Automated Landmark Detection for AR Based Craniofacial Surgical Assistance System

Sanghyun Byun[1], Muhammad Twaha Ibrahim[1], M. Gopi[1], Aditi Majumder[1]
Lohrasb R. Sayadi[2], Usama S. Hamdan[3], and Raj M. Vyas[2]

[1] Department of Computer Science, University of California - Irvine, Irvine, USA
[2] Department of Plastic Surgery, University of California - Irvine, Irvine, USA
[3] Global Smile Foundation

Abstract. The shape of the face of cleft lip patients varies significantly from a regular face due to the unique form and differing levels of severity of their condition. The first step in cleft lip repair requires surgeons to mark anthropometric landmarks that are used as a guide to conduct surgical incisions. These landmarks are different from the ones that are deemed important in a regular face and cannot be detected by existing facial landmark detection frameworks.

We propose a AI/ML based assistive tool that can automatically mark the anthropometric landmarks for cleft repair on the image of the cleft lip patient. We use a novel method for training a convolutional neural network that detects the anthropometric landmarks for patients with cleft lip without requiring a large number of images for training. By utilizing image ROI (region of interest) warp and direct regression, the proposed approach is able to accurately detect landmarks despite variation in the appearance of the cleft. Further, we show the significant improvement ROI warp has on the prediction of anthropometric landmarks used for cleft surgeries. We collaborate closely with reputed craniofacial surgeons to build our training datasets and validate the accuracy of our automated markings.

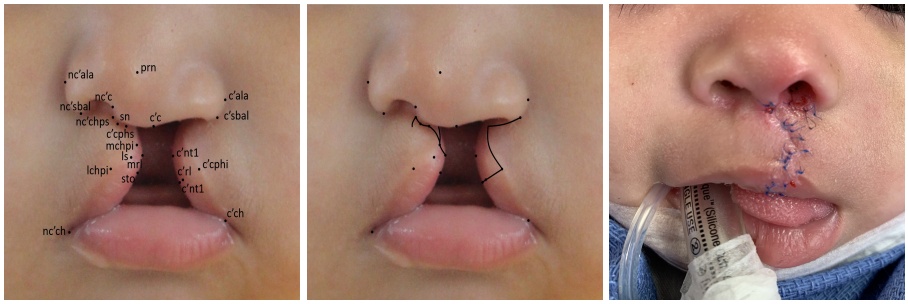
This tool is anticipated to have a tremendous impact on building surgical capacity for cleft repair surgeries, which has a huge shortage, in particular in rural areas, especially in emerging global areas of South America, Africa, and India.

Keywords: Convolutional Neural Network · Facial Landmark Detection · Keypoint Detection · Cleft Surgery · ROI Warp.

1 Introduction

Every year, around 195,000 babies globally are born with oral or facial clefts. 4.62 million people in the world today are living with an unrepaired cleft, which increases their chances of suffering from life-threatening problems like malnutrition, or death due to choking, by 2.15 times. Other life-impacting effects include speech impediment, deafness, malocclusion, gross facial deformity, and severe

psychological problems [30]. Therefore, children born with a cleft lip must undergo reconstructive surgeries that aim to stitch the lip and palette together. The first reconstructive surgery must happen before the 18th month when the facial tissue is soft, malleable, and therefore, more amenable to repair. As the child grows older and the face shape changes with age, they may require follow-up corrective surgeries up until the age of 14-15 years [12].



(a) cleft anthropometric landmark (b) Millard repair (c) post-surgery

Fig. 1: 21 cleft anthropometric landmarks from anterior view (a), surgery guideline used for Millard repair (b), and post surgery incision marks with blue stitches outlining the incision (c). *c'* stands for cleft side and *nc'* stands for non-cleft side.

Cleft surgery is one of the most challenging surgeries. Most of the time, surgery is performed on infants 3-6 months old where the surgical area is less than 4×4 cm. in size. Cleft lip deformations come in several levels of complexities and severities. Around 85% of surgeries utilize the rotation technique, named so due to the skin flap moving in a curved path during surgery. Successful planning and execution of the different kinds of incisions (e.g. Millard, Tennison-Randall, Mulliken) all start with a precision marking of 21 anatomical points of reference [26], called anthropometric landmarks, that are used to plan the incision (see Figure-1). Surgeons use these points to make measurements and guide the cleft repair surgery. They mark incisions using these points, and then during the surgery, they try to align the landmarks on either side of the cleft to reconstruct a balanced, symmetric, and aesthetically pleasing lip. Figure-1 show an example of techniques used.

Accurate markings of these points directly impact the quality and accuracy of the repair, which in turn determines the number of corrective surgeries required in the future. Cleft surgery is very sensitive to anthropometric landmarks. Incorrect marking can lead to an asymmetric reconstructed lip that requires further corrective surgeries, thereby increasing costs and discomfort to the patient. Despite years of experience, surgeons put in tremendous effort and time to mark

the 21 keypoints precisely due to their short-term and long-term impact on the surgical outcome. Getting to a level of accuracy that assures good outcome takes a lot of repetitive practice and is a skill built by the surgeon with years of experience.

Imagine an assistive tool that can collect data from expert surgeons during surgical planning and use an AI/ML-based method to predict the 21 anthropometric landmarks automatically. This assistive tool can be used in the following scenarios:

- Practicing surgeons can use this tool for skill building. Perform repetitive hands-on training on a large database of images even when they do not have access to an in-person trainer with direct feedback. They can even practice on a 3D mannequin, and the markings on its pictures can provide feedback on the accuracy they achieved.
- AR-based expert surgical assistance is becoming common in remote areas with low surgical capacity (see Figure-2). When a novice surgeon is performing a complex surgery with expert guidance using an AR application on a tablet, the predicted markings provide a great starting point. The accuracy of these predicted points provides a great foundation and reference for the remote expert to guide the surgeon. We are working closely with Dr. Raj Vyas and Dr. Ross Sayadi of the University of California, Irvine (UCI) and Children's Hospital of Orange County (CHOC) to provide such an assistive tool as part of a bigger effort on novel AR systems for remote surgical guidance [27, 29].

