Local Global Relational Network for Facial Action Units Recognition

Xuri Ge[1*], Pengcheng Wang[2*], Hu Han[3], Joemon M. Jose[1], Zhilong Ji[2], Zhongqin Wu[2], Xiao Liu[2]
[1]School of Computing Science, University of Glasgow, UK. [2]TAL Education Group, Beijing, China.
[3]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, CAS, Beijing, China.
x.ge.2@research.gla.ac.uk, wangpengcheng@tal.com, hanhu@ict.ac.cn, Joemon.Jose@glasgow.ac.uk,
{jizhilong,zhongqinwu,liuxiao15}@tal.com

***Abstract*— Many existing facial action units (AUs) recognition approaches often enhance the AU representation by combining local features from multiple independent branches, each corresponding to a different AU. However, such multi-branch combination-based methods usually neglect potential mutual assistance and exclusion relationship between AU branches or simply employ a pre-defined and fixed knowledge-graph as a prior. In addition, extracting features from pre-defined AU regions of regular shapes limits the representation ability. In this paper, we propose a novel Local Global Relational Network (LGRNet) for facial AU recognition. LGRNet mainly consists of two novel structures, *i.e.*, a skip-BiLSTM module which models the latent mutual assistance and exclusion relationship among local AU features from multiple branches to enhance the feature robustness, and a feature fusion&refining module which explores the complementarity between local AUs and the whole face in order to refine the local AU features to improve the discriminability. Experiments on the BP4D and DISFA AU datasets show that the proposed approach outperforms the state-of-the-art methods by a large margin.**

I. Introduction

Facial expression recognition has wide potential applications in diagnosing mental disease [26], improving e-learning experiences [24], detecting deception [3], face recognition and attribute estimation [10], [7], [6], assisting teaching in education [1], [27], *etc.* As a fundamental research problem, facial action units (AU) recognition is beneficial to facial expression recognition and analysis, and has received increasing attention in recent years. However, AU recognition is challenging because of the difficulty in identifying the subtle facial changes caused by AUs. Looking from biological perspective, the activation of AU corresponds to the movement of facial muscles, which inspired earlier works such as [34], [15] to design hand-crafted features to represent the appearance of different local facial regions. However, hand-crafted features are not discriminative enough to represent the facial morphology due to their shallow natures. Hence, in recent years deep learning based AU recognition methods have been studied to enhance the AU's feature representation.

Many existing automatic facial AU recognition methods aim to enhance the facial feature representation by combining local features from multiple independent branches, which are related to regions of different AUs. Some grid-based deep learning frameworks [16], [18] incorporate regional (patch-based) Convolutional Neural Network (CNN) features from a face with equal grids, as shown in Fig. 1 (a). For instance, the scheme in [23] combines local CNN features from equal partition grids by an LSTM [8]. However, dividing images into fixed grids leads to a number of issues: (i) it is difficult to focus exactly on the muscle area corresponding to each AU; (ii) ROIs for AUs with irregular shapes may not be well represented by grid-based features. To address the above issues, recent popular multi-branch combination-based methods [39], [30], [29] fuse global or local features from independent AU branches based on the corresponding muscle region detection, to refine the features for AUs with irregular regions, as shown in Fig. 1 (b). For instance, an end-to-end multi-branch framework in [28] is proposed to combine the features from independent branches for individual AU related muscle regions according to some predefined attention maps based on detected landmarks.

While the multi-branch combination based AU recognition methods show their effectiveness in local AU feature fusion,

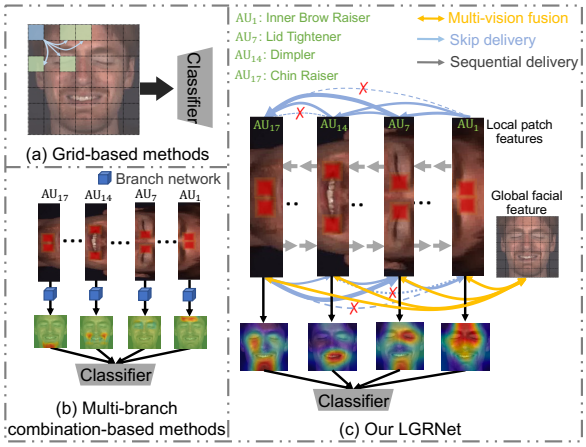

Fig. 1. Illustration of the different AU feature learning and classification schemes: (a) the traditional grid-based feature extraction and classification, (b) the multi-branch combination-based recognition methods, and (c) our LGRNet method. Compared with (a) and (b), our LGRNet exploits mutual assistance and mutual exclusion relations of local facial patch-based multiple branches via a novel bi-directional structure with skipping gates and refines their irregular representations by the global facial feature (best viewed in color).

