IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 12, 2019, accepted September 13, 2019, date of publication September 18, 2019, date of current version October 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942143

Automatic Facial Paralysis Evaluation Augmented by a Cascaded Encoder Network Structure

TING WANG[1], SHU ZHANG[2], LI'AN LIU[3], GENGKUN WU[1], AND JUNYU DONG[2], (Member, IEEE)

[1] Department of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[2] Department of Information Science and Engineering, Ocean University of China, Qingdao 266100, China
[3] Qingdao Hiser Medical Center, Qingdao 266034, China

Corresponding author: Shu Zhang (zhangshu@ouc.edu.cn)

This work was supported by National Natural Science Foundation of China (41906177), the Project of Shandong Province Higher Educational Science and Technology Program (J17KA076), the Scientific Research Foundation of Shandong University of Science and Technology for Recruited Talents (2017RCJJ047), China Postdoctoral Science Foundation Grant (2019M652476), Shandong Provincial Natural Science Foundation (ZR2018BD029), and Qingdao Postdoctoral Applied Research Project (2018yyyj65).

ABSTRACT Facial paralysis refers to a facial nerve disordering, with which people may lose the abilities to accurately control their facial muscles for certain facial performances. The diagnosis of such disordering is mainly based on the observation of patient's face in terms of the facial spatial information, such as facial asymmetry. Up to now, this area is still dominated by therapists' subjective examinations clinically. Therefore, automations for this task receive wide attentions in both academic and industrial fields. Recently, the deep learning based methods, the convolutional neural networks (CNNs) more specifically, demonstrate their competitive performance compared with traditional approaches in many areas. However, due to the lack of the structured/labelled facial paralysis data as training data, those deep learning based solutions are still not able to fully attach their superiorities to the facial paralysis evaluation tasks. Another essential aspect for automation in facial paralysis analysis is the facial spatial information extraction. Semantic segmentation is a better choice than traditional template-based facial landmark detection for analysing facial paralysis images, which contain faces in uncommon patterns. However, most existing semantic segmentation approaches are made for indoor or outdoor scene parsing. To this end, this paper presents a deep learning-based approach for automatic facial paralysis grading prediction. The proposed model utilizes a cascaded encoder structure, which explores the advantages of the facial semantic feature for facial spatial information extraction, and then benefits the facial paralysis assessment. A dual-stage cascaded training process is adopted to utilize a mixture of normal and paralysed faces as training data, which exports a well-trained deep neural network model for facial paralysis evaluation. Experiments are conducted in two aspects to demonstrate the performance of each components of the proposed model. Encouraging results are illustrated compared with several existing approaches in the related areas.

INDEX TERMS Facial nerve paralysis, dual-stage, cascaded neural network, facial attribute segmentation, paralysis grading prediction.

I. INTRODUCTION

The evaluation of the facial paralysis requires the measurements for the facial symmetry of patients while they are performing certain facial expressions. To achieve automations in such process, the spatial information related to facial attributes including eyes, nose, mouth and other facial key regions should be extracted and analysed robustly and automatically. In the past few years, the detection and localization of the facial attributes are mostly achieved based on the facial landmarks [1] extracted using algorithms such as CLM [2], SDM [3], AAM [4], [5], and many others [6]. However, for facial paralysis evaluations that may involve uncertain facial conditions or severe asymmetric facial expressions, those template-based facial landmark detections can fail easily. It poses challenges for automatic facial paralysis grading prediction tasks.

Another way to address the automatic facial paralysis analysis problem, is through full image analysis. Most recently, deep learning based methods, and more specifically the con-

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

volutional neural network (CNN) based approaches become popular for many tasks in different fields, such as remote sensing [7], brain-computer interaction [8] and semantic segmentation [9] among others. They demonstrate their promising performance compared against traditional methods by full image analysis for feature extraction. However, to accomplish the facial paralysis evaluation based on CNN, it needs a large number of labelled facial paralysis images as training data for deep model establishment. Unfortunately, there is still no such dataset publicly available. It brings obstacle to attach the state-of-the-art performance of the deep learning-based approaches to the automatic facial paralysis grading problem.

To tackle those problems, this paper explores the possibilities to utilize a deep neural network model based on a dual-stage cascaded encoder strategy for facial paralysis analysis. Researches show that the semantic segmentations for different facial attribute regions can contribute to facial symmetry measurement by providing accurate facial attribute shapes, locations and the geometric relationship between attributes. However, existing semantic segmentation solutions [9]–[11] are mostly focusing on the targets included in indoor or outdoor scene. Most recently, there are also a few researches start to set foot into the facial attribute segmentation area for normal human faces [12], [13]. This paper further extends the semantic segmentation into the field of facial paralysis analysis with limited facial paralysis data. The facial semantic features are utilized to contribute to the facial paralysis grading prediction. A cascaded training scheme presented in this paper can effectively address the insufficient training facial paralysis training data problem.

The contributions of this paper are as follows:

1) **A cascaded encoder strategy for facial paralysis grading prediction.** This paper introduces a dual-encoder structure to better utilize the facial semantic features beneficial to the facial paralysis grading task. The first encoder can produce semantic feature maps supporting the facial attribute segmentation. The facial semantic features can deliver rich facial spatial information, which is essential for facial paralysis evaluation.
2) **A cascaded training scheme to better utilize the limited number of facial paralysis data.** Unlike normal facial images that are abundant and publicly available in many open datasets, the facial paralysis images are less common and still no open accessible datasets for them. Overfitting can be easily encountered if only utilize those limited number of facial paralysis data as training data. A two-stage scheme can well address this problem by a cascaded training process.
3) **Experiments with real-world data demonstrate the competitive performance of the proposed model.** Encouraging performance are concluded with the visual and statistical comparisons against several existing approaches in the related fields.

The rest of the paper is organized as follows: Section II discusses the background researches related to the proposed study; Section III introduce the proposed model in two stages; Section IV discusses some details in the model training process; Section V demonstrates the performance of the proposed method compared with several existing approaches; and the paper is finally concluded in Section VI.

II. RELATED WORK

Facial paralysis refers to a facial nerve disordering, with which people have difficulties to properly control their facial muscles to accurately perform certain facial motions. It has various origins such as traumatic, idiopathic, congenital or toxic among others. The scales of the facial muscle control malfunctions can be variable. For example, Chevalier *et al.* [14], Brackmann and House [15] scales, or ‘‘Sydney’’ and ‘‘Sunnybrokk’’ [16] scales. No matter what grading standard is used, the facial paralysis evaluation still highly relies on the close observation of the facial spatial information by the therapists clinically. Thus, it motivates a range of automatic facial paralysis analysis researches that try to relieve the manpower required in those tasks.

