ALGRNet: Multi-relational Adaptive Facial Action Unit Modelling for Face Representation and Relevant Recognitions

Xuri Ge, Joemon M. Jose, Pengcheng Wang, Arunachalam Iyer, Xiao Liu, Hu Han

Abstract—Facial action units (AUs) represent the fundamental activities of a group of muscles, exhibiting subtle changes that are useful for various face analysis tasks. One practical application in real-life situations is the automatic estimation of facial paralysis. This involves analyzing the delicate changes in facial muscle regions and skin textures. It seems logical to assess the severity of facial paralysis by combining well-defined muscle regions (similar to AUs) symmetrically, thus creating a comprehensive facial representation. To this end, we have developed a new model to estimate the severity of facial paralysis automatically and is inspired by the facial action units (FAU) recognition that deals with rich, detailed facial appearance information, such as texture, muscle status, *etc.* Specifically, a novel **A**daptive **L**ocal-**G**lobal **R**elational **N**etwork (ALGRNet) is designed to adaptively mine the context of well-defined facial muscles and enhance the visual details of facial appearance and texture, which can be flexibly adapted to facial-based tasks, *e.g.,* FAU recognition and facial paralysis estimation. ALGRNet consists of three key structures: (i) an adaptive region learning module that identifies high-potential muscle response regions, (ii) a skip-BiLSTM that models the latent relationships among local regions, enabling better correlation between multiple regional lesion muscles and texture changes, and (iii) a feature fusion&refining module that explores the complementarity between the local and global aspects of the face. We have extensively evaluated ALGRNet to demonstrate its effectiveness using two widely recognized AU benchmarks, BP4D and DISFA. Furthermore, to assess the efficacy of FAUs in subsequent applications, we have investigated their application in the identification of facial paralysis. Experimental findings obtained from a facial paralysis benchmark, meticulously gathered and annotated by medical experts, underscore the potential of utilizing identified AU attributes to estimate the severity of facial paralysis.

Index Terms—Facial paralysis estimation, Facial action units detection, Facial action units, Skip-BiLSTM, Fusion&Refining

1 INTRODUCTION

Deep learning based facial analysis tasks, such as facial recognition and facial expression recognition, aim to extract facial visual features that capture the intricate facial appearance and texture information using well-crafted Convolutional Neural Networks (CNNs). Many existing methods [1], [2], [3], [4] directly extract a global facial representation from an entire face image through CNNs to perform subsequent recognition tasks. However, accurately localizing the relevant muscle regions that contribute significantly becomes challenging, thus hindering the utilization of potentially responsive muscle regions in specific facial analysis tasks, such as facial paralysis estimation.

Recently, facial action units (AUs) have been defined to represent the precise muscle activities that capture detailed facial information. Initially, AUs are used in the Facial Action Coding System (FACS) [5], which can manually code nearly any anatomically possible facial expression via different groups of specific AUs. However, these earlier methods relying on hand-crafted features, which have two significant defects: (i) shallow hand-crafted features lack discrimination in representing facial morphology, and (ii) existing AU-based applications focus primarily on emotion-related facial actions, disregarding other decision-making processes.

On the one hand, deep learning based AU recognition methods [6], [7], [8], [9], [10] have been explored to enhance the AU's feature representation for face analysis. To obtain rich and detailed facial representations, existing facial AU recognition methods [6], [8], [9], [10] combine local features from multiple independent AU branches, each corresponding to a separate AU patch. However, as shown in Fig. 1 (a), grid-based deep learning methods [6], [7] that divide the image into fixed grids fail to accurately correspond patches with AU muscle regions. Multi-branch combination-based methods [10], [11], [12] refine AU-related features by fusing global or local features from independent AU branches

We thank Professor Brian O'reilly for sharing part of the facial paralysis dataset. This research was supported in part by the National Key R&D Program of China (grant 2018AAA0102501), and Natural Science Foundation of China (grant 62176249). Xuri Ge's research was supported in part by China Scholarship Council (CSC) from the Ministry of Education of China (No. 202006310028).

- *Xuri Ge is with the School of Computing Science, University of Glasgow, Scotland, UK (e-mail: x.ge.2@research.gla.ac.uk).*
- *Joemon M. Jose is with the School of Computing Science, University of Glasgow, Scotland, UK (e-mail: joemon.jose@glasgow.ac.uk).*
- *Pengcheng Wang is with Tomorrow Advancing Life Education Group (TAL), Beijing 100080, China (e-mail: wangpengcheng2@tal.com).*
- *Arunachalam Iyer is with the Department of Otolaryngology and Head and Neck Surgery, University Hospital Monklands, Airdrie, Scotland, UK, and also with University of Glasgow, Scotland, UK (aruniyer-ent@gmail.com).*
- *Xiao Liu is with the Online Media Business Unit at Tencent, Beijing 100080, China (e-mail: ender.liux@gmail.com).*
- *Hu Han is the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, 100190, China, and University of the Chinese Academy of Sciences, Beijing 100049, China. (email: hanhu@ict.ac.cn).*

Manuscript received April 19, 2005; revised August 26, 2015.

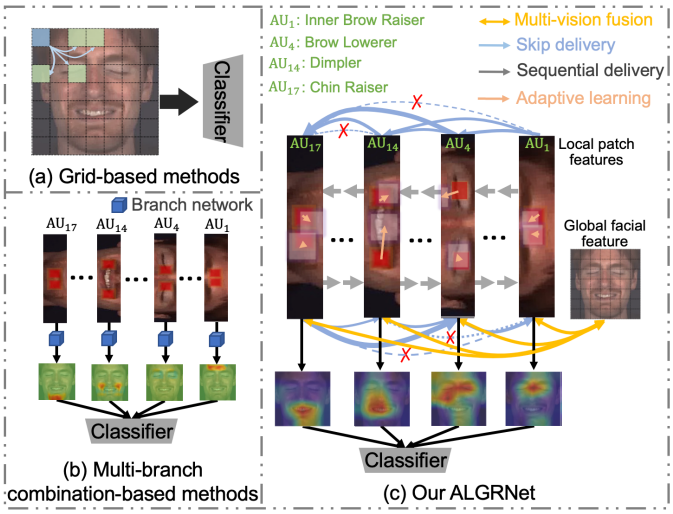


Fig. 1. Illustration of the different schemes for AU detection: (a) the traditional grid-based feature extraction and classification, (b) the popular multi-branch combination-based detection methods, and (c) ALGRNet method: ALGRNet, in comparison with (a) and (b), adaptively adjusts the AU areas in terms of different individuals based on detected landmarks, exploits mutual facilitation and inhibition of region-based multiple branches through a novel bidirectional structure with skipping gates and refines their irregular representations guided by the global facial feature.

based on pre-defined muscle regions, as shown in Fig. 1 (b). Nevertheless, these methods overlook the inter-relationship between multiple AU areas and the local-global context of each face. The multiple patches related to individual AUs may have a strong positive or negative latent correlation in different face states. Here, if multiple AUs jointly affect the target AU category, it is defined as a positive correlation (mutual assistance), otherwise a negative correlation (mutual exclusion). For example, adjacent AU2 ("Outer Brow Raiser") and AU7 ("Lid Tightener") will be activated simultaneously when scaring and non-adjacent AU6 ("Cheek Raiser") and AU12 ("Lip Corner Puller") will be activated simultaneously when smiling. Several recent works [13], [14], [15], [16] have focused on capturing the interactions among different AUs for local feature enhancement, considering the relationship of multiple facial patches to achieve better robustness than using a single patch. For instance, the studies in [17], [18], [19] incorporated AU knowledge-graph derived from statistical benchmarks to provide additional relational guidance for enhancing facial region representation. Another study [20] utilized the spectral perspective of graph convolutional network (GCN) to model the AU relationship, requiring an additional AU correlation reference extracted from EAC-Net [21]. Despite the improvement by the introduced AU relationship modelling, these methods rely on the prior knowledge of AU correlation to define a fixed graph to exploit useful information from correlated AUs. Other studies [16], [22] employed an adaptive graph to model AU relationships based on global features, but they overlooked the local-global feature interactions that enhance the distinguishability of AUs by exploiting the complementary global details. Furthermore, these methods ignored the physiological phenomenons that adjacent related muscles often exhibit high potential correlation due to muscle linkage, and the relationship between non-adjacent related muscles may vary across different expressions and individuals.