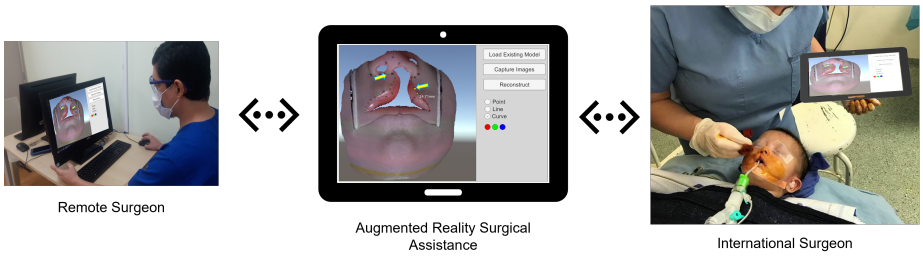


Fig. 2: Example of a remote surgery done through augmented reality proctoring on a tablet.

1.1 Our Contributions

In this paper, we propose a novel method that uses convolutional neural networks (CNNs) to develop an assistive tool that automatically detects and labels the

anthropometric landmarks used in cleft surgeries. The input to our tool is a database of images of cleft lip patients with anthropometric marks. Such data can easily be collected during the planning phase of any cleft repair surgery. Our main contributions are as follows:

- An end-to-end network for accurately detecting all 21 anthropometric landmarks in an image of a face with cleft lip deformity. We make use of transfer learning to avoid building an entirely new solution for the cleft lip. Instead, we leverage the large amount of work done on detecting general facial features and the large training datasets thereof. This allows us to achieve high-accuracy detection very efficiently with training datasets that are an order of magnitude smaller in size.
- We show that warping the input image to a pre-computed template face and extracting the predicted landmark coordinates from this ROI (region of interest)-warped space significantly improves precision.
- We also build an efficient GUI that can help surgeons mark the 21 anthropometric landmarks on a cleft lip image which is subsequently used for training.

Both the AI/ML-based automatic landmark prediction method and the GUI can be adapted for different kinds of surgeries in the future. We find, talking to our surgeon collaborators, that almost all surgeries involve such marking of key landmarks before planning the incisions.

2 Related Work

Recent advances in artificial intelligence and machine learning, especially in the subfield of face detection and recognition, can be used to assist surgeons with the marking process by automatically detecting the keypoints from a single capture of a face with a cleft lip. However, current state-of-the-art methods would not be able to detect landmarks accurately. This is because a cleft face has unique features that are not present in normal faces (see Figure-4). This makes it harder for convolutional models to extract correct feature maps. However, they can be leveraged to help us in detecting landmark points for faces with cleft lip deformities.

2.1 Regression Based Facial Landmark Detection

Regression-based methods are a more traditional approach to landmark detection and can be further divided into *direct regression* and *cascaded regression* models. Facial landmark detection through *direct regression* [3, 32, 33, 35, 38, 41] predict landmarks by passing an image through a backbone, then putting the output into a fully-connected network to predict the coordinates. Here, the backbone network can be any network (e.g. ResNet or HRNet) that can extract necessary features from a given image to predict the coordinates. *Cascaded regression* models [18, 21, 23, 39, 40, 43] take regression models a step further by using coarse-to-fine methods [43] to predict coordinates. These models use the

previously detected landmarks to iteratively update the coordinates, generating predicted landmark coordinates after a certain number of iterations. Though cascaded regression models tend to be more effective as they use iterative steps for prediction, they need a large training set to attain prior knowledge of predefined face shapes.

2.2 Heatmap Based Facial Landmark Detection

Inspired by fully-convolutional networks [20], heatmap-based methods aim to predict landmark coordinates without the use of a fully-connected layer. Therefore, most heatmap facial landmark detection models use coordinated prediction to generate a semantic map i.e. a heatmap that has Gaussian distributions around the predicted landmark coordinates [2,4,16,19,25,37]. Softmax function is then used to force the sum of elements to one. The coordinate is then extracted by the argmax function. As information gathered from a local view of an image does not give a full understanding of the image, Wei et al. [31] proposed an alternative *convolutional pose model* where the heatmap is generated in stages, slowly increasing the effective receptive field, an area of the image that the regressor focuses on. Although designed to be used for human pose estimation, it is often altered and trained to detect facial landmarks as well. When compared to regression models, heatmap facial landmark detection often shows superior results in terms of accuracy. However, as heatmaps are highly vulnerable to correlated features, in cases where there are only a few training images, heatmap-based methods will often exhibit noise, reducing the robustness and making it harder to assess the usefulness of the model.

2.3 One-Hot Facial Landmark Detection

Region-based Convolutional Neural Network (R-CNN) [11], a network built to detect objects in an image through region proposals, is also used for facial landmark detection. An initial CNN is used to extract rectangular region proposals, which are then used to classify each region using a separate classifier. To enhance this method to an even finer level of detection and classification, *Mask R-CNN* [13] was developed on top of Faster R-CNN [24]. This replaces the classification head with a region-of-interest pooling, reducing the overhead as well as allowing the model to give pixel-wise classification of the image.

2.4 Other Techniques for Facial Landmark Detection

Although most prior works put a heavy focus on the model architecture for the performance of their methods, other aspects of learning such as *loss function* and *image pre-processing* can affect the final results. While not studied as much, there are a few cases where such methods have shown to have a significant impact on the results. As all training sequences aim to reduce the value of their loss functions [7,22,34], choosing the right function is crucial to any machine learning

model. In the domain of facial landmark detection using regression techniques, Feng et al. [7] proposed a wing-shaped loss function to improve the accuracy of facial detection models by increasing the impact of small to medium errors using a combination of linear and non-linear parts, while Fard et al. [22] proposed an adaptive loss function using a difference between predicted landmarks and ground truth. In heatmap-based techniques, Yan et al. [34] calculates the difference between two probability distributions using Wasserstein distance to output its value. In the domain of image pre-processing, Zhao et al. [42] uses a correction network in the image pre-processing stage to enhance the result of the detection network. It corrects an image that has been warped by a fish-eye lens (e.g. the ones used with doorbell cameras for wider view angle). This is done by using two networks in sequence. Correction networks predict coefficients for their radial transformation equation, and alignment networks generate a projective transformation matrix.

