c) lip normalization for a control and modified image.

The transformation stage uses landmarks from 12 patients with repaired cleft lip, chosen for clearly displaying common non-aesthetic outcomes. Cleft lip landmarks are scaled and displaced before combining with control landmarks, partially retaining inherent features of the control subject. We will compare our lip transformation approach with a strategy presented in a previous study [23], which aims to control images to recreate lip asymmetry in patients with repaired CL. The strategy defined seven landmark transformations by observing patients' images, which were applied to the control subjects. This baseline method is referred to as *lip transformation v1* (LT-v1), and our proposed approach is referred to as *lip transformation v2* (LT-v2). We hypothesize that our proposed strategy will result in modified control images displaying more naturalistic lip shapes, as it not only draws inspiration from cleft lip shapes but also transfers the exact upper lip shape to control lip images.

3.2 Lip normalization

The objective of the lip normalization stage is to filter out abnormalities in the lips, aiming to make the lip's appearance more similar to patterns observed in control subjects while remaining congruent with the patient's anatomical structure. Figure 3 illustrates our lip normalization process, comprising two stages: (1) an image compression and decompression process, and (2) a face restoration stage. These processes contribute to filtering out facial anomalies and generating fine details for the facial image.

First, we utilized the pretrained ArcFace model [8] to generate facial features from the input image (top diagram in Fig. 3). The encoded representations of ArcFace leverage substantial information from the training distribution while preserving facial identity since it is trained for face recognition using over 493K identities from the IBUG-500K dataset. Subsequently, we employed the inverted pretrained ArcModel, as detailed in [8], to decode the representation and reconstruct the original facial image. For this compression

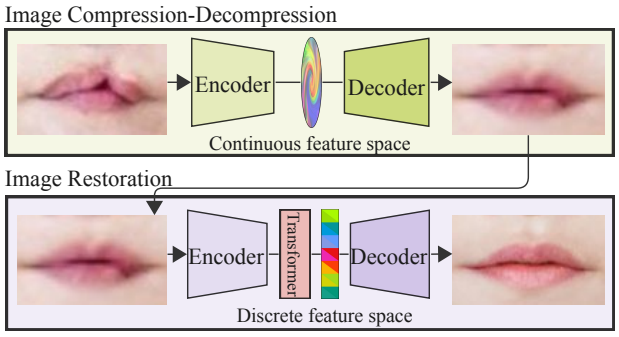


Figure 3: Lip normalization process consisting in an image compression and decompression process (top figure), followed by a face restoration process (bottom figure).

and decompression process, we utilized the InsightFace pipeline[2], based on the ArcFace model. This model was originally intended for swapping faces of different identities. However, we use this strategy to reconstruct the encoded representation of the same facial image. The encoder-decoder process of images containing lip abnormalities smoothens characteristics not represented in the training set, bringing us closer to our lip normalization goal. However, the pretrained ArcModel is limited to recovering images with a resolution of 128×128 pixels. The process results in low-resolution and slightly degraded images that maintain facial identity, and reduce face abnormalities, similar to previous CL normalization solutions [4, 6].

To address the resolution limitation, we rely on the face restoration model, CodeFormer[3] [31], which generates high-quality details in degraded input facial images without landmarks guidance (bottom diagram in Fig. 3). The CodeFormer model was originally trained for the self-reconstruction of high-quality images from the FFHQ dataset [13]. Its first stage results in a trained encoder, a learned codebook, and a decoder. In the second stage, low-quality images with varying levels of degradation are used, and the encoder is fine-tuned to learn features related to low-quality images. Nine self-attention blocks and a linear layer are trained to sequentially model the encoder representation and project the features into the previously learned codebook space. Finally, the projected features were passed through the frozen decoder, which outputs the restored image with fine-grained details. This face restoration model learned from 70,000 images of subjects, without reported facial deformities. Consequently, we expect the reconstructed images to adhere to the mean of the lip appearance across the population.

For the scope of our study, we cropped the lip area based on landmarks, although the normalization process was performed for the entire face. Figure 2 exemplifies the process of transforming a control lip image to introduce an abnormality seen in patients with repaired CL. We highlight how the lip normalization process on the modified lips removes the unilateral asymmetry in the upper lip. In contrast, the lip normalization applied to a control lip image without transformation results in a representation very similar to the original. Additionally, we draw attention to Figure 2c, which shows the lip normalization results of control and modified images. The high similarity between both results supports our goal of preserving the idiosyncratic characteristics of the individual after the normalization.

