2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Enhanced Facial Landmarks Detection for Patients with Repaired Cleft Lip and Palate

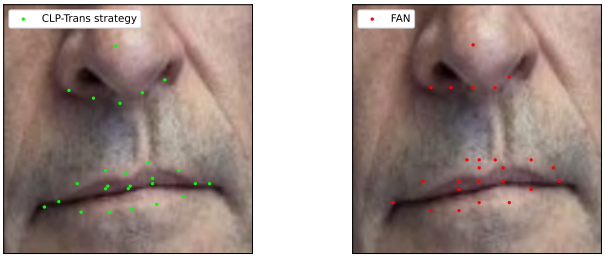
Karen Rosero[1], Ali N. Salman[1], Berrak Sisman[1], Rami R. Hallac[2,3], and Carlos Busso[1]
[1] Department of Electrical and Computer Engineering, The University of Texas at Dallas, TX 75080, USA
[2] Department of Plastic Surgery, University of Texas Southwestern Medical Center, TX 75390, USA
[3] Analytical Imaging and Modeling Center, Children's Health Dallas, TX 75235, USA

***Abstract*— *Cleft lip and palate* (CLP) is a congenital condition causing deformities in the oral and labial tissues. Post-surgery, patients often experience residual issues like facial asymmetry, and speech disorders. Tracking points in the orofacial area using a facial landmark detector (FLD) contributes to the assessment of speech development and movement impairments. However, off-the-shelf FLDs fail at delineating the lips of patients with repaired CLP. To address this need, our study introduces the *CLP-Trans* strategy, a domain transfer solution that employs tailor-made affine transformations to modify facial images sourced from publicly available datasets, which constitute our source domain, whereas images of patients with repaired CLP form our target domain. We aim to reduce distribution disparities between the source and target domains for FLD by simulating common outcomes of CLP repair surgery. The system utilizes a deep *convolutional neural network* (CNN) to learn from transformed images, therefore, preserving the privacy and facilitating the reproducibility of the findings. The strategy achieves statistically significant improvements in the *normalized mean square error* (NMSE), reducing it from 2.417 to 2.086 (i.e., 13.7% error reduction) by using the proposed strategy when evaluating images of patients with CLP.**

I. Introduction

Cleft lip and palate (CLP) are congenital conditions that emerge during fetal development and manifest as deformities of the oral and labial tissues in the orofacial area. The orofacial function encompasses complex, coordinated vital activities such as breathing, chewing, swallowing, and speaking. Additionally, the orofacial region plays a fundamental role in social interaction, involving emotional communication, facial expression, and appearance [22]. Therefore, deformities caused by CLP substantially compromise the aesthetic, morphological, and functional aspects of the orofacial function [8]. These congenital disorders affect a significant number of individuals globally at a rate of 0.45 in 1,000 [30], and at a rate of 1 in 1,000 in the USA [21]. The standard surgical approach for CLP involves the restoration of the lip, palate, and nose's anatomical structure, including the muscular components. The primary objective of optimal CLP surgical treatment is to reinstate the functional aspects of the lip and nose while simultaneously achieving maximal symmetry and aesthetic outcome for both structures [12]. However, depending on the severity level of the cleft, a variable degree of residual lip scarring, facial asymmetry, and speech disorders are expected after the first repair surgery [25], [8]. Then, multiple surgeries are often required.

Even though patients with CLP undergo speech evaluations on an annual basis, those with severe speech disorders are reevaluated as frequently as deemed necessary by qualified speech-language pathologists. Speech evaluation commonly relies on instrumental procedures such as videofluoroscopy, nasopharyngoscopy, nasometry, measurement of the pressure flow, and speech recordings [1]. Consequently, the diagnosis and evaluation of speech therapy can be a challenging task for physicians, and it can also be unpleasant for patients, especially when they are children. An automated and less invasive evaluation can contribute to a better quality of life for patients. In this context, automatic detection of facial landmarks in patients with cleft lip can be useful for analyzing movement impairments that negatively influence speech. However, current off-the-shelf automatic facial landmark detectors, such as dlib [16], MediaPipe [15], FAN [4], and OpenFace [2], are based on machine or deep learning models that are trained on non-cleft lip facial images. As these methods are not trained on CLP data, they fail when used on repaired CLP. These models tend to inaccurately delineate the lips of patients with repaired CLP by smoothening the points that diverge from a standard mouth shape, missing the key information needed to assess the effectiveness of the surgical procedure (Fig. 1).



(a) *CLP-Trans.* (b) FAN [4].

Fig. 1: Comparison of the facial landmarks extraction using our *CLP-Trans* strategy and FAN [4].

The automatic detection of facial landmarks in patients with CLP has primarily focused on providing early pathology diagnosis [37], [26] and presurgical guidance, including assessing the severity level of the cleft [19], [33]. Few studies have considered the automatic detection of facial landmarks in patients with CLP as a tool for treatment planning and diagnosis [18], [12]. However, these studies commonly rely on 2D or 3D images of patients with CLP that are subsequently manually annotated or corrected by specialists. Additionally, the collected images are not disclosed to preserve patient

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

privacy, which limits the reproducibility of the findings.

Motivated by the need to improve automatic *facial landmark detection* (FLD) for patients with repaired CLP, we propose the *CLP-Trans* strategy. This approach leverages a domain transfer solution that employs affine transformations to modify the facial appearance by manipulating specific geometric properties in the orofacial area while leaving the remaining facial structures unmodified. Facial images sourced from publicly available datasets constitute our source domain, whereas images of patients who underwent repaired CLP form our target domain. The objective is to modify these images to reduce distribution disparities between the source and target domains for FLD, simulating common outcomes of CLP repair surgery.

Subsequently, a deep convolutional neural network is trained using the modified images, preserving the privacy of patients with cleft lip. The approach is scalable since it relies on modifications of images of non-cleft lip subjects, allowing us to use existing datasets for FLD. The approach is appealing because it does not need data from patients of the target domain, for whom obtaining consent can be more challenging, relying on their parents or legal guardians, especially when treating young patients.