2.5 Medical Domain

Facial landmark detection is used widely in the medical field, especially in the field of plastic surgery, to analyze facial structures before suggesting treatment plans [5, 9, 10]. Freitas et al. [9] proposed a facial feature detection network extracting facial contour, contour simplification, and point localization from a side face profile image to be used for general facial plastic surgery (e.g. reconstructive nose surgery). AI/ML based landmark prediction for landmark detection in cleft-lip faces was suggested first by our collaborator Sayadi et al. [28] in a medical journal which is developed into a comprehensive method and system in this work.

3 Method

Our method seeks to detect 21 anthropometric cleft-lip landmarks (CLL) that surgeons use to guide the surgery (see Figure-1). These landmarks are located around the cleft and are unique to the cleft side, denoted by the prefix *c'*, as well as the non-cleft side, denoted by the prefix *nc'*. Five of these 21 landmarks that are least subjective to the surgeon preferences are: (a) *prn*: the tip of the nose; (b) *c'ala, nc'ala*: the wings of the nostrils on the cleft and non-cleft side, respectively; and (c) *c'ch, nc'ch*: the junction of the upper and lower lips on the cleft and non-cleft side, respectively.

Our method uses a network that consists of four main components: (a) detection of landmark points in the face outline of the input image to cut out a region of interest (ROI) that focuses on the cleft lip deformity, (b) creating a rectangular input image from just the ROI via a warp-and-crop; (c) detection of cleft landmarks in the warped and cropped image, and finally, (d) inverse cropping and warping on the detected landmarks to find their location in the original image. Figure-3 shows the outline of our network.

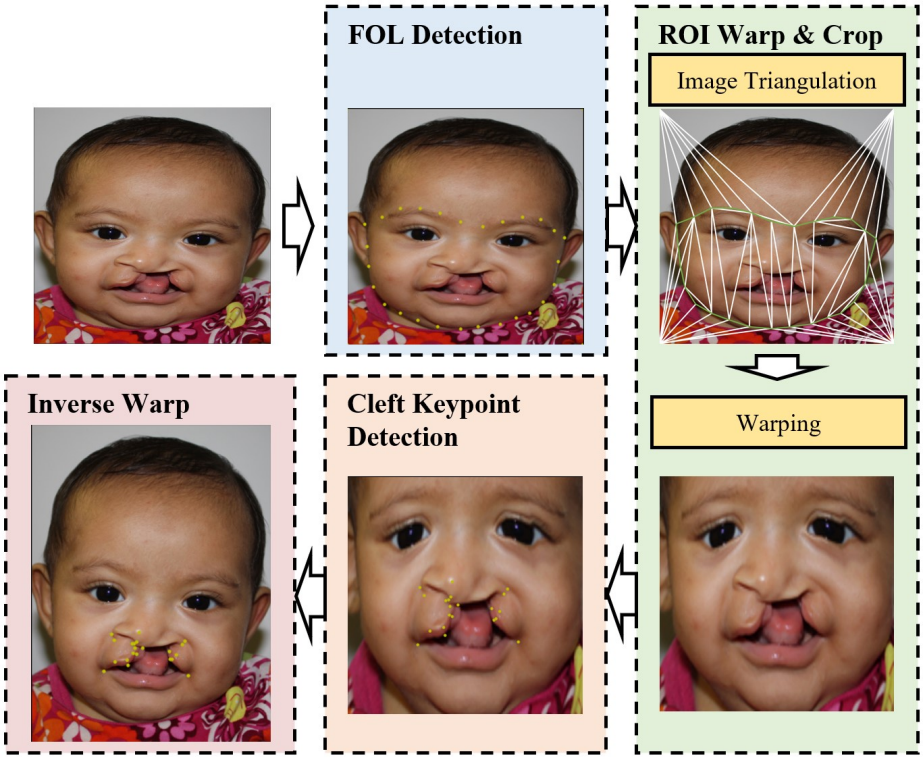
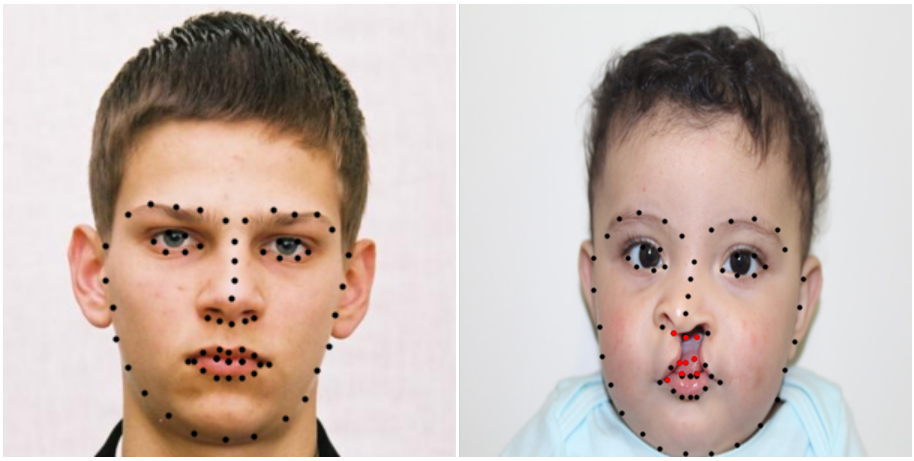


Fig. 3: Flowchart of proposed method. *FOL Detection*: The Facial Outline Landmarks (FOLs) are shown in yellow. *ROI Warp & Crop*: The ROI, drawn in green, is triangulated and is used to warp and crop the image. *Cleft Keypoint Detection*: The cleft landmarks are then detected on the warped image. *Inverse Warp*: Finally, the keypoints are inverse-warped to determine their correct locations in the original image.

3.1 Facial Outline Landmark (FOL) Detection

In order to detect the face outline that focuses on the cleft deformity, we use 27 landmarks along the mandible and the eyebrows of the face. These facial outline landmark (FOL) points were computed by taking the average of all training images' reference points. The face was cropped and resized into a 1024×1024 image, which was then passed through HRNet [17] to detect the landmarks along the mandible and eyebrows for each image. For each of these FOL points, the mean coordinate was calculated after excluding the outliers that were unusually far from the mean.