Partial Facial Action Units	
Description	**Facial Muscle**
Inner brow raiser	frontalis (pars medialis)
Outer brow raiser	frontalis (pars lateralis)
Brow lowerer	corrugator supercilii
Cheek raiser	orbicularis oculi
Nose wrinkler	levator labii superioris alaeque nasi
Upper lip raiser	levator labii superioris
Lip corner puller	zygomaticus major
Lip corner depressor	triangularis
Chin raiser	mentalis
Lips part	orbicularis oris

Fig. 2. The descriptions and corresponding facial muscles of the partial facial AUs. The first row of images is the definitions of AU centers based on the detected landmarks on facial AU detection methods [9], [10], [23] and the second is a facial paralysis patient with the detected bounding boxes of potential muscle lesions from [24]. It is clear to observe that the AU regions can cover most areas of potential muscle lesions.

On the other hand, AUs offer independent interpretation and accurate localization, making them valuable for various higher-order decision-making processes beyond facial expression recognition, such as mental disease diagnosis [25], depression analysis [26], and deception detection [27]. As depicted in Fig. 2, AUs capture fine-grained facial behaviours and possess inherent properties of symmetry and flexibility, inspiring exploration in higher-order decision-making tasks. For instance, previous works [26] evaluated the impact of depression on facial response using FACS-based [5] methods but relied on shallow hand-crafted AU features. Similarly, automated FACS-based systems analysed facial actions in neuropsychiatric patients [28] but overlooked the relational dependence and physiological phenomenon of natural linkages between multiple muscle groups. However, these inherent properties of AUs play crucial roles in real-world face analysis applications, including facial paralysis estimation, which remains underexplored. As shown in Fig. 2, facial paralysis is the temporary or permanent weakness or lack of movement affecting the muscles on one side of the face, where most AU regions can cover the potential muscle lesions. Therefore, AU-based automatic facial paralysis recognition can leverage rich facial representation and inherent properties of AUs in symmetry and flexibility by combining well-defined muscle regions (similar to AUs), one of the facial biometrics' challenging and meaningful applications.

Motivated by the aforementioned considerations, we present a novel approach utilizing a flexible and innovative model for automatically estimating facial paralysis based on a novel facial action unit (AU) detection. Our method introduces two key advancements. Firstly, we propose an adaptive local-global relational network (ALGRNet) that operates in an end-to-end manner. This network achieves exceptional performance in AU detection, highlighting its impressive representation abilities and its potential to be seamlessly applied to facial paralysis estimation. To accommodate facial variations across individuals, we introduce an adaptive region learning module that detects landmarks and corresponding offsets. This aspect becomes crucial due to individual facial differences, especially when dealing with symptoms of facial paralysis which often involve displacement changes such as muscle sagging. Drawing inspiration

from physiological phenomena, which suggest that adjacent related muscles tend to exhibit high potential correlation. In contrast, non-adjacent corresponding muscles may display variations in different expressions and individuals. To this end, we design a skip-BiLSTM to capture implicit interactive information exchange among patch-based branches (each AU corresponds to one branch) via multiple connections, *i.e.* sequential and skipping connections. These connections effectively capture the potential relationships of assistance and exclusion among the sequential branches, with the ability to adjust transfer within the BiLSTM [29] for adjacent patches, while distant patches are connected via skipping-type gates. As each AU branch is treated independently and equally, this skip connection method minimizes information loss compared to traditional BiLSTM. Subsequently, we introduce a novel feature fusion and refining module to enhance the local features obtained from the skip-BiLSTM, guided by global grid-based features. In contrast to previous basic feature fusion methods [10], [30], our gated fusion architecture in the feature fusion and refining module effectively supplements global information, including non-AU region information, for each local AU region. This is crucial because different AUs may prioritize different global information. Secondly, the features learned by ALGRNet can be utilized either for AU recognition through a multi-branch classification network or seamlessly integrated into a facial paralysis estimation classifier with minimal adjustments to the AU positions. We are the pioneers in investigating the effectiveness of an end-to-end deep learning-based AU detection model for predicting the severity of facial paralysis. The intrinsic characteristics of AUs, such as their ability to represent facial features vigorously, exhibit a degree of symmetry and flexibility, making them suitable for aiding in the automated diagnosis of patients with facial palsy. Existing methods [31], [32] have demonstrated the feasibility of this approach, but they lack robust AU recognition capabilities.

Our contributions can be summarized as follows:

- We propose a novel end-to-end AU detection model that combines adaptive local facial muscle features, their relationships, and local-global contexts to improve facial representation. This model offers flexibility and applicability in face paralysis diagnosis, which is a pioneering effort in developing a well-designed model for this purpose.
- A new adaptive region learning module is proposed to improve the accuracy of muscles corresponding to action units and accommodate symmetric muscle region biases due to individual or lesion differences, thereby further improving the robustness and flexibility of the model.
- We propose a novel skip-BiLSTM module based on the natural physiological phenomenons to improve the representation of local AUs by modelling the mutual assistance and exclusion relationships of individual AUs via multiple inter-muscular connections, *i.e.* sequential and skipping. And a new gated feature fusion&refining module, filtering information that contributes to the target AU, even non-defined AU areas, is further designed to facilitate more discriminative local AU feature generation.
- The proposed ALGRNet achieves new state-of-the-art on two AU detection benchmarks, *i.e.,* BP4D and DISFA, without any external data or pre-trained models in additional data. Notably, we achieve superior performance to baselines on a collected facial paralysis dataset (named ***FPara***), which validates the potential of our ALGRNet for facial paralysis estimation.

Compared to the AU detection method in our conference version [23], we propose a new adaptive region learning module in Section 3.2 to improve muscle regions' accuracy and accommodate symmetric muscle region biases due to lesions or individual differences. In particular, the adaptive region learning module contains learning of scaling factors to change the size of the corresponding muscle regions and offset learning to adjust landmark differences for different individuals and disease presentations slightly. This suggests that adaptive region learning could help the model to focus accurately on the muscle region changes corresponding to each AU and to obtain better robustness and generalization ability. In addition, we did not evaluate the generalizability and transferability of the AU detection presented in the previous version, unlike in this study. In this study, we innovatively explore and apply the proposed ALGRNet to facial paralysis estimation, which improves the effectiveness of facial paralysis recognition and estimation by focusing on activation features of multiple symmetrical muscle regions and global facial information. Specifically, we exploit a facial paralysis dataset that medical professionals annotate to four grades of facial paralysis degrees, *i.e.* normal, low, medium and high grade. For facial paralysis estimation, we focus on the muscle areas as newly defined ***PAUs*** that are preferred in the facial paralysis ratings to the AU predefined muscle regions in traditional AU detection tasks. Due to the independent interpretation of AUs, the flexible changes in AU positions do not affect the representation capability of the face features. Finally, we combine the multiple symmetrical muscle region features enhanced by the interaction, and the useful global information, to obtain the final facial features for the facial paralysis grade classification. To the best of our knowledge, there has yet to be any existing work in the literature on estimating facial paralysis using a well-designed AU detection model. Compared with our earlier work [23], we also provide more quantitative evaluations to show the effectiveness and transferability of our ALGRNet in facial paralysis estimation.

2 Related Work

2.1 Facial Action Units Detection

Automatic AU detection is a task that detects the movement of a set of facial muscles. Recently, patch-learning based methods are the most popular paradigms for AU detection [33], [34], [35], [36], [37]. For instance, [38] used CNNs and BiLSTM to extract and model the image regions for AUs, which are pre-selected by domain knowledge and facial geometry. However, all the above methods need to predefined the patch location first. To address these issues, [9] proposed to jointly estimate the location of landmarks and the presence of action units in an end-to-end framework,

where landmarks can also use to compute the attention map for each AU separately. Recent works [13], [14], [15], [16], [39] explicitly take into consideration the linkage relationship between different AUs for AU detection, which relies on action unit relationship modelling to help improve recognition accuracy. Typically, [40] exploited the relationships between AU labels via a dynamic Bayesian network. [18] embedded the relations among AUs through a predefined graph convolutional network (GCN). [17] integrated the prior knowledge from FACS into an offline graph, which can construct a knowledge graph coding the AU correlations. However, these methods require prior connections by counting the co-occurrence probabilities in different datasets. [16], [19], [22] applied an adaptive graph to model the relationships between AUs based on global features, ignoring local-global feature interactions.

