The Effect of Improving Facial Alignment Accuracy on the Video-based Detection of Neurological Diseases

Diego L. Guarin Member, IEEE, Andrea Bandini Member, IEEE, Aidan Dempster, Henry Wang, Siavash Rezaei, Yana Yunusova and Babak Taati Member, IEEE

Abstract— Background: Automatic facial landmark localization is an essential component in many computer vision applications, including video-based detection of neurological diseases. Machine learning models for facial landmarks localization are typically trained on faces of healthy individuals, and we found that model performance is inferior when applied to faces of people with neurological diseases. Fine-tuning pre-trained models with representative images improves performance on clinical populations significantly. However, questions related to the characteristics of the database used to fine-tune the model and the clinical impact of the improved model remain. Methods: We employed the Toronto NeuroFace dataset – a dataset consisting videos of Healthy Controls (HC), individuals Post-Stroke, and individuals with Amyotrophic Lateral Sclerosis performing speech and non-speech tasks with thousands of manually annotated frames - to fine-tune a well-known deep learning-based facial landmark localization model. The pre-trained and fine-tuned models were used to extract landmark-based facial features from videos, and the facial features were used to discriminate clinical groups from HC. Results: Fine-tuning a facial landmark localization model with a diverse database that includes HC and individuals with neurological disorders resulted in significantly improved performance for all groups. Our results also showed that fine-tuning the model with representative data greatly improved the ability of the subsequent classifier to classify clinical groups vs. HC from videos. Conclusions: Using a diverse database for model fine-tuning might result in better model performance for HC and clinical groups. We demonstrated that fine-tuning a model for landmark localization with representative data results in improved detection of neurological diseases.

DLG is supported by the Michael J. Fox Foundation for Parkinson's Research.

DLG is with the Biomedical Engineering Program, Department Of Biomedical And Chemical Engineering And Sciences. Florida Institute of Technology, 150 W. University Blvd. Melbourne, FL 32901 USA (email: dguarinlopez@fit.edu). DLG is a post-graduate affiliate of the Vector Institute for Artificial Intelligence, 661 University Ave Suite 710, Toronto, ON M5G 1M1 Canada

AB and SR are with KITE | Toronto Rehabilitation Institute, 550 University Ave., Toronto ON M5G 2A2 Canada

HW and AD are with the Department of Speech Language Pathology, University of Toronto, 500 University Ave., Toronto, ON M5G 1V7 Canada

YY is with the Department of Speech Language Pathology, University of Toronto, 500 University Ave., Toronto, ON M5G 1V7 Canada, and the Hurvitz Brain Sciences Program, Sunnybrook Research Institute, 2075 Bayview Ave., Tonton ON M4N 3M5 Canada

BT is with KITE | Toronto Rehabilitation Institute, 550 University Ave., Toronto ON M5G 2A2 Canada, and the Department of Computer Science, University of Toronto, 214 College St, Toronto ON M5T 3A1 Canada

Index Terms— Computer-aided Diagnosis, Neurological Diseases, Facial Landmarks Localization, Facial Alignment, Stroke, ALS.

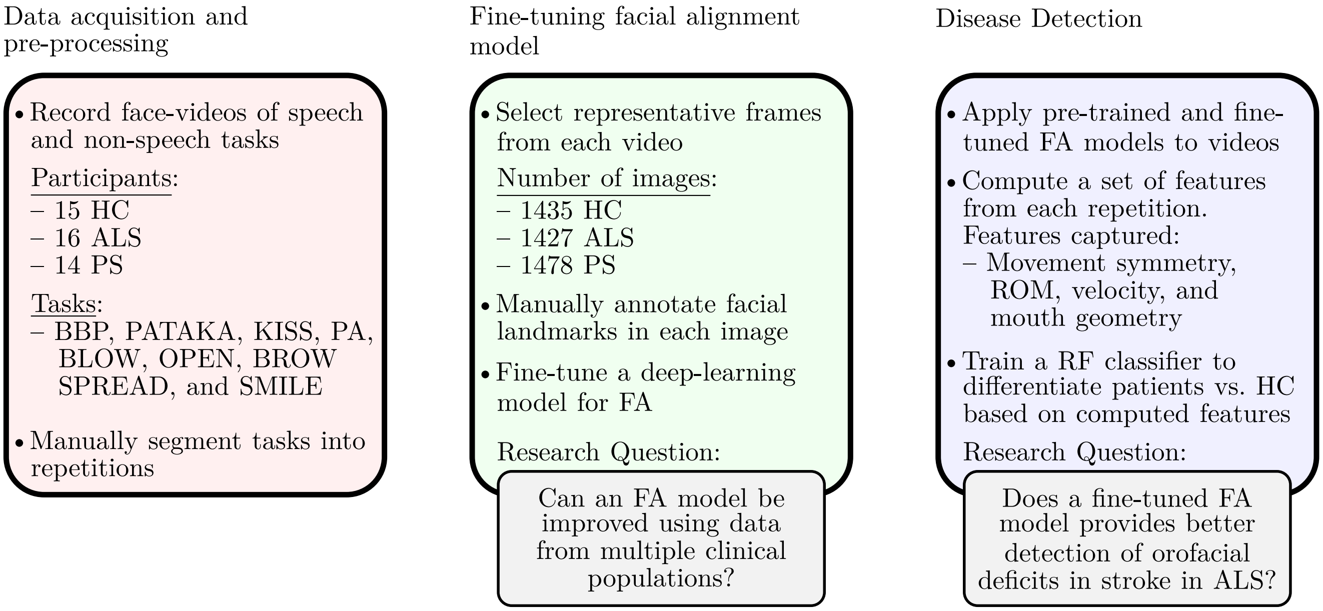
I. Introduction

FACIAL alignment (FA) refers to the use of machine learning models and algorithms for automatic localization of pre-defined landmarks in facial images [1]. FA is often an important first step in many computer vision applications including face recognition [2], [3], emotion detection [4]–[6], and human-computer interactions [7], [8]. In these applications, coordinates of detected facial landmarks are either used to compute a set of features or fed directly to a machine learning model (e.g. for pain detection [9]–[11], or detection of neurological diseases and motor disorders [12]–[29]) or used in pre-processing to align face images for a subsequent deep learning model (e.g. for face verification [30]).