A. AUTOMATIC FACIAL PARALYSIS ANALYSIS

People with facial nerve paralysis commonly are incapable of performing certain facial expressions accurately. Therefore, most existing facial paralysis evaluation methods target at identifying the spatial information like asymmetries demonstrated in the paralysed faces. For example, Hsu *et al.* [17] presented an deep network for facial palsy analysis. Their method relied on the facial landmark localization based on line segmentation strategy. Wang *et al.* [18] proposed an automatic facial paralysis degree evaluation combining the static and dynamic facial asymmetric features. They utilized facial landmark detection to establish a facial mesh to describe distinguishable facial regions for asymmetry calculation. Encouraging results were demonstrated. However, the performance highly relies on the accuracy of the results from the facial landmark detection algorithm. Sajid *et al.* [19] presented a solution for automatic facial paralysis evaluation. They utilize a CNN for paralysis scale prediction, which was trained by the augmented training data exported from a Generative adversarial Network (GAN). They addressed the problem of insufficient facial paralysis data as training data by a GAN-based image synthesis. However, using such synthetic training data can potential jeopardize the distribution modelling process for the real-world natural data. Samsudin *et al.* [20] also introduced an automatic method for this task. They used optical flow extracted in the images to provide objective paralysis evaluation on House-Brackmann grading system. The optical flow in their method described the facial movements, which were used to export the measurements of the facial asymmetry in terms of distance and area. However, as they mentioned in their paper, the unintentional head movements could produce incorrect estimation of the optical flow, which then led to some limitations in their method. More solutions for this area can be found [21], [22]. Most of them still highly relied on the traditional facial landmark detection algorithms for facial spatial information extraction. However, those facial landmark detection methods were less accurate when applied to facial paralysis images with uncommon facial geometries.

One way to address this issue, is through deep learning based semantic segmentation for facial attribute extraction.

B. LEARNING-BASED SEMANTIC SEGMENTATION FOR FACIAL ATTRIBUTES

The goal of the semantic segmentation is to estimate pixel-level label prediction for a image. It is treated as a dense classification problem that plays a very important role in image understanding. From the past few years, CNNs have demonstrated their advantages for semantic segmentation tasks. For example, MaskRCNN [11] tries to extend the Region-based Convolutional Neural Networks from object detection field [23], [24] into the semantic segmentation areas. It utilizes an intermediate network of Region Proposal Network (RPN) with attention mechanism [25], [26] to achieve instance-level semantic segmentation. It is the most successfully solution at that time, even for some applications at present. Recently, Fully Convolutional Neural Networks (FCNNs) [9], [27] become a group of popular architectures in this field due to their advantages in effective feature generation and end-to-end training. FCNNs only utilize ConNet layers (with basic components of convolution, pooling and activation function) for feature extraction. They can simultaneously simplify and speed up the learning and the inference for the semantic segmentation model. DeepLab [9] is another widely acknowledged solution for semantic segmentation up to now. It brings the concept of atrous/dilated convolution to the deep learning community. This concept can introduce larger field-of-view during the feature extraction without increasing the number of parameters or the amount of computations. It thus can produce a larger feature map, which is much beneficial to semantic segmentation tasks.

In the meanwhile, attention mechanism [25], [26] starts to become a helpful tool for semantic segmentation tasks. Attention mechanism tries to simulate human behaviours when conducting certain learning-based tasks. Human commonly focuses his attention on certain region of interest and ignores the surroundings that are far from this region. It can improve the efficiency for accomplishing the tasks. There are a range of researches [28]–[31] adopting this concept for a better performance. They tries to model long-range dependencies for distant pixels in a image to describe saliency region. Wang *et al.* [32] described this attention mechanism as a non-local mean, which is a kind of non-local neural network. Based on this sense, they also presented a spacetime non-local block concept to model the dependencies in sequential data like videos over long-tem time intervals.

The majority of those methods is targeting the indoor or outdoor scene parsing. To achieve the promising performance of those methods, a large number of training data in the target domain is required. However, in practice, the facial paralysis data is much less common. There is even no accessible dataset for such structured paralysed facial images.

There are also some researches proposed for facial feature semantic segmentation [12], [13]. However, they are only targeting the faces from normal people. The paralysed facial images can still pose challenges with uncommon patterns in the faces.

To address aforementioned problems, this paper proposes a dual-stage model and cascaded training strategy for automatic facial paralysis evaluation. The proposed model utilize the encoder trained by facial attribute semantic segmentation task to provide facial spatial information for facial paralysis analysis. The cascaded training can reduce the needs for a very large number of paralysed facial data.

III. METHODOLOGY

A. MODEL OVERVIEW

Inspired by the clinic approaches, the proposed automatic facial paralysis analysis is achieved by utilizing the spatial information delivered by the facial attributes when performing various facial expressions. Instead of the manual observation that is commonly conducted clinically, the proposed method introduces the deep learning into the facial attribute spatial information analysis task, which is then contributed to the facial paralysis evaluation. As shown in Fig. 1, the pipeline of the proposed method consists of two main stages: (**a**) the analysis of the facial attribute spatial information depicted by the semantically segmented facial regions in images; and (**b**) the facial paralysis evaluation based on the exported facial spatial information encoded the first stage.

Unlike some of the existing learning-based facial paralysis analysis methods that rely on an end-to-end training strategy, the proposed method utilizes a intermediate model, which extracts facial attributes though a multi-scale attention module enhanced semantic segmentation. The training process for this model can help to produce a latent feature map z that contains the higher-level spatial information supporting the facial attribute segmentation. As a condition map, this featurized spatial information is then imported to the following stage along with the input image for facial paralysis evaluation. z provides meaningful and measurable information for facial spatial information analysis, which is more robust than the vague representation of common features that are learnt in an end-to-end manner. To be more specific, the facial attribute segmentation result is not used to help the facial paralysis grading prediction, but the extracted facial spatial latent feature does, as demonstrated in Fig. 1. The proposed cascaded strategy thus can enhance the performance of facial paralysis evaluation.