The most relevant previous studies to our work are [9], [10], which combine AU detection and face alignment into a multi-branch network. Different from these methods, our proposed ALGRNet can adaptively adjust the target muscle region corresponding to each AU and utilizes the learned mutual assistance and exclusion relationships between the target muscle and other muscle regions to enhance the feature representation of the target AU. Doing so allows us to provide more robustness and interpretability than [10].

2.2 Facial Paralysis Estimation

Facial paralysis estimation has recently attracted extensive research attention [32], [41], [42], due to the significant psychological and functional impairment to patients. Nottingham system [43] is a widely accepted system for the clinical assessment of facial nerve function, which is similar to House-Brackmann (H-B) [44]. In addition, over twenty other methods of recognizing and assessing facial paralysis are available in the literature. However, these methods are about facial paralysis by medical professionals and are time-consuming and subjective. More recently, deep learning has been widely applied for facial analysis, including face recognition, face alignment, *etc*. [45] and [46] proposed two efficient quantitative assessments of facial paralysis based on the detected key points. [47] proposed to obtain the facial paralysis degree by calculating the changes in the surface areas of a specific facial region. [48] considered both static facial asymmetry and dynamic transformation factors in evaluating the degree of facial paralysis. However, most existing methods only use deep learning methods to pave the way for physical computation and do not directly model and predict the depth features of a face image. In addition, they exploit hand-crafted features and post-processing to obtain the final result. Although this increases the potential for interpretation, handcrafted features are not discriminative enough to represent facial morphology due to their shallow natures. In addition, we intend to trigger the signs that a patient may have the disease as early as possible so that the patient can be further diagnosed by a medical professional, which will facilitate the potential patient receiving treatment earlier. Recently there are many new approaches [1], [2], [3], [4] extracted facial appearances with high-level semantic features as input to the classifier via popular convolutional neural networks in an end-to-end model. However, while these methods extracted coarse facial representations with the help of robust convolutional neural networks, they lacked fine-grained information [49], [50] about accurate muscle regions.

In contrast to these existing methods, we utilize a novel end-to-end framework (ALGRNet) to predict the grade of facial paralysis, which takes into account local AU locations, features, inter-relationships, and local-global contexts.

3 Approach

The framework of the proposed ALGRNet is presented in Fig. 3, which can perform AU detection and facial paralysis estimation. ALGRNet is composed of four main modules, *i.e.,* adaptive region learning module (Subsection 3.2) for adaptive muscle region localisation, a skip-BiLSTM module (Subsection 3.3) for mutual assistance and exclusion relationship modelling, a feature fusion&refining module (Subsection 3.4) for refining features of irregular muscle regions, and a multi-classifier module (Subsection 3.1) for predicting the grade of facial paralysis.

3.1 Overview of ALGRNet

For AU detection or facial paralysis estimation, our method uses a multi-branch network [9], [10], [51], where each branch corresponds to a specific predefined AU or PAU (defined in Fig. 4). *Due to patient confidentiality, we display data for AU detection using a generic image in Fig. 3.*) In contrast to previous methods, we are exploiting the relationship between multiple AUs related to symmetrical muscle areas, which plays a crucial role in building a robust facial palsy detection model. In addition, due to the diversity of expression, lesion extent, and individual characteristics, we also attempted to learn adaptive muscle region offsets and scaling factors for each muscle region. To this end, we design three modules (adaptive region learning module, skip-BiLSTM module, and feature fusion&refining module) based on the established multi-branch network that can fully exploit inter-regional and local-global feature interactions.

We first adapt a hierarchical and multi-scale region learning network from [9] as our stem network, which extracts the grid-based global features and the local region features. However, unlike [9], we add two simple linear-based networks combined with the previous face alignment network, named adaptive region learning module (detailed in Section 3.2), to learn the offsets and scaling factors for each region adaptively. After that, local patches $A = \{A_1, A_2, ..., A_n\}$ are computed from the learned locations and their features $V = \{v_1, v_2, ..., v_n\}$ can be extracted through the stem network, where n is the numbers of selected patches. For the sake of simplicity, we do not repeat here the detailed structure of the stem network [9].

In our ALGRNet, and contrast to the traditional sequence spreading of LSTM, we design a novel skip-BiLSTM module (detailed in Section 3.3) to address the lack of sufficient delivery of local patch information between individual branches, which can transmit information in two ways (sequential delivery and skipping delivery) in both two directions (forward and backwards). The sequential delivery of information enables full exploration of the contextual relationships between adjacent patches. The skipping delivery

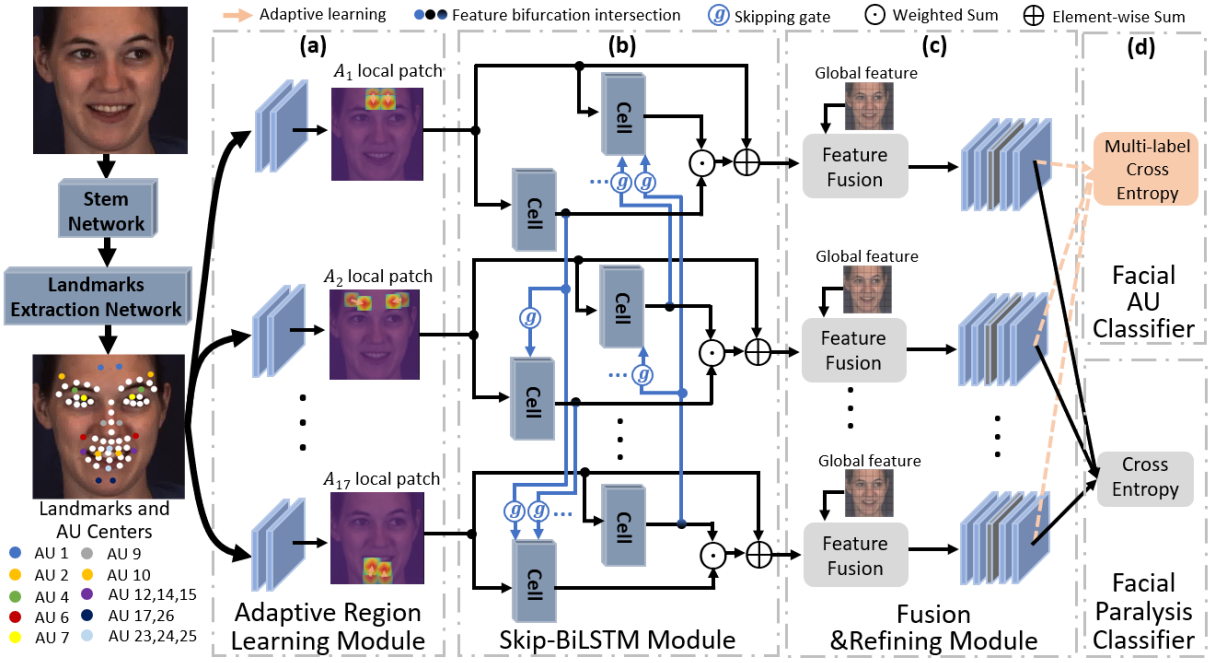


Fig. 3. The overall architecture of the proposed ALGRNet for facial paralysis estimation or AU detection. The location definition of salient muscle regions (as new PAUs) for facial paralysis estimation is detailed in the Fig. 4 and the definition used for AU detection is from [10]. We utilize a simple landmark localization network to detect the landmarks and two linear-based network to learn the offsets and scaling factor of AU centers, which are used to compute local AU patches. We then feed the features into the novel multi-branch network with a skip-BiLSTM module and a feature fusion&refining module, with each branch corresponding to an AU (each AU contains two relatively symmetrical muscle areas). The skip-BiLSTM module explores positive and negative relations among different AU branches by different information delivery options. And the feature fusion&refining module in each branch helps the local AU region to fit irregular shape guided by the global grid-based feature. Finally, a multi-label binary-classifier for AU detection is employed to predict individual AU activation probabilities and a multi-class classifier for facial paralysis estimation is used to predict the grade of facial paralysis.

highlights the interaction of information from non-adjacent related patches. After skip-BiLSTM, we get a set of the local patch features $S = \{s_1, s_2, ..., s_n\}$, which are expected to have all the valuable information from adjacent and non-adjacent patches.