We evaluate our model using three different test sets. First, we use the original test set of the source domain, as defined in the publicly available databases used to train our model. Second, we apply our *CLP-Trans* strategy on the provided test set to get a synthetic CLP test set that resembles the target domain. Lastly, we evaluate our model on repaired CLP images to ensure that our proposed method improves landmark detection on real data, not just the modified faces. A flowchart of the *CLP-Trans* strategy is shown in Fig. 2. We evaluate the results using the *normalized mean square error* (NMSE) between each fitted shape and the ground truth annotations. Using the *CLP-Trans* strategy, we achieved an NMSE of 2.086 on patients' images, compared with an NMSE of 2.417 obtained from the system without transformations, leading to a 13.7% error reduction. The contributions of this paper are as follows:

- A domain transfer solution that uses tailor-made transformations to recreate the orofacial asymmetry resulting from CLP repair surgery.
- Adapting a deep learning model to improve the detection of facial landmarks in patients with repaired CLP.
- Ensuring privacy protection for the images of patients with CLP, as the deep learning model is trained using images containing synthetic CLP conditions.
- Demonstration that the inclusion of transformed images resembling the repaired CLP condition enhances FLD even on unmodified images.

II. Related Works

In cleft lip and palate analysis, both 2D and 3D photographs have been used to develop FLD methods for various purposes. These objectives include early diagnosis of the pathology in fetuses, pre- and postoperative evaluation, cleft lip severity diagnosis, and surgical support. In this section, we present a review of noteworthy studies conducted in this area, mainly focusing on the advancements made in FLD for cleft lip and palate analysis in 2D.

Lee et al. study [18] closely aligns with our research, focusing on FLD in patients with repaired CLP. The study aimed to track crucial lip landmarks for diagnosing communication impairments and to plan appropriate treatments. They employed the *active appearance model* (AAM) [34] to identify 64 facial landmarks. During the training stage, a statistical facial model is obtained from images containing manually annotated landmarks. In the inference phase, first a face detector isolates regions containing faces. Then, the AAM extracts the facial landmarks. However, the study does not report specific performance metrics.

In the context of guiding incisions in the orofacial area for cleft lip and palate repair surgery, Li et al. [19] and Sayadi et al. [33] proposed the use of deep CNNs to predict landmarks on unrepaired cleft lip 2D images. Li et al. [19] fine-tuned a model pretrained on the FLD dataset Menpo [39], using 2,568 images of patients with unrepaired cleft lip. The goal was to predict the position of 12 landmarks, resulting in improved localization of surgical markers compared to state-of-the-art (SOTA) facial feature extraction methods. Similarly, Sayadi et al. [33] extended this approach to place 21 cleft anthropometric points in 2D real images and videos using the *high-resolution network* (HRNet) [38] and the mirroring data augmentation strategy.

The identification of facial landmarks in individuals with unrepaired cleft lip can aid in assigning severity grades to the cleft before surgery. McCullough et al. [23] experimented with five different CNN-based models trained on 800 cleft lip images with manually annotated landmarks specific to cleft lip repair surgery. Among these models, the MobileNet model [13] demonstrates the best performance in FLD.

Chen et al. [5], [6] proposed an alternative approach, employing image inpainting to generate non-cleft lip images and their corresponding landmarks based on cleft-lip images. The method involves training a convolutional-based generation network on facial images of individuals without cleft lip, masked for privacy preservation. During inference, the orofacial area of images of patients with unrepaired cleft lip is manually masked and processed, generating images representing plausible outcomes of cleft lip repair surgery along with their facial landmarks. Although this approach does not specifically improve the detection of cleft lip landmarks, it produces images that simulate common outcomes of the repair surgery.

In line with the motivation of Lee et al. [18], Hallac et al. [12] tracked the movement of 13 landmarks in video stereophotogrammetry recordings of 23 patients with repaired cleft lip, aiming to detect dynamic facial asymmetries. Although the work does not concentrate on improving FLD, it is relevant to note that the landmarks were automatically tracked using the DI4D view software, but required manual verification frame by frame to ensure accurate placement. We gain insights into the advancements made in cleft lip analysis and facial landmark extraction by reviewing these

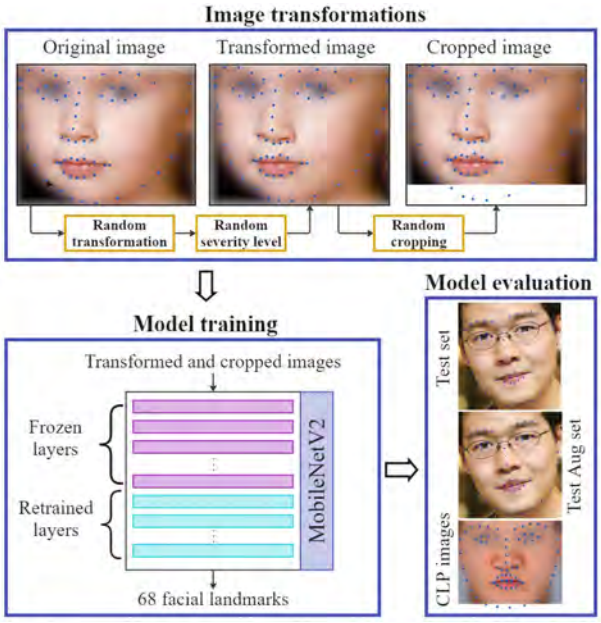


Fig. 2: Flowchart of the *CLP-Trans* strategy. The images have been blurred for privacy preservation. However, the images are not blurred to train the model.

related studies. It becomes evident that there is a need for a system capable of accurately identifying facial landmarks in patients with repaired cleft lip, which aligns with the focus of our work.

No established domain transfer methodologies exist for patients with CLP. Jin et al. [14] identified domain discrepancies affecting landmark detection in orthodontic diagnosis proposing a self-training strategy using unlabeled target domain data. However, privacy concerns arise as this approach requires patient data for model training.

III. Methodology

This section describes the image domain transfer foundation along with the image transformation details and the MobileNetV2 model adaptation, illustrated in Fig. 2.