*The co-first authors contribute equally. This work was supported in part by National Key R&D Program of China (No. 2020AAA0104500), China Scholarship Council (CSC) from the Ministry of Education of P.R. China (No. 202006310028), Natural Science Foundation of China (grants 61732004 and 61672496), and Youth Innovation Promotion Association CAS (grant 2018135).

there are still limitations in modeling their mutual relationship as well as the local-global context. On one hand, the multiple patches related to individual AUs, may have a strong positive or negative latent correlation in most expressions Here, if multiple AUs jointly affect the target AU category, it is defined as positive correlation (mutual assistance), otherwise 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. In addition, some AUs may not to be activated simultaneously, *e.g.*, we cannot simultaneously stretch our mouth ("AU20") and raise our cheek ("AU6"). Inspired by these biological phenomenons, we argue that capturing the interactive information delivery between patch-based branches, such as sequential/skipping delivery of adjacent/non-adjacent related regions, is important for enhancing the representation of AU features. On the other hand, the global face feature provides important cues to refine the limited regular patch features, which is important to deal with irregular muscle shapes. This is because the local AU patches may not cover the entire face, and other non-AU regions may also be activated due to muscle linkage. To the best of our knowledge, the above two key issues are left unexploited in the literature.

To address the above problems, we propose a novel LGRNet for facial AU recognition. In particular, we first extract the grid-based global feature by multi-layer CNNs and local AU features based on the detected facial landmarks. Then, we use spatially ordered AU branches as initialization to replace the conventional disordered AU branches, which is based on the muscle positions of AUs from top to bottom (see Fig. 1 (c)). This is because adjacent muscle areas have a natural potential for correlation from biology. We then design a skip-BiLSTM to capture the potential assistance and exclusion relations among these sequential branches, where the adjacent patches are adjustable transfer in BiLSTM [5] while the distant patches are connected via skipping-type gates. We argue that each AU branch is independent and equal, so such a skip connection manner can minimize the loss of information compared with traditional BiLSTM. Moreover, we design a novel feature fusion&refining module to refine the local features from skip-BiLSTM guided by global grid-based features. Different with previous feature fusion methods [4], our gated fusion architecture in feature fusion&refining module can appropriately supplement global information, even non-AU region information, for each local AU patches. It is very important because different AUs may focus on different global information. Finally, the features learned by LGRNet are fed to a multi-branch classification network for AU recognition.

The contributions of our LGRNet for facial AU recognition are as follows:

- We propose a skip-BiLSTM approach to model the mutual assistance and exclusion relationship of individual AUs, which leads to improved robustness in AU recognition;
- We propose a method for local AU feature refinement with the assistant of global grid-based features, making the local AU features more discriminative;
- The proposed LGRNet outperforms the state-of-the-art approaches for AU recognition on two benchmarks, *i.e.,* BP4D and DISFA.

II. Related Work

AU recognition has been studied for decades and several methods have been proposed for this problem. Most existing methods for AU recognition are based on patch learning [40], [38], [9], [13], [19]. For instance, [33] used sparse coding to recover facial expressions using the composition rules of different fixed patches for different AUs. [38] performed a patch selection approach, where patches for AUs were selected by group sparsity learning for structure learning with shallow representations. [9] proposed to use domain knowledge and facial geometry to pre-select a relevant image region for a particular AU and feed it to a convolutional and bi-directional Long Short-Term Memory (LSTM) neural network. However, all above methods need to predefined the patch location first. To address these issues, [40] proposed a set of adaptive ROI cropping nets, based on local convolutional neural network, to learn regional features separately. [28] jointed facial AU recognition and face alignment in an end-to-end framework, where the face alignment results can aid AUs to learn the irregular attention distribution of the ROI of AU patches. [23] leveraged the facial shape as a regularization term in order to learn person-independent AU features.

Recent works in facial AU recognition also pay attention to capture the interactions of different AUs for local feature enhancement with multi-label learning. On one hand, taking into account the relationship of multiple face patches can provide more robustness than using single patch. [22] embedded the relations among AUs through a predefined graph convolutional network (GCN). [12] incorporated AU knowledge-graph as an extra guidance for the enhancement of facial region representation. However, these methods need the prior connections by co-occurrence probability in different datasets. On the other hand, some approaches tried to apply the local relationship information into multi-label learning. For instance, [39] proposed a joint patch learning and multi-label learning method, in which the local regions of AUs are defined as patches centered around the facial landmarks.

In contrast to previous studies, LGRNet automatically models the relation structure of the facial AUs by the use of a contextual structure along with a skipping operation. The most relevant existing works to ours are [28], [29], which combine facial AU recognition and face alignment into a multiple independent branches network. Different from these methods, our LGRNet is capable of exploiting the learned correspondence of different AUs to enhance the target local AU, as well as considering other non-AU regions. Doing so allows us to provide more robustness than [29], which also improves the interpretability of the model.