We start by detecting FOL points since the presence of a cleft lip has almost no effect on the landmarks on the facial outline (see Figure-4). The deformity affects the nasolabial region most significantly, and therefore any heatmap-based FOL point detection method can still accurately detect landmarks along the



(a) normal face (b) cleft patient

Fig. 4: Comparison of detected FOL points on a sample face from the 300W dataset and cleft lip patient. Incorrectly detected landmarks with large errors on nasolabial region of cleft lip patients are marked in red.

mandible and eyebrows for cleft faces. Heatmap-based methods work better for this purpose than regression-based methods since they deliberately focus on regions near the landmark for prediction, allowing the FOL points with sufficient distance from the nasolabial region to be detected correctly.

We use HRNet [17] that has been trained over the 300W dataset to detect the FOL of the cleft face. This dataset consists of images of faces with varying poses, expressions, skin tones, and lighting conditions and has been labeled with 68 facial landmarks along the mandible, eyes, nose, and mouth. In our work, we detect all 68 facial landmarks but only retain the 27 points comprising the facial outline. Figure-4 compares the detected FOL points for a normal face and a cleft lip face. Note how the points on the cleft are incorrectly detected. However, the outline is still detected accurately.

3.2 ROI Warp and Crop

The detected FOL cut out a region of interest (ROI) in the image that focuses only on the cleft deformity. However, though we use front-face images, they may be somewhat different in scale and orientation. Therefore, we want to cast all images to a general template to make the subsequent detection of CLL points more robust. Therefore, we use piecewise affine transformation to warp the ROI to a template ROI. The template ROI is generated by taking the average of facial outlines of 50 images, following the concept presented by Felzenswalb et al. [6], where intended detection objects are given deformable templates in the form of triangular meshes. Piecewise affine transformation separates the image into

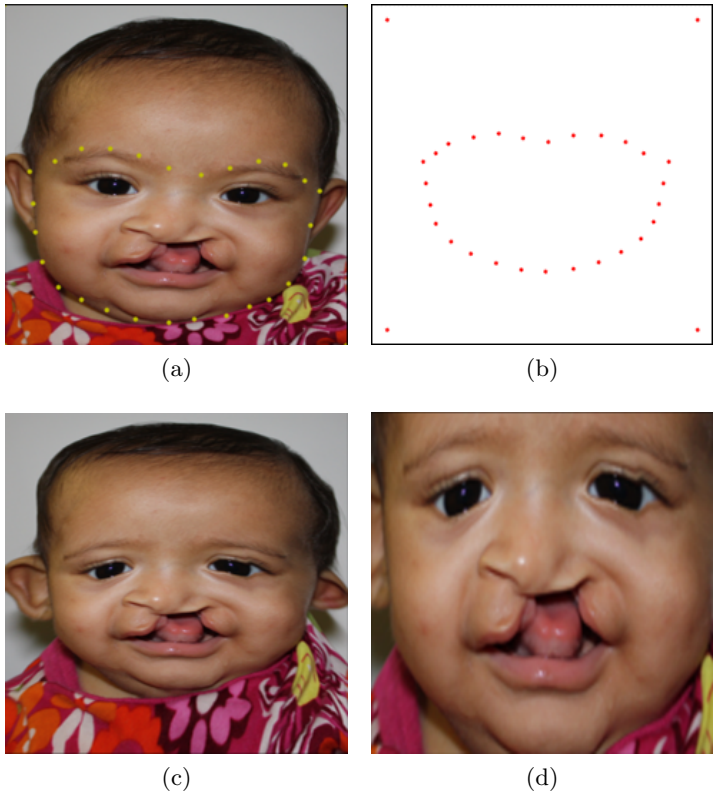


Fig. 5: (a) Input image of a cleft patient with the FOL points (in yellow) carving out the ROI. (b) The template ROI in red. (c) Warped image where the ROI of the input image is warped to the template ROI. (d) The warped image is cropped to retain only the minimal rectangle enclosing the warped ROI.

triangular segments for warping, allowing the warping of images using multiple reference points as shown in Figure-5. As shown by Ye et al. [36], due to the use of multiple anchor points for warping, piecewise affine transformation results in a much more controlled transformation compared to a general error-minimizing non-linear transformation. It also preserves the local shape of sub-regions in the area of cleft deformity. Finally, we crop the warped image to retain the minimal rectangle enclosing the warped ROI. Since our cleft lip dataset is small, and a large variation in the appearance occurs due to the severity and shape of the cleft as well as their scale and orientation, the previously mentioned warp-and-crop applied to each input image standardizes the appearance of the ROI for more accurate predictions. Figure-6 shows the triangulations used for warping for different images in greater detail.

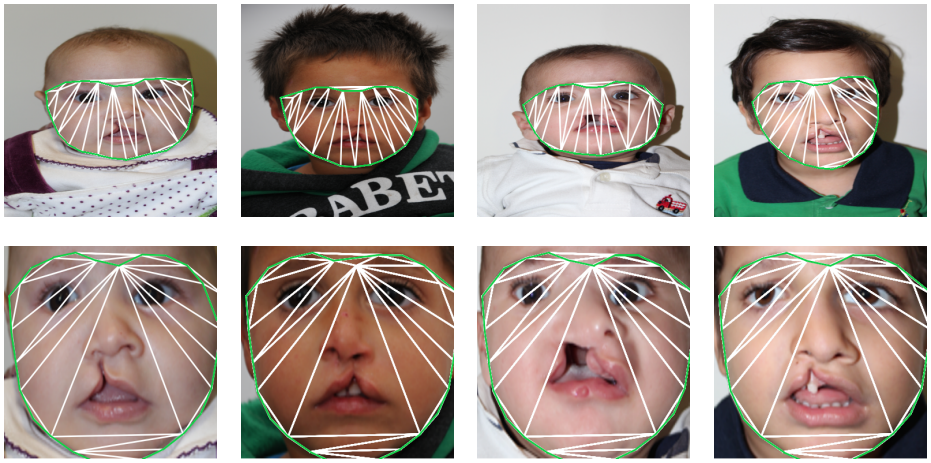


Fig. 6: Triangulations used for piecewise affine transform for different kinds of cleft lip patients. This demonstrates the generality of our method on multiple images with different attributes (e.g. skin tone color, illumination, type of cleft).