[2]https://github.com/deepinsight/insightface
[3]https://github.com/sczhou/CodeFormer

3.3 Siamese CNN for lip abnormality detection

We propose the use of a Siamese CNN architecture for the detection of lip abnormalities, where one branch generates a representation of the image before the normalization, and the other generates a representation after the normalization. Subsequently, these embeddings are concatenated and passed through a dense layer followed by a final classification layer to determine the presence of abnormalities in the lips. Figure 4 describes this process.

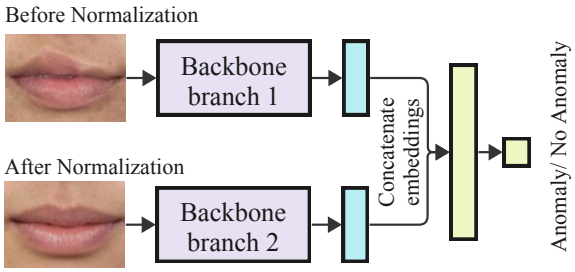


Figure 4: Siamese CNN for lip abnormality detection. The network processes the original and normalized images to determine the presence of any lip abnormality.

The backbone of our Siamese architecture is the MobileNetV2 network [25], renowned for its lightweight design suitable for resource-constrained environments such as smartphones. The MobileNetV2 is constructed with blocks of lightweight depthwise separable convolutions aimed at reducing the network's computational burden by facilitating feature reuse. The MobileNetV2 network consists of 156 layers, including convolutional layers, batch normalization, dropout, *rectified linear unit* (ReLU6), and pooling layers. The detailed architecture of MobileNetV2 can be found in Sandler et al. [25]. We adopted this backbone architecture for both branches of the Siamese CNN. Following partial finetuning, these branches are utilized to extract embeddings from the input images.

The Siamese CNN receives pairs of corresponding images, one before and one after lip normalization. The model is trained on pairs of control subjects, where minor differences are expected, and pairs of modified images of control subjects, where differences are more pronounced due to the removal of anomalies in the normalized image. Despite the subtle differences in control pairs, the model is anticipated to discern between normal variations and those indicative of abnormalities. Both branches of the backbone share weights and produce a 1,280D vector representation for each input. Upon concatenation, the dimensionality is reduced to 256 before passing through a classification layer to make the final determination regarding the presence of lip abnormalities.

3.4 Baseline CNN for lip abnormality detection

We contrast our proposed method with a baseline that lacks the normalized version of the lip image to evaluate the impact of incorporating two branches in the Siamese CNN network. The backbone of the baseline model is the same as the MobileNetV2 network to ensure a fair comparison. This baseline model is trained using control images and modified control images without lip normalization. Subsequently, the image embedding is condensed into a 256-dimensional vector, which is then processed by a final classification layer responsible for identifying the presence of lip abnormalities.

4 EXPERIMENTAL SETUP

This section outlines the datasets used for training both our Siamese CNN and the baseline model. We also describe the images of patients with repaired CL that we will use to evaluate the models on real data from our target population. Additionally, we provide an overview of the image preprocessing steps and training hyperparameters employed in the study.

4.1 Datasets

Considering the age variability of our target population, we need to include children, adolescents, and adult faces of different ethnicities to make our model robust to high variability in the testing set. We have selected two datasets: the *Chicago face dataset* (CFD) for adult faces and the *young labeled faces in the wild* (YLFW) dataset for children and adolescents. Further details are presented below.

4.1.1 Adult Faces. The Chicago face dataset, developed by the University of Chicago, comprises high-resolution, standardized frontal photographs of individuals of various genders, ethnicities, and ages ranging from 17 to 65. This dataset, available upon request for scientific research, includes three subsets. The main CFD subset [17] consists of images of 597 unique subjects representing Asian, Black, Latino, or White ethnic backgrounds. These subjects were photographed posing with a range of facial expressions, including neutral, happy (with an open or closed mouth), angry, and fearful expressions. The second subset, CFD-MR [18], expands upon the main CFD by including images of 88 unique subjects with multiracial ancestry. Similar to the main subset, these individuals were photographed displaying various facial expressions. The third subset, CFD-INDIA subset [16], follows the same pose pattern as the previous subsets and contains images of 142 unique individuals representing the Indian population. In total, 1,235 images from the Chicago Face Dataset were pre-processed for face detection and landmark placement using the dlib library [14].

4.1.2 Children Faces. To mitigate bias towards adult faces, we opted to incorporate images from the YLFW dataset [20]. YLFW was specifically designed for children's face recognition and offers a balanced representation across African, Asian, Caucasian, and Indian ethnicities within the age range of 0-18 years. The benchmark subset of YLFW, comprising 9,810 images from 3,069 unique identities, is available for research purposes. Given that these images were sourced from the website, they exhibit a wide range of head poses and facial expressions. For the scope of our study, we refined the dataset by selecting only frontal images with yaw and pitch angles inferior to 15°. Additionally, we applied exclusion criteria to filter out incomplete faces, lip occlusions, and facial expressions that do not display the natural shape of the lips. Following these criteria, we selected 456 frontal images for our proposed approach, ensuring consistency in preprocessing steps for face detection and landmark placement.