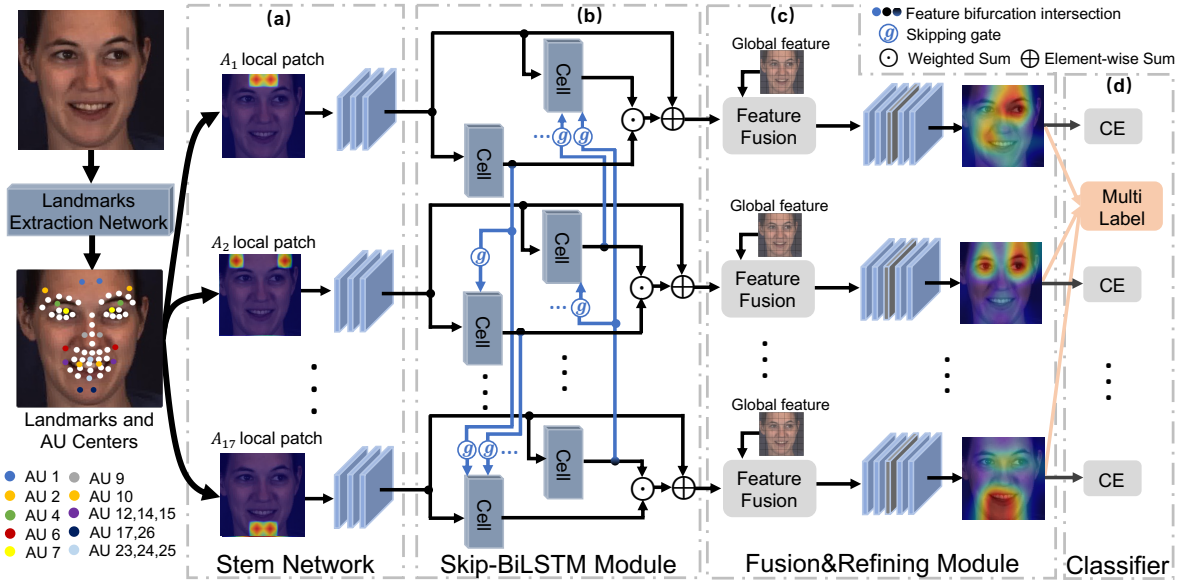

Fig. 2. The overall architecture of the proposed LGRNet for facial AU recognition. We utilize a simple but efficient landmark localization network to detect the landmarks of AU centers, which are used to compute local AU patches. The AU patches are fed into our multi-branch LGRNet, with one branch per AU, and a skip-BiLSTM module is proposed to model mutual assistance and exclusion relationship among different AU branches. Then, a feature fusion&refining module is designed to refine the local features with the assistance of the global grid-based features; so that they can capture more discriminative features for AU with irregular shapes. Finally, a multi-classifier is used to predict the activation probabilities of individual AUs (best viewed in color).

III. Approach

Fig. 2 shows the overall diagram of the proposed LGRNet approach, which consists of a skip-BiLSTM module for mutual assistance and exclusion modeling, a feature fusion&refining module for refining features of irregular AU regions, and a multi-classifier module for predicting the AU activation probability. We provide the details below.

A. Overview of LGRNet

Similar to [2], [28], we also employ a multi-branch network for the multi-label facial AU recognition task. However, different from previous approaches, we argue that exploiting the relationship among multiple patches plays a vital role in building a robust AU recognition model. In this way, we design two modules (skip-BiLSTM module and Fusion&Refining module) based on the foundation of existing multi-branch network, which can fully mine the local and global interactive relations among the AU patches and non-AU regions, and obtain a more discriminative and robust representation.

We first employ a region learning module and face alignment module simultaneously as our stem network from the widely used multi-branch network [28]. Given a face image I, we adapt these modules by getting the patch regions based on the extracted landmarks. In particular, [28] contains a hierarchical and multi-scale region learning network which can extract features from each local patch with different scales, thus obtaining multi-scale AU representations. Different from other complex face alignment methods, we utilize an efficient landmark extraction network similar to [29], including three convolutional layers connected to a max-pooling layer. Note that stem network is shared for all branches, which greatly reduces training costs and the complexity of network training. According to the learned landmarks, *i.e.*, $L = \{l_1, l_2, ..., l_m\}$, local patches $A = \{A_1, A_2, ..., A_n\}$ are calculated and their features $V = \{v_1, v_2, ..., v_n\}$ are learned via the stem network, where m and n are the numbers of landmarks and selected patches, respectively. For simplicity, we do not repeat the detailed structure of the stem network here.