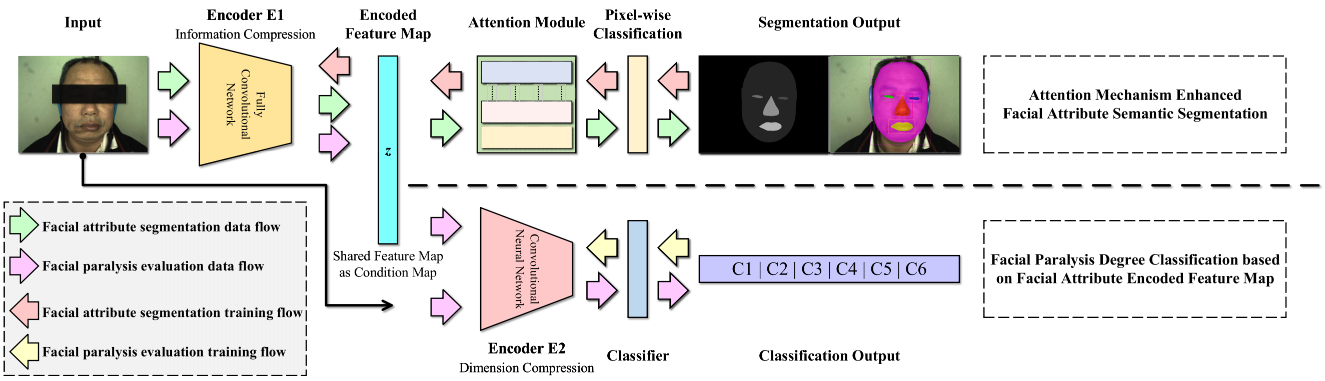

Input
Encoder E1
Information Compression
Fully Convolutional Network
Encoded Feature Map
z
Attention Module
Pixel-wise Classification
Segmentation Output
Attention Mechanism Enhanced Facial Attribute Semantic Segmentation
Facial attribute segmentation data flow
Facial paralysis evaluation data flow
Facial attribute segmentation training flow
Facial paralysis evaluation training flow
Shared Feature Map as Condition Map
Convolutional Neural Network
Encoder E2
Dimension Compression
Classifier
C1 | C2 | C3 | C4 | C5 | C6
Classification Output
Facial Paralysis Degree Classification based on Facial Attribute Encoded Feature Map

FIGURE 1. The pipeline of the proposed facial paralysis evaluation strategy. It consists of two components, a facial attribute semantic segmentation network, and a facial paralysis evaluation network. The two components share a same encoded feature map extracted by E1. This shared feature map contains the compressed spatial information of facial attributes. Please note that the funnel shape of E1 indicates the compression of information, while the funnel shape of E2 refers to the spatial dimension reduction.

B. ATTENTION ENHANCED SEMANTIC FACIAL FEATURE SEGMENTATION

The first stage in the pipeline is the semantic segmentation for facial attributes. As discussed in Sec. II, the CNN-based semantic segmentation has demonstrated its competitive performance in both outdoor and indoor scene parsing in the past few years. They commonly formulate a pair of an encoder and a pixel-wise classifier in various forms [9]–[11], [27], [28], [33]–[35]. The encoder extracts a set of feature maps to describe the semantic information, while the classifier transforms those feature maps into a pixel-by-pixel classification over the image. In this paper, we further extend the semantic segmentation into the field of facial attribute extraction. Since the spatial information are encoded in the latent features produced by the encoder, we can thus utilize those encoded features for facial spatial information analysis, and then achieve facial paralysis evaluation.

The architecture for the proposed model in this stage is shown in Fig. 2. A fully convolutional network (FCN) is utilized as an encoder to extract the latent spatial features for facial images. We introduce the dilated residual blocks in the encoder network for pixel-level facial feature predictions. Inspired by [9], the pretrained ResNet-101 without downsampling operations is adopted to provide a weight initialization for the encoder network. The atrous convolutions are also used to replace the traditional convolutional operations to preserve more details along the way across the layers.

Following the encoded feature map z, a multi-scale attention module is composed to produce the potential facial attribute locations from different levels. In classic solutions for semantic segmentation, the pixels that belong to one class can still have intra-class difference, especially for pixels that have large spatial displacements with each other. In the last few years, attention module [25] becomes widely adopted to address this problem due to its ability for modelling long-range dependencies. It is now successfully enhance the tasks such as text understanding [25], image generation [26], and image question answering [36] among others.

The attention module tries to model the pixel dependencies from distant pixels. It takes the feature map $x \in \mathbf{R}^{C\times H\times W}$ from the previous layers as the input. The feature map x is firstly transformed into $Q_{spl}(x)$, $K_{spl}(x)$ and $V_{spl}(x)$ using three Conv-BN-ReLU (Convolution, Batch Normalization and Rectified Linear Unit) blocks respectively, where $Q_{spl}(x) = W_q x$, $K_{spl}(x) = W_k x$, $V_{spl}(x) = W_v x$. We reshape the feature map $Q_{spl}(x)$, $K_{spl}(x)$ and $V_{spl}(x)$ from $\mathbf{R}^{C\times H\times W}$ to $\mathbf{R}^{C\times D}$ where $D = H \times W$. Then a weight $s_{j,i}$ indicate spatial attention on i^{th} location when predicting on j^{th} region.

$$s_{j,i} = \frac{exp[Q_{spl}(x)_i^T K_{spl}(x)_j]}{\Sigma_{i=1}^{D}\{exp[Q_{spl}(x)_i^T K_{spl}(x)_j]\}} \tag{1}$$

$s_{j,i}$ can form up an attention map s. Based on this attention map, a spatial attention feature map $A^{spl} = [A_1^{spl}, A_2^{spl}, \ldots, A_j^{spl}, \ldots, A_D^{spl}] \in \mathbf{R}^{C\times D}$ can be formulated

$$A_j^{spl} = \lambda_{spl}\, \Sigma_{i=1}^{D}(s_{j.i} V_{spl}(x)_i) \tag{2}$$

where a scale parameter λ_{spl} is adopted to enable the module assigning more weight to the non-local evidence. It is initialized by 0, and gradually learnt during the training process. The spatial attention feature map A^{spl} is achieved by a matrix multiplication between $V_{spl}(x)$ and transposed attention map s. S is then reshaped back from $\mathbf{R}^{C\times D}$ into $\mathbf{R}^{C\times H\times W}$.

To properly model the attentive information for unconstrained faces, a multi-scaled spatial attention strategy is adopted in the proposed method in n different scales, as shown in Fig. 2. For each scale, a certain sized convolution layer is employed in aforementioned Conv-BN-ReLU block. For demonstrative purpose, we use 3×3, 5×5, 7×7 sized convolutions for three scales in the model implementation in this paper.