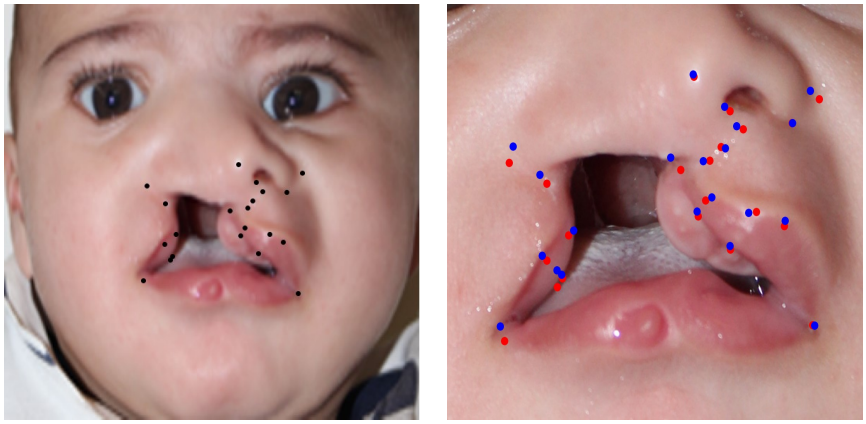
3.3 Cleft Lip Landmark (CLL) Detection

Following the warp, we proceed to detecting the cleft lip landmarks (CLL) using a small dataset of 200 images. As noted in Section-2, although heatmap regression has proven to produce slightly better results in facial landmark detection, our dataset is not sufficiently large for the model to reach convergence. Therefore, we use ResNet-50 [14] with direct regression trained over the Wingloss [7] function.

ResNet [14] proposes a solution to the vanishing gradient problem by introducing a residual block, which merges previous input with outputs to prevent harmful layers from affecting the result. Wingloss [7] is a loss function used in network training in which a wing-shaped function is used to control the linearity of the training. Feng et al. [8] have shown Wingloss to significantly improve the accuracy of facial landmark detection models using direct regression.

3.4 Inverse Image Warp

Finally, the predicted CLL points need to be inverse-warped to determine their locations in the original image domain. Figure-7 shows the results and compares the predictions with ground truth hand-marked by collaborating surgeons. We employ the same technique used for transforming ground truth to the warped image domain for creating an inverse warp.



(a) (b)

Fig. 7: Detected landmarks on warped ROI (a). In a view zoomed in on the cleft lip deformity (b), we show the predictions in red and the ground truth in blue. Note the precision alignment of red and blue points.

4 Implementation and Results

4.1 Dataset

Our dataset consists of 500 unilateral cleft lip photos provided to us by the Global Smile Foundation and is approved by the Institutional Review Boards, a group that has been formally designated to review and monitor biomedical research involving human subjects. Of the 500 images, approximately 200 frontal images of patients with cleft lips were labeled with 21 keypoints that are used for cleft surgery techniques, such as one shown in Figure-1. These images capture a frontal view of the entire face of patients and were carefully labeled by our surgeon collaborators who have deep experience in performing cleft lip repair. Almost all patients are less than 2 years which is the typical age range for cleft lip repair surgery. Excluding incompletely labeled images, we train our model on 150 images and evaluate our approach on the remaining 50.

Table 1: Interocular NME for cleft lip dataset

Method	***Test***	***Full***	***90%***	***80%***
Wingloss [7]	2.84e-2	2.81e-2	2.38e-2	2.12e-2
Wingloss - Warped	**2.35e-2**	**2.35e-2**	**1.84e-2**	**1.71e-2**
[1]Sayadi et al. [28]	N/A	3.87e-2	N/A	N/A

[1]Only Full NME is provided by Sayadi et al. [28]