4.1.3 Faces with repaired cleft lip. We gathered 395 images of patients who had undergone CL repair from various sources, including websites of CL-specialized surgeons, hospital websites, and nonprofit organizations dedicated to CL treatment. The selection of images adhered to the same criteria applied to the YLFW dataset. To ensure accurate landmark placement, we utilized the *facial landmark detectors* (FLD) developed by Rosero et al. [23] for the faces of individuals with CL. This FLD has been specifically tailored to improve landmark detection for our target population. Furthermore, the orofacial images of patients with repaired CL depicted in Figures 1 and 2 were provided by Children's Medical Center Dallas with the consent of the patients involved.

4.2 Images preprocessing

We employ the ZFace toolbox[4] [11] to filter images across all datasets, ensuring that head positions exhibit less than a 15° deviation in both yaw and pitch angles. This restriction is crucial as images with greater rotation may not fully capture the natural shape of the lips, a critical consideration for our study. Once selected, these images undergo an additional frontalization process leveraging the dense mesh of 512 3D facial landmarks provided by ZFace. This toolkit produces a normalized and frontalized version of the 3D mesh, along with its fit to the facial image. We subsequently apply affine transformations to small regions of the image, defined by the 512 landmarks, to align them with the frontalized mesh.

The lip transformation stage relies on the 68-point standard landmarks extracted using the dlib library [14]. In contrast, the lip normalization process does not require landmarks; instead, they operate on the entire face. After transformation and normalization, we crop the lip area based on the dlib mouth landmarks. Finally, the region of interest is reshaped into a $224 \times 224 \times 3$ dimension to ensure compatibility with the MobileNetV2 architecture.

4.3 Model hyperparameters

This study used PyTorch for implementation and training, leveraging the computational power of an NVIDIA GeForce RTX 4090 GPU. The training process involved a batch size of 16 samples and a learning rate set at 0.0001, with model optimization achieved through the use of the cross-entropy loss function and Adam optimizer. To prevent overfitting, a maximum training duration of 50 epochs was established, incorporating early-stopping criterion with patience of 5 epochs, continuously monitoring the development loss. For the specific task of lip abnormality detection, fine-tuning was conducted on the final dense and classification layers in conjunction with the last 98 layers from the MobileNetV2 architecture. The reason for modifying the last layers of the MobileNetV2 architecture is that these layers provide task-specific high-level features, leaving the early layers in the models frozen so they can extract general low-level features. The specific number of retrained layers was selected

[4] https://github.com/AffectAnalysisGroup/AFARtoolbox/tree/master/zface

based on the study of Rosero et al. [23] for improved performance. During training, the models are evaluated in the 'development' set to save the best model and stop the training process.

5 EXPERIMENTS AND RESULTS

This section presents the experiments and results conducted to demonstrate the superior accuracy of our proposed approach. Section 5.1 explores the improvement introduced by the use of lip normalization, Section 5.2 evaluates the lip transformation approach outlined in this study. Section 5.3 presents visual interpretations of the model's attention, Section 5.4 demonstrates the relevance of incorporating children's data into our model, and Section 5.5 analyzes the impact of the image restoration stage in the lip normalization process.

5.1 Siamese CNN vs Baseline CNN

In this experiment, we compare the performance of our Siamese CNN with lip transformation (Siamese LT-v2) against the Baseline CNN. Notably, the difference between these models arises from the utilization of two branches in the Siamese model, each dedicated to processing the input lip image before and after normalization, as opposed to the Baseline model, which processes the lip image without normalization via a single branch. Table 1 reports the mean cross-entropy losses and accuracies for six data subsets, with standard deviations calculated from seven runs with the same setup. The 'training' set comprises images of adults and children, with and without lip transformations. This ensures a balanced training dataset, with 50% of images exhibiting lip abnormalities and 50% serving as control images without lip anomalies. Similarly, the 'validation' and 'test' sets maintain the same distribution while ensuring different identities. The 'test cleft lip' set contains images of patients with repaired CL, presenting lip anomalies, for which the model is expected to classify all the samples as lip abnormalities. Additionally, we assess the models' performance on sets containing images of adults and children without lip anomalies related to CL repair. Therefore, perfect accuracy for the 'test children' and 'test adults' is achieved when all the samples are classified as normal.