In order to overcome the lack of adequate delivery of local patch information among individual patches, we design a novel skip-BiLSTM module (detailed in Section III-B), which can transmit information in two ways (sequential delivery or skipping delivery). The sequential delivery of information can fully explore the contextual relationship between adjacent patches. The skipping delivery focuses on the information interactive of non-adjacent related patches. Different from the traditional sequence spreading of LSTM, our skip-BiLSTM can directly calculate the correlation between a target AU and all previous AUs in the forward and backward directions. This is beneficial because there is little information loss during multi-branch transmission. After skip-BiLSTM, we get a set of local patch features $S = \{s_1, s_2, ..., s_n\}$, which are expected to have all the useful information from adjacent and non-adjacent patches.

Furthermore, we argue that the non-AU regions can be helpful for refining the local patch features and obtain salient micro-level features for the global face, which may be useful for handling irregular AU regions. Hence, we design a novel feature Fusion&Refining module (detailed in Section III-C), which can concentrate on the salient information from global facial feature G. Finally, the local patch features are integrated with global facial features as new patch-based

representation $R = \{r_1, r_2, ..., r_n\}$.

The face alignment and facial AU recognition are integrated into an end-to-end learning model. Our goal is to jointly learn all the parameters by minimizing both face alignment loss and facial AU recognition 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 [29]. In this paper, 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,

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

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 [28]. Moreover, we also employ a weighted multi-label Dice coefficient loss [21] to overcome the sample imbalance problem, which is formulated as:

$$\mathcal{L}_{dice} = \frac{1}{n}\sum_{i=1}^{n} w_i(1 - \frac{2p_i\hat{p}_i + \tau}{p_i^2\hat{p}_i^2 + \tau}), \quad (3)$$

where τ is the smooth term. Finally, the facial AU recognition loss is defined as:

$$\mathcal{L}_{au} = \mathcal{L}_{rec} + \mathcal{L}_{dice}, \quad (4)$$

Finally, the joint loss of our LGRNet is defined as:

$$\mathcal{L} = \mathcal{L}_{au} + \lambda\mathcal{L}_{align}. \quad (5)$$

where λ is a balancing parameter.

B. Skip-BiLSTM

Fig. 2 (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 [23], we regard the multiple patches as a sequence structure from top to bottom, which can transfer information by a Bi-directional LSTM based model [5] with our skipping-type gate. Different from the traditional BiLSTM, 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 can promote the transfer of relevant AUs information. The above process can be formulated as:

$$\overrightarrow{h_t} = \text{Cell}(\sum_{j=1}^{t-1}\overrightarrow{\hat{h}_j}, v_t), \quad (6)$$

$$\overrightarrow{\hat{h}_j} = \overrightarrow{h_j} f_j, \quad (7)$$

$$f_j = \sigma(\text{GAP}(W_j(\overrightarrow{h_j} v_t))), \quad (8)$$

where $\text{Cell}(\cdot)$ denotes the basic ConvLstm cell [31], σ is sigmoid function, and GAP denotes the global average pooling operation. W_j is the parameters of mapping function, in which we used Conv2D. For the backward delivery, we get the t-th patch feature which goes to the same forward process as:

$$\overleftarrow{h_t} = \text{Cell}(\sum_{j=t+1}^{n}\overleftarrow{\hat{h}_j}, v_t), \quad (9)$$

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:

$$s_t = v_t + (\overrightarrow{h_t} + \overleftarrow{h_t})/2, \quad (10)$$

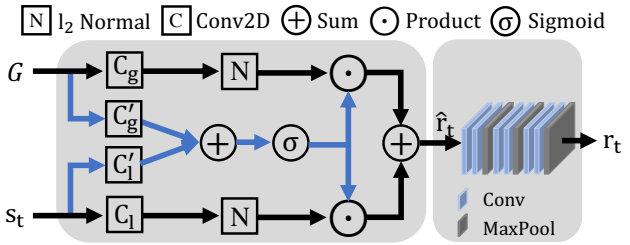

Fig. 3. The architecture of our feature fusion&refining module guided by global face feature.

C. Feature Fusion&Refining

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 AU branch because different AUs may focus on different global information. The grid-based global face feature G is extracted using a simple CNN with the same structure as the face alignment network [29]. As shown in Fig. 3, 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:

$$\alpha = \sigma(\mathrm{C}_g^{'} G + \mathrm{C}_l^{'} s_t), \quad (11)$$

$$\hat{r}_t = \alpha \odot ||C_g G||_2 \oplus (1 - \alpha) \odot ||C_l s_t||_2, \quad (12)$$

where σ is the sigmoid function, and $||\cdot||$ denotes the l_2-normalization. $\mathrm{C}_*^{'}$ 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. 3. F&R module contains

TABLE I

COMPARISONS OF AU RECOGNITION FOR 8 AUS ON DISFA IN TERMS OF F1-FRAME SCORE (IN %).

