

Deep Learning-Based Assessment of Lip Symmetry for Patients With Repaired Cleft Lip

Karen Rosero, MS^{1,2} , Ali N. Salman, PhD¹, Lucas M. Harrison, MD³ , Alex A. Kane, MD³, Carlos Busso, PhD^{1,2}, and Rami R. Hallac, PhD^{3,4} 

The Cleft Palate Craniofacial Journal
2025, Vol. 62(2) 289-299
© 2025, American Cleft Palate
Craniofacial Association



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10556656241312730
journals.sagepub.com/home/cpc



Abstract

Objective: Post-surgical lip symmetry assessment is a key indicator of cleft repair success. Traditional methods rely on distances between anatomical landmarks, which are impractical for video analysis and overlook texture and appearance. We propose an artificial intelligence (AI) approach to automate this process, analyzing lateral lip morphology for a quantitative symmetry evaluation.

Design: We utilize contrastive learning to quantify lip symmetry by measuring the similarity between the representations of the sides, which is subsequently used to classify the severity of asymmetry. Our model does not require patient images for training. Instead, we introduce dissimilarities in face images from open datasets using two methods: temporal misalignment for video frames and face transformations to simulate lip asymmetry observed in the target population. The model differentiates the left and right image representations to assess asymmetry. We evaluated our model on 146 images of patients with repaired cleft lip.

Results: The deep learning model trained with face transformations categorized patient images into five asymmetry levels, achieving a weighted accuracy of 75% and a Pearson correlation of 0.31 with medical expert human evaluations. The model utilizing temporal misalignment achieved a weighted accuracy of 69% and a Pearson correlation of 0.27 for the same classification task.

Conclusions: We propose an automated approach for assessing lip asymmetry in patients with repaired cleft lip by transforming facial images of control subjects to train a deep learning model, eliminating manual anatomical landmarks. Our promising results provide a more efficient and objective tool for evaluating surgical outcomes.

Keywords

artificial intelligence, cleft lip and palate, facial cleft, facial esthetics, facial morphology, lip form, assessment.

Introduction

The human face generally exhibits a bilateral symmetry pattern, where the soft tissues and skeletal structures are aligned according to a symmetrical vertical plane. While perfect symmetry is often considered an ideal standard for facial esthetics and functionality, minor degrees of asymmetry are common and can develop during growth.¹⁻³ However, congenital conditions such as craniosynostosis, hemifacial microsomia, or *cleft lip and palate* (CL/P), as well as acquired conditions such as Bell's palsy, tumors, or stroke, can result in pronounced facial asymmetry.⁴ Such abnormal facial asymmetry detrimentally affects facial esthetics, psychological self-perception, and social interactions.⁵ It may also impair orofacial functions that require precise coordination.⁶

Estimating facial symmetry in patients with CL/P is a fundamental aspect of anatomical shape analysis, particularly for

evaluating the outcomes of surgical repair. This study addresses three primary limitations in facial symmetry analysis for patients with repaired cleft lip. First, traditional methods for assessing lip asymmetry rely exclusively on distances between anatomically oriented landmarks, focusing on shape alone and

¹ Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX, USA

² Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

³ Department of Plastic Surgery, University of Texas Southwestern Medical Center, Dallas, TX, USA

⁴ Analytical Imaging and Modeling Center, Children's Health, Dallas, TX, USA

Corresponding Author:

Karen Rosero, Language Technologies Institute, Office 571 I, 5000 Forbes Ave, Pittsburgh, PA 15213, USA.

Email: kroseroj@andrews.cmu.edu

neglecting texture and appearance. Our prior research shows that computational tools for automated landmark placement struggle with facial features that deviate from typical forms, such as asymmetric lip shapes.^{7,8} To address this problem, we propose using *convolutional neural networks* (CNNs) to analyze additional patterns in lip images—such as edges, color, and texture—thereby moving beyond purely measurement-based approaches that may be prone to inaccuracies.

The second motivation is to build the infrastructure for dynamic lip/articulatory symmetry assessment. To recognize the dynamic nature of the face, recent technological advancements have introduced 3D video capture systems that allow for more comprehensive assessments of symmetry using pseudo-landmarks generated from 3D meshes.^{3,9} However, the widespread clinical adoption of these advanced techniques is often limited by the high cost and limited availability of the necessary imaging systems.^{10,11} Therefore, our third motivation aligns with developing a machine learning algorithm to assess facial asymmetry using 2D images captured under standard conditions with any camera-equipped device. To enhance feasibility and scalability, we eliminated the need for manual landmark placement, which typically requires the specialized expertise of experienced surgeons. Our algorithm analyzes overall lip characteristics unlike traditional methods that measure distances between corresponding landmarks.

Our *deep learning* (DL) model performance was evaluated on 146 patient images exhibiting varying degrees of asymmetry. One medical practitioner and four individuals familiar with CL/P rated these images using the *Cleft Aesthetic Rating Scale* (CARS),¹² which assesses nasolabial symmetry based on features such as vermilion border continuity, scar tissue, and philtrum shortening. We report the correlation between our DL-based assessment and the human perceptual ratings obtained through CARS.

Methods

This study adhered to ethical standards and received approval from the *Institutional Review Board* (IRB) at The University of Texas at Dallas under protocol number IRB-24-829, with sponsorship from the University of Texas Southwestern Medical Center and the Children's Analytical Imaging and Modeling Center Research Program. Informed consent was obtained from all participants, and measures were implemented to ensure confidentiality and protect participant privacy throughout the study.

A key challenge in applying DL techniques to CL/P conditions is the requirement for large datasets, as DL models depend heavily on extensive data for effective training. A small number of images is insufficient to train a DL model with robust generalization to unseen images. Furthermore, using patient data to train *artificial intelligence* (AI) systems raises privacy concerns, which are currently a topic of ongoing debate.^{13,14} To address this problem, we utilized facial images from open datasets of individuals without

facial conditions. These images were transformed to introduce asymmetries between the left and right sides of the lips, mimicking those seen in our target population. We applied two specific transformations: one replicating the asymmetry observed in patients with unsuccessful CL/P surgeries (CLP transformation), and another introducing asymmetry by temporally misaligning frames extracted from videos of control subjects. These approaches overcome data limitations, enabling the development of a DL-based tool for assessing lip symmetry in patients with repaired CL/P.

This section describes every stage of the process of our lip symmetry assessment approach, including the image transformations, the intuition and operation of our DL model, and the perceptive evaluation to compare our results with human ratings. These techniques are applied to publicly available datasets containing facial images of individuals without reported or evident lip asymmetry.