Along with the multi-scaled spatial attentions, a channel attention map is also computed to explore the inter-channel semantic dependencies. As shown in the bottom block in Fig. 2, the input feature $x_{n+1} \in \mathbf{R}^{C\times H\times W}$ from

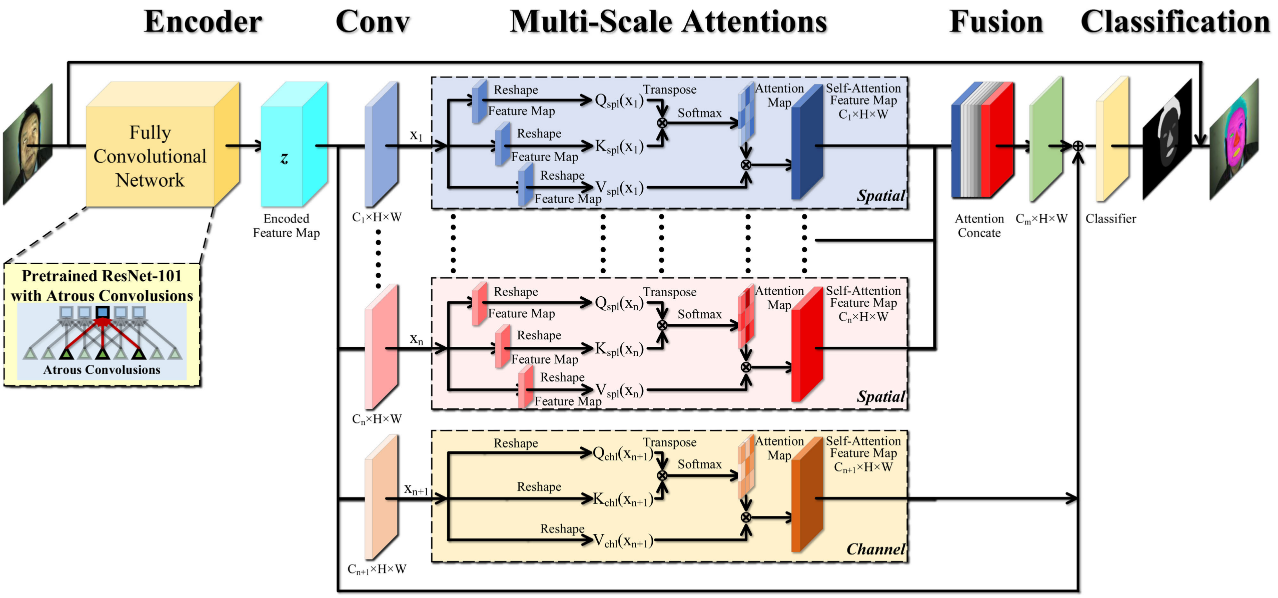


FIGURE 2. The architecture of the facial feature semantic segmentation adopting a multi-scale attention module.

previous layers is firstly reshaped into three feature maps $Q_{chl}(x_{n+1})$, $K_{chl}(x_{n+1})$ and $V_{chl}(x_{n+1})$, where $Q_{chl}(x_{n+1}) \in \mathbf{R}^{C\times D}$, $K_{chl}(x_{n+1}) \in \mathbf{R}^{C\times D}$, $V_{chl}(x_{n+1}) \in \mathbf{R}^{C\times D}$. Then a matrix multiplication between $K_{chl}(x_{n+1})$ and transposed $Q_{chl}(x_{n+1})$ along with a softmax layer can produce a channel attention map $c \in \mathbf{R}^{C\times C}$

$$c_{j,i} = \frac{exp[K_{chl}(x_{n+1})_i Q_{chl}(x_{n+1})_j^T]}{\Sigma_{i=1}^{C}\{exp[K_{chl}(x_{n+1})_i Q_{chl}(x_{n+1})_j^T]\}} \tag{3}$$

where $c_{j,i}$ models the i^{th} channel's impact on predicting the j^{th} channel. A channel attention feature map A^{chl} is then obtained though a matrix multiplication between the transposed c and $V_{chl}(x_{n+1})$.

$$A_j^{chl} = \lambda_{chl}\,\Sigma_{i=1}^{C}(c_{j.i}V_{chl}(x_{n+1})) \tag{4}$$

where λ_{chl} has an initial value of 0, and then gradually learnt by training. The multi-scaled spatial attention feature maps are fused through a concatenation followed by a convolution. Please note that the atrous convolution is adopted in the model to enhance the fine detail in the extracted features. A pixel-wise sum among the fused spatial attention feature map, the channel attention feature map and the original encoded feature map z from the FCN is inputted into a classifier for pixel-wise semantic predictions.

C. CNN AUGMENTED FACIAL PARALYSIS EVALUATION

After z is properly modelled by training the facial attribute semantic segmentation network, the second network can be constructed followed by a classifier for facial paralysis analysis. As shown in Fig. 1, the data flow for facial paralysis evaluation goes through a cascaded encoder structure. The first encoder **E1** is responsible for extracting facial spatial information that describes the facial attribute segmentations. The second encoder is used to export the measurable facial paralysis features based on the facial spatial information. We follow the structure from VGG16 to construct **E2**. The pretrained VGG16 contributes to the parameter initialization for the encoder **E2** when training the facial paralysis evaluation model. The weights in **E2** is fixed during this training process. A classifier using a softmax layer after a set of fully connected layers transforms the features exported by **E2** into facial paralysis gradings.

Clinically, according to House-Brackmann grading system, the facial paralysis is assessed using 6 scales, as shown in Table 1 [37]. During the assessment, the therapist asks facial paralysis patient to perform 5 typical facial expressions including closing eyes, raising eyebrows, opening mouth, wrinkling nose and puffing cheeks. The therapist carefully exams the patient's face when the expressions reach the maximum intensity. Inspiredly, the patient's facial expression images are imported into the proposed model, and then the 6-point based paralysis scale is exported as a grading classification result.

IV. MODEL TRAINING DETAILS

The facial paralysis images are always less common than normal facial images. There are a range of facial image datasets available publicly, yet no facial paralysis dataset can be found open accessible. This situation makes it hard or even impossible to extract competitive performance from an end-to-end facial paralysis grading prediction model training. To this end, a cascaded training scheme is proposed. As illustrated in Fig. 1, the proposed facial paralysis evaluation model utilizes a cascaded encoder strategy, and can produce a 6-point scale facial paralysis assessment prediction using a softmax layer that follows a set of fully connection layers.

TABLE 1. Facial paralysis grading system by house & Brackmann.

	Description	Gross	At rest	Motion	
Grade I	Normal	Normal.	Symmetry.	Normal facial function	
Grade II	Slight Dysfunction	Slight weakness noticeable on close inspection; may have very slight synkinesis.	Normal symmetry and tone.	Forehead Eye Mouth	- moderate to good function; - complete closure with minimum effort; - slight asymmetry.
Grade III	Moderate Dysfunction	Obvious but not disfiguring difference between two sides; noticeable but not severe synkinesis, contracture, and/or hemi-facial spasm.	Normal symmetry and tone.	Forehead Eye Mouth	- slight to moderate movement; - complete closure with effort; - slightly weak with maximum effort.
Grade IV	Moderate Severe Dysfunction	Obvious weakness and/or disfiguring asymmetry.	Normal symmetry and tone.	Forehead Eye Mouth	- none; - incomplete closure; - asymmetric with maximum effort.
Grade V	Severe Dysfunction	Only barely perceptible motion.	Asymmetry.	Forehead Eye Mouth	- none; - incomplete closure; - slight movement.
Grade VI	Total Paralysis	No perceptible motion.	Asymmetry.	No movement	

Those two cascaded encoders are trained separately, as two types of training flow arrows demonstrated in Fig. 1.