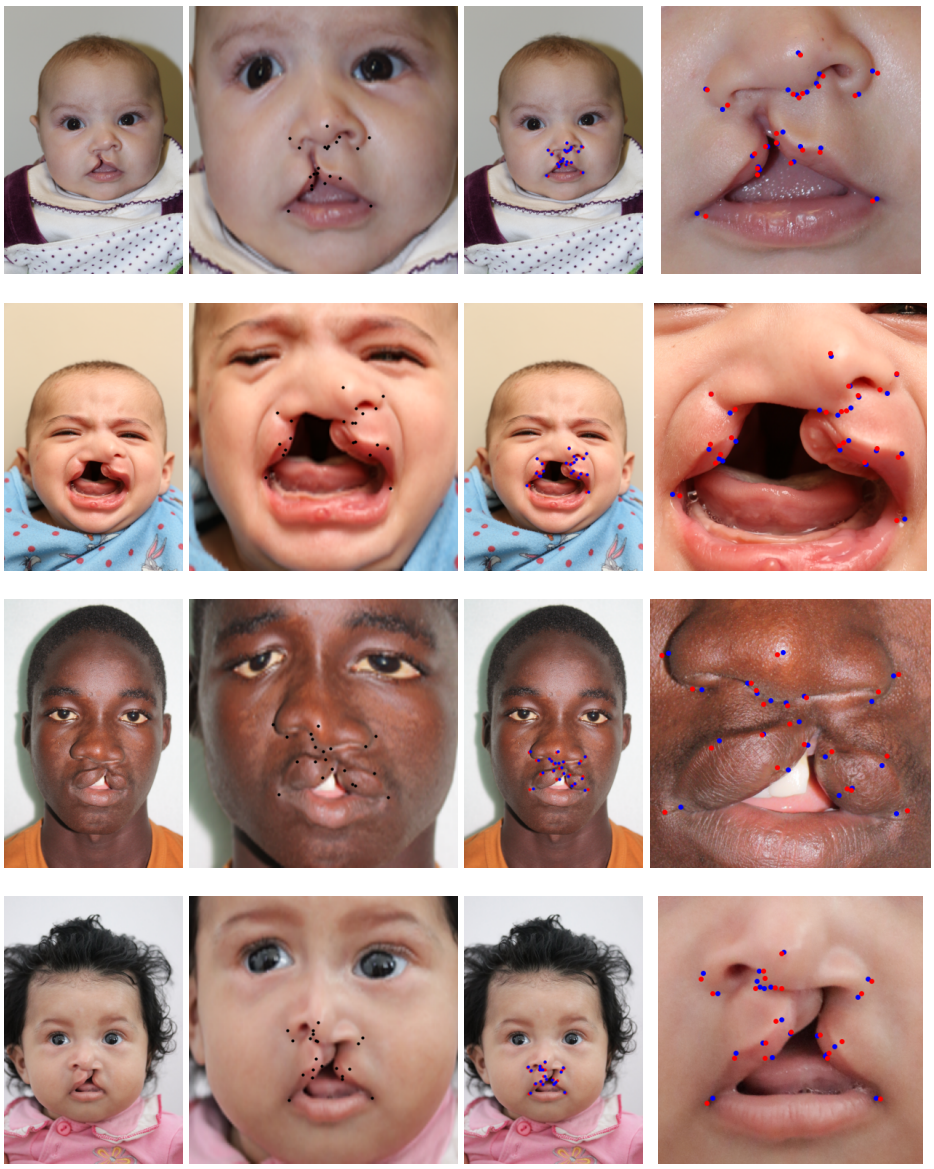


Fig. 8: Cleft anthropometric landmark predictions. Original image (left), predictions on warped ROI (left-middle), predictions on original image (right-middle), zoomed in predictions on original image (right).

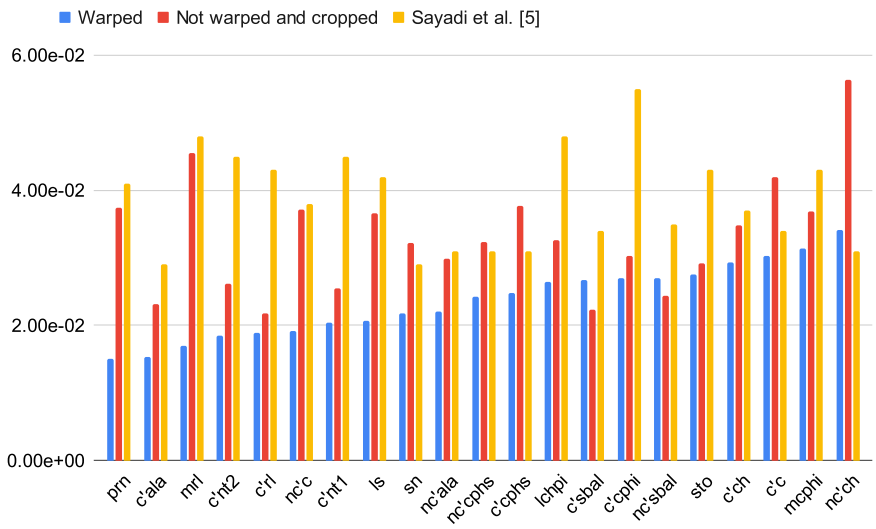


Fig. 9: NME for all 21 cleft anthropological landmark predictions with Warped Wingloss, without Warp-and-Crop Wingloss [7], and Sayadi et al. [28].

4.2 Evaluation Metric

In this work, we measure the accuracy of the detected keypoints by reporting the interocular Normalized Mean Error (NME), ϵ_i, computed as:

$$\epsilon_i = \frac{1}{V_i} \sum_{j \in K} v_{ij} \frac{d_{ij}}{L_i}, \tag{1}$$

where K is the set of all 21 keypoints, d_{ij} is the Euclidean distance (in pixels) between the j-th detected keypoint in image i and its ground truth location, v_{ij} is a binary variable that is 1 if the keypoint is visible in image i and 0 otherwise, L_i is the interocular distance i.e. the normalized Euclidean distance between the centers of the pupils and $V_i = \sum_{j \in K} v_{ij}$, the number of unoccluded points in image i.

4.3 Training

We train two networks separately: one where the input images have been ROI-warped and one where the input images are not ROI-warped. Both networks are trained on 150 training images.

4.4 Results

The test dataset consists of 50 cleft images of various face shapes, skin tones, and severity of cleft lip to best analyze the accuracy of the networks with minimal

Table 2: Interocular NME Comparison for All 21 Landmarks.

Name	Wingloss [7]	Warped-Wingloss	[1]Sayadi et al. [28]
prn	3.75e-02	**1.51e-02**	4.10e-02
c'ala	2.32e-02	**1.53e-02**	2.90e-02
mrl	4.55e-02	**1.70e-02**	4.80e-02
c'nt2	2.61e-02	**1.85e-02**	4.50e-02
c'rl	2.18e-02	**1.89e-02**	4.30e-02
nc'c	3.71e-02	**1.92e-02**	3.80e-02
c'nt1	2.55e-02	**2.04e-02**	4.50e-02
ls	3.66e-02	**2.06e-02**	4.20e-02
sn	3.22e-02	**2.18e-02**	2.90e-02
nc'ala	2.99e-02	**2.21e-02**	3.10e-02
nc'cphs	3.23e-02	**2.42e-02**	3.10e-02
c'cphs	3.76e-02	**2.48e-02**	3.10e-02
lchpi	3.26e-02	**2.64e-02**	4.80e-02
c'sbal	**2.24e-02**	2.67e-02	3.40e-02
c'cphi	3.03e-02	**2.70e-02**	5.50e-02
nc'sbal	**2.44e-02**	2.70e-02	3.50e-02
sto	2.92e-02	**2.75e-02**	4.30e-02
c'ch	3.48e-02	**2.93e-02**	3.70e-02
c'c	4.19e-02	**3.02e-02**	3.40e-02
nc'cphi	3.69e-02	**3.14e-02**	4.30e-02
nc'ch	5.64e-02	3.41e-02	**3.10e-02**

[1]Sayadi et al. [28] values are approximated from provided graph.

bias. The full set is evaluated using all 200 marked images. Figure-8 shows our predicted points compared with ground truth for some of these images. Note that despite having a large variation in the cleft lip deformity, our method is able to predict the CLLs accurately. Table-1 lists the average NMEs for our dataset both with and without the warp-and-crop. Figure-9 compares our method to one that does not perform the warp-and-crop and also with those reported by Sayadi et al. [28]. Our network provides superior performance with warp-and-crop and has a higher NME without warp-and-crop. Additionally, both versions of our network perform better than Sayadi et al. [28] on the full dataset (see Table-1).