Method	AU Index								Avg.
	1	2	4	6	9	12	25	26	
EAC-Net [14]	41.5	26.4	66.4	**50.7**	**80.5**	**89.3**	88.9	15.6	57.4
JAA-Net [28]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
LP-Net [23]	29.9	24.7	**72.7**	46.8	49.6	72.9	93.8	65.0	56.9
ARL [30]	43.9	42.1	63.6	41.8	40.0	76.2	**95.2**	66.8	58.7
JÂA-Net [29]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
LGRNet (Ours)	**62.6**	**64.4**	72.5	46.6	48.8	75.7	94.4	**73.0**	**67.3**

TABLE II

COMPARISONS OF AU RECOGNITION FOR 12 AUS ON BP4D IN TERMS OF F1-FRAME SCORE (IN %).

Method	AU Index												Avg.
	1	2	4	6	7	10	12	14	15	17	23	24	
EAC-Net [14]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
MLCR [22]	42.4	36.9	48.1	77.5	**77.6**	83.6	85.8	61.0	43.7	**63.2**	42.1	55.6	59.8
JAA-Net [28]	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
LP-Net [23]	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 [30]	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 [29]	**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
LGRNet (Ours)	50.8	47.1	57.8	77.6	77.4	**84.9**	**88.2**	**66.4**	**49.8**	61.5	46.8	52.3	**63.4**

three blocks. Each block consists of two convolutional layers and a maxpooling layer. Then multi-patch features R are sent to the multi-label binary classifier to calculate the occurrence probabilities of individual AUs.

IV. EXPERIMENTS

A. Dataset and Implementation Detail

Dataset. We evaluate the effectiveness of the proposed approach on the popular BP4D [37] and DISFA [20] datasets. **BP4D** consists of 328 facial videos from 41 participants (23 females and 18 males) who were involved in 8 sessions. Similar to [13], [30], [29], 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 [13], [29]. In comparison to BP4D, the experimental protocol and lighting conditions deliver DISFA to be a more challenging dataset. Following the experiment setting of [29], we evaluated the model using the 3-fold subject-exclusive cross-validation protocol.

Training strategy. Our model is trained on a single NVIDIA Tesla V100 GPU with 32 GB memory. The whole network is trained with the default initializer of PyTorch [25] with the stochastic gradient descent (SGD) solver, a Nesterov momentum [32] 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 2 epochs. Maximum epoch number is set to 20. To enhance the diversity of training data, aligned faces are further randomly cropped into 176×176 and horizontally flipped. Regarding face alignment network and stem network, we set the value of the general parameters to be the same with [29]. The filters for the convolutional layers in refining architecture are used 3×3 convolutional filters with a stride 1 and a padding 1. In our paper, all of the mapping Conv2D operations are used 1×1 convolutional filters with a stride 1 and a padding 1. The dimensionality of hidden state in ConvLstm cell is set to 64 and the filters for the convolutional layers in ConvLstm cell are used 3×3 convolutional filters with a stride 1 and a padding 1. λ is set to 0.5 for the jointly optimizing of face alignment and facial AU recognition. For comparison purpose, the numbers of AUs are 12 (as in [13], [30], [29]) and 8 (as in [13], [29]) for BP4D and DISFA respectively. During training, each frame is annotated with 49 landmarks detected and calculated by SDM [35].

Performance Metric. For all methods, F1 score for all the AUs on BP4D and DISFA are calculated and then averaged (denoted as **Avg.**) for comparison.

B. Comparison with State-of-the-art Methods

We compare our proposed LGRNet with several baselines on the BP4D and DSIFA datasets in Table II and Table I, including EAC-Net [14], MLCR [22], JAANet [28], LP-Net [23], ARL [30], and JÂA-Net [29]. Note that, the best and second best results are shown using bold and underline, respectively. The performances of the baselines in Table I and II are the reported results. We omit models [11], [36], [17] that require additional annotated data.

Quantitative comparison on DISFA: AU recognition results by different methods on DISFA are shown in Table I, where the proposed LGRNet shows clear improvements for

TABLE III

EFFECTIVENESS OF KEY COMPONENTS OF LGRNET EVALUATED ON DISFA IN TERMS OF F1-FRAME SCORE (IN %).