Many of these applications rely on pre-trained FA models for localization of facial landmarks. These models are typically trained using large databases of manually or semi-automatically annotated facial images [1]. These databases often consist of thousands of photographs with a large variety of poses, expressions, illumination, backgrounds, and scales. Thus, pre-trained FA models are designed to provide accurate facial landmark localization under general conditions [31]–[36].

Challenges remain when applying pre-trained FA models to photographs from clinical populations [28], [37]. For instance, we recently demonstrated that pre-trained FA models perform better in healthy, young individuals as compared to patients with neurological diseases such as Alzheimer's disease, stroke, and ALS, and motor disorders such as facial palsy [23], [28], [29], [37]–[39]. This phenomenon, known as algorithmic bias, is attributed to the lack of representative data in the databases used to train FA models [29], [37], [40]–[42]

One approach to mitigate the bias in a pre-trained FA model is to fine-tune the model using representative data applying transfer learning techniques [43], [44]. Recently, we demonstrated that fine-tuning a pre-trained, deep-neural network-based FA model with a handful of manually annotated photographs of patients with facial



Data acquisition and pre-processing

- Record face-videos of speech and non-speech tasks

Participants:
– 15 HC
– 16 ALS
– 14 PS

Tasks:
– BBP, PATAKA, KISS, PA, BLOW, OPEN, BROW SPREAD, and SMILE

- Manually segment tasks into repetitions

Fine-tuning facial alignment model

- Select representative frames from each video

Number of images:
– 1435 HC
– 1427 ALS
– 1478 PS

- Manually annotate facial landmarks in each image
- Fine-tune a deep-learning model for FA

Research Question:

Can an FA model be improved using data from multiple clinical populations?

Disease Detection

- Apply pre-trained and fine-tuned FA models to videos
- Compute a set of features from each repetition. Features captured:
  – Movement symmetry, ROM, velocity, and mouth geometry
- Train a RF classifier to differentiate patients vs. HC based on computed features

Research Question:

Does a fine-tuned FA model provides better detection of orofacial deficits in stroke in ALS?

Fig. 1. Graphical summary of the methods and research questions investigated on this paper. The Fig. depicts three stages described in this paper, including data acquisition and pre-processing, fine-tuning of Facial Alignment models, and automatic detection of neurological diseases. HC: Healthy Controls, ALS: people with Amyotrophic Lateral Sclerosis, PS: people Post Stroke, FA: Facial Alignment. See text for a detailed description of the different tasks

palsy significantly improved the model performance on individuals with the same clinical condition. Further, we demonstrated that it is possible to eliminate the pre-trained FA model bias against individuals with facial palsy by fine-tuning the model with 320 manually annotated photographs from patients [29]. Similarly, we observed a significant improvement in FA model performance in older adults with dementia after fine-tuning a deep-neural network-based model with 688 manually annotated representative photographs [37]. Improved model performance of a pre-trained FA model was also observed in individuals PS and in individuals with ALS after fine-tuning the model with 1371 and 920 manually annotated photographs, respectively [39]. Furthermore, we observed that fine-tuning an FA model with 1015 images of age-matched healthy controls (HC) improved the model performance when applied to photographs of individuals with neurological diseases significantly. However, the improvement was lower than when fine-tuning the model with representative clinical data [39].

Our previous results showed that fine-tuning an FA model with representative clinical data improved the model performance on that clinical population. They also showed that fine-tuning an FA model with data form age-matched HC improved the model performance on clinical populations recorded under the same conditions. Thus, the logical next step is to determine if fine-tuning a model with images from patients of multiple clinical groups and aged-matched HC also leads to improved model performance on the clinical and non-clinical groups. This research questions might have important clinical implications because collecting data from age-matched HC is typically straightforward, whereas collecting patient's data is often difficult and time-consuming, specially for rate diseases such as ALS, and while there are many differences in the way that each individual is affected by a neurological disease, there are also many similarities in the way that these diseases manifest in the orofacial musculature and function, e.g., muscle weakness, facial asymmetries. Based on these observations, we hypothesized that fine-tuning an FA model with data from multiple patient populations and age-matched HC can improve model performance on all clinical and non-clinical groups.

Furthermore, despite significant efforts to improve FA models performance on clinical populations, there is no quantitative evidence that the improved accuracy in landmark localization leads to an improved computer-aided diagnosis of neurological diseases from video based monitoring. We have shown that by using pre-trained FA models is possible to differentiate aged-matched HC from individuals PS with an accuracy of 87% [12], and age-matched HC from individuals with ALS with an accuracy close to 89% [14] using videos of speech and non-speech tasks. Based on these result, and the improved performance provided by fine-tuned FA models on clinical populations, we hypothesized that better diagnosis of neurological diseases from video based monitoring would be achieved by applying FA models fine-tuned with representative data as compared to pre-trained FA models.

The specific objectives of this paper are to: i) fine-tune a deep-neural network-based FA model with a database of manually annotated photographs of patients from different clinical populations with neurological diseases affecting the orofacial function, and age-matched healthy controls; and ii) assess the influence of the fine-tuned FA model on the computer-aided diagnosis of neurological diseases from video based monitoring. For these goals, we used facial videos from healthy controls and indivuals from two clinical populations – stroke survivors and individuals with ALS – performing a set of speech and non-speech tasks commonly used during clinical orofacial examinations [45], [46]. A subset of video frames were manually annotated and used to fine-tune a well-known pre-trained FA model. The pre-trained and fine-tuned FA models were used to estimate landmark-based facial features, and these features were used to automatically differentiate the clinical groups from HC.

Fig. 1 summarises the methods and research questions investigated on this paper. The diagram presents the three stages of our pipeline, including data acquisition and preprocessing, fine-tuning a FA model with representative data, and automatic detection of neurological disease from landmark-based facial features and video-based monitoring.