A. Image Domain Transfer

In the field of transfer learning, the scenario where labeled data is available solely from a source domain while lacking it from the target domain is termed transductive transfer learning. Additionally, encountering a situation where source and target data vary in domain but are intended for the same task presents a domain transfer problem [27]. In this study, there is plenty of labeled data within the source domain (FLD datasets). However, it is imperative not to use data from the target domain (patients with repaired CLP) for training the deep learning model. Instead, our approach involves the transformation of images from the source domain to generate synthetic images resembling those found in the target domain.

B. Image Transformations for Recreated CLP Outcomes

The primary surgical closure of the cleft is usually initiated within the first 12 months of age, with the aim of achieving normal speech and swallowing function [1]. However,

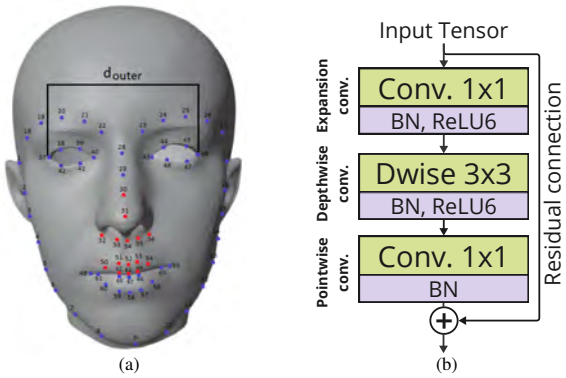


Fig. 3: (a) The 68 facial landmarks proposed by Sagonas et al. [29] used as standard. The points shown in red are the ones modified in the transformation process. The interocular distance d_{outer} is measured between the outer points of the eyes. (b) MobileNetV2 building block.

different severity levels of deformities in the orofacial area of repaired CLP are a potential complication that may require secondary surgical procedures [32]. Since secondary procedures can be performed at various stages from infancy through adulthood, there is a high probability that the patients with repaired CLP start speech therapy before having the corrective surgery to reduce lip and nose asymmetry. As a result, automatic facial landmark detectors that could assist in speech therapy may inaccurately place landmarks in the orofacial area of patients with CLP. Therefore, we propose a tailored technique inspired by common outcomes of the initial CLP repair surgery to transform facial images of FLD datasets into images that resemble repaired CLP. We consider seven transformations: unilateral upper lip asymmetry with and without Cupid's bow modification, bilateral asymmetry of the upper lip, nose deviation, which can appear along with lip asymmetry, absent Cupid's bow, and thin upper lip. We hypothesize that the use of these transformations on images without CLP can improve the automatic FLD on images of patients with repaired CLP.

We rely on the 68 facial landmarks (Fig. 3a) proposed by Sagonas et al. [29] in the standard format for FLD. First, we triangulate all the points to create smaller regions where the transformation will be applied. Subsequently, we identify the points in the orofacial area that will be modified to resemble the outcomes of CLP repair surgery (red points in Fig. 3a). Table I provides details of all the transformations, including the specific modified points and the corresponding limit values. These limits were selected by experimentation, such that the minimum value introduces a slight modification and the maximum value produces a severe modification while preventing from deforming nearby regions. Then, these limit values were mapped into the range $(0-1)$ to standardize the severity level applied in both transformations: those based on the rhinal ala-columella distance and those based on the thickness of the upper lip.

The transformations related to modifications of the upper lip are determined by the *thickness of the upper lip* (TUL), which is calculated as the mean distance between points $(51, 62)$ and $(53, 64)$, as shown in Fig. 3a. An exception is

the transformation resulting in an absent Cupid's bow, which involves displacing point 52 to the same height as the *cupid's bow peaks* (CBPs). Regarding the nose transformations, we measure the distance between the point in the rhinal ala (e.g. point 32) and the rhinal columella (e.g. point 33) to define the *rhinal ala-columella* (RAC) distance suitable for the nose transformation of the subject. Once the source (i.e., original location) and target (i.e., selected level of anomaly) points are defined, an affine transformation is applied to the corresponding regions.

TABLE I: Transformations that simulate common outcomes of the CLP first repair procedure.

Transformation	Position	Points	Limits†
Unilateral asymmetry	Left	51	0.3-1.2 TUL
	Right	53	
Unilateral asymmetry + Cupid's bow modification	Left	51,52	
	Right	52,53	
Bilateral asymmetry	Center	51,52,53	
Nose deviation	Left	30,31,32, 33,34	0.3-0.8 RAC
	Right	30,31,34, 35,36	
Nose deviation with lip asymmetry	Left	30,31,32, 33,34, 51	
	Right	30,31,34, 35,36, 53	
Absent Cupid's bow	Center	52	Height of CBP
Thin lip	Upper lip	50,51,52, 53,54	Half of the TUL

† The limit values are then mapped into severity levels ranging from $(0-1)$.

In the training process with the *CLP-Trans* strategy, one transformation per image is randomly selected and applied to the facial image and its landmarks. It is important to note that not all images within a batch receive the same transformation.

Fig. 4 demonstrates the process for five of the transformations. The first row of Fig. 4 illustrates the transformations applied to an image of a subject without CLP. The figure shows how the transformations modify different points in the orofacial area. The second row shows images of patients with CLP, as examples of the CLP repair surgery outcomes.

C. Deep Convolutional Neural Network

Our proposed approach is built upon the MobileNetV2 network [31], chosen for its lightweight architecture that can generate predictions even in resource-constrained environments. This implementation could allow physicians to execute our face-based solution during therapy on resource-constrained devices such as smartphones. As illustrated in Figure 3b, the building block of MobileNetV2 comprises lightweight depthwise separable convolutions. These convolutions divide each standard convolution into three parts: an expansion layer, a depthwise convolution, and a pointwise convolution. This division helps in reducing the network's computational load. The expansion layer takes a low-dimensional tensor from the preceding block and augments the output channels by a factor defined as the expansion factor $\boldsymbol{t}$. Subsequently, the depthwise convolution filters the input and diminishes the spatial dimensions if the block has a stride of 2. Finally, the projection layer employs a pointwise convolution to decrease the number of channels in the intermediate feature maps. This strategy creates a linear bottleneck that encourages the reuse of features. Moreover, the inclusion of a residual connection, which adds the input of the building block to its output, assists in maintaining a smooth flow of gradients throughout the network.