All metrics present statistically significant difference between the Siamese LT-v2 and baseline models, except for the 'training' and 'test adults' subsets, for which values are comparable. Specifically, for evaluation conducted on the 'test' set, the Siamese LT-v2 model exhibited a 4% increase in accuracy on unseen subjects. Notably, the most substantial enhancement is observed within the 'test cleft lip' subset, which encompasses lip images from our target population. The Siamese LT-v2 model outperforms the Baseline model accuracy by 29%. These results demonstrate the efficacy of incorporating a second branch for processing the normalized version of the input image, thereby enhancing the robustness and generalization capability of the model to real data within our target population. The Siamese LT-v2 model learns more meaningful representations from images before and after normalization compared to the Baseline, which exclusively learns from lip images without normalization.

Furthermore, we evaluate the system performance on unseen control data extracted from the general 'test' set, comprising just adults or children without lip transformation. We highlight the fact that control images from adults or children are anticipated to exhibit minimal lip abnormalities; thus, the images before and after lip normalization should closely resemble each other. As shown in 'test children' and 'test adults' sets in Table 1, Siamese LT-v2 achieves an accuracy of 0.89 for children and 0.95 for adults without lip abnormalities, demonstrating that it is not biased towards the detection of lip abnormalities only. The Baseline model shows comparable accuracy in children and adult images.

We evaluated the impact of replacing the MobileNetV2 backbone in our Siamese LT-v2 model and baseline with a simpler three-block CNN. For Siamese LT-v2, this change reduced accuracy on the cleft lip test set from 0.89 to 0.62, while the baseline accuracy dropped from 0.60 to 0.53. These results highlight the necessity of a more complex model for effectively extracting high-level features.

Table 1: Comparison of the Siamese CNN approach with lip normalization and the Baseline CNN model containing a single branch for lip abnormality classification.

Set	Experiment	Loss ↓	Accuracy ↑
Training	Siamese LT-v2	0.08* ± 0.04	0.97 ± 0.01
	Baseline CNN	0.15 ± 0.03	0.95 ± 0.01
Development	Siamese LT-v2	0.10* ± 0.03	0.97* ± 0.004
	Baseline CNN	0.18 ± 0.02	0.93 ± 0.01
Test	Siamese LT-v2	0.16* ± 0.03	0.96* ± 0.01
	Baseline CNN	0.22 ± 0.05	0.92 ± 0.02
Test Cleft Lip	Siamese LT-v2	0.38* ± 0.10	**0.89* ± 0.03**
	Baseline CNN	1.23 ± 0.25	**0.60 ± 0.06**
Test Children	Siamese LT-v2	0.33* ± 0.15	0.89* ± 0.05
	Baseline CNN	0.43 ± 0.15	0.85 ± 0.06
Test Adults	Siamese LT-v2	0.26±0.07	0.95 ± 0.02
	Baseline CNN	0.27 ± 0.12	0.91 ± 0.06

The arrows ↑ ↓ indicate if the metric improves by increasing or decreasing, respectively. Applying a one-tailed two-sample proportion t-test with p-value < 0.01, (*) indicates that Siamese LT-v2 is significantly better than the Baseline.

5.2 Comparison of the lip transformation approaches

The lip transformation technique aims to recreate lip abnormalities associated with the lip shape commonly found in patients with repaired CL. This method enables us to train deep-learning models to discern patterns of lip abnormality without directly utilizing patient images. Rosero et al. [23] tailored these lip transformations based on values selected by experimentation to deform the lip shape (LT-v1). The modifications were applied based on measurements of the upper lip of control subjects. We propose an improved lip transformation technique that distorts the upper lip shape of control subjects based on lip landmarks extracted from images of patients with repaired CL (LT-v2). This enhancement yields modified control images displaying more naturalistic lip shapes.

Figure 5 presents a comparison of the accuracy between the lip transformation techniques, evaluated on the same data subsets of our prior experiment and trained using our Siamese CNN. Mean

values across seven runs are shown on top with the standard deviation as error bar. On the 'training' and 'development' sets, using LT-v1 leads to better performance than LT-v2 with 2% gain in accuracy. Learning using images generated with LT-v1 is comparatively easier than using images generated with LT-v2, which includes more subtle changes. However, the real benefit of using LT-v2 is observed when we compare the results on real CL images (i.e., the 'test cleft lip' set). Our LT-v2 strategy clearly outperforms LT-v1, exhibiting a 11% increase in accuracy. The subtle changes created with LT-v2 better represent the target anomalies observed in patients with repaired CL. Likewise, we observe comparable accuracy in the subsets 'test children' and 'test adults' samples.