II. Materials and Methods

A. Toronto NeuroFace dataset

The Toronto NeuroFace dataset [39] – a novel and open-access dataset for facial analysis in individuals with neurological diseases – was used in this study. Here, we provide a brief description of the database, experimental setup, and tasks.

Participants: Forty-five participants are included in this dataset: 16 individuals with amyotrophic lateral sclerosis (ALS, 8 female), 14 individuals post stroke (PS, 5 female), and 15 age-matched healthy-controls (HC, 7 females). All participants were cognitively unimpaired at the time of recording as demonstrated by a Montreal Cognitive Assessment score ≥ 26 [47], and passed a hearing screening. Table I presents the demographics and clinical summary of the participants. The study was approved by the Research Ethics Boards at the Sunnybrook Research Institute and University Health Network: Toronto Rehabilitation Institute. All participants signed informed consent according to the requirements of the Declaration of Helsinki.

1) Experimental Setup: Participants were seated in front of an Intel RealSense™ depth camera (SR300 or D400) with a face-to-camera distance between 30 cm and 60 cm. A continuous light source was placed adjacent to the camera to provide uniform illumination. Participants were asked to look directly at the camera and were recorded during the execution of standard speech and non-speech tasks used during clinical orofacial examinations. A video comprised of color (RBG) and depth information was recorded for each task. Both streams were recorded synchronously at approximately 50 frames per second at VGA resolution (640×480 pixels). A total of 332 videos were included in the database: 108 from HC participants, 113 from individuals PS, and 111 from individuals with ALS.

2) Tasks: Participants were asked to perform a set of speech and non-speech tasks commonly used during clinical orofacial examinations [45], [46]. The tasks included 10 repetitions of the sentence ”Buy Bobby a Puppy” at a comfortable speaking rate and loudness (BBP); repetitions of the syllable /pa/ as fast as possible on a single breath (PA); repetitions of the syllables /pataka/ as fast as possible on a single breath (PATAKA); puckering the lips 5 times (BLOW); pretend to kiss a baby 5 times (KISS); maximum opening of the mount 5 times (OPEN); pretending to smile with tight lips 5 times (SPREAD); making a big smile 5 times (SMILE); raising the eyebrows 5 times (BROW); and maintaining a neutral facial expression with eyes open and mouth closed for 20 s (REST). Participant were encouraged to take breaks between tasks to prevent fatigue; however, not all participants were able to perform all tasks.

TABLE I

Demographics and clinical information for the three participant groups: Healthy Controls (HC), Post-Stroke (PS), and Amyotrophic Lateral Sclerosis (ALS). Presented values a are mean and range.

	Age (years)	Duration* (days)	ALSFRS - R Total	ALSFRS - R Bulbar
HC	58.3 [19 - 78]	-	-	-
PS	64.0 [21 - 89]	579.8 [2 - 3262]	-	-
ALS	61.8 [45 - 75]	689.4 [176 - 1640]	35.6 [26 - 40]	9.4 [6 - 12]

* Duration indicates days since stroke for PS group, and days since diagnosis for ALS group.
ALSFRS-R: ALS Functional Rating Scale - Revised

B. Fine-tuning an FA model with representative data

We fine-tuned a well-known, deep-learning-based model for FA using manually annotated representative clinical data [28], [29], [37], [39]. Next, we briefly describe the model, the manual annotation procedure, and the approach used to fine-tune the model and evaluate its performance.

1) Pre-trained FA model: The pre-trained FA model corresponds to the Facial Alignment Network (FAN), a deep-learning-based model trained with more than 230,000 photographs [35]. The FAN model consist of an initial face detection stage that returns a $256\times256 pixelsimagecenteredaroundtheface.Theface - centeredimageisthendown - sampledintoasetof256featuremapsofdimensions64\times64$, and passed into four stacked hourglass networks, an architecture commonly used for facial landmark localization [48]–[50], that transforms the feature maps into a set of 68 heat-maps. Each heat-map provides the estimated position of a facial landmarks. The pre-trained FAN model and Python API are freely available online (https://github.com/1adrianb/face-alignment).

2) Manual annotations: A set of 4340 video frames (1435 for HC, 1478 for PS, and 1427 for ALS) were extracted from the videos. Extracted frames were intended to capture a wide range of facial gestures during task

execution. Additional details regarding frame selection can be found in [39].

The locations of 68 facial landmarks described by the Multi-PIE 2D configuration [51], and defining the eyebrows, eyes, nose, mouth, and jawline, were manually localized by a trained annotator in each extracted frame. Manually annotated facial landmarks were considered as the ground truth positions.

3) Fine-tuning the FAN model with representative data: The parameters of the first four stages of the pre-trained FAN model were frozen and not modified furing the fine-tuning process. The parameters of last hourglass network were updated using the Toronto NeuroFace dataset. Optimization algorithm and hyper-parameters were the same to those used by Bulat and Tzimiropoulos to train the FAN model [35]. Our training algorithm used a recently introduced loss function, adaptive wing-loss, which improves model performance by penalizing small errors more than the traditional squared-loss [36], [52], [53]

Twelve participants from each group were randomly selected and used to train the model. Data from the remaining participants were used to test the model performance by computing the accuracy in landmark localization. Accuracy was computed in terms of the Root-Mean-Squared Error (RMSE) between manually annotated and model predicted landmark positions normalized by the intercanthal distance (NRMSE) [33].

4) Statistical analysis: Statistical differences between results yielded by the pre-trained and fine-tuned models and ground truth landmark position were evaluated using the t-test (statistical significance was considered at $p < 0.01$) and the standardized mean difference (SMD), computed as

$$SMD = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

were μ_1, μ_2, s_1, s_2, n_1, and n_2 are the mean, standard deviation, and number of elements of the difference between pre-trained FAN model predictions and ground truth landmark positions; and μ_2, s_2, and n_2 are the mean, standard deviation, and number of elements of the difference between the fine-tuned FAN model predictions and ground truth landmark positions. For $0 < SMD < 0.5$, the difference between groups is considered to be small; for $0.5 \leq SMD < 0.8$, the difference between groups is considered to be medium; and for $SMD \geq 0.8$, the difference between groups is considered to be large [54], [55].