We adapted the MobileNetV2 architecture, which consists of a total of 17 building blocks, referred to as bottlenecks. These sequential bottlenecks are preceded by a standard 3×3 convolution with 32 channels and followed by a regular 1x1 convolution, an average pooling layer, and a dense layer. Additionally, we modified the original output classification layer of MobileNetV2 to accommodate the FLD task. Initially, the model had a dense layer with dimension set to 1,000. Our implementation uses a dense layer with dimension $2N$, enabling the prediction of 68 2D points, where $N = 68$ facial landmarks expressed in Cartesian coordinates. In essence, the task transitioned from a classification task to a regression task. MobileNetV2 encompasses a total of 53 convolutional layers throughout its architecture. However, it contains 156 layers if we consider all batch normalization, dropout, rectified linear unit (ReLU6), and pooling layers. Further details about the MobileNetV2 architecture can be found in [31].

We use the MobileNetV2 pre-trained on ImageNet-1k [9] for image classification as our initial model. Then, we conducted experiments by reusing the learned weights of a specific number of layers and retraining a set of layers for our landmark detection task. The number of trainable layers serves as a hyperparameter, which will be discussed in Section V.

IV. Experimental Settings

The *CLP-Trans* strategy was implemented in PyTorch and executed in an NVIDIA GeForce RTX 4090 GPU. This section provides further details about the datasets, the preprocessing procedure, the selection of hyperparameters, and performance metrics.

A. Datasets for Facial Landmark Detection

The 300 Faces In-The-Wild Challenge (300W-Challenge) [29] standardized the landmark configuration of existing datasets to provide a manually corrected set of 68 points expressed in Cartesian coordinates for automatic facial points detection. The datasets for facial alignment contain images with different resolutions, subjects' identities, head poses, facial expressions, lighting conditions, and partial occlusions [29]. Additionally, they share a common characteristic of having been collected from websites on the Internet, resulting in images taken under unconstrained conditions. Four datasets are available for research purposes and can be downloaded from the Intelligent Behavior Understanding Group website [1]:

[1] https://ibug.doc.ic.ac.uk/resources/300-W/

(a) Original image from the HELEN set. (b) Repaired unilateral CLP. (c) Repaired bilateral CLP. (d) Nose asymmetry after CLP repair surgery. (e) Absent Cupid's bow after repair surgery. (f) Thin lip after repair surgery.

Fig. 4: Image (a) displays the original picture before transformations. The top images (b)-(f) imitate the outcomes of CLP repair surgery, while the bottom images show the orofacial area of patients with repaired CLP. The figure shows the resemblance of the deformity between the modified images at the top and the images of patients with CLP at the bottom.

- *Annotated faces in-the-wild* (AFW) dataset [41]: 337 facial images were extracted from 205 images. The dataset originally considered 6 landmarks, however, we employ the 68 points re-annotated for the 300W-Challenge.
- HELEN dataset [17]: contains 2,330 portrait images of approximate face size of 500×500 pixels. The database has 2,000 images for the train set, and 330 images for test set. The original 194 points were replaced by the 68 standard landmarks of the 300W-Challenge.
- *Labeled face parts in-the-wild* (LFPW) dataset [3]: contains 1,035 images. The database defines 811 images for the train set, and 224 images for the testing set. The images may contain occlusions and theatrical makeup.
- IBUG dataset [29]: contains 135 images of highly expressive faces.

Since the test split of HELEN and LFPW were predefined, we randomly select 20% of the training samples to form the development set. This resulted in 1,600 images in the HELEN training set and 400 images in the development set, while LFPW had 647 images in the training set and 164 in the development set. We keep the original test set unmodified. For the AFW and IBUG datasets, which were not initially divided, we separated them into three parts: 70% for training, 15% for development, and 15% for testing. For the AFW corpus, we use 235 images for the training set, 50 for the development set, and 52 for the test sets. For the IBUG corpus, we use 95 images for the training set, 20 for the development set, and 20 for the testing set. Across all databases, the sets included a total of 2,577 images for training, 634 images for development, and 626 images for testing.

B. Images of Patients with Repaired CLP

We evaluate the *CLP-Trans* strategy with images of real CLP patients as well. We obtained 123 images from the internet, prioritizing websites of surgeons specialized in CLP [10], [7], hospital websites [20], and non-profit organizations for CLP treatment [35], [36]. Some images were also collected using target queries on search engines (e.g., "repaired cleft lip patient"). The 68 facial landmarks were first placed using the dlib face detector and landmarks estimator [16] as a starting point. We then performed manual corrections, with particular attention to the orofacial area. The dlib toolkit failed to detect a face in six images. For these images, we manually cropped the faces, placing the facial landmarks following the standard format shown in Fig. 3a.

C. Preprocessing Approach

To crop the images of the 300W-Challenge datasets, we rely on the provided bounding boxes, expanding them by 10% of the image dimension on each side. Then, the images are resized to 224×224 pixels to ensure compatibility with the pretrained weights of the MobileNetV2 model. The landmarks are mapped to the new dimension. The next step is the stage of image transformation, followed by a normalization step. The images of patients with repaired CLP are cropped based on the dlib face detector and undergo the same resizing and normalization procedure.

D. Hyperparameters' Selection

The *CLP-Trans* strategy is implemented with a batch size of 16 samples, a learning rate of 0.0001, the *mean square error* (MSE) loss function, and Adam optimizer. All experiments were trained for a maximum of 400 epochs with an early-stopping patience of 20 epochs, while monitoring the development loss. We determine some hyperparameters used in this study by analyzing the experimental results. In Section V, we demonstrate the performance of the *CLP-Trans* strategy under different numbers of retrained layers in the MobileNetV2 architecture, different severity levels, static or dynamic transformations of CLP, and diverse augmentation strategies (random crop and multiple augmentations).

E. Performance Evaluation

We estimate the performance of the predicted landmarks with the MSE using two approaches. The first approach adheres to the metric described in [29], which computes the point-to-point MSE and normalizes it by the interocular distance d_{outer}. We estimate d_{outer} as the Euclidean distance between the outer points of each eye (points 37 and 46 in Fig. 3a). Then, the NMSE point-to-point (NMSE_{ptp}) is computed for each facial image, and the median value for the whole set is reported as the final metric. The second approach, used by Dong et al. [11], sums the RMSE of all landmarks and normalizes the result by d_{outer} instead of the point-to-point

difference. The final NMSE for all the 2D points in the face ($NMSE_{face}$) is reported as the mean value of the entire set.