Methods	Setting		AU Index								**Avg.**
	S-B	F&R	1	2	4	6	9	12	25	26	
w/o full			47.1	61.1	66.3	44.7	52.2	74.9	92.2	66.2	63.1
w/o F&R	√		62.6	64.2	72.4	42.3	49.9	**76.1**	93.5	72.6	66.7
w/o S-B		√	58.7	**65.2**	**73.5**	43.9	**53.5**	72.2	94.1	64.7	65.7
w/ Bi		√	61.1	58.4	70.9	45.5	47.9	74.9	92.5	70.8	65.2
LGRNet	√	√	**62.6**	64.4	72.5	**46.6**	48.8	75.7	**94.4**	**73.0**	**67.3**

TABLE IV

EFFECTIVENESS OF KEY COMPONENTS OF LGRNET EVALUATED ON BP4D IN TERMS OF F1-FRAME SCORE (IN %).

Methods	Setting		AU Index											**Avg.**
	S-B	F&R	1	2	4	6	7	10	12	14	15	17	23	24	
w/o full			50.1	47.1	54.3	77.3	75.1	82.5	88.1	61.7	44.9	62.7	45.2	49.9	61.6
w/o F&R	√		50.4	46.9	53.4	**79.0**	77.4	84.7	87.4	63.0	45.3	**63.3**	47.0	**55.7**	62.8
w/o S-B		√	**51.3**	47.6	56.3	78.2	76.2	83.7	88.1	64.4	49.1	61.9	46.1	49.8	62.7
w/ Bi		√	50.7	**50.0**	55.2	77.0	75.7	84.1	88.2	63.4	49.1	62.3	**47.3**	52.0	62.9
LGRNet	√	√	50.8	47.1	**57.8**	77.6	**77.4**	**84.9**	**88.2**	**66.4**	**49.8**	61.5	46.8	52.3	**63.4**

TABLE V

MEAN ERROR (%) RESULTS OF DIFFERENT FACE ALIGNMENT MODELS ON DISFA AND BP4D (LOWER IS BETTER).

Methods	DISFA	BP4D
JÂA-Net	4.02	3.80
LGRNet	3.68	3.34

several AUs annotated in DISFA compared with the state-of-the-art methods. Specifically, compared with the state-of-the-art method JÂA-Net [29], our LGRNet achieves 3.8% improvements in terms of average F1 score and also achieves significantly outperforms for all AUs annotated in DISFA. Furthermore, we achieve the best performance in terms of average F1 score compared with all baselines.

Quantitative comparison on BP4D: AU recognition results by different methods on BP4D are shown in Table II, where the proposed LGRNet outperforms the state-of-the-art methods with impressive margins. LGRNet achieves 1.0% higher average F1 score compared with JÂA-Net. Furthermore, LGRNet achieves the best or second-best recognition performance for most of the 12 AUs annotated in BP4D compared with the state-of-the-art methods.

Experimental results of our LGRNet demonstrate its effectiveness in improving AU recognition accuracy on BP4D and DISFA, as well as good generalization ability.

C. Ablation Studies

We perform detailed ablation studies on DISFA and BP4D to investigate the effectiveness of each component of our proposed LGRNet for facial AU recognition.

1) Effects of skip-BiLSTM: In Table III and IV, LGRNet decreases absolutely by 1.6% and 0.7% in terms of average F1 score when removing skip-BiLSTM (indicated by w/o S-B) on DISFA and BP4D, respectively. Furthermore, in order to fully verify the effectiveness of our skipping operation, we replace skip-BiLSTM with the basic BiLSTM [5] (indicated by w/ Bi) for information delivery between multiple branches in our LGRNet (also with Fusion&Refining module), LGRNet achieves lower average F1 scores of 65.2% and 62.9% on DISFA and BP4D, respectively. These observations indicate that a rough definition of the relationship between AUs from top to bottom may not be the best way to simulate the real relationship between AUs. And skipping operation can significantly boost the performance, which suggests that our skipping-type gates play a vital role in our model. In addition, these results also indicate that using skip-BiLSTM to model the mutual assistance and exclusion relationship between AUs is effective for improve AU recognition accuracy.

2) Effects of feature fusion&refining module: Without fusion&refining module (indicated by w/o F&R in Table III and IV), we directly conduct classification over the output of skip-BiLSTM. Great AU recognition performance degradation can be observed, *i.e.*, 0.6% average F1 score drop on DISFA and BP4D. It suggests that the proposed fusion&refining module, which refines local AU feature guided by grid-based global feature, plays a vital role in our model.