C. Video-based diagnosis of neurological diseases

Diagnosis of neurological diseases was achieved by 1) manually segmenting the tasks by repetition, 2) applying an FA models to localize the 68 facial landmarks in each video frame, 3) reconstructing the 3D, real world coordinates of the 68 facial landmark in each video frame, 4) extracting landmarks-based facial features from each repetition, and 5) using a classification algorithm to detect the presence or absence of the disease based on the extracted features. Next, we describe these steps in detail and provide a brief description of the landmarks-based features used in this study.

1) Participants and tasks: Nine participants declined for their data to be shared publicly, so their recordings were not used for further analysis. Thus, data from 36 participants were used for video-based diagnosis of neurological diseases, 11 individuals with ALS (7 female), 14 individuals PS (4 female), and 11 HC (4 female). Furthermore, only tasks common to all 36 participants were used for video-based diagnosis, analyzed tasks include: BBP, OPEN, SPREAD, and REST. Finally, video recordings were the participant did not look directly at the camera during task execution were removed from the analysis. A total of 138 videos were included for the video-based diagnosis of neurological diseases: 40 from HC participants, 54 from individuals PS, and 44 from individuals with ALS.

2) Tasks segmentation: All tasks, except REST, were manually segmented into individual repetitions by a trained observer; the observer identified the beginning and end of each repetition using the audio or video recordings.

3) Face alignment: Pre-trained and fine-tuned FAN models were used to automatically estimate the position of the 68 facial landmarks in each video frame. The time-series containing the $[x_{2d},\ y_{2d}]$ coordinates for each landmarks were smoothed using a 5-points median filter.

4) Reconstruction of 3D coordinates: Color and depth streams were aligned using the camera extrinsic parameters. Afterwards, the real world coordinates (in mm) for each landmark were computed using a pinhole camera model with the depth information provided by the depth sensor (z_{3d}), the 2d coordinates ($[x_{2d},\ y_{2d}]$), and the intrinsic parameters provided by the camera manufacturer. This procedure resulted in a set of $[x_{3d},\ y_{3d},\ z_{3d}]$ coordinates for each landmark. The origin of the 3D coordinate system was the center of the IR camera, and the x, y, and z axes were along the lateral, vertical, and frontal directions, respectively.

5) Feature extraction: For each repetition of each task, a set of features were extracted using the 3D coordinates of selected landmarks. Different features were extracted to separate individuals PS from healthy controls, and individuals with ALS from healthy controls. Features to identify individuals PS measured mouth range of motion and velocity, and facial symmetry (13 features). These features have been previously described and validated [12]. Features to identify individuals with ALS measured mouth range of motion and velocity, overall movement of the lower lip, mouth symmetry, and the overall roundness of lips during movement (11 features). These features have been previously described and validated [14].

6) Classification: Disease detection was performed on a task by task basis using a random forest (RF) classification algorithm. Twelve classification tests were conducted, by combining data from: Two diseases (HC vs. PS, and HC

TABLE II
SUMMARY OF THE TWELVE RF CLASSIFIERS TRAINED TO DIFFERENTIATE BETWEEN HEALTHY CONTROLS (HC) AND AMYOTROPHIC LATERAL SCLEROSIS (ALS) PATIENTS, AND HEALTHY CONTROLS (HC) AND POST-STROKE (PS) PATIENTS, FROM VIDEOS OF PARTICIPANTS PERFORMING ON SPEECH TASK (BBP), AND TWO NON-SPEECH TASKS (OPEN AND SPREAD).

Classification	Task	FA model
HC vs. ALS	BBP	Pre-trained
HC vs. ALS	BBP	Fine-tuned
HC vs. ALS	OPEN	Pre-trained
HC vs. ALS	OPEN	Fine-tuned
HC vs. ALS	SPREAD	Pre-trained
HC vs. ALS	SPREAD	Fine-tuned
HC vs. PS	BBP	Pre-trained
HC vs. PS	BBP	Fine-tuned
HC vs. PS	OPEN	Pre-trained
HC vs. PS	OPEN	Fine-tuned
HC vs. PS	SPREAD	Pre-trained
HC vs. PS	SPREAD	Fine-tuned

vs. ALS), three tasks (BBP, OPEN, and SPREAD), and two FA models (pre-trained and fine-tuned). Table II summarises the different RF classifiers trained in this study. The output of the RF model was a probability (a value between 0 and 1) that each repetition was performed by an individual suffering from a neurological disease. The probability that a task was performed by an individual suffering from a neurological disease was considered as the average probability from all repetitions of the same task.

Classification performance was evaluated using leave-one-subject-out cross-validation (LOSO-CV). For each fold of the LOSO-CV, all the repetitions belonging to a single participant were used as the test set, and the RF classifier was trained with the repetitions from the other participants. Performance was evaluated using the receiver operating characteristic (ROC) curve and the corresponding area under the ROC curve (AU-ROC).

III. RESULTS

A. FA model fine-tuning

Fig. 2 presents the cumulative distribution of the NRMSE between the ground truth landmarks position and the results yielded by the pre-trained FAN model with blue lines, and fine-tuned FAN model orange lines. Fig. 2 A) present the results obtained for HC, B) for individuals PS, C) for and individuals with ALS.