V. Results

This section present the outcomes obtained from various configurations used to evaluate the *CLP-Trans* strategy. We compare different implementations to determine the best approach, with results presented in the order that system improvements were introduced. Table II compiles the results of all experiments to facilitate the comparison of the performance metrics (we will discuss these results in this section). Regarding the performance metrics, we document the MSE loss of the 68 points on the face (MSE_{face}), the MSE loss of the 15 points in the orofacial area (MSE_{orof}), and the MSE loss of the 53 remaining points in the face (MSE_{rest}). We also report the normalized metric $NMSE_{face}$ described in Section IV-E. These metrics allow us to compare the facial landmark errors in the orofacial area, regions not related to the orofacial area, and the entire face. Therefore, we can analyze if the proposed transformations have unintentional artifacts in the target orofacial region and outside the target orofacial region.

We use several data subsets to report the metrics. The 'Training' set can contain transformed images depending on the experiment. The sets labeled as 'Development' and 'Test' are never transformed; the 'Test Mod' subset contains the same samples as the 'Test' set, but its images were transformed into synthetic images with repaired CLP. Finally, the 'CLP images' subset contains 136 images of patients with repaired CLP, as detailed in Section IV-B.

A. Model Trained Without Transformations

As an initial reference, we train the facial landmark system without CLP transformations to evaluate its performance as we change the number of retrained layers of the adapted MobileNetV2 model. The model adapted for FLD contains a set of pre-trained layers followed by a randomly initialized dense layer for regression. We need to retrain a specific number of layers to adapt the model to the FLD task. Since the layers closer to the input are responsible for the extraction of low-level features of images such as contours, edges, angles, and colors, we prioritize reusing these first layers as they can be considered task-agnostic. Therefore, we experimented with retraining the latest layers, which are responsible for extracting high-level features. Fig. 5 shows the variation of the $NMSE_{face}$ metric when a different number of layers is retrained for our task. The 'Test Mod' set, which contains faces with synthetic repaired CLPs, reached its best performance when retraining 96 layers of the adapted MobileNetV2 architecture. Therefore, we implement the rest of the experiments by retraining the last 96 layers of the model.

All metrics associated with the best result of this experiment are shown in the column No-Trans of Table II. Note that the performance of the system decreases when evaluating it on the 'Test Mod' subset, which contains images modified to resemble the outcomes of CLP repair surgery, showing the necessity of our domain transfer approach.

TABLE II: Performance metrics of the *CLP-Trans* system under different experimental conditions.

	Exp.	Metrics			
		MSE_{face}	MSE_{orof}	MSE_{rest}	$NMSE_{face}$
Training	No-Trans	16.259	5.447	10.812	2.991
	St-Trans	17.973	6.326	11.646	3.151
	Dyn-Trans	14.840	5.445	9.396	2.860
	Cr-DA	18.849	6.508	12.342	3.083
	Mul-DA	12.892	4.989	7.903	2.614
Development	No-Trans	15.734	4.132	12.024	2.813
	St-Trans	16.143	4.129	12.014	2.891
	Dyn-Trans	14.678	3.592	11.086	2.778
	Cr-DA	15.320	3.986	11.334	2.735
	Mul-DA	13.739	3.305	10.434	2.564
Test	No-Trans	17.339	4.454	12.885	**2.860**
	St-Trans	17.041	4.369	12.672	**2.839**
	Dyn-Trans	17.112	4.591	12.520	**2.807***
	Cr-DA	15.826	3.951	11.874	**2.758***
	Mul-DA	14.500	3.715	10.785	**2.571****
Test Mod	No-Trans	23.277	5.317	17.960	**3.369**
	St-Trans	21.475	5.358	16.117	**3.258**
	Dyn-Trans	20.898	5.003	15.895	**3.159***
	Cr-DA	19.076	4.547	14.529	**3.057***
	Mul-DA	17.370	4.017	13.353	**2.829****
CLP images	No-Trans	15.279	**3.097**	12.182	**2.417**
	St-Trans	16.062	**2.902**	13.160	2.478
	Dyn-Trans	13.862	3.039	10.823	**2.263***
	Cr-DA	12.365	2.834	9.531	**2.222***
	Mul-DA	10.802	2.580	8.221	**2.086****

The asterisk (*) indicates that the approach is significantly better than the model without transformations (No-Trans), using a one-tailed two-sample proportion t-test with p-value < 0.05. For the double asterisk (**), the same t-test ensures a p-value< 0.01.

B. Static Transformations

In these experiments, the image transformation type is randomly selected for every sample of the 'Training' set. However, the severity level of the transformation was set in a static way, such that all modified samples undergo a transformation with the same severity. Fig. 6b shows the $NMSE_{face}$ metric for different values of static severity levels. The best performance of the St-Trans experiment is achieved with a maximum severity level of 0.3, for which we obtain $NMSE_{face} = 3.258$ in the 'Test Mod' set. This result is better than the best result without transformations (No-Trans) experiment in Fig. 6a). All the metrics associated with the best result of this experiment (St-Trans) are shown in Table II. Although the St-Trans experiment did not improve the $NMSE_{face}$ metric on real CLP images, we draw attention to the MSE_{orof} metric, which improves for the 'CLP images' set.

Training our model with modifications with the same CLP severity level in all samples did not improve the performance for the 'CLP images' set. This finding aligns with the nature of the CLP condition, as the severity of the cleft is not the same for all patients, and subsequently, the repair surgery outcome can also vary depending on the initial severity of the cleft.