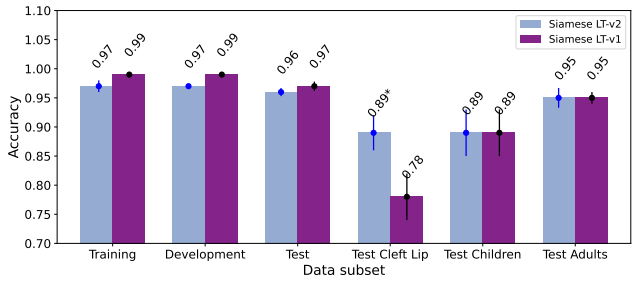
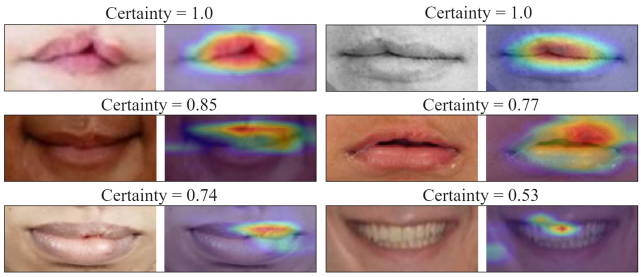


Figure 5: Accuracy comparison between Siamese LT-v1 and Siamese LT-v2. The asterisk (*) represents a statistically significant difference between the models when applying a one-tailed two-sample proportion t-test with p-value < 0.01.

5.3 Features visualization

This section aims at visualizing the specific areas of focus within an image that the model relies on for prediction. We employed the *class activation mapping* (CAM) technique [30]. This technique generates heatmaps that accentuate the regions deemed most crucial for class prediction by extracting feature maps of the model from the last convolutional layer after the global average pooling layer. Figure 6a illustrates six examples evaluated by our Siamese LT-v2 model on images of patients with repaired CL from the 'test cleft lip' subset. These images were correctly identified as containing lip abnormalities. Based on their respective heatmaps (shown to the right), the model emphasizes regions displaying abnormal lip shapes or scar tissue. The certainty of the model's prediction is indicated at the top of each example. Images showing more severe or noticeable sequelae are recognized with higher certainty. The heatmaps of the first five examples in Figure 6a attribute significant importance to regions containing abnormalities. Conversely, in the sixth example, which has a low certainty (i.e., 0.53), lip abnormality is imperceptible even to the human eye, despite the patient belonging to our target domain. In such cases, the model might shift its focus to differences in teeth, as the lips do not exhibit apparent abnormalities. We observe that 'hard samples', where abnormalities are not evident, can lead to misclassifications by our model.

Figure 6b depicts four examples of images of patients with repaired CL classified as not having lip abnormalities, each with varying certainties. In the first example, based on the heatmap,



(a) Samples correctly classified as containing lip abnormalities.

(b) Samples misclassified as lacking lip abnormalities.

Figure 6: CAM technique [30] applied to the Siamese LT-v2 model evaluated on images of patients with repaired CL. The color progression in the heatmaps ranges from blue, denoting regions with low weights, to red, highlighting areas with high weights.

the model correctly emphasizes the upper lip, where anomalies are commonly found; however, as no anomalies were detected, the sample was classified as normal. We emphasize that the outcome of CL surgery can be highly satisfactory in some cases. Thus, it is understandable that our model misclassifies such samples as normal. In the last example of Figure 6b, the heatmap primarily highlights the upper lip, where an abnormality is identifiable on the left side. Despite being misclassified, the model's certainty is 0.57, indicating that even in error cases, its certainty is affected. Therefore, samples with low certainty can be cross-checked by an expert.

5.4 Importance of inclusion of children's facial images

We conducted an additional experiment to analyze the importance of adding children's images alongside adult samples in training our Siamese CNN approach. Specifically, we excluded the children samples from the 'training' and 'development' subsets to train the Siamese CNN model without children data (referred to as Siamese w/o C). Table 2 presents the corresponding metrics alongside those achieved by our proposed approach, the Siamese LT-v2 model. Similar to the experiment detailed in Section 5.2, the Siamese w/o C model exhibits superior metrics on the 'training' subset since the model is trained exclusively on adult faces, potentially leading to overfitting to this population. Consequently, when evaluated on the 'test' set, which includes images of both adults and children, the Siamese w/o C model experiences a performance decrease due to its limited generalization to an unseen age population that can present slightly different facial characteristics.