Temporal Misalignment

The first strategy in this study is to introduce temporal misalignment to create dissimilarities between the left and right sides of the lips in subjects displaying natural symmetry. We randomly select two non-consecutive frames from a video where the subject reads a pre-defined sentence. Each frame is processed using the *dlib* toolkit¹⁵ for automatic facial landmark detection (indicated by blue dots in Figure 1). Since the subjects in the images do not have lip problems, the facial landmark algorithm performs well in most images. Lip landmarks guided the cropping of the *region of interest* (ROI), and central points relative to the cupid's bow are used to split the image into left and right sides. One side of each image is selected, resized to $224 \times 224 \times 3$ pixels, and paired with the corresponding side from a different frame to create misaligned or “negative” pairs. This incongruence, sometimes significant, helps the model learn to identify dissimilarities between the sides.

For comparison, our DL model also needs “positive” pairs to differentiate between aligned and misaligned pairs. The left and right images are obtained from a random video frame for positive pairs. These images are processed similarly, with face landmarks extracted and ROI cropped. The sides of the lips are split and resized to $224 \times 224 \times 3$ pixels, as shown in Figure 1b. The model was trained to recognize symmetric lip representations in these positive pairs, where no apparent dissimilarities were expected.

We utilized videos from the American speakers in the CRSS-4English-14 corpus,¹⁶ which consists of recordings of 105 subjects (55 females and 50 males) with no reported or evident facial conditions. These videos were collected by the Center for Robust Speech Systems at The University of Texas at Dallas. We selected five videos of each subject and processed them to extract video frames as 2D images. The selected videos contain read speech recordings with consistent prosody and frontal facial orientation. The linguistic content of these videos was chosen to ensure a broad representation of English phonemes, thereby capturing a diverse range of lip articulations.

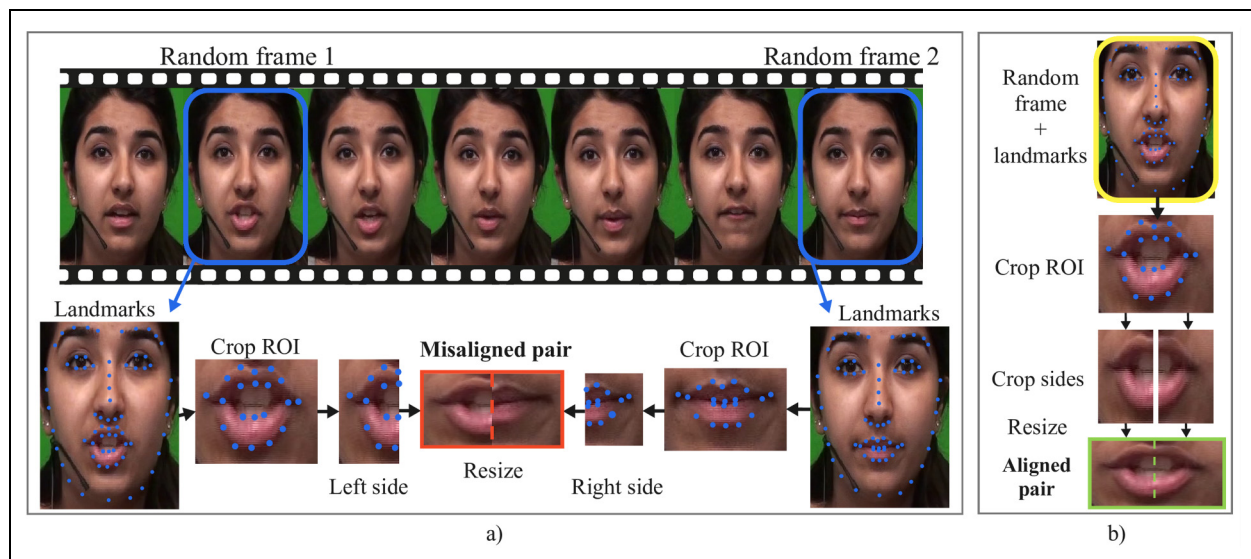


Figure 1. Image transformation based on temporal misalignment to introduce dissimilarity between left and right sides on lips of control subjects. The video frames shown in the figure were obtained from the CRSS-4English-14 corpus. (a) An example of a temporally misaligned pair, which combines the lip sides of two pairs obtained from the same frame. (b) An aligned or matched pair obtained from the same frame.

CLP Transformation

We adopt the technique proposed in our previous work by Rosero et al,⁷ in which images of control subjects are transformed to recreate the lip and nose configuration of patients with repaired CL/P. They defined seven transformations applied to the orofacial area, from which we selected the transformation for unilateral upper lip asymmetry due to its suitability to our lateral asymmetry assessment goal. The process starts by using the dlib toolkit as our landmark detector. Figure 2 illustrates the process of triangulating the control image into small regions based on facial landmarks. Specifically, areas around the lips are warped to match patients with residual unilateral asymmetry lip format. After the CLP transformation, the displaced lip points are used to crop the ROI on both the original control image and the transformed one. We also apply erosion filters to the ROI to remove small-scale noise from images.

We use frontal facial images from the *Chicago Face Dataset* (CFD)^{17–19} and the *Young Labeled Faces in the Wild* (YLFW) dataset²⁰ for this technique. The datasets contain images of subjects from diverse ethnicities, genders, and ages without evident or reported facial conditions. Licenses for research purposes are available. Our DL model learns dissimilar characteristics from asymmetric pairs obtained using CLP transformation.

For both techniques—temporal misalignment and CLP transformation—positive and negative image pairs are cropped using the same strategy. This strategy is based on lip landmarks, with the lip sides split according to points relative to the cupid's bow. It is important to note that the model learns from both modified and unmodified control images. We do not use data from individuals with CL/P. The images from our target population are used only to evaluate our model.

Deep Learning Model

DL is a subfield of AI built to gather knowledge by learning patterns directly from data.²¹ This work mainly relies on CNNs to build our model.

Image Processing Using Convolutional Neural Networks. Medical diagnosis supported by the analysis of images generated in clinical practice has gained attention due to advancements in medical imaging equipment and computational techniques, such as CNNs.^{8,22,23} Inspired by the operation of the visual cortex of mammals, artificial neurons in convolutional layers recognize image patterns, such as edges, colors, or texture. When grouped together, artificial neurons form trainable filters optimized to produce meaningful image representations. Convolutional layers extract information from small regions or *patches* all over the image, as shown in Figure 3a. By condensing the information of a small patch into a single number, these layers reduce the dimensionality of the input while keeping the most relevant information.