Finally, when we simultaneously drop the skip-BiLSTM and fusion&refining modules (indicated by w/o full in Table III and IV), the average F1 score of our method reduces

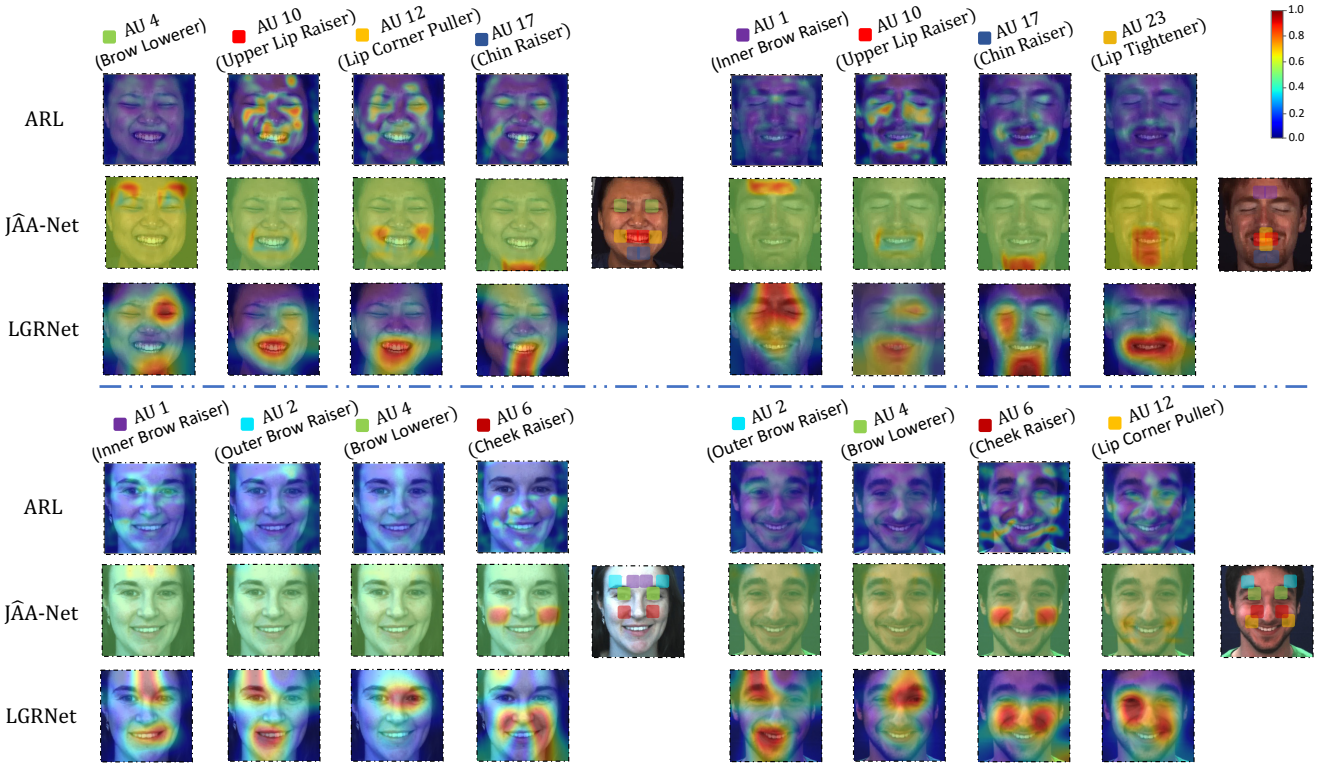

Fig. 4. Visual comparisons of the predicted heatmaps by different methods for four examples from the BP4D dataset (in the first row) and DISFA dataset (in the second row). The learned weights are visualized with different colors in the color bar, which are overlaid on the images. "■" denote the different muscle regions of corresponding AUs with different colors, which are calculated by the detected landmarks. Note that the "■" of different colors in the original image on the right are only for a clearer distinction between different AUs (best viewed in color).

from 67.3% to 63.1% on DISFA, and from 63.4% to 61.6% on BP4D, respectively. We have observed that considering both of skip-BiLSTM and fusion&refining modules can significantly boost the performance of facial AU recognition.

D. Results for Face Alignment

We jointly take face alignment network into our LGRNet via auxiliary training, which can provide effective muscle regions based on the detected landmarks corresponding to each AU. Table V shows the mean error results of our LGRNet and baseline method JÂA-Net [29] on DISFA and BP4D. Compared with JÂA-Net, our LGRNet achieves competitive 3.68 and 3.34 mean error on DISFA and BP4D respectively, which indicates the effectiveness of our joint training.

E. Visualization of Results

To better understand the effectiveness of our proposed model, we visualize the learned heatmaps of LGRNet (the outputs of F&R module) and other methods, corresponding to different AUs, as shown in Fig. 4. Four examples 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 LGRNet, local patches not only concentrate on their own regions, but can also establish a positive correlation with other patches as well as other non-AU regions. Different from the excessive localization of JÂA-Net [29] and the bad influence of unrelated regions of ARL [28], our LGRNet accurately captures potential mutual assistance (in red) and mutual exclusion (in blue) relationships of the local patch feature for each AU and other assistance AUs, as well as non-AU regions in global face, which can improve the discriminative ability of each AU. The heatmaps of the same AU category in different examples are roughly consistent, but there are also differences due to individual differences. This reveals that our LGRNet can learn certain rules in different datasets and automatically make adjustments based on different samples, compared with the predefined GCN methods [12], [22].