Furthermore, to deal with irregular muscle areas, a novel feature fusion&refining module (detailed in Section 3.4) is developed to refine the local patches to obtain salient micro-level features for the global facial feature G. Finally, the new patch-based representations $R = \{r_1, r_2, ..., r_n\}$ for AUs are obtained by integrating local muscle features and global facial features.

This work integrates face alignment and AU detection (or facial paralysis estimation) into an end-to-end learning model. We aim to learn all the parameters jointly by minimizing face alignment loss and facial paralysis estimation loss (or facial AU detection loss) over the training set. The face alignment loss is defined as:

$$\mathcal{L}_{align} = \frac{1}{2d_o^2}\sum_{i=1}^{m}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2], \quad (1)$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote the ground-truth (GT) coordinate and corresponding predicted coordinate of the i-th facial landmark, and d_o is the ground-truth inter-ocular distance for normalization.

In this paper, following [10], we also regard facial AU recognition as a multi-label binary classification task. It can be formulated as a supervised classification training objective as follows,

$$\mathcal{L}_{au} = -\frac{1}{n}\sum_{i=1}^{n} w_i[p_i \log \hat{p}_i + (1 - p_i)\log(1 - \hat{p}_i)], \quad (2)$$

where p_i denotes the GT probability of occurrence for the i-th AU, which is 1 if occurrence and 0 otherwise, and $\hat{p}_i$ denotes the predicted probability of occurrence. w_i is the data balance weights, which is employed in [9]. Moreover, the loss of facial paralysis estimation is formulated as:

$$\mathcal{L}_{par} = -w_i q \mathrm{Log}(\hat{q}), \quad (3)$$

where q and $\hat{q}$ are the label and predicted probability for the facial paralysis grades, respectively. w_i is the data balance weights obtained by counting the different classes in the training set. Finally, we optimize the whole end-to-end network by minimizing the joint loss function $\mathcal{L} = \mathcal{L}_{au}(\text{or } \mathcal{L}_{par}) + \lambda \mathcal{L}_{align}$ over the training set.

3.2 Adaptive Region Learning Module

Instead of the predefined muscle regions based on landmarks, and given the nonconformity of facial areas, especially with facial palsy, we use two simple, fully connected networks to adaptively learn the offsets and scaling factors for all muscle regions, respectively. Specially, we utilize an efficient landmark extraction network after the stem network to extract the landmarks $L = \{l_1, l_2, ..., l_m\}$ (m is the numbers of landmarks) similar to [10], including three convolutional layers connected to a max-pooling layer. Simultaneously, two networks containing two fully-connected layers are used to get the adaptive offsets $O =$

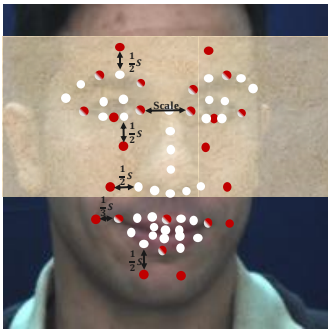


Fig. 4. New definitions for the 12 locations of muscle centers of facial paralysis estimation, which are marked in red or mixed red. The detected landmarks are marked in white or mixed red. "Scale" denotes the distance between two inner eye corners.

$\{o_1, o_2, ..., o_{2n}\}$ and scaling factors $E = \{e_1, e_2, ..., e_n\}$, respectively. According to the learned landmarks, offsets and scaling factors, local patches A are calculated. In particular, we first use the same rules in [10] to get the locations of target muscle area centers based on the detected landmarks and then update the locations by adding the learned offsets. Please note that, we change the predefined muscle region centers according to clinical diagnosis experience to better fit the high-potential lesion regions of facial paralysis, as shown in Fig. 4[1], based on the detected landmarks when we apply ALGRNet on facial paralysis estimation. When defining the new salient muscle regions as new ***PAUs*** (note that each PAU contains two muscle regions and the number of PAUs is the same with AUs.), we maintain its roughly symmetrical distribution on faces. *For a clearer description, we do not distinguish between AUs and PAUs in the follow-up, and default to PAUs for facial paralysis estimation.* Different from [10], we make the scaling factor E learnable rather than a fixed value, where e_i is the width ratio between the region of AU_i and whole feature map. After that, we generate an approximate Gaussian attention distribution for each region following [9]. Finally, based on the learned regions, local patch features V are extracted via the stem network.

3.3 Skip-BiLSTM

Fig. 3 (b) shows the detailed structure of our skip-BiLSTM module for contextual and skipping relationship learning. Specifically, we extract a set of local patch features $V = \{v_1, v_2, ..., v_n\}$ from the stem network, and feed them to skip-BiLSTM. Distinct from the prior works [8], we regard the multiple patches as a sequence structure from top to bottom, which can transfer information by a Bi-directional LSTM based model [29] with our skipping-type gate. Different from the traditional BiLSTM or tree-LSTM [52], [53], our skip-BiLSTM can directly calculate the correlation between a target AU and all other AUs. For the t-th patch ($t > 1$), the extracted feature v_t is used to learn the weights with forward hidden states $H = \{h_1, ..., h_{t-1}\}$ by the skipping-type gates, which can determine the correlation coefficient between past AUs and current AU. And then the new states $\hat{H} = \{\hat{h_1}, ..., \hat{h}_{t-1}\}$ and v_t are fed into the t-th forward cell in the skip-BiLSTM to learn the association weights, which

1. Due to patient confidentiality agreements, we cannot show real patients with facial palsy. This example image is from BP4D.

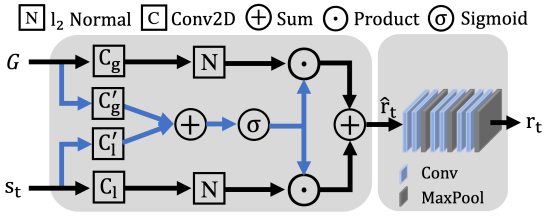


Fig. 5. The architecture of our feature fusion&refining module guided by global face feature.

can promote the transfer of relevant AUs information. The above process can be formulated as:

$$\overrightarrow{h_t} = \text{Cell}(\sum_{j=1}^{t-1} \overrightarrow{\hat{h}_j}, v_t), \tag{4}$$

$$\overrightarrow{\hat{h}_j} = \overrightarrow{h_j} f_j, \tag{5}$$

$$f_j = \sigma(\text{GAP}(W_j(\overrightarrow{h_j} v_t))), \tag{6}$$

where $\text{Cell}(\cdot)$ indicates the basic ConvLstm cell [54], and GAP denotes the global average pooling operation. W_j is the parameters of mapping function, in which we used Conv2D. σ denotes sigmoid function. We obtain the t-th patch feature for backward delivery, which follows the identical forward method as:

$$\overleftarrow{h_t} = \text{Cell}(\sum_{j=t+1}^{n} \overleftarrow{\hat{h}_j}, v_t), \tag{7}$$

In order to fully promote the information interactive among individual AUs, the final representation for each patch is computed as the average of the hidden vectors in both directions, as well as the original patch feature:

$$s_t = v_t + (\overrightarrow{h_t} + \overleftarrow{h_t})/2, \tag{8}$$

3.4 Feature Fusion&Refining Module

To exploit the useful global face feature, we design a gated fusion architecture and a refining architecture (F&R) that can selectively balance the relative importance of local patches and global face grids. We add these two architectures on each local branch because different local muscles may focus on different global details. The grid-based global face feature G is extracted using a simple CNN with the same structure as the face alignment network [10]. As shown in Fig. 5, after obtaining the learned t-th local patch feature, it is fused with the grid-based global feature G by the fusion architecture, which can be formulated as:

$$\alpha = \sigma(\text{C}_g^{'} G + \text{C}_l^{'} s_t), \tag{9}$$

$$\hat{r}_t = \alpha \odot ||C_g G||_2 \oplus (1 - \alpha) \odot ||C_l s_t||_2, \tag{10}$$

where σ is the sigmoid function, and $||\cdot||$ denotes the l_2-normalization. $\text{C}_*^{'}$ and C_* denote the Conv2D operation. $\oplus$ denotes the element-wise weighted sum of $||C_g G||_2$ and $||C_l s_t||_2$ according to the learned gate vector α.

The final local fusion feature s_t for t-th patch refined by our F&R module is shown in Fig. 5. F&R module contains three blocks. Each block consists of two convolutional layers and a maxpooling layer. Then multi-patch features R are fused into a multi-class classifier for the paralysis

TABLE 1
Overview information of our collected facial paralysis dataset.