As several convolutional layers are appended, more complex features are extracted. Therefore, CNNs include several convolutional layers along with a few linear layers that combine the 2D characteristics into a 1D vector representation. Figure 3b illustrates our Siamese CNN model. This structure is called Siamese since it contains two identical branches that process each side of the input pair. The vector representation of each image can be used for various tasks of computer vision, such as image classification or segmentation. In our model, we will use it for symmetry estimation.

Several well-established CNN models have shown impressive results on computer vision tasks.^{24–26} We have considered the MobileNetV2 model²⁶ as the backbone for our CNN branches due to its light computational burden and high performance

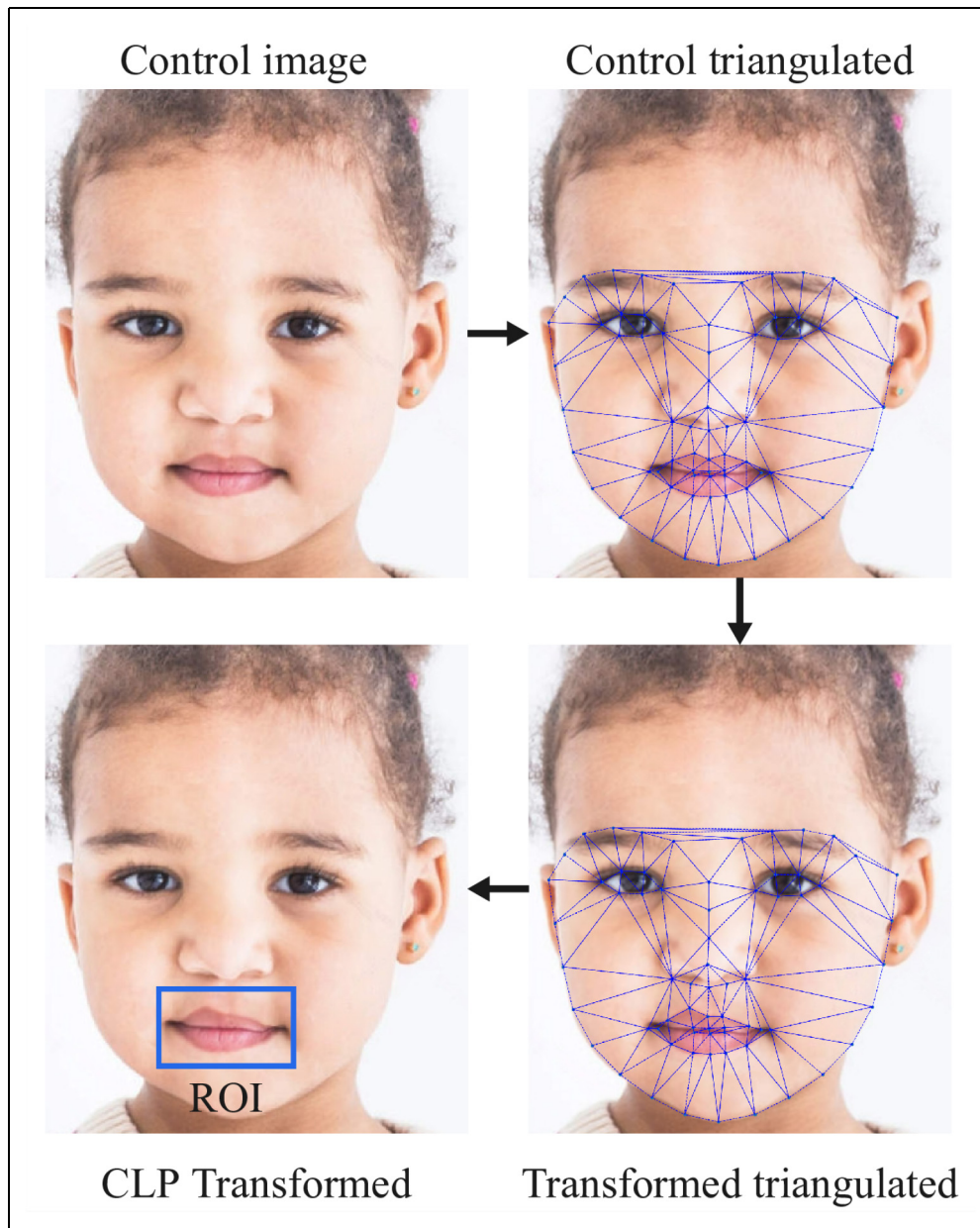


Figure 2. Stages of the CLP transformation, in which a control image of the Young Labeled Faces in the Wild (YLFW) dataset is triangulated to warp regions in the lip area to simulate unilateral CLP asymmetry.

across a variety of tasks. The study of Sandler et al²⁶ provides further technical details about the MobileNetV2 model.

Contrastive Learning. The parameters in our Siamese CNN are optimized for the training data with the guidance of a loss function. Our design adopts the contrastive loss, which encourages the model to learn closer representations for positive pairs while distant representations for negative pairs.²⁷ The contrastive loss is given by:

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^N (y_i \cdot D^2 + (1 - y_i) \cdot \max(0, \text{margin} - D)^2),$$

where D is the Euclidean distance between the lateral representations of a pair, guiding the quantification of lip symmetry. The binary label y_i indicates whether the pair is positive ($y_i = 1$) or negative ($y_i = 0$). The number of training pairs is denoted by N , and the margin parameter enforces the separation of representations of negative pairs from those of positive pairs.

As shown in Figure 3c, the Siamese CNN branches predict two 1D representations, which are subsequently used to compute D , aiding in the assessment of differences between left and right lip representations. Positive pairs are expected to have low D values, while negative pairs are expected to