When evaluating images of patients with repaired CL, the accuracy decreased 6% using the Siamese w/o C model instead of the Siamese LT-v2 model. The drop in performance is more severe in the 'test children' set, where the accuracy reduces from 89% (Siamese

LT-v2) to 62% (Siamese w/o C). This drop in performance corresponds to an absolute reduction of 27%. Given that the Siamese w/o C model is exclusively trained with images from adults, a higher accuracy is expected when evaluating the models on unseen adult images without lip abnormalities. However, we observe that the performances of both models are similar. All these results validate the use of child data to increase the variability in our models.

Table 2: Ablation study on the exclusion of children's data to train our best model.

Set	Experiment	Loss ↓	Accuracy ↑
Training	Siamese LT-v2	0.08 ± 0.04	0.97 ± 0.01
	Siamese w/o C	0.05 ± 0.02	0.99 ± 0.0001
Development	Siamese LT-v2	0.10 ± 0.03	0.97 ± 0.004
	Siamese w/o C	0.13 ± 0.01	0.96 ± 0.01
Test	Siamese LT-v2	0.16* ± 0.03	0.96* ± 0.01
	Siamese w/o C	0.32 ± 0.02	0.91 ± 0.01
Test Cleft Lip	Siamese LT-v2	0.38 ± 0.10	0.89* ± 0.03
	Siamese w/o C	0.42 ± 0.10	0.83 ± 0.02
Test Children	Siamese LT-v2	0.33* ± 0.15	**0.89* ± 0.05**
	Siamese w/o C	1.37 ± 0.14	**0.62 ± 0.05**
Test Adults	Siamese LT-v2	0.26±0.07	0.95 ± 0.02
	Siamese w/o C	0.17 ± 0.02	0.96 ± 0.01

The arrows ↑ ↓ indicate if the metric improves by increasing or decreasing, respectively. Applying a one-tailed two-sample proportion t-test with p-value < 0.01, (*) indicates that Siamese LT-v2 is significantly better than Siamese w/o C.

5.5 Impact of the image restoration stage

This section evaluates the impact of the image restoration stage within our two-stage lip normalization approach, explained in section 3.2. We omitted the image restoration stage while only keeping the image compression-decompression stage. With this modification in the lip normalization procedure, we train our Siamese LT-v2 model, denoted as 'Siamese LT-v2: No restoration' in Figure 7. We observe a statistically significant decrease across all test sets upon the removal of the image restoration stage. Specifically, our target 'test cleft lip' set experienced a decrease of 10% in accuracy. This result highlights the importance of incorporating both stages within the lip normalization approach.

6 CONCLUSIONS

This paper introduced a Siamese CNN for detecting lip abnormalities, achieving an accuracy of 89% when evaluated on patients with repaired cleft lip. Lip abnormality detection is enhanced by incorporating a lip normalization stage, which filters out anomalies to align the appearance more closely with patterns observed in control subjects. The normalization process involves two stages: image compression-decompression, which slightly degrades image resolution while smoothing lip characteristics not represented in the control data, and a face restoration model that generates high-quality details in the degraded input facial images without landmark guidance.

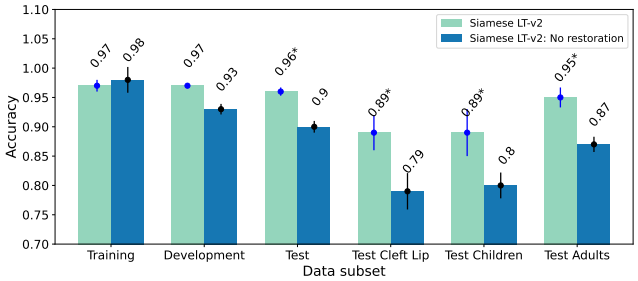


Figure 7: Accuracy comparison between our best Siamese LT-v2 approach and a variant lacking the image restoration stage in our lip normalization technique. The asterisk (*) represents a statistically significant difference between the models when applying a one-tailed two-sample proportion t-test with p-value < 0.01.

The use of a lip transformation technique to recreate lip abnormalities associated with the lip shape of patients with repaired CL enables the training of deep learning models without relying on patient data. This advancement not only enhances privacy protection but also overcomes the challenge of accessing a large volume of patient images. Our lip transformation distorts the upper lip shape of control subjects based on lip landmarks extracted from images of patients with repaired CL, resulting in an 11% accuracy improvement compared to a previous approach that tailored lip landmark displacements through experimentation. We additionally offer a visual interpretation of the areas used by our top-performing model to predict lip abnormality, demonstrating that the model analyzes the upper lip's abnormal regions with precision. Furthermore, we highlight the importance of incorporating data from both adults and children to train our model, as excluding control children's data adversely affects the model's generalization on unseen control children samples. Lastly, we analyze the impact of the image restoration stage within the lip normalization procedure. We observed that using only the image compression-decompression stage negatively impacts the performance of our Siamese LT-v2 model, reducing its accuracy in 10%.