Grade	Normal	Low	Medium	High
Num. of Video	20	29	20	20
Num. of Frame	9049	16970	11019	10547

grade estimation or sent to a multi-label binary classifier to calculate the occurrence probabilities of individual AUs for AU detection.

4 EXPERIMENTS

4.1 Dataset

We evaluate the effectiveness of the proposed approach for facial AU detection on popular BP4D [56] and DISFA [57] datasets. **BP4D** consists of 328 facial videos from 41 participants (23 females and 18 males) who were involved in 8 sessions. Similar to [10], [12], we consider 12 AUs and 140K valid frames with labels. **DISFA** consists of 27 participants (12 females and 15 males). Each participant has a video of $4,845$ frames. We also limited the number of AUs to 8, similar to [9], [10]. Compared to BP4D, the experimental protocol and lighting conditions deliver DISFA to be a more challenging dataset. Following the experiment setting of [10], we evaluated the model using the 3-fold subject-exclusive cross-validation protocol.

To evaluate the effectiveness of our ALGRNet for facial paralysis severity estimation, we exploited a facial paralysis dataset from the NHS, Scotland, named ***FPara*** (the details in Table 1), which consists of 89 videos of facial paralysis patients performing various types of facial paralysis exercises inline with the House-Brackmann (H-B) scale [44]. Each video consisted of facial paralysis patients performing exercises, such as raising eyebrows, closing eyes gently, closing eyes tightly, scrunching up their face and smiling, *etc.* Please note that all videos do not include patient rest time and remove some pauses, thus ensuring that our frame-based classification method can be fully applied. They were part of a previous study on facial paralysis with patient consent for research [58]. These videos are assigned an H-B scale from 1 to 6, one normal and six severe with no body movements. We then further split into four grades, such as normal (H-B score=1), low (H-B score=2), medium (3≤H-B score≤4) and high (5≤H-B score≤6) grades. FPara data is summarised in Table 1. All facial paralysis grades are evaluated using subject-exclusive 3-fold cross-validation, where two folds (about 80%) are used for training, and the remaining one is used for testing (about 20%).

4.2 Implementation Detail

Our model is trained on a single NVIDIA Tesla V100 GPU with 32 GB memory. The whole network is trained using PyTorch [59] with the stochastic gradient descent (SGD) solver, a Nesterov momentum [60] of 0.9 and a weight decay of 0.0005. The learning rate is set to 0.01 initially, with a decay rate of 0.5 every two epochs. The maximum epoch number is set to 20. Aligned faces are further randomly cropped into 176×176 and horizontally flipped to enhance the diversity of training data. Regarding the face alignment network and stem network, we set the value of the general parameters to be the same with [10]. The filters for the convolutional layers in refining architecture are used 3×3 convolutional filters with a stride one and a padding 1. In our paper, all of the mapping Conv2D operations used 1×1 convolutional filters with a stride one and a padding 1. The dimensionality of the hidden state in ConvLstm cell is set to 64. The filters for the convolutional layers in ConvLstm cell are the same as the refining architecture. λ is set to 0.5 for jointly optimizing AU detection (or facial paralysis estimation) and face alignment. The ground-truth annotations of 49 landmarks of training data is detected by SDM [61]. Different from JÂA-Net [10], we averaged the predicted probability of the local information and the integrated information as the final predicted activation probability for each AU, rather than simply using the integrated information of all the AUs. The main difference between our ALGRNet applying to facial palsy and AU detection lies in the final classifier, where facial paralysis is four categories, and AU detection is two categories per AU.

4.3 Performance Metric.

We evaluate the performance of all methods in terms of the F1 score (%), which has been widely used for classification. The F1-frame score is the harmonic mean of the Precision P and Recall R, calculated by $\mathrm{F1} = 2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$. For comparison, we calculate the F1 score for all facial paralysis grades on FPara and for all the AUs on DISFA and BP4D and then average them (denoted as **Avg.**) separately with "%" omitted.

4.4 Overall Performance of Facial AU Detection

We compare the proposed ALGRNet for AU detection with several single-image based baselines in Table 2 and Table 3. The performances of the baselines in Table 3 and 2 are their reported results.

For a more comprehensive display, we also show methods (marked with $*$) [16], [22], [55] that use additional data, such as ImageNet [62] and VGGFace2 [63], *etc*, for pre-training. Since our stem network only consists of a few simple convolutional layers, even if we pre-trained on additional datasets, it is unfair compared to pre-training on deeper feature extraction networks, such as ResNet50 [64]. Our results are still excellent compared with theirs, which demonstrates the superiority and effectiveness of our proposed learning scheme. We omit the need for additional modal inputs and non-frame-based models [65], [66].

Quantitative comparison on BP4D: We report the performance comparisons between our ALGRNet and baselines on BP4D in Table 2. As it can be observed, our ALGRNet significantly outperforms all the other methods in terms of F1-frame score and achieves the first and second places for most of the 12 AUs annotated in BP4D. Our ALGRNet achieves 1.1% higher average F1-frame score compared with the latest state-of-the-art method JÂA-Net.

Quantitative comparison on DISFA: We also report the performance of our proposed ALGRNet on DISFA. Table 3 shows the performance of our ALGRNet is the best in terms of average F1 score compared with all baselines. And

TABLE 2
Performance comparisons on F1-frame score of diverse AU detection for 12 AUs on BP4D. All values are in %. * means the method employed pretrained model on additional dataset, such as ImageNet and VGGFace2, *etc*, so we do not compare. The first and second places are marked with the bold font and underline, respectively.

Method	AU Index												Avg.
	1	2	4	6	7	10	12	14	15	17	23	24	
DSIN [51]	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.8	38.8	41.6	58.9
LP-Net [8]	46.9	45.3	55.6	77.1	76.7	83.8	87.2	63.3	45.3	60.5	**48.1**	54.2	61.0
ARL [12]	45.8	39.8	55.1	75.7	**77.2**	82.3	86.6	58.8	47.6	62.1	47.4	**55.4**	61.1
JÂA-Net [10]	**53.8**	47.8	**58.2**	**78.5**	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
HMP-PS* [16]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
DML* [55]	52.6	44.9	56.2	79.8	80.4	85.2	88.3	65.6	51.7	59.4	47.3	49.2	63.4
ALGRNet (Ours)	51.2	**48.2**	57.3	77.9	76.4	**84.9**	**88.2**	**64.8**	**50.8**	**62.8**	47.6	51.9	**63.5**

TABLE 3
Performance comparisons on F1-frame score of diverse AU detection for 8 AUs on DISFA. All values are in %. The first and second places are marked with the bold font and underline, respectively.

Method	AU Index								Avg.
	1	2	4	6	9	12	25	26	
DSIN [51]	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
LP-Net [8]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
ARL [12]	43.9	42.1	63.6	41.8	40.0	**76.2**	**95.2**	66.8	58.7
JÂA-Net [10]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
HMP-PS* [16]	21.8	48.5	53.6	56.0	58.7	57.4	55.9	56.9	61.0
DML* [55]	62.9	65.8	71.3	51.4	45.9	76.0	92.1	50.2	64.4
ALGRNet (Ours)	**63.8**	**65.4**	**73.6**	44.5	**54.1**	74.0	94.7	**69.9**	**67.5**

TABLE 4
Performance comparisons on F1-frame score (in %) of diverse facial paralysis estimation for 4 grades on FPara.

Method	Facial Paralysis Grades				Avg.
	Normal	Low	Medium	High	
ResNet18 [64]	99.8	50.7	47.7	67.9	66.5
ResNet50 [64]	99.9	53.9	54.7	71.4	70.0
Transformer-based [67]	100	**63.0**	58.6	68.7	72.6
JÂA-Net [10]	100	55.9	62.8	72.5	72.8
ALGRNet (Ours)	**100**	55.9	**72.1**	**73.2**	**75.4**

our approach significantly outperforms all other methods for most of the 8 AUs annotated in DISFA. Compared with the existing end-to-end feature learning and multi-label classification methods DSIN [51] and ARL [12], the average F1-frame score of our proposed ALGRNet get 13.9% and 8.8% higher, respectively. Moreover, compared with the multi-branch combination-based state-of-the-art method JÂANet [10], our ALGRNet achieves 4.0% improvements in terms of average F1-frame score.