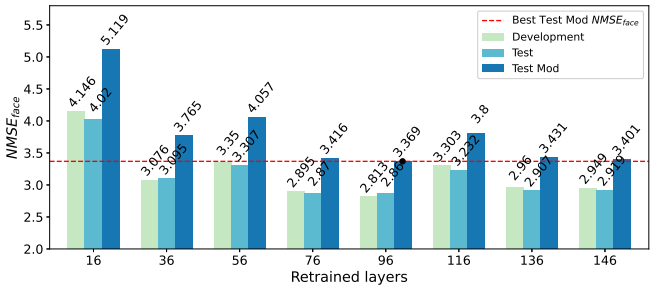
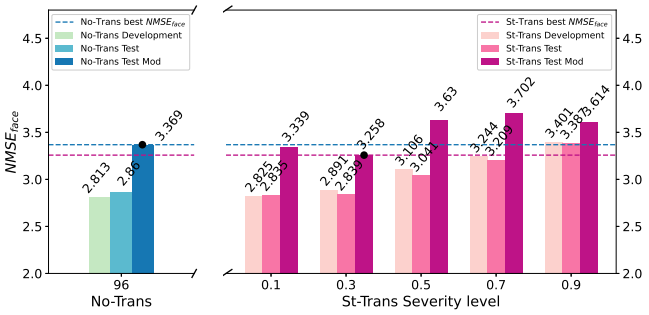


Fig. 5: Performance of the FLD system trained without CLP transformations while retraining a different number of layers. The best $NMSE_{face}$ performance is marked with a dashed red line.



(a) Best result No-Trans. (b) Performance of St-Trans.

Fig. 6: Comparison between the *CLP-Trans* strategy trained without transformations (No-Trans) and the static severity levels of CLP transformations (St-Trans).

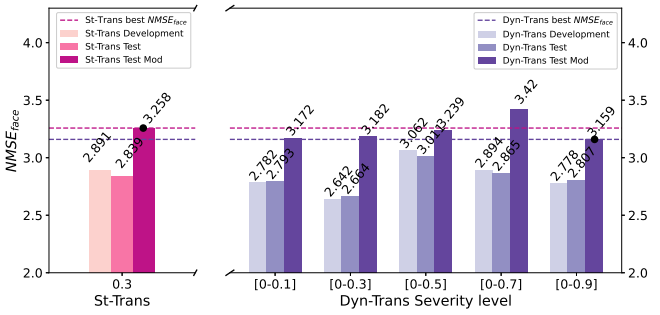
C. Dynamic Transformations

We evaluate a dynamic strategy for the CLP transformations to address the limitations observed in Section V-B. The approach, referred to as Dyn-Trans, randomly selects the severity level for every sample such that it does not surpass a maximum severity level. Fig. 7b shows the $NMSE_{face}$ metric for different values of the dynamic severity levels. While a maximum severity level of 0.9 resulted in the best performance, severities of 0.1 and 0.3 also produced comparable results. Fig. 7a shows the best result for the model with static transformation St-Trans to facilitate the comparison between both experiments. The dynamic strategy leads to an improvement in the results. Table II shows all the metrics associated with the best result of Dyn-Trans. We highlight that the $NMSE_{face}$ metric improved for all subsets.

D. Random Crop Data Augmentation

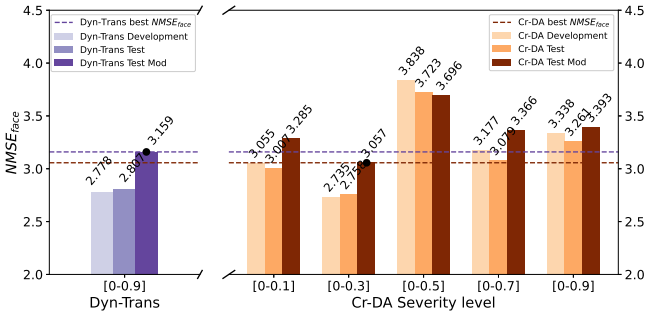
The next step in our implementation was to adapt the well-established data augmentation technique of randomly cropping images to our FLD task. Based on the ground-truth bounding box used to crop the facial image, we crop 20% off the dimension of one of the sides at a time. Then, the landmarks were displaced to match the cropped image. This experiment, referred to as Cr-DA, is applied along with the dynamic CLP transformation strategy using different values for the maximum severity level.

Fig. 8b shows the performance of the $NMSE_{face}$ metric for different dynamic severity levels of CLP, along with random cropping. The best result of the Cr-DA experiment is better



(a) Best result St-Trans. (b) Performance of Dyn-Trans.

Fig. 7: Comparison between the best result of the St-Trans experiment and the performance of Dyn-Trans. Using a dynamic severity level with a maximum severity of 0.9 leads to better performance than the results using the St-Trans setting.



(a) Best result Dyn-Trans. (b) Performance of Cr-DA.

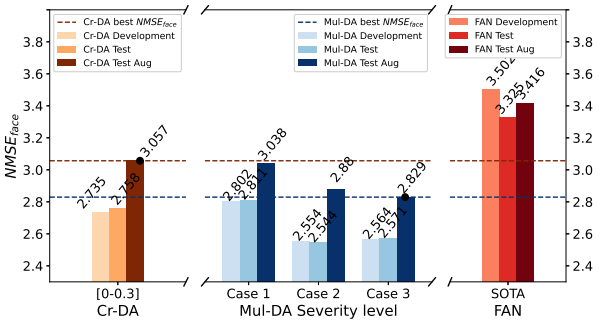
Fig. 8: Comparison between the best result of the Dyn-Trans experiment and the performance of the Cr-DA experiment, which includes random cropping along with dynamic severity levels of CLP transformations.

than the performance of the Dyn-Trans experiment (Fig. 8a), demonstrating the benefits of using a traditional data augmentation technique for computer vision, along with the CLP transformation technique proposed in this work. Table II shows all the metrics for the Cr-DA setting. The performance improves on the images of patients with CLP as well.

E. Multiple CLP Severity Levels

The last step in our implementation consists of a multiple augmentation approach, referred to as Mul-DA. This method triplicates the 'Training' set by using three different maximum severity levels for our *CLP-Trans* strategy. The approach is applied to each training subset using three implementations. Case 1 sets the maximum severity levels to 0.5, 0.3, and 0.1. Case 2 sets the maximum severity levels to 0.7, 0.5, and 0.3. Case 3 sets the maximum severity levels to 0.9, 0.7, and 0.5. This approach also uses the random crop augmentation strategy discussed in Section V-D.