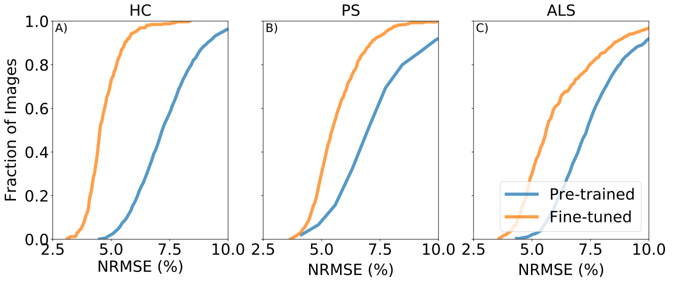


Fig. 2. Cumulative distribution of the NRMSE between the ground truth landmark positions and the results yielded by the Pre-trained and Fine-tuned FAN models for A) Healtly Controls (HC), B) patients Post-Stroke (PS), and C) patients with amyotrophic lateral sclerosis (ALS). Landmarks defining the jaw were not considered for the results presented here.

TABLE III
MEAN ± STANDARD DEVIATION OF THE NRMSE, AND THE RESULTING STATISTICAL ANALYSIS OF THE DIFFERENCE BETWEEN THE GROUND TRUTH LANDMARK POSITIONS AND THE POSITIONS YIELDED BY THE PRE-TRAINED AND FINE-TUNED FAN MODELS FOR HEALTLY CONTROLS (HC), INDIVIDUALS POST-STROKE (PS), AND INDIVIDUALS WITH AMYOTROPHIC LATERAL SCLEROSIS (ALS). LANDMARKS DEFINING THE JAW WERE NOT CONSIDERED FOR THESE RESULTS.

	HC	PS	ALS
Pre-trained Model	7.4 ± 1.4	7.7 ± 1.6	7.9 ± 2.5
Fine-tuned Model	4.7 ± 0.7	6.2 ± 1.7	5.7 ± 1.1
t-test	$t = 32.6$ $p \ll 0.01$	$t = 14.1$ $p \ll 0.01$	$t = 20.1$ $p \ll 0.01$
SMD	2.0	0.9	1.0

Table III summarized the results of Fig. 2 and demonstrates that fine-tuning the FAN model with a database composed of manually annotated images from HC participants, individuals with ALS, and individuals PS improved the model performance for all groups significantly. In particular, for the HC participants, there was a large, significant improvement in the NRMSE. Fine-tuning the FA model reduced the NRMSE from 7.4 ± 1.4 % to 4.7 ± 0.7% ($t = 32.6 \ - \ p \ll 0.01$, $SMD = 2.0$). Similarly, for individuals PS, there was a large, significant improvement in the NRMSE. Fine-tuning the FA model reduced the NRMSE from 7.7 ± 1.6 % to 6.2 ± 1.7% ($t = 14.1 \ - \ p \ll 0.01$, $SMD = 0.9$). Finally, for individuals with ALS, there was a large, significant improvement in the NRMSE. Fine-tuning the FA model reduced the NRMSE from 7.9 ± 2.5 % to 5.7 ± 1.1% ($t = 20.1 \ - \ p \ll 0.01$, $SMD = 1.0$).

Furthermore, as table III demonstrates the pre-trained FAN model yielded lower NRMSE for HC than for patients. In particular, there was a small, but significant difference in the NRMSE yielded for HC participants vs. individuals PS ($t = -4.5 \ - \ p \ll 0.01$, $SMD = 0.2$); and a small, but significant difference in the NRMSE yielded for HC participants vs. individuals with ALS ($t = -7.3 \ - \ p \ll 0.01$, $SMD = 0.3$). Finally, as table III shows, fine-tuning the FAN model with representative data increased the difference in the NRMSE yielded for HC and patients. In particular, there was a large, significant difference in the NRMSE yielded for HC participants vs. individuals PS ($t = -14.1 - p \ll 0.01$, $SMD = 1.1$); and a large, significant difference in the NRMSE yielded for HC participants vs. individuals with ALS ($t = -14.5 \ - \ p \ll 0.01$, $SMD = 1.0$).

B. Automatic detection of neurological diseases

1) Detection of stroke from video-based monitoring: Fig. 3 presents the ROC curve of RF classifiers trained to detect individuals PS and described in Table II. Results obtained

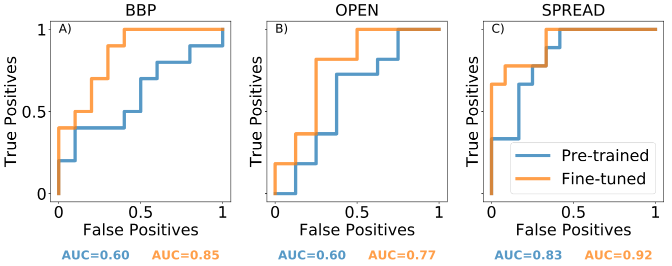


Fig. 3. ROC curves demonstrating the ability of the RF classifiers to distinguish between HC participants and PS patients. Classifiers were trained with features obtained with landmarks yielded by the pre-trained and fine-tuned FAN models applied to videos of participants performing A) BBP, B) OPEN and C) SPREAD tasks.

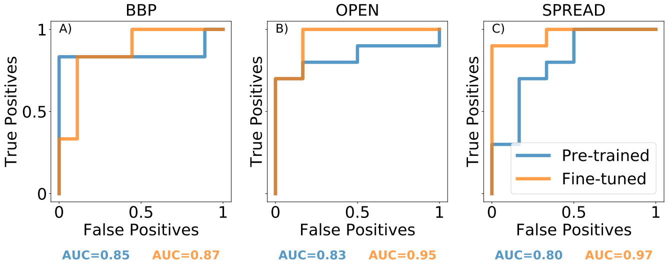


Fig. 4. ROC curves demonstrating the ability of the RF classifiers to distinguish between HC participants and ALS patients. Classifiers were trained with features obtained with landmarks yielded by the pre-trained and fine-tuned FAN models applied to videos of participants performing A) BBP, B) OPEN and C) SPREAD tasks.

from facial landmarks-based features yielded by the pre-trained FA model are presented in blue lines, and those yielded by the fine-tuned FA model are presented with orange lines. Fig. 3 A) present the results for BBP task, B) for OPEN task, and C) for SPREAD task.