4.5 Overall Performance of Facial Paralysis Estimation

Different from facial AU detection, the existing deep-learning-based facial paralysis estimation methods are rare, so we apply currently popular deep-learning classification methods, such as the ResNet [64] and Transformer [68], on our collected facial paralysis dataset (FPara). Besides, we also compare it with the state-of-the-art AU detection approach, JÂA-Net [10]. Specially, we evaluate the following methods:

- ResNet18 and ResNet50 [64]: These methods use different depth layers based on ResNet to model the input face images, which are similar to [69].
- Transformer-based method [67]: This baseline is motivated from self-attention and uses the Transformer [68] architecture. The output of the Transformer-based encoder [67] is treated as the latent representation for the input of the multi-label AU classifier.
- JÂA-Net [10]: This is a recently proposed multi-branch combination-based AU detection method, which can extract precise local muscle features thanks to a joint facial alignment network.

The first and second places are marked with bold font and "_", respectively.

Quantitative comparison on the collected FPara: Facial paralysis estimation results by different methods on our FPara are shown in Table 4. It has been shown that our ALGRNet outperforms all its competitors with impressive margins. Specifically, JÂANet is the latest state-of-the-art method that combines AU detection and face alignment into an end-to-end multi-label multi-branch network. Compared to the facial paralysis estimation model based on the state-of-the-art AU detection method JÂA-Net [10], our ALGRNet achieves 2.6% improvements in terms of average F1 score. The main reason lies in our ALGRNet overcomes the problem of non-transferable information between branches in the JÂA-Net and adaptively adjusts the muscle regions corresponding to the AUs. Moreover, the average F1 score of our ALGRNet gets 2.8% higher compared to the popular

TABLE 5
Ablation study of ALGRNet for 8 AUs on DISFA and for 4 grades on FPara. All values are in %.

Methods	Setting			AU Index								Avg.	Paralysis Grade				Avg.
	S-B	F&R	Ada	1	2	4	6	9	12	25	26		Nor.	Low	Med.	Hig.	
w/o full				47.1	61.1	66.3	44.7	52.2	74.9	92.2	66.2	63.1	99.8	54.6	64.1	70.9	72.3
w/o F&R	√			62.6	64.2	72.4	42.3	49.9	**76.1**	93.5	72.6	66.7	100	54.7	66.2	72.6	73.3
w/o S-B		√		58.7	65.2	73.5	43.9	53.5	72.2	94.1	64.7	65.7	99.9	55.1	65.3	71.3	72.9
w/ Bi		√		61.1	58.4	70.9	45.5	47.9	74.9	92.5	70.8	65.2	99.8	57.1	67.3	72.8	74.3
w/o Ada	√	√		62.6	64.4	72.5	**46.6**	48.8	75.7	94.4	**73.0**	67.3	100	**57.8**	68.7	72.0	74.6
ALGRNet	√	√	√	**63.8**	**65.4**	**73.6**	44.5	**54.1**	74.0	**94.7**	69.9	**67.5**	**100**	55.9	**72.1**	**73.2**	**75.4**

Transformer-based approach [67], although slightly inferior in low-grade recognition. We notice that our method is lower than that of [67] on the set with low facial paralysis grade. The main reason is due to the significantly imbalanced data distribution of the facial paralysis dataset, *i.e.*, much more patients with low grade than those with medium- and high-grade. In this case, the average F1 score can better reflect a model's performance over the whole grade range.

The eventual experimental results of our ALGRNet demonstrate that it is successful in boosting AU detection accuracy on BP4D and DISFA and having high generalization ability on our new facial paralysis dataset.

4.6 Ablative Analysis

To fully examine the impact of our proposed adaptive region learning module, skip-BiLSTM module and feature fusion&refining module, we conduct detailed ablative studies to compare different variants of ALGRNet for AU detection on DISFA and facial paralysis estimation on FPara.

4.6.1 Effects of adaptive region learning module

To cancel out the adaptive region learning (indicated w/o Ada), we follow the same experiment setting as [10] (It means each scaling factor e is set to 0.14.) to predefined muscle region based on the detected landmarks for each AU/PAU. In Table 5, ALGRNet decreases its F1 score to 74.6% and 67.3% on the collected FPara and DISFA respectively. Our whole ALGRNet may show slightly lower accuracy than the method without using adaptive region learning. This is because of the severe data imbalance issues of individual classes. After using adaptive region learning, our method may sacrifice the accuracies of a few AUs (or grades) while improving the overall accuracy.

4.6.2 Effects of skip-BiLSTM

In Table 5, when the skip-BiLSTM module is removed (indicated by w/o S-B), ALGRNet (without adaptive region learning module) shows an absolute decrease of 1.7% and 1.6% in the average F1 score for facial paralysis estimation on FPara and AU detection on DISFA, respectively. In addition, to explicitly validate the effectiveness of our skipping operation, we use the basic BiLSTM [29] (indicated by w/ Bi) instead of skip-BiLSTM for information sequential transfer across different branches in the ALGRNet (also with Fusion&Refining module), ALGRNet obtains lower average F1 scores of 74.3% and 65.2% on FPara and DISFA, respectively. The performance reduction verifies that roughly

TABLE 6
Mean error (lower is better) results of different face alignment models on BP4D, DISFA and FPara. All values are in %.

Methods	BP4D	DISFA	FPara
JÂA-Net	3.80	3.87	**5.15**
ALGRNet	**3.78**	**3.29**	5.18

defining the relationships between branches related to AU symmetry regions from top to bottom may not be the best way to model the real relationships between AUs. Notably, skipping operation can significantly improve performance, suggesting that our skip-type gates play an important role in our model.

4.6.3 Effects of feature fusion&refining module

Without the fusion&refining module (indicated by w/o F&R in Table 5 for facial paralysis estimation and AU detection, respectively), we directly conduct classification over the output of skip-BiLSTM. The average F1 score drops significantly from 74.6% to 73.3% on FPara and from 67.3% to 66.7% on DISFA, due to the lack of supplementary information from the global face for each patch. In addition, we simply fuse the global features to the local AU features following [9], [10], due to the lack of effective information filtering, the average F1 score drops from 74.6% to 73.9% on FPara and from 67.3% to 66.9% on DISFA. This suggests that the refined local region features from the proposed fusion&refining module, guided by the grid-based global features, significantly contribute to our model.

Finally, after simultaneously removing all the proposed adaptive region learning module, skip-BiLSTM and fusion&refining module (marked by w/o full in Table 5), a significant performance degradation in facial paralysis estimation and AU detection can be observed, *i.e.*, a 3.1% drop on FPara and a 4.4% drop on DISFA in terms of average F1 score. This sufficiently demonstrates that the potential mutual assistance and exclusion relationships between the adaptive AU patches, complemented by the global facial features, can significantly improve the performance of facial AU detection. Furthermore, for facial paralysis estimation, the adaptive local-global interaction based on symmetrical muscles (PAUs) greatly enhances the semantic representation of facial context, obtaining accurate semantic information from potential lesion regions and contextual relational help from the global face.

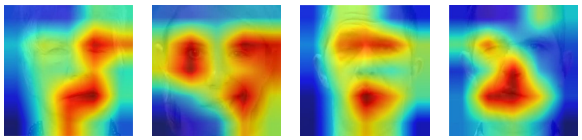


Fig. 6. Class activation maps that show the discriminative regions for different patients with different expressions on FPara datasets. Due to patient confidentiality agreements, we process patient images with strong transparency.