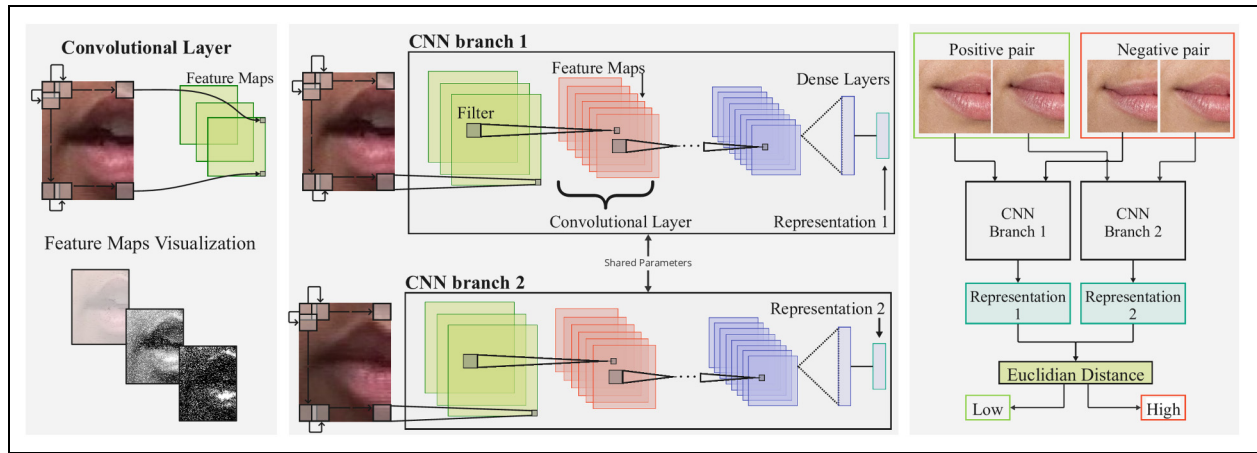


Figure 3. Stages of the Siamese CNN model, which is trained with positive and negative pairs obtained with the temporal misalignment and CLP transformations. (a) A single convolutional layer processing small patches all over the image. (b) The Siamese model contains two identical CNN branches that process each side of the input pair with a series of convolutional layers. Each branch learns a vector that represents the input images. (c) The Siamese model trained with positive and negative pairs, which result in high and low similarity.

have high values. The parameters of the Siamese CNN are optimized by minimizing the loss function $\mathcal{L}$ during training.

The Euclidean distance D is the quantitative measure of lip symmetry that guides the classification of pairs into positive and negative categories, with accuracy and precision metrics reported for this task. Moreover, D will be utilized to categorize images of patients with repaired CL/P based on the severity of lip asymmetry, as described in the Lip Symmetry Evaluation section.

Lip Symmetry Evaluation. We collected 146 frontal images of patients with repaired unilateral CL/P. These images were sourced from various online platforms that do not prohibit their use¹. These images are used solely to evaluate our system. They are neither used to train our Siamese CNN nor released or presented in this manuscript.

Lip symmetry was evaluated by human raters using the CARS,¹² a five-point photographic scale for evaluating nasolabial appearance in patients with repaired unilateral cleft lip. The scale originally analyzes nasal and lip characteristics of symmetry. However, for the scope of our study, we limited the analysis to lip-specific features: vermilion border continuity, scar tissue, and philtrum shortening. On this scale, a score of 1 represents a high symmetric appearance, while a score of 5 indicates a significant asymmetry. Only the ROI (the lips) was shown to the raters. One medical practitioner and four subjects familiar with the CL/P condition rated the images. To assess the consistency of human ratings of lip symmetry, we calculated the Fleiss' kappa, a statistical measure of inter-rater reliability. This metric quantifies the level of agreement among multiple raters, assigning categorical ratings to a set of images.²⁸

To correlate human-perceived symmetry with our system's predictions, we used the Euclidean distance D , calculated by our Siamese CNN, as a measure of lip symmetry. For each image, the trained Siamese CNN generates a D value,

normalized under the same conditions as the training set and rounded to the nearest category in CARS (1-5). We calculate the Pearson correlation between the CARS scores and the normalized Euclidean distances predicted by our model as an indicator of prediction association.

The metrics reported for these 146 images include the weighted accuracy and the weighted average precision. Accuracy is commonly used for multi-class classification tasks in DL. Precision, however, is more relevant for binary classification tasks such as detection. These tasks are common in the medical field, where controlling *false positives* (FP) is crucial. For this reason, we report the precision per class, considering FP for all wrongly classified samples. Then, we calculate the weighted average of the precision for each class, considering as FP all samples that are incorrectly classified. Subsequently, we calculate the weighted average of the precision across all classes. These weighted metrics also account for the severity of misclassifications. Specifically, the penalty for misclassification increases with the disparity between the predicted and actual classes. For example, the model is penalized more heavily for misclassifying an image from scale 1 as scale 5. In contrast, a misclassification of scale 2 as scale 1 incurs a smaller penalty, reflecting the similarities between images in adjacent classes.

Experiments

For both experiments, the DL model was trained using a subset of pairs, referred to as the “training” set. A separate group of “validation” pairs was used to select the optimal model configuration, while a different set of “testing” pairs was used to evaluate the system post-training. All these sets contain transformed or misaligned images of control subjects, and pairs were not shared across sets. Only the “CLP test” set contains images of patients with repaired CL/P.

Experiment 1: Siamese CNN With CLP Transformation. Images extracted from the CFD and YLFW datasets underwent CLP transformation and preprocessing steps, which involved cropping and splitting the images based on automatic facial landmarks. We used 2298 left-right lip pairs to train the Siamese CNN, 578 pairs for validation, and 510 pairs for testing, all obtained from control images. Each subset contained an equal proportion of negative and positive pairs. Both branches of the Siamese CNN were initialized with pre-trained parameters obtained from the ImageNet database.²⁹ We expect this starting model to be able to characterize primitive image patterns such as borders, edges, and textures. The initialized model was then further trained (fine-tuned) with the lip pairs to adapt the parameters for the specific task of symmetry assessment.

Experiment 2: Siamese CNN With Temporal Misalignment. For this experiment, we transform control images from the CRSS-4English-14 dataset to form negative pairs using the temporal misalignment technique. Extracting images from videos results in more training images than the CLP transformation experiment. We use 128,000 pairs for training, 29,000 for validation, and 31,000 for testing. As in the last experiment, we use an equal amount of positive and negative pairs from all subsets of data. We also use the same Siamese CNN model initialized with pre-trained parameters for object characteristic extraction and fine-tuned to perform the lip symmetry assessment.

Results

Siamese CNN With CLP Transformation

We report the model's accuracy and precision in differentiating between positive and negative pairs with the CLP transformation approach. Negative pairs underwent the CLP transformation, while positive pairs did not. The Siamese CNN model was optimized over 12 iterations of the training set, and the resulting metrics are presented in the row "CLP Transf." of Table 1.

During training, the model achieved an accuracy of 98.42% on this binary classification task, demonstrating its ability to automatically distinguish characteristics from both sides of

the lips and determine when they are dissimilar enough to be classified as a negative pair. The precision on the same training set was similarly high at 98.09%. A higher performance in the training set compared to the validation set is acceptable, as the model iteratively learns to perform the symmetry assessment task using the training data. During this process, the model adjusts its parameters to optimize performance on the validation set. Training stopped when validation metrics showed no improvement over three consecutive iterations, and the parameters of the best-performing model were retained. The characteristics learned from the training pairs enabled the model to achieve an accuracy of 85.35% and a precision of 86.51% on the validation set. While the validation set guided the training process and informed decisions about when to stop, the test set remained unseen throughout training. Consequently, the test set accuracy of 87.65% and precision of 86.67% are strong indicators of the model's ability to generalize and perform effectively on new, unseen samples.