Fig. 9b shows the results for the Mul-DA approaches. The best performance is achieved with the severity levels of Case 3, which exceeded the best result of the Cr-DA experiment (Fig. 9a). Table II reports the rest of the metrics for this evaluation, which indicate clear improvements in the images of real patients with CLP. This setting is the best result of our *CLP-Trans* strategy. We compare this model with FAN [4], a SOTA system for FLD (Fig. 9c), demonstrating the superior performance of our system across all subsets.



(a) Best result of Cr-DA. (b) Performance of Mul-DA. (c) FAN performance.

Fig. 9: Comparison between the best result of the Cr-DA experiment, the performance of our best result (Mul-DA experiment), and the performance of the SOTA facial landmark detector FAN [4].

We applied a one-tailed two-sample proportion t-test to the $\text{NMSE}_{\text{face}}$ metric to evaluate the statistical significance of the improvements. Each experiment that introduced a transformation or data augmentation technique (St-Trans, Dyn-Trans, Cr-DA, Mul-DA) was individually compared with the approach without transformations (No-Trans). We assert significance when p-value < 0.05. As shown in Table II, all experiments resulted in statistical significance, except for the St-Trans experiment, which, despite not satisfying the threshold, improved the metrics. Furthermore, we point out that the Mul-DA experiment satisfied a p-value < 0.01, demonstrating that our best experiment resulted in an even more significant improvement.

Columns MSE_{orof} and MSE_{rest} in Table II show the performance on the points placed in the orofacial area and the rest of the face, respectively. Although our main concern is to improve the placement of orofacial points, all experiments increased the landmark detection of the remaining points on the face as well. Moreover, the MSE_{orof} metric increased even when the model is evaluated on non-transformed images. This result demonstrates that the *CLP-Trans* strategy is not biased to have a good performance exclusively in facial images of patients with repaired CLP.

F. Comparison with Benchmarks for Facial Alignment

We evaluate the best implementation of the *CLP-Trans* strategy in the 300W test set, which is a well-known benchmark set used by studies on FLD [29]. This benchmark set comprises 600 facial images taken from indoor and outdoor spaces in equal proportions. No transformations were applied to the 300W test set, as we aimed to demonstrate that training our system with the *CLP-Trans* strategy does not negatively impact performance on faces without CLP.

We can make a close comparison between our work and the performance of the *style aggregated network* (SAN) system [11], since it was trained on the same landmark detection datasets (AFW, HELEN, LFPW, IBUG) used in our study. SAN uses data augmentation techniques to modify the color, lighting, and style of the images. SAN achieved an $\text{NMSE}_{\text{face}}$ of 3.980 on the 300W test set, while our best *CLP-Trans* system achieved an $\text{NMSE}_{\text{face}}$ of 3.411.

We also compare our work with the winner of the 300W-Challenge [40] that used a cascade of four CNNs, where each network refines the prediction of a specific face region. The approach achieved an NMSE_{ptp} of 0.0205 on the 300W test set. Our system reached a value of 0.046 for the same metric and test set. An important difference between the models is that the 300W-Challenge incorporates two additional databases that we did not use in this study: XM2VTS database [24] (2,360 images) and FRGC-V2 database [28] (4,950 images). The results in this section are important since we need to maintain high accuracies for individuals without CLP since some repaired CLP cases result in almost non-visible asymmetry or scars.

VI. Conclusions

This study introduced the *CLP-Trans* strategy, aimed at enhancing FLD for patients with CLP. The proposed approach modified facial images without CLP to mimic common outcomes of CLP first repair surgery in the orofacial region, eliminating the need to collect facial images of patients. Training a deep learning model that improves FLD for patients with CLP without relying on their photos represents a significant step forward in protecting their privacy while continuing to create computational tools that can assist in speech evaluation.

The deep convolutional neural network, MobileNetV2, was adapted and partially retrained for this task. A total of 96 layers were retrained and five experiments were conducted to assess the effectiveness of different transformations and data augmentation techniques.

Employing a dynamic severity level, randomly selected within the bounds of a maximum parameter, resulted in enhanced detection performance across all three test sets. The CLP transformation technique was further enhanced by incorporating randomly cropped facial images and introducing multiple severity levels, effectively triplicating the number of training images. As a result of these efforts, our best approach achieved a significant reduction in $\text{NMSE}_{\text{face}}$ from 2.417 to 2.086 through the implementation of the described domain transfer technique. Furthermore, the *CLP-Trans* strategy implemented with the best setting was evaluated on the 300W-Challenge dataset, demonstrating its competitive performance compared to standard FLD methods.

For future research directions, we consider applying our *CLP-Trans* strategy along with domain adaptation techniques on the feature-level. We also plan to explore the use of guided generative neural networks to replicate the outcomes of repair surgery on patients with CLP. These potential avenues of research hold the promise of further improving FLD accuracy and advancing the application of the *CLP-Trans* strategy in the field of orofacial analysis for speech treatment of patients with CLP.

VII. 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] American Cleft Palate-Craniofacial Association. Parameters for evaluation and treatment of patients with cleft lip/palate or other craniofacial anomalies. *The Cleft Palate Craniofacial Journal*, 2000.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. OpenFace: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV 2016)*, pages 1–10, Lake Placid, NY, USA, March 2016.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 545–552, Colorado Springs, CO, USA, June 2011.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *IEEE International Conference on Computer Vision (ICCV 2017)*, pages 1021–1030, Venice, Italy, October 2017.

[5] S. Chen, A. Atapour-Abarghouei, E. Ho, and H. Shum. INCLG: Inpainting for non-cleft lip generation with a multi-task image processing network. *Software Impacts*, 17:100517, September 2023.

[6] S. Chen, A. Atapour-Abarghouei, J. Kerby, E. Ho, D. Sainsbury, S. Butterworth, and H. H. Shum. A feasibility study on image inpainting for non-cleft lip generation from patients with cleft lip. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2022)*, pages 1–4, Ioannina, Greece, September 2022.

[7] L. Cuadros. Gallery of Cleft Memories. `http://www.nmcleft.org/gallery/`, 2023. Accessed: April 11, 2023.