As Fig. 3 shows, fine-tuning the FA model with representative data improved the ability of the RF classifier to distinguish between HC participants and individuals PS for all tasks and measured by the AUC of the ROC. In particular, for the BBP task, the AUC of the ROC curve improved from 0.60 to 0.85 by fine-tuning the FA model. For the OPEN task, the AUC of the ROC curve improved from 0.60 to 0.77 by fine-tuning the FA model. And finally, for the SPREAD task, the AUC of the ROC curve improved from 0.83 to 0.92 by fine-tuning the FA model.

2) *Detection of ALS from video-based monitoring*: Fig. 4 presents the ROC curve of RF classifiers trained to detect individuals with ALS and described in Table II. Results obtained from facial landmarks-based features yielded by the pre-trained FA model are presented in blue lines, and those yielded by the fine-tuned FA model are presented with orange lines. Fig. 4 A) present the results for BBP task, B) for OPEN task, and C) for SPREAD task.

As Fig. 4 shows, fine-tuning the FA model with representative data improved the ability of the RF classifier to distinguish between HC participants and individuals with ALS for all tasks and measured by the AUC of the ROC. In particular, for the BBP task, the AUC of the ROC curve improved from 0.87 to 0.87 by fine-tuning the FA model. For the OPEN task, the AUC of the ROC curve improved from 0.83 to 0.95 by fine-tuning the FA model. And finally, for the SPREAD task, the AUC of the ROC curve improved from 0.80 to 0.97 by fine-tuning the FA model.

IV. Discussion

Video-based, automatic detection of neurological diseases can revolutionize the diagnosis and monitoring of neurological conditions such as stroke and ALS, as it facilitates objective assessment of disease status potentially across various residential and clinical settings. In our previous work, we demonstrated the feasibility of applying FA technology for disease detection in individuals PS [12], individuals with ALS [14], and individuals suffering from Parkinson's disease [22]. Furthermore, we showed that fine-tuning FA models with representative, clinical data can improve the model performance in older adults with dementia [37], individuals with facial palsy [29], PS, and ALS [39]. Based on these results, we hypothesized that is possible to fine-tune an FA model with data from multiple clinical groups to obtain a FA model that yields improved results across multiple clinical populations, and that the improved FA model would lead to an improved ability to detect the disease from video-based monitoring. The results presented here support these hypotheses, and represent an important step towards the translation of FA technology for diagnosis of neurological diseases from laboratory-based methods to clinically useful tools.

A. Fine-tuning FA model with diverse dataset

After fine-tuning the deep-learning based FA model with a database of manually annotated video frames from HC participants, individuals with ALS, and individuals PS we observed a significant improvement in model performance for all three groups. However, the FA model performance improvement was greater for HC participants than for the clinical groups; the mean NRMSE improved by 36.5% for HC participants, 27.2% for ALS participants, and 19.5% for the PS group. Furthermore, we observed that fine-tuning the FA model with the Toronto NeuroFace dataset magnified the FA model's bias against clinical populations as measured by the standard mean difference. The pre-trained FA model showed a small (but statistically significant) difference in the NRMSE obtained for HC and patients whereas the FA model showed a large difference in the NRMSE obtained for HC and patients.

The results fit well with our understanding on how deep neural networks for FA learn from new data. After fine-tuning an FA model with representative images, the model gains additional information about 1) subjects' pose and expressions, 2) images illumination and background, 3) the differences in manual annotations between our database and the original database used for training and pre-trained the model, and 4) facial abnormalities induced by the disease [29]. In this case, all the videos were recorded under similar conditions and landmarks were manually localized by the same annotator. Thus, the model learns about the first three aspects from all the training images; in contrast, the FA model gains information about disease specific facial abnormalities from a small subset of the images in the database, likely justifying the sharp differences in performance between HC and patients observed with the fine-tuned model.

Results presented here agreed with our original hypothesis that is possible to obtain a more general FA model for clinical populations by fine-tuning the model with images from people with multiple neurological diseases. They also showed that including a large number of images from healthy controls in the database might help to teach the model about aspects such as illumination, background, and manual annotations. This is an important observation as there is typically abundant data from age-matching healthy controls available, but collecting patients' data can be challenging. Our results suggest that including images of healthy controls and patients recorded under similar conditions in the databases used to fine-tuning FA model might be beneficial.

B. Detection of neurological diseases

An important contribution of this study was to demonstrate that fine-tuning FA models with representative data can lead to improved detection of neurological diseases from video-based monitoring. Our classification results showed that landmarks-based facial features yielded by a fine-tuned FA model provided better detection of individuals PS and individuals with ALS for all the orofacial tasks analyzed in this study.

The tasks SPREAD, BBP, and OPEN provided the best, middle, and worst classification results, respectively, for detection of individuals PS using landmarks-based facial features. These results agree with our understanding of the typical sequelae associated with cerebrovascular accidents. First, stroke survivors typically develop unilateral facial paralysis [56], which is characterized by decreased facial symmetry during movement, and affects the patients' ability to smile [57]. Second, speech movements are commonly affected in stroke [58]. And finally, the facial paralysis observed in stroke survivors does not typically affect the jaw muscles used to open and close the mouth [59].

The tasks SPREAD, OPEN, and BBP provided the best, middle, and worst classification results, respectively, for detection of individuals with ALS using landmarks-based facial features. However, SPREAD and OPEN provided similar classification results (AUC of the ROC equal to 0.95 and 0.97 respectively). These results might be related to the fact that the speech task (BBP) involves more complex facial movements that the non-speech tasks (OPEN and SPREAD) so that the simple feature set used for disease detection might not be able to successfully capture the differences between patients and HC during the execution of the speech task.

Comparing the results of both classification tasks directly is not possible because they used different feature sets. Nevertheless, we observed better classification performance in the detection of individuals with ALS than in the detection of individuals PS for all tasks.

C. limitations

This study has two main limitations. Firstly, data from HC, individuals PS, and with ALS were recorded under tightly controlled pose, background, and illumination conditions. These laboratory conditions might be difficult to reproduce in more natural setting such as home recordings. Thus, it is likely that the FA model fine-tuned with the Toronto NeuroFace dataset will yield higher landmark localization error when applied to photographs and videos recorded under different conditions.