While the metrics reported for lip symmetry as a binary task provide valuable insight into the model's training performance, it is not appropriate to classify images of patients with repaired CL/P strictly as positive or negative pairs. This is because not all patients in our dataset exhibit upper lip asymmetry, reflecting the variability observed in real-world outcomes. Therefore, for evaluating images of patients in the CLP test set, we use the Euclidean distance D predicted by the model as a measure of symmetry. Figure 4a provides a graphical representation of the predicted D values for all pairs in both the test set and the CLP test set. This visualization highlights how the model organizes these samples prior to classification. Notably, negative pairs of the test set (red dots) are dispersed toward higher D values, with a mean of 18.15, while positive pairs of the test set (green dots) are concentrated at lower D values, with a mean of 2.32. This distinct separation demonstrates that our model effectively differentiates between image representations, maximizing the distinction between positive and negative pairs. The distances D for the CLP test set are also shown in Figure 4a as blue dots. These predictions are distributed across a range of values corresponding to both positive and negative samples in the test set. This spread reflects the variability in surgical outcomes among patients with repaired CL/P, as not all outcomes result in upper lip asymmetry.

The Euclidean distance D serves as a quantitative measure for guiding the classification of CLP test set images into the five classes defined by the CARS scale, rather than the binary classification (positive or negative) applied to other data subsets that exclude patient images. The weighted accuracy and precision metrics for the CLP test set are reported in Table 1. When aligning the predicted values with the five asymmetry levels defined in CARS and rated by a medical practitioner, the Siamese CNN model achieved a weighted accuracy of 75.34% and a weighted precision of 63.53%. While the accuracy of 75.34% is lower than the binary classification accuracy, this decrease is expected due to the increased complexity of the 5-class task. The CARS scale captures subtle variations in lip symmetry among patients with repaired CL/P,

Table 1. Siamese CNN Accuracy (%) and Precision (%) to Differentiate Positive and Negative Pairs on the Training, Validation and Testing Sets Obtained by Transforming CFD and YLFW Databases.

Experiment	Metric	Training	Validation	Test	CLP test
CLP Transf.	Accuracy	98.42	85.35	87.65	75.34
	Precision	98.09	86.51	86.67	63.53
Temp. Misal.	Accuracy	97.12	91.57	92.62	69.41
	Precision	84.52	83.94	85.19	70.04

The weighted accuracy (%) and precision (%) to classify patient images into a five-point scale of asymmetry is reported on the CLP test set.

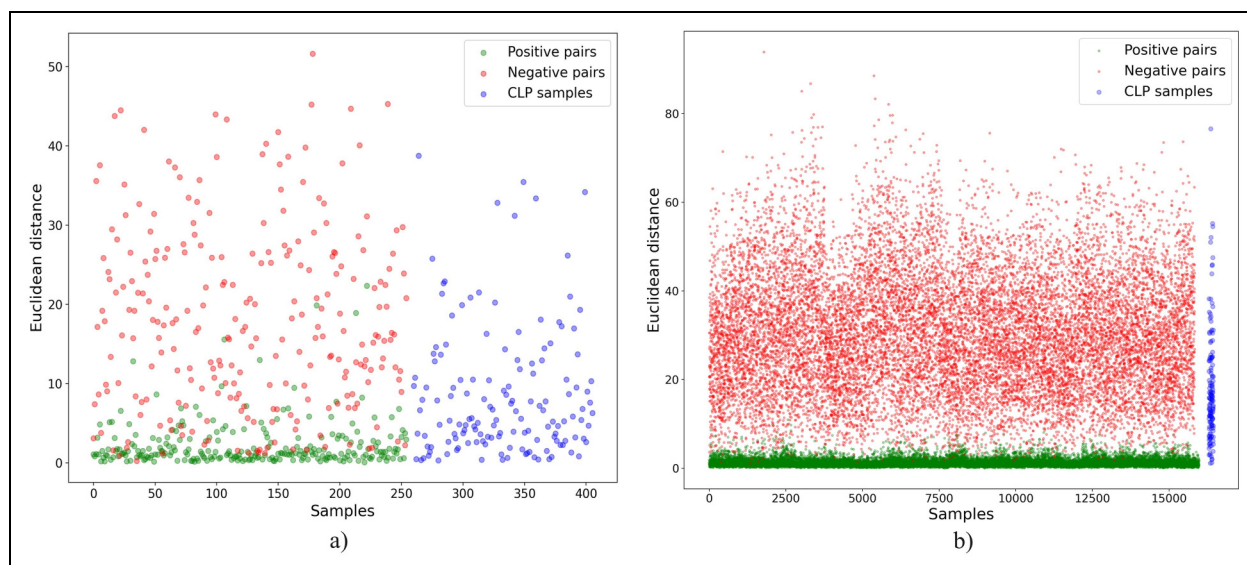


Figure 4. Visual representation of the Euclidean distance predicted by the Siamese CNN on the test sets of the two different image transformations and the CLP test set. (a) Siamese CNN model predictions using the CLP transformation technique. (b) Siamese CNN model predictions using the temporal misalignment technique.

making fine-grained distinctions necessary. Furthermore, an analysis of the dataset reveals that 88% of the images belong to classes with low severity (1, 2, and 3), which dominate the dataset. Conversely, classes 4 and 5, representing more severe cases, are underrepresented. This class imbalance adds another layer of difficulty to the classification task. Notably, a random guess in this scenario would achieve an accuracy of only 20%, highlighting the system's capability to identify meaningful patterns. The weighted precision of 63.53% reflects the increased likelihood of FP in a multi-class setting. However, this result remains competitive, given the inherent subjectivity of assessing lip symmetry.

We calculate the Pearson correlation between CARS human assessments and the Euclidean distances predicted by our model as an indicator of prediction association. Therefore, we normalize and map D values for the CLP test set to match the CARS scale, ranging from 1 to 5. The Pearson correlation is reported separately for ratings by a medical practitioner and by external evaluators familiar with the CL/P condition. The medical practitioner provides the surgeon's perspective, while non-medical subjects, representing the general population, offer judgments regarding facial features that deviate from normality at a first glance.