In the first training stage, the facial attribute segmentation model is trained using a mixture of training data composed of both normal faces and paralysed faces. The normal faces are from a publicly available facial dataset of *Annotated Facial Landmarks in the Wild* (AFLW) [38]. This dataset contains around 25k facial images with the facial attribute markup ground truth, which can be utilized for model training. We have also collected around 12k facial paralysis images to form up a facial paralysis dataset, and manually marked up the facial features in the images. This dataset is also utilized in the first training stage. The equipment used for the collection of this dataset is shown in Fig. 3.[1] We employ 30 epochs for AFLW dataset, and 20 epochs for our collected facial paralysis dataset during the training process. More specifically, around 2/3 of data in all those datasets are used for training, and the model validation and testing individually consume around 1/6 of data from the datasets.

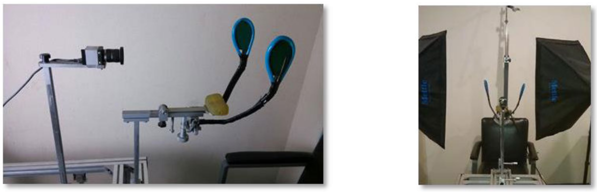


FIGURE 3. The equipment used for facial paralysis data collection, including a camera, a device to restrain head pose, and a pair of soft illumination light sources.

In the second training stage, the facial paralysis evaluation model is trained using our collected facial paralysis dataset only. During the second training stage, the weights for **E1** is fixed. Since the collected facial paralysis dataset only contains 500 patients with 5 expressions for each and around 5 images for each expression, a data augmentation is utilized by randomly rotating and mirroring some of the images. The data proportions for model training, validation and testing are 2/3, 1/6 and 1/6 respectively. Since the two encoders are training separately, we used all the facial paralysis data to train and test both encoders in two stages individually. Particularly, the separation of the paralysis training data is done according to the subject, which means no subjects in the testing set are included in the training or validation set.

We also adopted a poly learning rate strategy with base rate at 0.0002 for a better performance inspired by [9]. Adam solver is used to fulfil the training process.

V. EXPERIMENTS AND EVALUATIONS

To properly evaluate the performance of the proposed method, the experiments are divided into two parts: (**a**) the model effectiveness for facial attribute segmentation; and (**b**) the performance evaluation for facial paralysis grading prediction.

A. FACIAL ATTRIBUTE SEGMENTATION MODEL

The first stage of the proposed model is to semantically segment the image for facial attributes. By achieving this goal an encoder **E1** is constructed to extract the compressed facial semantic feature maps. The multi-scale attention module provides the assistance to highlight the regions of interests when locating the facial features. This **E1** can be helpful for the facial paralysis evaluation task, which relies on the analysis of the facial spatial information. The performance of the model in this stage is evaluated using a large scale of facial images mixed by normal and paralysed faces, including ALFW [38], LFW [44] and the collected facial paralysis dataset. The ground truth of the facial attributes are manually annotated by hired workers. The performance comparisons of the proposed method against a range of existing approaches are illustrated in Table 2. The performance metrics follow the same calculations as in [13].

With the ablation studies included, experiments are also conducted on the proposed method without attention module in the first stage to demonstrate How the different component techniques proposed in the work contribute to the final results. As it can be observed from Table 2, the classification error can be significantly reduced by the employment of the attention module. The attention maps extracted by the attention module as intermediate layers in the first stage

[1] This facial paralysis dataset is currently private according to an agreement with the subjects. We are working on a plan to conditionally release this dataset with consent from the subjects.