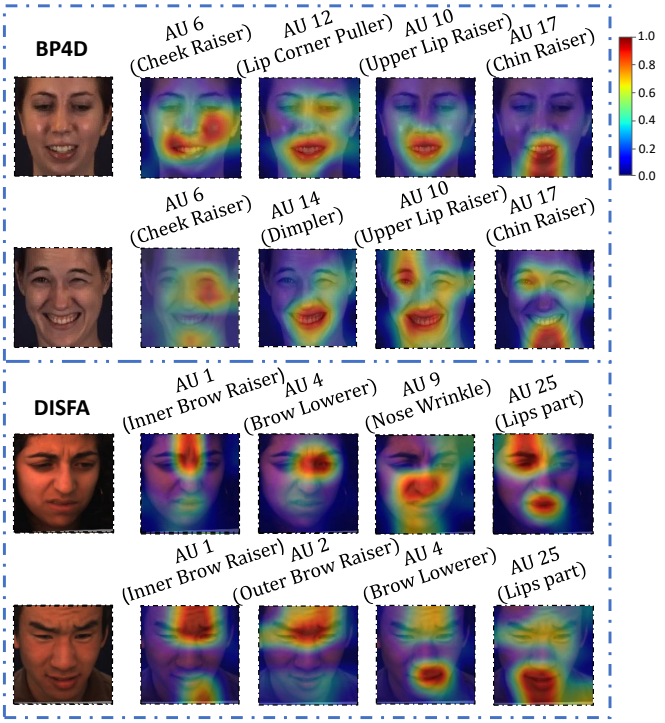


Fig. 7. Class activation maps that show the discriminative regions for different AUs in terms of different expressions and individuals on DISFA and BP4D datasets.

4.7 Results for Face Alignment

We integrate face alignment and facial paralysis estimation into our end-to-end ALGRNet, which can benefit each other as they are coherently related. For example, detected landmarks can help the model focus on the exact location of regions with high probability of muscle lesions as PAU patches. As shown in Table 6, compared with baseline method JÂA-Net [10], our ALGRNet performs comparably to baseline on FPara and better on BP4D and DISFA. The robustness of the adaptive region learning module allows our ALGRNet to outperform JÂA-Net in facial paralysis estimation and AU detection, even if sometimes with slightly lower landmark detection accuracy.

4.8 Visualization of Results

Fig. 6 shows four examples of the learned class activation maps of ALGRNet (the input of classifier) corresponding to different patients. It suggests that our method can mine the relationship between related muscle regions while accurately locating the muscle regions where the underlying disease occurs, thus enhancing the contextual detail of the face representation. For a clearer and adequate display, four examples of the learned class activation maps of ALGRNet (the outputs of F&R module) from two different datasets are given, two of which are from BP4D and two are from DISFA, containing visualization results of different genders with different AU categories. Through the learning of ALGRNet, not only the concerned AU regions can be accurately located, but also the positive (in red) or negative (in blue) correlation with other AU areas can be established and other details of the global face can be supplemented. This obviously improves the flaws of the excessive localisation of JÂA-Net [10] and the negative influence of unrelated regions of ARL [9]. In addition, it also adapts well to irregular muscle areas for different AUs. The heatmaps for the same AU category in the different examples are broadly consistent but also vary slightly by the individual, demonstrating that our ALGRNet can learn certain rules across different datasets and adaptively adjust to different samples.

5 Discussion

ALGRNet, an advanced facial representation stem network based on adaptive facial action units with multi-relational modelling, offers several notable advantages. Firstly, ALGRNet demonstrates outstanding performance in AU detection, showcasing its remarkable facial representation capabilities. This enables its application in a wide range of higher-order decision-making processes. Secondly, we have demonstrated that the features learned by ALGRNet can be effectively utilized either for AU recognition through a multi-branch classification network or seamlessly integrated into a facial paralysis estimation classifier with minimal adjustments to the AU positions. Through identifying symmetrical AUs, we have developed an effective facial palsy detector. This pioneering work explores the effectiveness of an end-to-end deep learning-based AU detection model in predicting the severity of facial paralysis. Facial paralysis is a debilitating condition that affects numerous individuals worldwide. Experimental findings from a meticulously gathered and annotated facial paralysis benchmark, conducted by medical experts, highlight the potential of utilizing identified AU attributes to estimate the severity of facial paralysis.

Limistations. One limitation of this study is the need to train the two tasks (AU and facial paralysis) separately using distinct datasets, rather than combining them within a single model. This is because there is still not known dataset with simultaneous AU and face paralysis annotations. An additional limitation is the low accuracy of a few categories due to category imbalance, although we have achieved the best overall results. In this case, the average F1 score can better reflect a model's performance over the whole grade range.

6 Conclusion

This paper introduces ALGRNet, an innovative adaptive local-global relational network designed for detecting facial action units and also in estimating the severity of facial paralysis through AU detection. ALGRNet capitalizes on the precision and adaptability of muscle region localization and leverages the comprehensive facial semantic feature representation offered by AU detection models. By

harnessing the interactive relationships and interplay between adaptive and symmetrical muscle regions, ALGRNet effectively captures the dynamic nature of these regions across various expressions and individual characteristics. ALGRNet employs a skip-BiLSTM mechanism to facilitate efficient information exchange, allowing for seamless transfer of local muscle features while modelling the potential assistance and exclusion relationships among AU branches. Furthermore, a novel feature fusion and refining module is incorporated into each branch, promoting the synergy between local features and grid-based global features while accommodating irregular muscle regions. We substantiate the effectiveness of our approach by conducting comprehensive experiments on two widely utilized benchmarks for AU detection. Furthermore, we have successfully applied AU detection to the detection of facial paralysis by identifying symmetrical Action Units (PAUs). Our experiments on a benchmark specifically designed for facial paralysis estimation highlighted the remarkable superiority of our method in accurately estimating the severity of facial paralysis.

REFERENCES

[1] 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, 2018.

[2] G. Storey, R. Jiang, S. Keogh, A. Bouridane, and C.-T. Li, "3dpalsynet: a facial palsy grading and motion recognition framework using fully 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 121 655–121 664, 2019.

[3] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, "Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2325–2332, 2020.

[4] S. M. Hossain, Z. Jamal, A. A. Noshin, and M. M. Khan, "Comparative study of deep learning algorithms for the detection of facial paralysis," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2022, pp. 0368–0377.

[5] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[6] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, "Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis," in *European Conference on Computer Vision*, 2014, pp. 151–166.

[7] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[8] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan, "Local relationship learning with person-specific shape regularization for facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 917–11 926.

[9] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *European Conference on Computer Vision*, 2018, pp. 705–720.

[10] ——, "Jaa-net: joint facial action unit detection and face alignment via adaptive attention," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 321–340, 2021.

[11] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit and holistic expression recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3931–3946, 2016.

[12] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," *IEEE transactions on Affective Computing*, 2019.

[13] Y. Wu and Q. Ji, "Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3400–3408.

[14] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[15] S. Wang, G. Peng, and Q. Ji, "Exploring domain knowledge for facial expression-assisted action unit activation recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 640–652, 2018.

[16] T. Song, Z. Cui, W. Zheng, and Q. Ji, "Hybrid message passing with performance-driven structures for facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6267–6276.

[17] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8594–8601.

[18] X. Niu, H. Han, S. Shan, and X. Chen, "Multi-label co-regularization for semi-supervised facial action unit recognition," in *Advances in Neural Information Processing Systems*, 2019, pp. 909–919.

[19] T. Song, Z. Cui, Y. Wang, W. Zheng, and Q. Ji, "Dynamic probabilistic graph convolution for facial action unit intensity estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4845–4854.

[20] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *MultiMedia Modeling*. Springer, 2020, pp. 489–501.

[21] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net: Deep nets with enhancing and cropping for facial action unit detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2583–2596, 2018.

[22] T. Song, L. Chen, W. Zheng, and Q. Ji, "Uncertain graph neural networks for facial action unit detection," in *AAAI Conference on Artificial Intelligence*, 2021, p. 5993–6001.

[23] X. Ge, P. Wan, H. Han, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Local global relational network for facial action units recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021, pp. 01–08.

[24] G.-S. J. Hsu, W.-F. Huang, and J.-H. Kang, "Hierarchical network for facial palsy detection." in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 580–586.

[25] D. R. Rubinow and R. M. Post, "Impaired recognition of affect in facial expression in depressed patients," *Biological Psychiatry*, vol. 31, no. 9, pp. 947–953, 1992.

[26] L. I. Reed, M. A. Sayette, and J. F. Cohn, "Impact of depression on response to comedy: a dynamic facial coding analysis." *Journal of Abnormal Psychology*, vol. 116, no. 4, p. 804, 2007.

[27] R. S. Feldman, L. Jenkins, and O. Popoola, "Detection of deception in adults and children via facial expressions," *Child Development*, pp. 350–355, 1979.

[28] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of Neuroscience Methods*, vol. 200, no. 2, pp. 237–256, 2011.

[29] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[30] X. Ge, F. Chen, C. Shen, and R. Ji, "Colloquial image captioning," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2019, pp. 356–361.