The Pearson correlation for the CLP test set is 0.37 when compared with human ratings by non-medical practitioners and 0.31 when compared to expert ratings. These values indicate an acceptable correlation with human ratings of lip symmetry. A related study on the automatic esthetic outcome assessment using a plane of symmetry for patients with repaired CL/P reported a Pearson correlation of 0.236 between predicted scores and human ratings.³⁰ Furthermore, the challenge of assessing lip symmetry in patients who have undergone cleft lip repair surgery is reflected in the inter-rater

agreement. For the 146 frontal images of patients with repaired unilateral CL/P, the Fleiss' kappa was 0.25, indicating fair agreement and underscoring the subjectivity of the task. Similar findings were observed in the study of Bakaki et al,³⁰ where human-assessed esthetic scores achieved a correlation of 0.457.

We anticipate that increasing the number of medical practitioners participating in the annotation process, rather than including non-medical evaluators familiar with the CL/P condition, could enhance inter-rater agreement in future studies. This improvement would likely provide more consistent and clinically relevant benchmarks for evaluating lip symmetry.

Siamese CNN With Temporal Misalignment

We report the model's accuracy and precision in differentiating between positive and negative pairs with the temporal misalignment approach. The Siamese CNN model was optimized over 15 iterations of the training set, and the resulting metrics are shown in Table 1 for the temporal misalignment experiment listed in row "*Temp. Misal.*"

Similar to the CLP transformation experiment, we obtain high accuracy on the training and validation subsets (ie, 97.12% and 91.57%, respectively), and an acceptable reduction of precision in the same subsets (ie, 84.52% and 83.94%) due to the presence of FP. The model's accuracy in differentiating between positive and negative pairs on unseen pairs of the testing set reached a value of 92.62%, while the precision reached a value of 85.19%. Analyzing the metrics for the training, validation, and testing sets on the binary task of classifying positive and negative pairs, we identified a slightly superior accuracy using the temporal misalignment approach with a slightly inferior but still competitive precision

to perform the task. However, the decision for the best approach should consider the CLP test set.

Figure 4b presents the predictions D for positive and negative pairs obtained with temporal misalignment in the test set. With more samples in the plot, the Siamese CNN trained with temporal misalignment also spreads out representations of negative pairs towards higher D values with a mean of 30.31, while keeping representations of positive pairs in lower D values with a mean of 1.35. Notably, the clusters of this experiment are a bit more disentangled than the CLP transformation experiment, which explains its superior accuracy of 92.62% on the testing set.

As in the CLP transformation experiment, we map the obtained predictions to the CARS scores ranging from 1 to 5 to evaluate the prediction association with human ratings. The Pearson correlation with expert medical ratings was 0.29, while the correlation with non-medical subjects was 0.27. The model predictions and the medical practitioner ratings were used to compute the weighted accuracy and precision, which resulted in a value of 69.41% and 62.20%, respectively. The reduced correlation and weighted metrics observed with the temporal misalignment transformation on the CLP test set can be attributed to the different nature of the transformations. While the CLP transformation specifically modifies the upper lip contour, the temporal misalignment approach introduces variations across the entire lip configuration, as illustrated in Figure 1a. The CARS scale, which evaluates the symmetry and continuity of the vermilion border, naturally aligns more closely with the CLP transformation results, given their shared focus on upper lip asymmetries. In contrast, the model trained with temporal misalignment learned to detect differences throughout the entire lip region, not just the upper lip. Moreover, the CLP transformation was trained on images predominantly depicting neutral facial expressions without teeth exposure. In contrast, the images used for the temporal misalignment approach were extracted from videos in which subjects were speaking, resulting in non-neutral lip configurations and visible teeth. These factors could influence the model's decision to classify an image pair as dissimilar for differences not only in the upper lip.

Discussion

This work presented a strategy to assess lip symmetry in patients with repaired CL/P without relying on landmarks for measurements. We demonstrate the feasibility of transforming facial images of control subjects to replicate the residual lip asymmetry seen in cleft lip repair surgery. The model trained with the CLP transformation technique better correlated with human ratings than the temporal misalignment technique. Therefore, our future work will focus on refining the CLP transformation approach to enhance its accuracy and applicability. Nevertheless, the temporal misalignment technique may be promising for other conditions affecting facial symmetry, particularly those that alter the overall lip configuration rather than the lip shape itself, such as facial palsy, facial paralysis, or Ramsay Hunt syndrome.

In parallel with our approach, advancements in image processing and computer vision have led to the development of automatic face landmark detectors for the general population.^{15,31–33} Cortes et al³⁴ used 2D images of CL/P patients collected from the internet and processed them with an automatic landmarking tool¹⁵ to rank images into three severity levels. However, a study by Rosero et al⁷ showed that these tools often fail on faces with anatomical deviations from the norm, as they were trained on faces without significant anomalies. Consequently, the automatic placement of landmarks in patients with repaired or unrepaired CL/P is often inaccurate, potentially leading to errors in calculating lateral symmetry based solely on these measurements. While these landmarks can be used to define a ROI—in this case, the lips—we propose an alternative approach that analyzes the left and right sides of the lips to provide an asymmetry metric based on image analysis, rather than relying exclusively on landmark-based measurement.

Our strategy enables the training of DL models using transformed images, addressing the limitations of previous approaches. Traditional machine learning models, often trained with fewer than a hundred images, were less complex and unable to extract highly intricate features from images. This limitation increased the risk of the models memorizing the training samples, leading to poor generalization to unseen images during evaluation. Our approach overcomes this challenge with an effective strategy that leverages large datasets and sophisticated image transformation techniques, allowing the models to learn complex tasks with greater accuracy and applicability.

Wu et al³⁵ and Al-Rudainy et al³⁶ compared features computed from image patches rather than using landmark distances, to estimate unrepaired cleft lip severity. However, their approaches relied on 3D stereophotogrammetry mesh to conduct a patch-based analysis. In contrast, our system evaluates 2D images of patients using a pre-trained model that can be deployed on-site in various clinics without relying on expensive equipment. Another advantage of this methodology is the protection of patient privacy. Our model does not use images of patients to learn the task. Instead, it learns from transformed images extracted from open datasets of subjects without conditions affecting the face. Therefore, patients do not need to consent to sharing their facial images, as their data will be protected by the medical center's data system.

Evaluating surgical outcomes, focusing on symmetry, is fundamental for advancing cleft care. Surgical success is judged by more than restoring function. Facial esthetics and symmetry are key determinants of satisfaction for patients and their families. As technology evolves, the integration of precise and objective assessment tools will further empower surgeons to refine their techniques and identify areas for improvement. The symmetry and continuity of the upper lip, Cupid's bow, and vermilion alignment are critical parameters for comparing postoperative outcomes of various cheiloplasty techniques, including Pfeifer's, Millard's, and Tennison's methods.^{37,38} However, traditional symmetry assessments are often time-consuming, subjective, and influenced by unconscious bias. In contrast, objective metrics based on image

features analysis, such as the symmetry estimation utilized in this study, allow for precise and reproducible outcomes analysis, contributing to evidence-based practice. Therefore, developing standardized protocols for assessing repairs may provide objective comparability across different lip repair techniques.