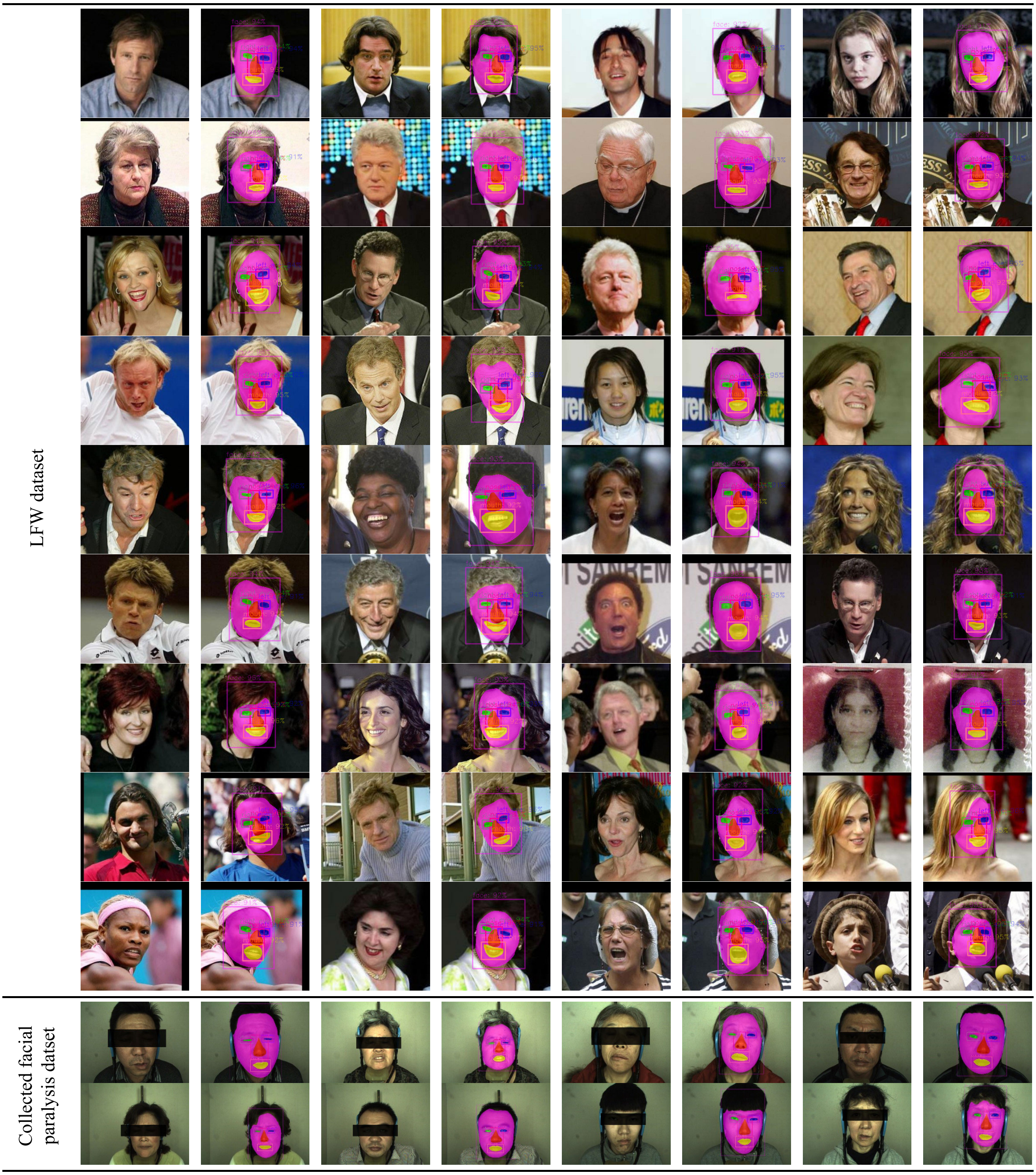


FIGURE 4. Samples of the visual demonstrations for facial attribute segmentations by the proposed method on the datasets of *Labeled Faces in the Wild* (LFW) [44] and our collected facial paralysis dataset. Those facial images contain the faces in different facial expressions (including exaggerated ones), different head poses, different illuminations and different occlusions. The proposed method can capture their facial attributes accurately.

are demonstrated in Fig. 5. The attention is visualized as a heatmap overlapping with the original image. The learnt feature map can provide efficient guidance for the facial attribute segmentation.

The experiment results are also visually demonstrated in Fig. 4. Those tested facial images are from both *Labeled Faces in the Wild* (LFW) dataset [44] and our collected facial paralysis dataset. As it can be observed, the faces in those

TABLE 2. Performance comparisons for facial attribute segmentation.

	Classification Errors (%)
Kumar *et al.* [39]	18.88
Zhang *et al.* [40]	15.00
Liu *et al.* [41]	12.70
Wang *et al.* [42]	12.00
Zhong *et al.* [43]	10.02
Kalayeh *et al.* [13]	8.84
Ours without attention module	12.32
Ours	8.72

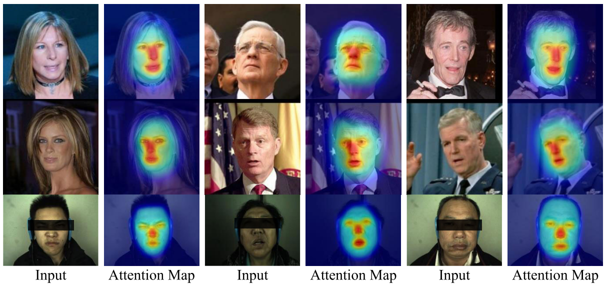


FIGURE 5. The visual demonstrations of the extracted attention map in the first stage of the proposed model. As it can be observed, the visual attention will efficiently guide the semantic segmentation towards a positive prediction.

images are in arbitrary sizes. Various facial expressions are also included. The multi-scaled attention module can perceive those facial features in different scales. In the meanwhile, the proposed model captures the facial attributes accurately even for the faces with various facial expressions.

We also visualize how the training loss is changing along with the increase in the number of training epoches. Following the standard semantic segmentation training strategy, the cross entropy loss is utilized in the training process. The changing curve is demonstrated in Fig. 6. As illustrated, with about 50 epoches, the parameters of the model can be sufficiently optimized.

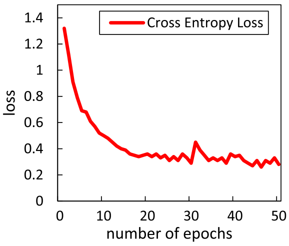


FIGURE 6. The training process for the first stage encoder produces a parameter optimization convergence using no more than 50 epochs with training data mixed by both normal faces and paralysed faces.

B. FACIAL PARALYSIS GRADING MODEL

To make proper performance comparisons with existing methods, we compute four standard evaluation metrics as described in [19], [45]. If TP, TN, FP and FN stand for True Positive, True Negative, False Positive and False Negative respectively, then those four performance measurements **Accuracy**, **Recall**, **Confidence** and **Dice** are as follows:

$$\textbf{Accuracy} = \frac{TP - FN}{TP + FN + FP + TN} \tag{5}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\textbf{Confidence} = \frac{TP}{TP + FP}$$

$$\textbf{Dice} = \frac{TP}{TP + (FN + FP)/2} \tag{7}$$

$$= 2 \times \frac{\textbf{Confidence} \times \textbf{Recall}}{\textbf{Confidence} + \textbf{Recall}} \tag{8}$$

where **Recall** refers to as true positive rate (tpr), and **Confidence** denotes the true positive accuracy (tpa). Considering all four metrics rather than only an accuracy can help to reduce the evaluation bias to an extent. To demonstrate the performance of the proposed model in the task of facial paralysis grading prediction, two types of the training strategy are evaluated:

1) **End-to-End Model Training.** With the traditional end-to-end training strategy, the training process for facial attribute segmentation is omitted. The training for facial paralysis grading model optimizes both two encoders of **E1** and **E2**. For this part of the experiments, only the facial paralysis dataset is utilized to train the model.
2) **Cascaded Model Training.** According to the proposed cascaded training scheme, the training of the facial attribute segmentation model and the training for facial paralysis grading model are two separated processes. The weights for **E1** is inherited from the facial attribute segmentation model and remains fixed during the training for facial paralysis grading model. Training details are discussed in Section IV.

Fig. 7 illustrates the performance comparisons of the proposed model between the end-to-end training strategy and the cascaded training strategy. The cascaded training scheme optimizes the first encoder **E1** using abundant facial images as the training data. The limited number of labelled facial paralysis images are only utilized to train the second encoder **E2**. This two-stage training process leads to a better performance than all-in-one training strategy, which lacks a sufficient number of training samples.