Second, participants were asked to look straight at the camera during task execution. We observed that maintaining this posture was difficult for some participants (both HC and patients) and they continuously turned their bodies or heads and looked away from the camera. Videos where the participant did not directly face the camera were not used in classification analysis as it was difficult to compare the differences in left and right facial movements. To alleviate this experimental limitation, we are developing a custom software application to provide real-time feedback on the participants pose. We believe that such visual feedback will help participants to maintain the correct head pose during task execution.

V. Conclusions

We demonstrated that fine-tuning a deep learning-based FA model with a diverse database composed of manually annotated facial images from healthy controls and individuals with multiple neurological disorders that affect orofacial movements, produces a FA model with significantly improved performance for all clinical and non-clinical groups. We also demonstrated that using the fine-tuned FA model results in better disease detection from video-based monitoring as compared to the results provided by a pre-trained FA model. These results provide some important guidelines for fine-tuning FA models to improve their performance in clinical populations, and validate the clinical importance of fine-tuning FA models with representative data when applying this technology for automatic monitoring and assessment of neurological diseases.

References

[1] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," International Journal of Computer Vision, vol. 127, no. 2, pp. 115–142, 2019.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[4] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 356–370, 2018.

[5] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5562–5570.

[6] R. Walecki, V. Pavlovic, B. Schuller, M. Pantic et al., "Deep structured learning for facial action unit intensity estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3405–3414.

[7] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: development and applications to human computer interaction." in 2003 Conference on computer vision and pattern recognition workshop, vol. 5. IEEE, 2003, pp. 53–53.

[8] M. C. Fysh and M. Bindemann, "Human–computer interaction in face matching," Cognitive science, vol. 42, no. 5, pp. 1714–1732, 2018.

[9] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face – pain expression recognition using active appearance models," Image and Vision Computing, vol. 27, no. 12, pp. 1788 – 1796, 2009.

[10] P. Dutta et al., "Facial pain expression recognition in real-time videos," Journal of healthcare engineering, vol. 2018, 2018.

[11] M. Lee, L. Kennedy, A. Girgensohn, L. Wilcox, J. S. E. Lee, C. W. Tan, and B. L. Sng, "Pain intensity estimation from mobile video using 2d and 3d facial keypoints," arXiv preprint arXiv:2006.12246, 2020.

[12] A. Bandini, J. R. Green, B. Richburg, and Y. Yunusova, "Automatic detection of orofacial impairment in stroke." in Interspeech, 2018, pp. 1711–1715.

[13] N. Eichler, H. Hel-Or, I. Shimshoni, D. Itah, B. Gross, and S. Raz, "3d motion capture system for assessing patient motion during fugl-meyer stroke rehabilitation testing," IET Computer Vision, vol. 12, no. 7, pp. 963–975, 2018.

[14] A. Bandini, J. R. Green, B. Taati, S. Orlandi, L. Zinman, and Y. Yunusova, "Automatic detection of amyotrophic lateral sclerosis (als) from video-based analysis of facial movements: speech and non-speech tasks." in 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 150–157.

[15] Y. Yunusova, E. K. Plowman, J. R. Green, C. Barnett, and P. Bede, "Clinical measures of bulbar dysfunction in als," Frontiers in neurology, vol. 10, 2019.

[16] A. Bandini, J. R. Green, J. Wang, T. F. Campbell, L. Zinman, and Y. Yunusova, "Kinematic features of jaw and lips distinguish symptomatic from presymptomatic stages of bulbar decline in amyotrophic lateral sclerosis," Journal of Speech, Language, and Hearing Research, vol. 61, no. 5, pp. 1118–1129, 2018.

[17] J. Wang, P. V. Kothalkar, M. Kim, A. Bandini, B. Cao, Y. Yunusova, T. F. Campbell, D. Heitzman, and J. R. Green, "Automatic prediction of intelligible speaking rate for individuals with als from speech acoustic and articulatory samples," International journal of speech-language pathology, vol. 20, no. 6, pp. 669–679, 2018.

[18] A. Bandini, J. R. Green, L. Zinman, and Y. Yunusova, "Classification of bulbar als from kinematic features of the jaw and lips: Towards computer-mediated assessment." in INTERSPEECH, 2017, pp. 1819–1823.

[19] A. Bandini, F. Giovannelli, S. Orlandi, S. D. Barbagallo, M. Cincotta, P. Vanni, R. Chiaramonti, A. Borgheresi, G. Zaccara, and C. Manfredi, "Automatic identification of dysprosody in idiopathic parkinson's disease," Biomedical Signal Processing and Control, vol. 17, pp. 47–54, 2015.

[20] A. Bandini, S. Orlandi, F. Giovannelli, A. Felici, M. Cincotta, D. Clemente, P. Vanni, G. Zaccara, and C. Manfredi, "Markerless analysis of articulatory movements in patients with parkinson's disease," Journal of Voice, vol. 30, no. 6, pp. 766–e1, 2016.

[21] A. Bandini, S. Orlandi, H. J. Escalante, F. Giovannelli, M. Cincotta, C. A. Reyes-Garcia, P. Vanni, G. Zaccara, and C. Manfredi, "Analysis of facial expressions in parkinson's disease through video-based automatic methods," Journal of neuroscience methods, vol. 281, pp. 7–20, 2017.

[22] D. L. Guarin, A. Dempster, A. Bandini, Y. Yunusova, and B. Taati, "Estimation of orofacial kinematics in parkinson's disease: Comparison of 2d and 3d markerless systems for motion tracking," arXiv preprint arXiv:2003.08048, 2020.

[23] D. L. Guarin, J. Dusseldorp, T. A. Hadlock, and N. Jowett, "A machine learning approach for automated facial measurements in facial palsy," JAMA facial plastic surgery, vol. 20, no. 4, pp. 335–337, 2018.