[31] G. Barrios Dell'Olio and M. Sra, "Farapy: An augmented reality feedback system for facial paralysis using action unit intensity estimation," in *The 34th Annual ACM Symposium on User Interface Software and Technology*, 2021, pp. 1027–1038.

[32] A. Gaber, M. F. Taher, M. Abdel Wahed, N. M. Shalaby, and S. Gaber, "Comprehensive assessment of facial paralysis based on facial animation units," *Plos One*, vol. 17, no. 12, p. e0277297, 2022.

[33] Y. Zhu, F. De la Torre, J. F. Cohn, and Y.-J. Zhang, "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 79–91, 2011.

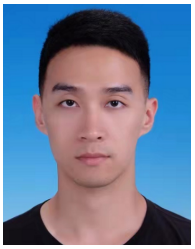
[34] C. Tang, W. Zheng, J. Yan, Q. Li, Y. Li, T. Zhang, and Z. Cui, "View-independent facial action unit detection," in *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2017, pp. 878–882.

[35] C. Ma, L. Chen, and J. Yong, "Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection," *Neurocomputing*, vol. 355, pp. 35–47, 2019.

[36] I. Ntinou, E. Sanchez, A. Bulat, M. Valstar, and Y. Tzimiropoulos, "A transfer learning approach to heatmap regression for action unit intensity estimation," *IEEE Transactions on Affective Computing*, 2021.
[37] X. Ge, J. M. Jose, S. Xu, X. Liu, and H. Han, "Mgrr-net: Multi-level graph relational reasoning network for facial action units detection," *arXiv preprint arXiv:2204.01349*, 2022.
[38] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–8.
[39] Z. Shao, Y. Zhou, H. Zhu, W.-L. Du, R. Yao, and H. Chen, "Facial action unit recognition by prior and adaptive attention," *Electronics*, vol. 11, no. 19, p. 3047, 2022.
[40] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.
[41] T. H. Ngo, M. Seo, N. Matsushiro, W. Xiong, and Y.-W. Chen, "Quantitative analysis of facial paralysis based on limited-orientation modified circular gabor filters," in *International Conference on Pattern Recognition*. IEEE, 2016, pp. 349–354.
[42] Z. Guo, W. Li, J. Dai, J. Xiang, and G. Dan, "Facial imaging and landmark detection technique for objective assessment of unilateral peripheral facial paralysis," *Enterprise Information Systems*, vol. 16, no. 10-11, pp. 1556–1572, 2022.
[43] G. E. Murty, J. P. Diver, P. J. Kelly, G. O'Donoghue, and P. J. Bradley, "The nottingham system: objective assessment of facial nerve function in the clinic," *Otolaryngology—Head and Neck Surgery*, vol. 110, no. 2, pp. 156–161, 1994.
[44] W. House, "Facial nerve grading system," *Otolaryngol Head Neck Surg*, vol. 93, pp. 184–193, 1985.
[45] J. Dong, L. Ma, Q. Li, S. Wang, L.-a. Liu, Y. Lin, and M. Jian, "An approach for quantitative evaluation of the degree of facial paralysis based on salient point detection," in *International Symposium on Intelligent Information Technology Application Workshops*. IEEE, 2008, pp. 483–486.
[46] J. Barbosa, K. Lee, S. Lee, B. Lodhi, J.-G. Cho, W.-K. Seo, and J. Kang, "Efficient quantitative assessment of facial paralysis using iris segmentation and active contour-based key points detection with hybrid classifier," *BMC medical imaging*, vol. 16, no. 1, pp. 1–18, 2016.
[47] G. Cheng, J. Dong, S. Wang, H. Qu *et al.*, "Evaluation of facial paralysis degree based on regions," in *Third International Conference on Knowledge Discovery and Data Mining*. IEEE, 2010, pp. 514–517.
[48] T. Wang, S. Zhang, J. Dong, L. Liu, and H. Yu, "Automatic evaluation of the degree of facial nerve paralysis," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11 893–11 908, 2016.
[49] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognition*, vol. 130, p. 108792, 2022.
[50] Z. Zha, H. Tang, Y. Sun, and J. Tang, "Boosting few-shot fine-grained recognition with background suppression and foreground alignment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
[51] C. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," in *European Conference on Computer Vision*, 2018, pp. 298–313.
[52] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *The Association for Computer Linguistics*, 2015, pp. 1556–1566.
[53] X. Ge, F. Chen, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Structured multi-modal feature embedding and alignment for image-sentence retrieval," in *ACM International Conference on Multimedia*, 2021, pp. 5185–5193.
[54] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, pp. 802–810, 2015.
[55] S. Wang, Y. Chang, and C. Wang, "Dual learning for joint facial landmark detection and action unit recognition," *IEEE Transactions on Affective Computing*, 2021.
[56] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.
[57] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.
[58] B. F. O'Reilly, J. J. Soraghan, S. McGrenary, and S. He, "Objective method of assessing and presenting the house-brackmann and regional grades of facial palsy by production of a facogram," *Otology & Neurotology*, vol. 31, no. 3, pp. 486–491, 2010.
[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026–8037.
[60] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, 2013, pp. 1139–1147.
[61] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.
[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[63] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 67–74.
[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[65] H. Yang, T. Wang, and L. Yin, "Adaptive multimodal fusion for facial action units recognition," in *ACM International Conference on Multimedia*, 2020, pp. 2982–2990.
[66] P. Liu, Z. Zhang, H. Yang, and L. Yin, "Multi-modality empowered network for facial action unit detection," in *IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 2175–2184.
[67] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46.
[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[69] 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.

Xuri Ge received the M.S. degree in computer science from the School of Information Science and Engineering, Xiamen University, China, in 2020. He is currently pursuing the Ph.D. degree with the school of computer science, University of Glasgow, Scotland, UK. His current research interests include computer vision, medical image processing, and multimedia processing.

Joemon M. Jose is a Professor at the School of Computing Science, University of Glasgow, Scotland and a member of the Information Retrieval group. His research focuses around the following three themes: (i) Social Media Analytics; (ii) Multi-modal interaction for information retrieval; (iii) Multimedia mining and search. He has published over 300 papers with more than 9200 Google Scholar citations, and an h-index of 48. He leads the Multimedia Information Retrieval group which investigates research issues related to the above themes.



Pengcheng Wang is currently a Researcher with Tomorrow Advancing Life Education Group, Beijing, China. His research interests include the applied AI, such as intelligent multimedia processing, and computer vision. As a Key Team Member, he achieved the best performance in various competitions, such as the EmotioNet facial expression recognition challenge.

Arunachalam Iyer is a Honorary Associate Clinical Professor of Surgery, University of Glasgow, Scotland, UK. He is also a consultant ENT surgeon, University Hospital Monklands, UK. He is currently working as a Consultant ENT surgeon & Otologist at Lanarkshire and a council member of the British society of Otology and the section of Otology, Royal society of Medicine. He also teaches at the Temporal bone course at University of Glasgow and is involved in training junior doctors.

Xiao Liu received the Ph.D. degree in computer science from Zhejiang University in 2015. He worked at Baidu from 2015 to 2019 and at Tomorrow Advancing Life Education Group from 2019 to 2021. He is currently a Researcher with the Online Media Business Unit at Tencent, Beijing, China. His research interests include the applied AI, such as intelligent multimedia processing, computer vision, and learning systems. His research results have expounded in more than 40 publications at journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, CVPR, ICCV, ECCV, AAAI, and MM. As a Key Team Member, he achieved the best performance in various competitions, such as the ActivityNet challenges, NTIRE super resolution challenge, and EmotioNet facial expression recognition challenge.

Hu Han (Member, IEEE) is a Professor at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). His research interests include computer vision, pattern recognition, biometrics, and medical image analysis. He has published more than 70 papers with more than 6000 Google Scholar citations. He was a recipient of the IEEE Signal Processing Society Best Paper Award (2020), ICCV2021 Human-centric Trustworthy Computer Vision Best Paper Award, IEEE FG2019 Best Poster Presentation Award, and 2016/2018 CCBR Best Student/Poster Award. He is/was the Associate Editor of Pattern Recognition, IJCAI2021 SPC, ICPR2020 Area Chair, ISBI2022 Session Chair, and VALSE LAC. He has organized a number of special sessions and workshops of "vision based vital sign analysis and health monitoring" in ICCV2021, CVPR2020, FG2019/20/21.