Moreover, our system evaluates 2D images of patients using a pre-trained model that can be deployed on-site in various clinics without sharing patient data or uploading it to the cloud. This is feasible because the system can run on local devices. The architecture of our model is lightweight and can function on devices with lower computational power than those used for training. Consequently, once fully trained, the model can be adapted to run on smartphones. We envision our system as a complementary tool for medical practitioners, enabling the estimation of asymmetry without requiring manual landmark placement or distance measurements between corresponding facial points.

We selected frontal images with neutral facial expressions for our CLP test set. However, our images exhibit varying lighting conditions, slight head rotations, and different camera resolutions. Standardizing a setup for collecting patient images to test our models would likely improve the performance of our metrics on unseen data. We are also committed to exploring improved transformation techniques to continue using images of control subjects instead of patient images for training DL models. Image generation techniques, such as diffusion models,³⁹ have demonstrated their potential to produce realistic facial photos. They could be considered for future work targeting the generation of faces displaying lip asymmetry due to residual scarring.

While our study demonstrates the potential of automatic tools for lip symmetry assessment, especially using the CLP transformation approach, a few limitations must be acknowledged. First, both approaches developed in this work require frontal facial images, as lateral head rotations adversely impact symmetry prediction. The datasets used for both static images and video analysis consist exclusively of frontal faces without occlusions, which is a prerequisite for our system's operation. Furthermore, the system's performance currently relies on non-patient data. To simulate the facial features of our target population, we transformed publicly available datasets; however, this approach may not fully capture the variability present in real patient data. We anticipate that incorporating anonymized patient images into the training process will improve further the system's accuracy and generalizability.

Conclusions

This study presents a DL approach for assessing lip symmetry in patients with repaired CL/P. The CLP transformation, which simulates unilateral lip asymmetry, demonstrated a stronger correlation with human ratings and achieved a weighted categorization accuracy of 75.34% in assessing lip symmetry. The temporal misalignment approach, while less effective, still provided valuable insights. Our models were trained on

transformed control images, preserving patient privacy and allowing deployment in clinical settings through a system designed to run locally on devices, including smartphones, without relying on cloud-based data processing.

Acknowledgments

The authors 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.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval and Informed Consent Statements

Data collection for this study was approved by the UTD Institutional Review Board in IRB-24-829, and all the rater subjects freely consented to participate in the study.

Funding

This study was funded by the University of Texas Southwestern Medical Center and the Children's Analytical Imaging and Modeling Center Research Program through a UTSW Interagency Cooperation Contract.

ORCID iDs

Karen Rosero  <https://orcid.org/0000-0002-8118-4213>

Lucas M. Harrison  <https://orcid.org/0000-0002-3168-1879>

Rami R. Hallac  <https://orcid.org/0000-0001-9025-399X>

Note

1. <https://www.craniofacialteamtexas.com/>;
<https://www.elegantplasticsurgery.com/>;
<https://facesurgeon.in/>;
<https://www.hdofs.com/>;
<https://helpusgivesmiles.org/>;
<https://www.lwbkids.org/>;
<https://lawplasticsurgery.com/>;
<https://www.sargentercraniofacial.com/>;
<http://www.nmcleft.org/>

References

1. Ferrario VF, Sforza C, Poggio CE, et al. Distance from symmetry: a three-dimensional evaluation of facial asymmetry. *JOMS*. 1994;52(11):1126-1132. doi:10.1016/0278-2391(94)90528-2
2. Nkenke E, Benz M, Maier T, et al. Relative en- and exophthalmometry in zygomatic fractures comparing optical non-contact, non-ionizing 3D imaging to the Hertel instrument and computed tomography. *JOMS*. 2003;31(6):1010-5182. doi:10.1016/j.jcms.2003.07.001
3. Nkenke E, Lehner B, Kramer M, et al. Determination of facial symmetry in unilateral cleft lip and palate patients from three-dimensional data: technical report and assessment of measurement errors. *Cleft Palate-Craniofacial J*. 2006;43(2):129-137. doi:10.1597/04-138.1