[24] J. J. Greene, J. Tavares, D. L. Guarin, N. Jowett, and T. Hadlock, "Surgical refinement following free gracilis transfer for smile reanimation," Annals of plastic surgery, vol. 81, no. 3, pp. 329–334, 2018.

[25] J. J. Greene, J. Tavares, D. L. Guarin, and T. Hadlock, "Clinician and automated assessments of facial function following eyelid weight placement," JAMA Facial Plastic Surgery, vol. 21, no. 5, pp. 387–392, 2019.

[26] J. R. Dusseldorp, M. M. van Veen, D. L. Guarin, O. Quatela, N. Jowett, and T. A. Hadlock, "Spontaneity assessment in dually innervated gracilis smile reanimation surgery," JAMA facial plastic surgery, vol. 21, no. 6, pp. 551–557, 2019.

[27] J. J. Greene, D. L. Guarin, J. Tavares, E. Fortier, M. Robinson, J. Dusseldorp, O. Quatela, N. Jowett, and T. Hadlock, "The spectrum of facial palsy: The meei facial palsy photo and video standard set," The Laryngoscope, vol. 130, no. 1, pp. 32–37, 2020.

[28] D. L. Guarin, Y. Yunusova, B. Taati, J. R. Dusseldorp, S. Mohan, J. Tavares, M. M. van Veen, E. Fortier, T. A. Hadlock, and N. Jowett, "Toward an automatic system for computer-aided assessment in facial palsy," Facial Plastic Surgery & Aesthetic Medicine, vol. 22, no. 1, pp. 42–49, 2020.

[29] D. L. Guarin, B. Taati, T. Hadlock, and Y. Yunusova, "Automatic facial landmark localization in clinical populations–improving model performance with a small dataset," Journal of Neuroengineering and Rehabilitation, 2020 - submitted for publication.

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[31] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in 2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, 2011, pp. 2144–2151.

[32] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.

[33] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 50–58.

[34] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," Image and vision computing, vol. 47, pp. 3–18, 2016.

[35] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.

[36] Z.-H. Feng, J. Kittler, M. Awais, and X.-J. Wu, "Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks," International Journal of Computer Vision, pp. 1–20, 2019.

[37] B. Taati, S. Zhao, A. B. Ashraf, A. Asgarian, M. E. Browne, K. M. Prkachin, A. Mihailidis, and T. Hadjistavropoulos, "Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia," IEEE Access, vol. 7, pp. 25 527–25 534, 2019.

[38] A. Asgarian, S. Zhao, A. B. Ashraf, M. Erin Browne, K. M. Prkachin, A. Mihailidis, T. Hadjistavropoulos, and B. Taati, "Limitations and biases in facial landmark detection d an empirical study on older adults with dementia," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 28–36.

[39] A. Bandini, S. Rezaei, D. L. Guarın, M. Kulkarni, D. Lim, M. I. Boulos, L. Zinman, Y. Yunusova, and B. Taati, "A new dataset for facial motion analysis in individuals with neurological disorders," IEEE Journal of Biomedical and Health Informatics, 2020 - submitted for publication.

[40] T. Panch, H. Mattie, and R. Atun, "Artificial intelligence and algorithmic bias: implications for health systems," Journal of Global Health, vol. 9, no. 2, 2019.

[41] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO). IEEE, 2017, pp. 1–7.

[42] A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[43] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2799–2804. [Online]. Available: https://www.aclweb.org/anthology/D18-1302

[44] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, "Mitigating gender bias in natural language processing: Literature review," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1630–1640. [Online]. Available: https://www.aclweb.org/anthology/P19-1159

[45] J. R. Duffy, "Motor speech disorders: clues to neurologic diagnosis," in Parkinson's Disease and Movement Disorders. Springer, 2000, pp. 35–53.

[46] Y. Yunusova, J. R. Green, J. Wang, G. Pattee, and L. Zinman, "A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (als)," JoVE (Journal of Visualized Experiments), no. 48, p. e2422, 2011.

[47] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," Journal of the American Geriatrics Society, vol. 53, no. 4, pp. 695–699, 2005.

[48] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 79–87.

[49] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 88–97.

[50] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," Image and Vision Computing, vol. 47, pp. 27–35, 2016.

[51] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," Image and Vision Computing, vol. 28, no. 5, pp. 807–813, 2010.

[52] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2235–2245.

[53] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6971–6981.

[54] J. Cohen, Statistical power analysis for the behavioral sciences. Routledge, 2013.

[55] S. S. Sawilowsky, "New effect size rules of thumb," Journal of Modern Applied Statistical Methods, vol. 8, no. 2, p. 26, 2009.

[56] R. F. Baugh, G. J. Basura, L. E. Ishii, S. R. Schwartz, C. M. Drumheller, R. Burkholder, N. A. Deckard, C. Dawson, C. Driscoll, M. B. Gillespie et al., "Clinical practice guideline: Bell's palsy," Otolaryngology–Head and Neck Surgery, vol. 149, no. 3_suppl, pp. S1–S27, 2013.

[57] M. M. van Veen, J. Tavares-Brito, B. M. van Veen, J. R. Dusseldorp, P. M. Werker, P. U. Dijkstra, and T. A. Hadlock, "Association of regional facial dysfunction with facial palsy–related quality of life," JAMA Facial Plastic Surgery, vol. 21, no. 1, pp. 32–37, 2019.

[58] H. L. Flowers, F. L. Silver, J. Fang, E. Rochon, and R. Martino, "The incidence, co-occurrence, and predictors of dysphagia, dysarthria, and aphasia after first-ever acute ischemic stroke," Journal of communication disorders, vol. 46, no. 3, pp. 238–248, 2013.

[59] R. T. Manktelow, L. R. Tomat, R. M. Zuker, and M. Chang, "Smile reconstruction in adults with free muscle transfer innervated by the masseter motor nerve: effectiveness and cerebral adaptation," Plastic and reconstructive surgery, vol. 118, no. 4, pp. 885–899, 2006.