Table-2 shows the NME for all 21 keypoints detected by the proposed method with and without warp-and-crop and compares them to the results reported by Sayadi et al. [28]. Our algorithm accurately and precisely predicts the anthropometric landmarks well within the boundaries set forth by benchmarks [1, 44]. The landmarks *c'ala*, *nc'ala*, and *prn*, all on the nose, have the lowest NME whereas *c'ch*, the cleft-side lip corner, was close to the median NME. *nc'ch*, the lip corner on the non-cleft side, showed the highest NME. Overall, our network, both with and without warp-and-crop, detects all keypoints, except *nc'ch*, with a lower NME than Sayadi et al. [28].

(a) with apparatus (b) tilted face (c) adult face

Fig. 10: Examples of failure cases. (a) FOLs could not be detected for images of patients with surgical apparatus, (b) images where the camera is oriented more than 10° from the frontal view, and (c) cleft lip deformity in adult faces.

Table-3 notes the average computation time for the major components of the network: FOL detection, ROI warp, and cleft keypoint detection. Our machine used for computation was equipped with 11900K processor, RTX 4090 GPU, and 128GB 3600MHz DDR4. Note that inverse warp is omitted from the Table-3 as it takes less than 1 ms.

Table 3: Computation time on network components.

FOL Detection	**ROI Warp**	**Cleft Keypoint Detection**
358 ms	661 ms	950 ms

*Computation time calculated on 11900K with RTX 4090.
*Inverse warp not shown as computation takes less than 1 ms.

4.5 Discussion and Limitations

The results show that our network has a lower error for certain landmarks but a higher error for others. For example, *c'ch*, *c'ala*, and *nc'ala* have the lowest NME, whereas *c'ch* and *nc'ch* have the highest NME. *c'ala*, *nc'ala*, and *prn*, all on the nose, have the lowest NME as they are anchor points that are the least subjective to surgeon preferences. In comparison, *c'ch* and *nc'ch*, the corners of the mouth, have a larger region of acceptable marking, and therefore, it is harder for our model to pinpoint the exact location. In marking the anthropometric landmarks, *c'cphi*, *c'nt1*, *c'nt2*, and *c'rl* are most subjective to surgeon preferences. However,

as all images were marked by Dr. Ross Sayadi, they did not show high NME in our study.

Benefiting from ROI warp, our network can accurately detect cleft landmarks on images at a camera angle between -10° and 10°. Any camera angle beyond this range results in incorrect detections, as our data does not contain enough example images to train the network. Despite thorough training, certain image features have been shown to cause performance drops and failure (see Figure-10). Such cases are (i) facial occlusion by apparatus, e.g. due to breathing tubes, band-aids, etc., (ii) non-frontal views of the cleft, (iii) adult patients with cleft-lip and (iv) patients with eyes closed. We suspect the small size and variety of our dataset to be the root cause of most of the failure cases. As our dataset consists mainly of full frontal face view images of patients less than two years of age with eyes open, our network is more likely to fail on images not meeting these criteria.

The proposed method initially detects FOL for warp-and-crop. This means that the network would fail to detect keypoints in an image where the facial boundary is not visible. The trained model also shows significant performance drops in subnasal views of the cleft. This is because the dataset used for training does not have any subnasal images. However, as long as the angle is not severely skewed, the warp-and-crop step of the proposed method corrects the image enough for the detection to operate correctly.

5 Conclusion

In summary, we have presented the first method to accurately detect cleft lip landmarks using an AI/ML based training method. In order to enhance the accuracy of this detection, we have used a warp-and-crop method that standardizes the image to counterbalance the facial deformations caused by the cleft-lip condition. Not only does it improve the performance of detection, but it also allows an end-to-end detection of cleft landmarks. With the help of ResNet-50 and Wingloss function, we optimize the network for use in cleft-lip surgeries. The empirical results show that the proposed method outperforms prior methods significantly. In the future, we would like to enhance this technique in the following ways.

- First, we would like our technique to be robust to different camera angles by expanding our training data set to non-frontal views.
- In addition, we would like to use temporal coherence and GPUs to enhance the performance of the prediction to get it near real-time so that the predicted markings stick to the face in AR based video surgical assistance sessions. Finally, we are also building a Spatially Augmented Reality (SAR) system that
- Spatially augmented reality (SAR) based system exist today that use a projector and RGBD camera (e.g. Azure Kinect) to illuminate surgical guidance marks directly on the surgical site. A remote expert marks the landmark

points or lines on the 3D model of the surgical site (captured by a structured light scan) using a GUI that shows up on-site in the surgical area in real-time [15]. We would like to extend our predictions to 3D models (instead of 2D images) to be used as a starting point on SAR systems. VR headset-based systems can also benefit from this.
– We would like to forge new directions by adapting the same technique for critical landmark detection for other surgeries as well.

Acknowledgment

This work was supported in part by The American Society of Maxillofacial Surgeons. We thank Dr. Raj Vyas and Dr. Ross Sayadi for the large amount of time spent with us on numerous discussions, marking of cleft lip images for creating data sets and helping us understand the key importance of the anthropometric landmarks. We thank Global Smile Foundation for providing us with the valuable cleft lip datasets.

References

1. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. pp. 545–552 (06 2011). https://doi.org/10.1109/CVPR.2011.5995602
2. Bulat, A., Sanchez, E., Tzimiropoulos, G.: Subpixel heatmap regression for facial landmark localization. CoRR **abs/2111.02360** (2021)
3. Bulat, A., Tzimiropoulos, G.: Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. vol. 9914 (09 2016). https://doi.org/10.1007/978-3-319-48881-3_43
4. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans (12 2017). https://doi.org/10.1109/CVPR.2018.00019
5. Chandaliya, P., Nain, N.: Plasticgan: Holistic generative adversarial network on face plastic and aesthetic surgery. Multimedia Tools and Applications **81**, 1–22 (09 2022). https://doi.org/10.1007/s11042-022-12865-5
6. Felzenszwalb, P.: Representation and detection of deformable shapes. IEEE transactions on pattern analysis and machine intelligence **27**, 208–20 (03 2005). https://doi.org/10.1109/TPAMI.2005.35
7. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks (06 2018). https://doi.org/10.1109/CVPR.2018.00238
8. Feng, Z.H., Kittler, J., Awais, M., Wu, X.J.: Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. International Journal of Computer Vision **128** (09 2020). https://doi.org/10.1007/s11263-019-01275-0
9. Freitas, R., Aires, K., Campelo, V.: Automatic location of facial landmarks for plastic surgery procedures. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics **2014**, 1444–1449 (10 2014). https://doi.org/10.1109/smc.2014.6974118