4. Chojdak-Lukasiewicz J, Paradowski B. Facial asymmetry: a narrative review of the most common neurological causes. *Symmetry (Basel)*. 2022;14(4):737. doi:10.3390/sym14040737
5. Tobiasen JM, Hiebert JM. Combined effects of severity of cleft impairment and facial attractiveness on social perception: an experimental study. *Cleft Palate-Craniofacial J*. 1993;30(1):82-86. doi:10.1597/1545-1569_1993_030_0082_ceosoc_2.3.co_2
6. de Souza Freitas JA, das Neves LT, de Almeida ALPF, et al. Rehabilitative treatment of cleft lip and palate: experience of the Hospital for Rehabilitation of Craniofacial Anomalies/USP (HRAC/USP)-part 1: overall aspects. *J Appl Oral Sci*. 2012;20(1):9-15. doi:10.1590/S1678-77572012000100003
7. Rosero K, Salman AN, Sisman B, et al. Enhanced facial landmarks detection for patients with repaired cleft lip and palate. *Paper presented at 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*. May 27–31, 2024; Istanbul, Turkey. Accessed July 12, 2024. <https://ieeexplore.ieee.org/abstract/document/10582022>.
8. Rosero K, Salman AN, Hallac RR, et al. Lip abnormality detection for patients with repaired cleft lip and palate: a lip normalization approach. *Paper presented at 26th International Conference on Multimodal Interaction (ICMI 2024)*. November 4–8, 2024; San José, Costa Rica. Accessed November 25, 2024. <https://dl.acm.org/doi/abs/10.1145/3678957.3685726>.
9. Wu J, Heike C, Birgfeld C, et al. Measuring symmetry in children with unrepaired cleft lip: defining a standard for the three-dimensional midfacial reference plane. *Cleft Palate-Craniofac J*. 2016;53(6):695-704. doi:10.1597/15-053
10. Hallac RR, Feng J, Kane AA, et al. Dynamic facial asymmetry in patients with repaired cleft lip using 4D imaging (video stereophotogrammetry). *J Craniomaxillofac Surg*. 2017;45(1):8-12. doi:10.1016/j.jcms.2016.11.005
11. Gattani S, Ju X, Gillgrass T, et al. An innovative assessment of the dynamics of facial movements in surgically managed unilateral cleft lip and palate using 4D imaging. *Cleft Palate-Craniofac J*. 2020;57(9):1125-1133. doi:10.1177/1055665620924871
12. Mosmuller DGM, Mennes LM, Prah C, et al. The development of the cleft aesthetic rating scale: a new rating scale for the assessment of nasolabial appearance in complete unilateral cleft lip and palate patients. *Cleft Palate-Craniofac J*. 2017;54(5):555-561. doi:10.1597/15-274
13. Lotan E, Tschider C, Sodickson DK, et al. Medical imaging and privacy in the era of artificial intelligence: myth, fallacy, and the future. *J Am Coll Radiol*. 2020;17(9):1159-1162. doi:10.1016/j.jacr.2020.04.007
14. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics*. 2021;22(1):122. doi:10.1186/s12910-021-00687-3
15. King DE. Dlib-ml: a machine learning toolkit. *JMLR*. 2009;10:1755-1758. doi:10.5555/1577069.1755843
16. Tao F, Busso C. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Speech Commun*. 2019;113:25-35. doi:10.1007/s11633-019-1175
17. Ma DS, Correll J, Wittenbrink B. The Chicago face database: a free stimulus set of faces and norming data. *Behav Res Methods*. 2015;47(4):1122-1135. doi:10.3758/s13428-014-0532-5
18. Ma DS, Kantner J, Wittenbrink B. Chicago face database: multi-racial expansion. *Behav Res Methods*. 2021;53(3):1289-1300. doi:10.3758/s13428-020-01482-5
19. Lakshmi A, Wittenbrink B, Correll J, et al. The India face set: international and cultural boundaries impact face impressions and perceptions of category membership. *Front Psychol*. 2021;12:627678. doi:10.3389/fpsyg.2021.627678
20. Medvedev I, Shadmam F, Gonçalves N. Young Labeled Faces in the Wild (YLFW): a dataset for children faces recognition. *Paper presented at 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*. May 27–31, 2024; Istanbul, Turkey. Accessed July 13, 2024. <https://ieeexplore.ieee.org/document/10582021>.
21. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. 1st ed. MIT Press; 2016.
22. Anwar SM, Majid M, Qayyum A, et al. Medical image analysis using convolutional neural networks: a review. *J Med Syst*. 2018;42:1-13. doi:10.1007/s10916-018-1088-1
23. Rosero K, Salman AN, Busso C, et al. A tailored machine learning approach for cleft lip symmetry analysis. *Abstract presented at the American Cleft Palate Craniofacial Association Annual Meeting (ACPA 2024)*; April 10–13, 2024; Denver, CO.
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *Paper presented at International Conference on Learning Representations (ICLR 2015)*. May 7–9, 2015; San Diego, CA. Accessed April 15, 2024. <https://arxiv.org/abs/1409.1556>.
25. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Paper presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. June 7–12, 2015; Boston, MA. Accessed March 1, 2024. <https://ieeexplore.ieee.org/document/7298594>.
26. Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks. *Paper presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. June 18–23, 2018; Salt Lake City, UT. Accessed March 23, 2024. <https://ieeexplore.ieee.org/document/8578572>.
27. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In *Paper presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*. June 20–25, 2005; San Diego, CA. Accessed March 21, 2024. <https://ieeexplore.ieee.org/document/1467314>.
28. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378-382. doi:10.1037/h0031619
29. Deng JM, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. *Paper presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*. June 20–25, 2009; Miami, FL. Accessed June 16, 2024. <https://ieeexplore.ieee.org/document/5206848>.
30. Bakaki P, Richard B, Pereira E, et al. Key landmarks detection of cleft lip-repaired partially occluded facial images for aesthetics outcome assessment. *Paper presented at International Conference on Image Analysis and Processing*. May 23–27, 2022; Lecce, Italy. Accessed November 18, 2024. https://link.springer.com/chapter/10.1007/978-3-031-06430-2_60.
31. Baltrusaitis T, Robinson P, Morency LP. OpenFace: an open source facial behavior analysis toolkit. *Paper presented at IEEE Winter Conference on Applications of Computer Vision (WACV 2016)*. March 7–10, 2016; Lake Placid, NY. Accessed January 16, 2024. <https://ieeexplore.ieee.org/document/7477553>.
32. Bulat A, Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). *Paper presented at IEEE International Conference on*

- Computer Vision (ICCV 2017)*; October 22–29, 2017; Venice, Italy. Accessed December 5, 2023. <https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.116>.
33. Kartynnik Y, Ablavatski A, Grishchenko I, et al. Real-time facial surface geometry from monocular video on mobile GPUs. *Paper presented at CVPR Workshop on Computer Vision for Augmented and Virtual Reality (CVPR 2019)*. June 17, 2019; Long Beach, CA. Accessed November 8, 2023. <https://research.google/pubs/real-time-facial-surface-geometry-from-monocular-video-on-mobile-gpus/>.
 34. Cortés G, Villalobos F, Flores M. Asymmetry level in cleft lip children using dendrite morphological neural network. *Paper presented at 11th Mexican Conference on Pattern Recognition (MCPR 2019)*. June 29, 2019; Querétaro, Mexico. Accessed January 8, 2024. https://link.springer.com/chapter/10.1007/978-3-030-21077-9_22#citeas.
 35. Wu J, Tse R, Shapiro LG. Learning to rank the severity of unrepaired cleft lip nasal deformity on 3D mesh data. *Paper presented at 22nd IEEE International Conference on Pattern Recognition (ICPR 2014)*. August 24–28, 2014; Stockholm, Sweden. Accessed February 10, 2024. <https://ieeexplore.ieee.org/document/6976799>.
 36. Al-Rudainy D, Ju X, Stanton S, et al. Assessment of regional asymmetry of the face before and after surgical correction of unilateral cleft lip. *J Craniomaxillofac Surg*. 2018;46(6):974-978. doi:10.1016/j.jcms.2018.03.023
 37. Baek RM, Myung Y, Park I, et al. A new all-purpose bilateral cleft lip repair: bilateral cheiloplasty suitable for most conditions. *J Plast Reconstr Aesthet Surg*. 2018;71(4):537-545. doi: 10.1016/j.bjps.2017.09.016
 38. Kumar RVK, Reddy YS. Cheiloplasty by Pfeifer's technique. *J Cleft Lip Palate Craniofacial Anomalies*. 2017;4(Suppl 1): S113-S117. doi: 10.4103/jclpca.jclpca_59_17
 39. Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. *Paper presented at IEEE International Conference on Computer Vision (CVF 2023)*. October 1–6, 2023; Paris, France. Accessed February 18, 2024. <https://ieeexplore.ieee.org/document/10377881>.