For our future work, we envision our system as a first stage for model selection, wherein models specialized in facial landmark detection for patients with lip abnormalities can be chosen over generic models for landmark detection that are prone to error in our target population. We also plan to expand the model selection for audio solutions by incorporating a specialized automatic speech recognition system that can be selected for patients with repaired CL who present speech disorders.

ACKNOWLEDGMENTS

We gratefully acknowledge the financial support provided by the University of Texas Southwestern Medical Center and the Children's Analytical Imaging and Modeling Center Research Program.

REFERENCES

[1] S. Agarwal, R. R. Hallac, R. Mishra, C. Li, O. Daescu, and A. Kane. 2018. Image Based Detection of Craniofacial Abnormalities using Feature Extraction by

Classical Convolutional Neural Network. In *IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2018)*. Las Vegas, NV, USA, 1–6. https://doi.org/10.1109/ICCABS.2018.8541948

[2] R. Alansari, C. Bedos, and P. Allison. 2014. Living with cleft lip and palate: the treatment journey. *The Cleft palate-craniofacial journal* 51, 2 (March 2014), 222–229. https://doi.org/10.1597/12-255

[3] W. Ali, M. Imran, M. U. Yaseen, K. Aurangzeb, N. Ashraf, and S. Aslam. 2023. A transfer learning approach for facial paralysis severity detection. *IEEE Access* 11 (November 2023), 127492–127508. https://doi.org/10.1109/ACCESS.2023.3330242

[4] O. Boyaci, E. Serpedin, and M. A. Stotland. 2020. Personalized quantification of facial normality: a machine learning approach. *Scientific Reports* 10 (December 2020), 21375. https://doi.org/10.1038/s41598-020-78180-x

[5] K. Caughlin, E. Duran-Sierra, S. Cheng, R. Cuenca, B. Ahmed, J. Ji, V.V. Yakovlev, M. Martinez, M. Al-Khalil, H. Al-Enazi, J. A. Jo, and C. Busso. 2021. End-to-End Neural Network for Feature Extraction and Cancer Diagnosis of In Vivo Fluorescence Lifetime Images of Oral Lesions. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2021)*. Guadalajara, Mexico, 3894–3897. https://doi.org/10.1109/EMBC46164.2021.9629739

[6] S. Chen, A. Atapour-Abarghouei, E.S.L. Ho, and H.P.H. Shum. 2023. INCLG: Inpainting for non-cleft lip generation with a multi-task image processing network. *Software Impacts* 17 (September 2023), 100517. https://doi.org/10.1016/j.simpa.2023.100517

[7] J.A. de Souza Freitas, L.T. das Neves, A.L.P.F. de Almeida, D.G. Garib, I.K. Trindade-Suedam, R.Y.F. Yaedú, R.d.C.M.C. Lauris, S. Soares, T.M. Oliveira, and J.H.N. Pinto. 2012. Rehabilitative treatment of cleft lip and palate: experience of the Hospital for Rehabilitation of Craniofacial Anomalies/USP (HRAC/USP)-Part 1: overall aspects. *Journal of Applied Oral Science* 20, 1 (February 2012), 9–15. https://doi.org/10.1590/S1678-77572012000100003

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. Long Beach, CA, USA, 4685–4694. https://doi.org/10.1109/CVPR.2019.00482

[9] A. Hayajneh, M. Shaqfeh, E. Serpedin, and M.A. Stotland. 2023. Unsupervised anomaly appraisal of cleft faces using a StyleGAN2-based model adaptation technique. *PLOS ONE* 18, 8 (August 2023), 1–21. https://doi.org/10.1371/journal.pone.0288228

[10] J.-C. Hou, M. Thonnat, F. Bartolomei, and A. McGonigal. 2022. Automated video analysis of emotion and dystonia in epileptic seizures. *Epilepsy Research* 184 (August 2022), 106953. https://doi.org/10.1016/j.eplepsyres.2022.106953

[11] L.A. Jeni, J.F. Cohn, and T. Kanade. 2017. Dense 3D face alignment from 2D video for real-time use. *Image and Vision Computing* 58 (February 2017), 13–24. https://doi.org/10.1016/j.imavis.2016.05.009

[12] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. 2020. Training Generative Adversarial Networks with Limited Data. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Vol. 33. Virtual, 12104–12114. https://doi.org/10.48550/arXiv.2006.06676

[13] T. Karras, S. Laine, and T. Aila. 2021. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (December 2021), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

[14] D.E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (July 2009), 1755–1758.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS 2012)*, Vol. 25. Lake Tahoe, CA, USA, 1097–1105.