10. Freitas, R.T., Aires, K.R.T., Campelo, V.E.S.: Locating facial landmarks towards plastic surgery. In: 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images. pp. 219–225 (2015). https://doi.org/10.1109/SIBGRAPI.2015.40
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (11 2013). https://doi.org/10.1109/CVPR.2014.81
12. Guerrero, C.: Cleft lip and palate surgery: 30 years follow-up. Annals of maxillofacial surgery **2**, 153–157 (03 2012). https://doi.org/10.4103/2231-0746.101342
13. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**, 1–1 (06 2018). https://doi.org/10.1109/TPAMI.2018.2844175
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. pp. 770–778 (06 2016). https://doi.org/10.1109/CVPR.2016.90
15. Ibrahim, M.T., Gopi, M., Vyas, R., Sayadi, L.R., Majumder, A.: Projector illuminated precise stencils on surgical sites. IEEE Conference on Virtual Reality and 3D User Interfaces (2023)
16. Jackson, A., Valstar, M., Tzimiropoulos, G.: A cnn cascade for landmark guided semantic part segmentation (10 2016)
17. Ke, S., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. pp. 5686–5696 (06 2019). https://doi.org/10.1109/CVPR.2019.00584
18. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment. pp. 2034–2043 (07 2017). https://doi.org/10.1109/CVPRW.2017.254
19. Lan, X., Hu, Q., Cheng, J.: HIH: towards more accurate face alignment via heatmap in heatmap. CoRR **abs/2104.03100** (2021)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Arxiv **79** (11 2014)
21. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. pp. 3691–3700 (07 2017). https://doi.org/10.1109/CVPR.2017.393
22. Pourramezan Fard, A., Mahoor, M.: Acr loss: Adaptive coordinate-based regression loss for face alignment. pp. 1807–1814 (08 2022). https://doi.org/10.1109/ICPR56361.2022.9956683
23. Ranjan, R., Patel, V., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (03 2016). https://doi.org/10.1109/TPAMI.2017.2781233
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (06 2015). https://doi.org/10.1109/TPAMI.2016.2577031
25. Robinson, J.P., Li, Y., Zhang, N., Fu, Y., Tulyakov, S.: Laplace landmark localization. CoRR **abs/1903.11633** (2019)
26. Rossell Perry, P.: A 20-year experience in unilateral cleft lip repair: From millard to the triple unilimb z-plasty technique. Indian Journal of Plastic Surgery **49**, 340 (09 2016). https://doi.org/10.4103/0970-0358.197226
27. Sayadi, L., Chopan, M., Sayadi, J., Samai, A., Arora, J., Anand, S., Evans, G., Widgerow, A., Vyas, R.: Operating room stencil: A novel mobile application for surgical planning. Plastic and Reconstructive Surgery - Global Open **9**, e3807 (09 2021). https://doi.org/10.1097/GOX.0000000000003807

28. Sayadi, L., Hamdan, U., Zhangli, Q., Hu, J., Vyas, R.: Harnessing the power of artificial intelligence to teach cleft lip surgery. Plastic and Reconstructive Surgery - Global Open **10**, e4451 (07 2022). https://doi.org/10.1097/GOX.0000000000004451
29. Vyas, R., Sayadi, L., Bendit, D., Hamdan, U.: Using virtual augmented reality to remotely proctor overseas surgical outreach: Building long-term international capacity and sustainability. Plastic & Reconstructive Surgery **146**, 622e–629e (11 2020). https://doi.org/10.1097/PRS.0000000000007293
30. Vyas, T., Gupta, P., Kumar, S., Gupta, R., Gupta, T., Singh, H.: Cleft of lip and palate: A review. Journal of Family Medicine and Primary Care **9**, 2621 (06 2020). https://doi.org/10.4103/jfmpc.jfmpc_472_20
31. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. CoRR **abs/1602.00134** (2016)
32. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. pp. 2129–2138 (06 2018). https://doi.org/10.1109/CVPR.2018.00227
33. Wu, Y., Hassner, T., Kim, K., Medioni, G., Natarajan, P.: Facial landmark detection with tweaked convolutional neural networks. Arxiv **PP** (01 2015). https://doi.org/10.1109/TPAMI.2017.2787130
34. Yan, Y., Duffner, S., Phutane, P., Berthelier, A., Blanc, C., Garcia, C., Chateau, T.: 2d wasserstein loss for robust facial landmark detection. Pattern Recognition **116** (03 2021). https://doi.org/10.1016/j.patcog.2021.107945
35. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. pp. 2025–2033 (07 2017). https://doi.org/10.1109/CVPRW.2017.253
36. Ye, Y., Shan, J., Bruzzone, L., Shen, L.: Robust registration of multimodal remote sensing images based on structural similarity. IEEE Transactions on Geoscience and Remote Sensing **PP**, 1–18 (02 2017). https://doi.org/10.1109/TGRS.2017.2656380
37. Yin, S., Wang, S., Chen, X., Chen, E., Liang, C.: Attentive one-dimensional heatmap regression for facial landmark detection and tracking. pp. 538–546 (10 2020). https://doi.org/10.1145/3394171.3413509
38. Zadeh, A., Lim, Y., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for 3d facial landmark detection. pp. 2519–2528 (10 2017). https://doi.org/10.1109/ICCVW.2017.296
39. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. pp. 1–16 (09 2014). https://doi.org/10.1007/978-3-319-10605-2_1
40. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23** (04 2016). https://doi.org/10.1109/LSP.2016.2603342
41. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**, 1–1 (01 2015). https://doi.org/10.1109/TPAMI.2015.2469286
42. Zhao, H., Ying, X., Shi, Y., Tong, X., Wen, J., Zha, H.: Rdcface: Radial distortion correction for face recognition. pp. 7718–7727 (06 2020). https://doi.org/10.1109/CVPR42600.2020.00774
43. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. pp. 386–391 (12 2013). https://doi.org/10.1109/ICCVW.2013.58

44. Çeliktutan, O., Ulukaya, S., Sankur, B.: A comparative study of face landmarking techniques. EURASIP Journal on Image and Video Processing **2013** (03 2013). https://doi.org/10.1186/1687-5281-2013-13