Table 3 demonstrates further performance comparisons of the proposed method with other existing state-of-the-art approaches. As observed, due to the insufficient training data of facial paralysis images, the end-to-end training strategy for the proposed method can result in an acceptable but lowest performance among all solutions. Although the dual-encoder structure in the proposed method can produce useful semantic features for facial images, it however also introduces a large

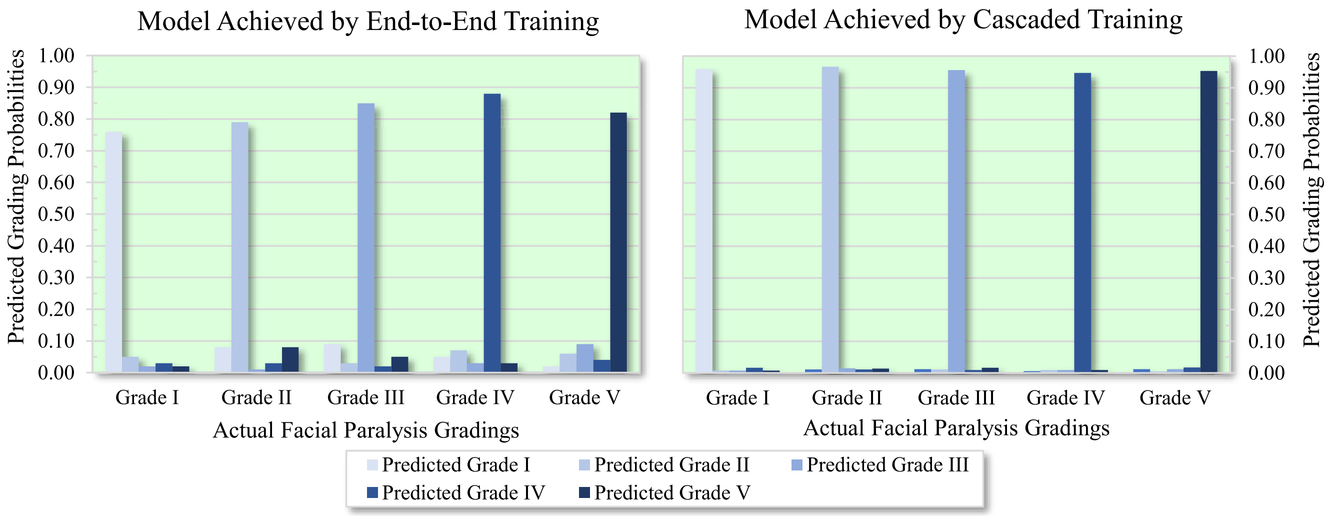


FIGURE 7. Percentages for predicted labels for all five facial paralysis gradings. They actually indicate the metrics in terms of *TP* (the correctly predicted probability) and *FP* (the incorrectly predicted probabilities for other grading categories). As observed, the cascaded training strategy can significantly enhance the performance for the proposed model.

TABLE 3. Performance comparisons for facial grading prediction in terms of four metrics.

	Average **Accuracy** (%)	**Recall** (*tpr*) (%)	**Confidence** (*tpa*) (%)	**Dice** (%)
Insu *et al.* [46]	89.00	88.49	89.23	88.00
Hyun *et al.* [47]	88.90	90.07	87.11	86.66
Muhammad *et al.* [19]	92.60	93.14	92.91	93.00
End-to-End Training (Our Model)	82.90	82.53	83.41	82.96
Cascaded Training (Our Model)	95.60	95.90	95.75	95.82

number of additional parameters that bring more burden to the training process. On the other hand, the dual-encoder of **E1** and **E2** along with the cascaded training scheme can significantly enhance the paralysis grading prediction, and push the performance to the best of all.

C. LIMITATIONS AND FUTURE WORKS

Since the facial paralysis analysis using the proposed model should pass facial image through two cascaded encoder networks, it is still hard to reach real-time computation for images above 720p resolution. Under our testing configurations, implemented on a machine with an intel i7 quad core CPU and a Nvidia RTX 2080 GPU, it takes about 0.4 seconds to process one image in average. Although it can be acceptable for clinical applications of facial paralysis evaluation, this computational cost can still be a potential limitation for the proposed solution. Future work will be conducted to establish more connections between two encoders, such as weight sharing, to reduce computations during the process. A low computational costly model can be easily deployed onto potable device such as smart phones or tablets, which can support the applications such as personal medical care or remote diagnosis among others.

VI. CONCLUSION

This paper presents a solution for automatic facial paralysis evaluation. A cascaded encoder structure is proposed in the paper. The first encoder is trained with the task of facial attribute semantic segmentation based on the training data mixed from both normal faces and paralysed faces. It thus can produce rich facial spatial information, which is essential for facial paralysis evaluation. The second encoder is trained with the facial paralysis grading prediction task using paralysed facial images as training data. It can export the facial paralysis features from the input facial images. This cascaded training process relieves the requirement for a large amount of facial paralysis data to construct the prediction model. Practically, those facial paralysis data are much less common than images in other domains. Experiments are conducted to evaluate the performance of each component in the proposed model. Encouraging results are visually and statistically demonstrated compared with several existing methods in related areas.

REFERENCES

[1] A. F. Abate, P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, "Near real-time three axis head pose estimation without training," *IEEE Access*, vol. 7, pp. 64256–64265, 2019.

[2] A. Zadeh, Y. C. Lim, T. Baltrušaitis, and L. P. Morency, "Convolutional experts constrained local model for 3D facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Oct. 2017, pp. 2519–2528.

[3] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 532–539.

[4] J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Fast and exact newton and bidirectional fitting of active appearance models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1040–1053, Feb. 2017.

[5] G. Tzimiropoulos and M. Pantic, "Fast algorithms for fitting active appearance models to unconstrained images," *Int. J. Comput. Vis.*, vol. 122, no. 1, pp. 17–33, 2017.

[6] C. Liu, L. Feng, H. Wang, and B. Wu, "Face Alignment via Multi-Regressors Collaborative Optimization," *IEEE Access*, vol. 7, pp. 4101–4112, 2019.

[7] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.

[8] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. Zhang, "A survey on deep learning based brain computer interface: Recent advances and new frontiers," May 2019, *arXiv:1905.04149*. [Online]. Available: https://arxiv.org/abs/1905.04149

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[10] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "MaskLab: Instance segmentation by refining object detection with semantic and direction features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4013–4022.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[12] H. Kim, J. Park, H. Kim, E. Hwang, and S. Rho, "Robust facial landmark extraction scheme using multiple convolutional neural networks," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3221–3238, 2018.

[13] M. M. Kalayeh, B. Gong, and M. Shah, "Improving facial attribute prediction using semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6942–6950.

[14] A.-M. Chevalier, A. Miranda, M. Lacôte, and J. P. Bleton, *Clinical Evaluation of Muscle Function*. London, U.K.: Churchill Livingstone, 1987.

[15] D. E. Brackmann and J. W. House, "Facial nerve grading system," *Otolaryngol.-Head Neck Surg.*, vol. 93, no. 2, pp. 146–147, Apr. 1985.