[8] J. de Souza Freitas, L. das Neves, A. de Almeida, D. Garib, I. Trindade-Suedam, R. Yaedú, R. Lauris, S. Soares, T. Oliveira, and J. Pinto. 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):9–15, February 2012.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, USA, June 2009.

[10] C. Derderian. Plastic and Reconstructive Surgery. `https://www.drderderian.com/`, 2022. Accessed: April 10, 2023.

[11] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 379–388, Salt Lake City, UT, USA, June 2018.

[12] R. Hallac, J. Feng, A. Kane, and J. Seaward. Dynamic facial asymmetry in patients with repaired cleft lip using 4D imaging (video stereophotogrammetry). *Journal of Cranio-Maxillofacial Surgery*, 45(1):8–12, January 2017.

[13] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv e-prints (arXiv:1704.04861)*, pages 1–9, April 2017.

[14] H. Jin, H. Che, and H. Chen. Unsupervised domain adaptation for anatomical landmark detection. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2023)*, volume 14220 of *Lecture Notes in Computer Science*, pages 695–705. Springer, Cham, Vancouver, BC, Canada, October 2023.

[15] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann. Real-time facial surface geometry from monocular video on mobile GPUs. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality (CV4ARVR 2019)*, pages 1–4, Long Beach, CA, June 2019.

[16] D. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, July 2009.

[17] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang. Interactive facial feature localization. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *European Conference on Computer Vision (ECCV 2012)*, volume 7574 of *Lecture Notes in Computer Science*, pages 679–692. Springer Berlin Heidelberg, Florence, Italy, October 2012.

[18] N. Lee, C. Heaston, A. Rey, T. Hartman, and C. Trotman. Facial landmark extraction for lip tracking of patients with cleft lip using active appearance model. In C. Stephanidis, editor, *International Conference on Human-Computer Interaction (HCI 2011)*, volume 174 of *Communications in Computer and Information Science (CCIS)*, pages 350–354. Springer-Verlag Berlin Heidelberg, Orlando, FL, USA, July 2011.

[19] Y. Li, J. Cheng, H. Mei, H. Ma, Z. Chen, and Y. Li. CLPNet: Cleft lip and palate surgery support with deep learning. In *IEEE Engineering in Medicine and Biology Society (EMBC 2019)*, pages 3666–3672, Berlin, Germany, July 2019.

[20] C. H. S. Louis. Cleft Palate and Craniofacial Institute. `https://www.stlouischildrens.org/conditions-treatments/plastic-surgery/photo-gallery`, 2023. Accessed: April 11, 2023.

[21] C. Mai, J. Isenburg, M. Canfield, R. Meyer, A. Correa, C. Alverson, P. Lupo, T. Riehle-Colarusso, S. Cho, D. Aggarwal, and R. Kirby. National population-based estimates for major birth defects, 2010-2014. *Birth Defects Research*, 111(18):1420–1435, October 2019.

[22] N. Mariano, M. Sano, V. Curvêllo, A. De Almeida, K. Neppelenbroek, T. M. Oliveira, and S. Soares. Impact of orofacial dysfunction on the quality of life of adult patients with cleft lip and palate. *The Cleft Palate-Craniofacial Journal*, 55(8):1138–1144, March 2018.

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

[24] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *International conference on audio and video-based biometric person authentication (AVBPA 1999)*, pages 965–966, Washington, DC, USA, March 1999.

[25] K. Millar, A. Bell, A. Bowman, D. Brown, T.-W. Lo, P. Siebert, D. Simmons, and A. Ayoub. 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):150–157, March 2013.

[26] S. Moos, F. Marcolin, S. Tornincasa, E. Vezzetti, M. Violante, G. Fracastoro, D. Speranza, and F. Padula. Cleft lip pathology diagnosis and foetal landmark extraction via 3D geometrical analysis. *International Journal on Interactive Design and Manufacturing*, 11:1–18, February 2017.

[27] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[28] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, pages 947–954, San Diego, CA, USA, June 2005.

[29] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, March 2016.

[30] N. Salari, N. Darvishi, M. Heydari, S. Bokaee, F. Darvishi, and M. Mohammadi. Global prevalence of cleft palate, cleft lip and cleft palate and lip: A comprehensive systematic review and meta-analysis. *Journal of Stomatology, Oral and Maxillofacial Surgery*, 123(2):110–120, April 2022.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 4510–4520, Salt Lake City, UT, USA, June 2018.

[32] S. Sarrami, A. Skochdopole, A. Ferry, E. Buchanan, L. Hollier Jr, and R. Dempsey. Revisional techniques for secondary cleft lip deformities. *Seminars in Plastic Surgery*, 35(02):65–71, May 2021.

[33] L. Sayadi, U. Hamdan, Q. Zhangli, J. Hu, and R. Vyas. Harnessing the power of artificial intelligence to teach cleft lip surgery. *Plastic and Reconstructive Surgery Global Open*, 10(7):e4451, July 2022.

[34] K. Seshadri and M. Savvides. Robust modified active shape model for automatic facial landmark annotation of frontal faces. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BATS 2009)*, pages 1–8, Washington, DC, USA, September 2009.

[35] O. Smile. Operation Smile: Stories You Make Possible. `https://www.operationsmile.org/stories-you-make-possible`. Accessed: June 11, 2023.

[36] S. Train. Smile Train: Every smile matters. Every smile deserves to be seen. `https://www.smiletrain.org/`. Accessed: June 15, 2023.

[37] E. Vezzetti, D. Speranza, F. Marcolin, and G. Fracastoro. Diagnosing cleft lip pathology in 3D ultrasound: A landmarking-based approach. *Image Analysis and Stereology*, 35(1):53–65, March 2016.

[38] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 43(10):3349–3364, October 2021.

[39] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The Menpo facial landmark localisation challenge: A step towards the solution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017)*, pages 2116–2125, Honolulu, HI, USA, July 2017.

[40] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *IEEE International Conference on Computer Vision Workshops (ICCVW 2013))*, pages 386–391, Sydney, NSW, Australia, March 2013.

[41] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pages 2879–2886, Providence, RI, USA, June 2012.