V. Conclusion

In this work, we study the problem of facial action units recognition and propose a novel multi-branch multi-label based approach namely LGRNet. The proposed approach enables efficient information delivery via a novel skip-BiLSTM and models the potential mutual assistance and exclusion relationships among spatially ordered branches for local AU features. LGRNet also consists of a feature fusion&refining module that exploits complementarity between local AU feature and grid-based global feature to obtain refined local AU features. Extensive experimental evaluations on two widely used AU recognition benchmarks show that our LGRNet is able to learn more robustness and discriminative features for facial AU recognition.

REFERENCES

[1] M. N. Butt and M. Iqbal. Teachers' perception regarding facial expressions as an effective teaching tool. *Contemporary Issues in Education Research*, 4(2):11–14, 2011.

[2] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *ECCV*, pages 298–313, 2018.

[3] R. S. Feldman, L. Jenkins, and O. Popoola. Detection of deception in adults and children via facial expressions. *Child development*, pages 350–355, 1979.

[4] X. Ge, F. Chen, C. Shen, and R. Ji. Colloquial image captioning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 356–361. IEEE, 2019.

[5] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[6] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609, 2018.

[7] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Transactions on Information Forensics and Security*, 8(1):191–204, 2013.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[9] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *WACV*, pages 1–8, 2016.

[10] B. Klare and A. K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *International Conference on Pattern Recognition*, pages 1513–1516, 2010.

[11] N. N. Lakshminarayana, N. Sankaran, S. Setlur, and V. Govindaraju. Multimodal deep feature aggregation for facial action unit recognition using visible images and physiological signals. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE, 2019.

[12] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, volume 33, pages 8594–8601, 2019.

[13] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *CVPR*, pages 1841–1850, 2017.

[14] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *Trans. Pattern Anal. Mach. Intell.*, 40(11):2583–2596, 2018.

[15] Y. Li, S. Wang, Y. Zhao, and Q. Ji. Simultaneous facial feature tracking and facial expression recognition. *Trans. Image Process.*, 22(7):2559–2573, 2013.

[16] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812, 2014.

[17] P. Liu, Z. Zhang, H. Yang, and L. Yin. Multi-modality empowered network for facial action unit detection. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2175–2184. IEEE, 2019.

[18] 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 *ECCV*, pages 151–166, 2014.

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

[20] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Trans. Affective Comput.*, 4(2):151–160, 2013.

[21] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571, 2016.

[22] X. Niu, H. Han, S. Shan, and X. Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *NeurIPS*, pages 909–919, 2019.

[23] 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 *CVPR*, pages 11917–11926, 2019.

[24] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen. Automatic engagement prediction with gap feature. In *ICMI*, pages 599–603, 2018.

[25] 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 *NerIPS*, pages 8026–8037, 2019.

[26] D. R. Rubinow and R. M. Post. Impaired recognition of affect in facial expression in depressed patients. *Biological psychiatry*, 31(9):947–953, 1992.

[27] M. Sathik and S. G. Jonathan. Effect of facial expressions on student's comprehension recognition in virtual educational environments. *SpringerPlus*, 2(1):1–9, 2013.

[28] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *ECCV*, pages 705–720, 2018.

[29] Z. Shao, Z. Liu, J. Cai, and L. Ma. Jaa-net: Joint facial action unit detection and face alignment via adaptive attention. *IJCV*, 2020.

[30] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma. Facial action unit detection using attention and relation learning. *Trans. Affective Comput.*, 2019.

[31] 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. *NeurIPS*, pages 802–810, 2015.

[32] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013.

[33] S. Taheri, Q. Qiu, and R. Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *Trans. Image Process.*, 23(8):3590–3603, 2014.

[34] Y. Tong and Q. Ji. Learning bayesian networks with qualitative constraints. In *CVPR*, pages 1–8, 2008.

[35] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[36] H. Yang, T. Wang, and L. Yin. Adaptive multimodal fusion for facial action units recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2982–2990, 2020.

[37] 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*, 32(10):692–706, 2014.

[38] 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 detection. In *CVPR*, pages 2207–2216, 2015.

[39] 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. *Trans. Image Process.*, 25(8):3931–3946, 2016.

[40] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas. Learning multiscale active facial patches for expression analysis. *Trans. Cybern.*, 45(8):1499–1510, 2014.