[16] A. Lakshmi, B. Wittenbrink, J. Correll, and D.S. Ma. 2021. The India Face Set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in Psychology* 12 (February 2021), 627678. https://doi.org/10.3389/fpsyg.2021.627678

[17] D.S. Ma, J. Correll, and B. Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47 (December 2015), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

[18] D.S. Ma, J. Kantner, and B. Wittenbrink. 2021. Chicago face database: Multiracial expansion. *Behavior Research Methods* 53 (June 2021), 1289–1300. https://doi.org/10.3758/s13428-020-01482-5

[19] M. McCullough, S. Ly, A. Auslander, C. Yao, A. Campbell, S. Scherer, and W.P. Magee III. 2021. Convolutional neural network models for automatic preoperative severity assessment in unilateral cleft lip. *Plastic and Reconstructive Surgery* 148, 1 (July 2021), 162–169. https://doi.org/10.1097/PRS.0000000000008063

[20] I. Medvedev, F. Shadmand, and N. Gonçalves. 2024. Young Labeled Faces in the Wild (YLFW): A Dataset for Children Faces Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*. Istanbul, Turkiye, 1–10. https://doi.org/10.1109/FG59268.2024.10582021

[21] K. Millar, A. Bell, A. Bowman, D. Brown, T.-W. Lo, P. Siebert, D. Simmons, and A. Ayoub. 2013. Psychological status as a function of residual scarring and facial asymmetry after surgical repair of cleft lip and palate. *The Cleft Palate-Craniofacial Journal* 50, 2 (March 2013), 150–157. https://doi.org/10.1597/10-222

[22] J. Olveres, G. González, F. Torres, J. C. Moreno-Tagle, E. Carbajal-Degante, Alejandro Valencia-Rodríguez, N. Méndez-Sánchez, and B. Escalante-Ramírez. 2021. What is new in computer vision and artificial intelligence in medical image analysis applications. *Quantitative imaging in medicine and surgery* 11, 8 (August 2021), 3830–3853. https://doi.org/10.21037/qims-20-1151

[23] K. Rosero, A. Salman, B. Sisman, R. Hallac, and C. Busso. 2024. Enhanced Facial Landmarks Detection for Patients with Repaired Cleft Lip and Palate. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*. Istanbul, Turkey, 1–10. https://doi.org/10.1109/FG59268.2024.10582022

[24] K. Rosero, A. N. Salman, C. Busso, and R. Hallac. 2024. A Tailored Machine Learning Approach for Cleft Lip Symmetry Analysis. In *The American Cleft Palate Craniofacial Association (ACPA 2024)*. Denver, CO.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. Salt Lake City, UT, USA, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[26] V.P. Sharma, H. Bella, M.M. Cadier, R.W. Pigott, T.E.E. Goodacre, and B.M. Richard. 2012. Outcomes in facial aesthetics in cleft lip and palate surgery: a systematic review. *Journal of Plastic, Reconstructive & Aesthetic Surgery* 65, 9 (September 2012), 1233–1245. https://doi.org/10.1016/j.bjps.2012.04.001

[27] J.A. Thompson, P.C. Heaton, C.M.L. Kelton, and T.J. Sitzman. 2017. National estimates of and risk factors for inpatient revision surgeries for orofacial clefts. *The Cleft Palate-Craniofacial Journal* 54, 1 (January 2017), 60–69. https://doi.org/10.1597/15-206

[28] C.-A. Trotman, C. Phillips, G.K. Essick, J.J. Faraway, S.M. Barlow, H. Wolfgang Losken, J. van Aalst, and L. Rogers. 2007. Functional outcomes of cleft lip surgery. Part I: study design and surgeon ratings of lip disability and need for lip revision. *The Cleft Palate Craniofacial Journal* 44, 6 (November 2007), 598–606. https://doi.org/10.1597/06-124.1

[29] Y. Yang and X. Guo. 2020. Generative landmark guided face inpainting. In *Pattern Recognition and Computer Vision (PRCV 2020)*, Y. Peng, Q. Liu, H. Lu, Z. Sun, C. Liu, X. Chen, H. Zha, and J. Yang (Eds.). Lecture Notes in Computer Science, Vol. 12305. Springer Berlin Heidelberg, Nanjing, China, 14–26. https://doi.org/10.1007/978-3-030-60633-6_2

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. Vegas, NV, USA, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

[31] S. Zhou, K. Chan, C. Li, and C. Change Loy. 2022. Towards robust blind face restoration with codebook lookup transformer. In *Conference on Neural Information Processing Systems (NeurIPS 2022)*, Vol. 35. New Orleans, LA, 30599–30611. https://doi.org/10.48550/arXiv.2206.11253