[16] S. E. Coulson, G. R. Croxson, R. D. Adams, and N. J. O'dwyer, "Reliability of the 'Sydney,' 'Sunnybrook,' and 'House Brackmann' facial grading systems to assess voluntary movement and synkinesis after facial nerve paralysis," *Otolaryngol.-Head Neck Surg.*, vol. 132, no. 4, pp. 543–549, 2005.

[17] G.-S. J. Hsu, J.-H. Kang, and W.-F. Huang, "Deep hierarchical network with line segment learning for quantitative analysis of facial palsy," *IEEE Access*, vol. 7, pp. 4833–4842, 2019.

[18] T. Wang, S. Zhang, J. Dong, L. Liu, and H. Yu, "Automatic evaluation of the degree of facial nerve paralysis," *Multimedia Tools Appl.*, vol. 75, no. 19, pp. 11893–11908, 2016.

[19] M. Sajid, T. Shafique, M. J. A. Baig, I. Riaz, S. Amin, and S. Manzoor, "Automatic grading of palsy using asymmetrical facial features: A study complemented by new solutions," *Symmetry*, vol. 10, no. 7, p. 242, Jun. 2018.

[20] W. S. W. Samsudin, R. Samad, K. Sundaraj, M. Z. Ahmad, and D. Pebrianti, "Regional assessment of facial nerve paralysis using optical flow method," in *Proc. 10th Nat. Tech. Seminar Underwater Syst. Technol.*, Z. M. Zain, H. Ahmad, D. Pebrianti, M. Mustafa, N. R. H. Abdullah, R. Samad, and M. M. Noh, Eds. Singapore: Springer, 2019, pp. 505–514.

[21] A. Song, Z. Wu, X. Ding, Q. Hu, and X. Di, "Neurologist standard classification of facial nerve paralysis with deep neural networks," *Future Internet*, vol. 10, no. 11, p. 111, 2018.

[22] F. Xie, Y. Ma, Z. Pan, X. Guo, J. Liu, and G. Gao, "Degree evaluation of facial nerve paralysis by combining LBP and Gabor features," in *Proc. 2nd Int. Symp. Image Comput. Digit. Med.*, Oct. 2018, pp. 143–147.

[23] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[24] X. Wang, A. Shrivastava, and A. Gupta, "A-Fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2606–2615.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. New York, NY, USA: Curran Associates, 2017, pp. 5998–6008.

[26] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," May 2018, *arXiv:1805.08318*. [Online]. Available: https://arxiv.org/abs/1805.08318

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[28] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," Sep. 2018, *arXiv:1809.02983*. [Online]. Available: https://arxiv.org/abs/1809.02983

[29] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *Proc. BMVC*, 2018, pp. 1–13.

[30] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[31] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6656–6664.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[33] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[36] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[37] K. M. de Oliveira Fonseca, A. M. Mourão, A. R. Motta, and L. C. C. Vicente, "Scales of degree of facial paralysis: Analysis of agreement," *Brazilian J. Otorhinolaryngol.*, vol. 81, no. 3, pp. 288–293, 2015.

[38] M. Köstinger, 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 *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

[39] N. Kumar, P. Belhumeur, and S. Nayar, "FaceTracer: A search engine for large collections of images with faces," in *Proc. Eur. Conf. Comput. Vis.*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Germany: Springer, 2008, pp. 340–353.

[40] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1637–1644.

[41] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[42] J. Wang, Y. Cheng, and R. S. Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2295–2304.

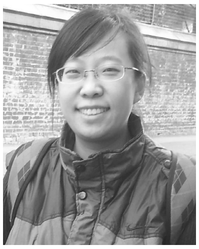
[43] Y. Zhong, J. Sullivan, and H. Li, "Leveraging mid-level deep representations for predicting face attributes in the wild," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3239–3243.

[44] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A Survey," in *Advances in Face Detection and Facial Image Analysis*, M. Kawulok, M. E. Celebi, and B. Smolka, Eds. Cham, Switzerland: Springer, 2016, pp. 189–248.

[45] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[46] I. Song, N. Y. Yen, J. Vong, J. Diederich, and P. Yellowlees, "Profiling bell's palsy based on House-Brackmann score," *J. Artif. Intell. Soft Comput. Res.*, vol. 3, no. 1, pp. 41–50, Dec. 2014.

[47] H. S. Kim, S. Y. Kim, Y. H. Kim, and K. S. Park, "A smartphone-based automatic diagnosis system for facial nerve palsy," *Sensors*, vol. 15, no. 10, pp. 26756–26768, Oct. 2015.



TING WANG received the Ph.D. degree and the master's degree in computer application technologies from the Ocean University of China, Qingdao, China. She is currently a Lecturer of computer science with the Shandong University of Science and Technology, Qingdao. Her main research interests include image processing, facial recognition, and machine learning. She is a member of the China Computer Federation.

SHU ZHANG received the Ph.D. degree in computer application technologies from the Ocean University of China, Qingdao, China, where he is currently a Lecturer. He was a Research Associate with the University of Portsmouth, Portsmouth, U.K. His main research interests include computer vision, feature analysis, facial modeling, 3D reconstruction, video processing, underwater image analysis, and deep learning among others. He is a member of the China Computer Federation.

LI'AN LIU received the Ph.D. degree in medical science. He is currently a Professor, a Chief Physician, a Master Supervisor, and the Director of the Acupuncture and Moxibustion Department, Qingdao Hiser Medical Center, Qingdao Traditional Chinese Medicine Hospital. His research interests include medicine, nerve system, facial paralysis, and rehabilitation. He is also the Vice President of the Shandong Acupuncture Association, the Vice President and the Secretary General of the Qingdao Acupuncture Association, and the Councilor of the Chinese Acupuncture Association. He is also a Panel Member of the National Natural Science Foundation of China.

GENGKUN WU received the B.E. degree from the College of Computer Science and Engineering, Shandong University of Science and Technology, China, in 2010, and the Ph.D. degree from the School of Computer Science and Technology, Ocean University of China, in 2015. He was a Postdoctoral Researcher with Zhejiang University, from 2015 to 2017. He is currently a Lecturer with the College of Computer Science and Engineering, Shandong University of Science and Technology. His research interests include computer simulation, modeling and optimization, ocean wave modeling and rendering, and calculation of surface electromagnetic scattering coefficient.

JUNYU DONG received the B.Sc. and M.Sc. degrees in applied mathematics from the Ocean University of China, Qingdao, China, and the Ph.D. degree in image processing from the School of Math and Computer Sciences, Heriot-Watt University, Edinburgh, U.K. He is currently the Dean of the College of Information Science and Engineering and a Professor and the Head of the Department of Computer Science and Technology, Ocean University of China. His main research interests include texture perception and analysis, 3D reconstruction, video analysis, and underwater image analysis.