Region of Interest Based Graph Convolution: A Heatmap Regression Approach for Action Unit Detection

Zheng Zhang
Department of Computer Science,
Binghamton University
zzhang27@binghamton.edu

Taoyue Wang
Department of Computer Science,
Binghamton University
twang61@binghamton.edu

Lijun Yin
Department of Computer Science,
Binghamton University
lijun@cs.binghamton.edu

ABSTRACT

Machine vision of human facial expressions has been studied for decades, from prototypical expressions to Action Units (AUs), from hand-crafted to deep features, from multi-class to multi-label classifications. Since the widely adopted deep networks lack interpretation on learnt representations, human prior knowledge cannot be effectively imposed and examined. On the other hand, AU is a human defined concept. In order to align with this idea, a finer level of network design is desired. In this paper, we first extend the heatmaps to ROI maps, encoding the location of both positive and negative occurred AUs, then employ a well-designed backbone network to regress it. In this way, AU detection is performed in two stages, key regions localization and occurrence classification. To prompt the spatial dependency among ROIs, we utilize graph convolution for feature refinement. The decomposition of similarity matrix is supervised by AU labels. This novel framework is evaluated on two benchmark databases (BP4D and DISFA) for AU detection. The experimental results are superior to the state-of-the-art algorithms and baseline models, demonstrating the effectiveness of our proposed method.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

AU Detection; Graph Convolution; Heatmap Regression

ACM Reference Format:
Zheng Zhang, Taoyue Wang, and Lijun Yin. 2020. Region of Interest Based Graph Convolution: A Heatmap Regression Approach for Action Unit Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413674

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413674

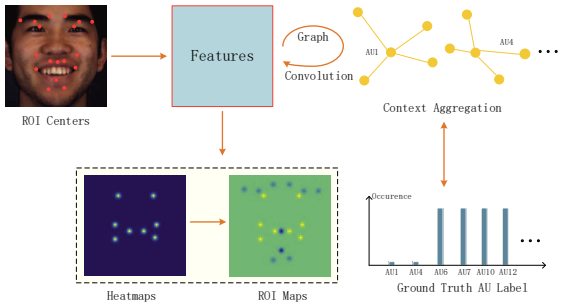


Figure 1: With the definition of centroids, AU detection can be converted into heatmap regression task. The idea of heatmap is extended to ROI map including absent AUs (better viewed in color). Features are refined with graph convolution. We impose additional supervision for the spectral decomposition process of graph similarity matrix.

1 INTRODUCTION

Facial action unit (AU) [6] analysis has been essential for understanding human emotions. Different from prototypical facial expression recognition (FER), AUs characterize the facial muscle movements. By various combinations of AUs, we can obtain a rich set of facial expressions.

From the machine learning perspective, understanding facial expressions is no longer limited to the multi-class classification problem [1, 13, 20, 27]. In many recent AU occurrence studies [12, 19, 23, 36], given the input images, researchers employ Convolution Neural Network (CNN) to extract deep features and feed them into classifiers to get the predictions of target AUs. The network output is for a group of AUs since learning shared features is more parameter efficient. Joint study of multiple AUs could also leverage their relationship, which makes it a multi-label classification problem. Deep features usually have a much lower spatial resolution compared to the one of original input. They are highly abstracted which prevent us from understanding the contribution of individual location in terms of specific AU. The interpretability is further hindered by global pooling of the features.

As depicted in Fig. 1, we propose to model AU detection as a regression problem and construct ROI maps based on heatmap [25] concept. Therefore, AU detection becomes a process of joint localization and classification. The motivations behind come from two facts: (1) We are able to impose prior knowledge on the location of AUs which better supervises the feature learning; (2) Heatmap

regression has been proven effective in facial landmark detection and human pose estimation tasks. Many well structured networks, like stacked hourglass [18], simple pose [32], HRNet [24, 28] have been designed for extracting task agnostic deep representations. Compared to direct optimization of coordinates/values, which is a highly non-linear object, heatmap serves as a strong supervision to preserve the structure of input. We argue that this concept can also benefit AU related tasks.

Given the advantage of heatmap regression, there still exists three major limitations. First, the generation of ROI maps constricts the receptive fields into several human defined ROIs which will lose long range context information. Such context information would be good complements of studying ROIs. Second, since AU detection task doesn't require the exact location of maximum activation, the most intuitive substitute of conventional heatmap decoding protocol is to check if the mean or maximum over each channel of ROI maps larger than some threshold. This solution is not only lack of robustness, but would be distracted by activations of non-related regions which are sub-optimal. Lastly, each AU is assigned to one channel. There is no explicit modeling of spatial relationships. To mitigate this problem, we employ graph convolution [9, 31] to model the spatial dependency of AU due to its outstanding power of relationship modeling. Many existing works have shown the necessity by applying graphs for learning spatial relationships for image analysis [3, 33] as well as AU detection [10]. Regular convolutional operations are not sufficient to capture long range semantic and spatial relationships between objects of an image. As stated in [16], even after hundreds of convolutions, the receptive field of the units of a network is severely limited. On the other hand, as stated in [15], graph encodes dependencies between regions, such dependencies are of much longer range than those captured by local convolutions.

Graph neural network [31] takes graph as input and is designed to learn features of non-euclidean data structure. Since the image itself is in grid structure, we come up with the question about how to define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. To find the relationship, each spatial location in the feature map can be considered as one node in the graph. Edge is induced between every pair of the nodes. Correspondingly, the similarity matrix A is interpreted as how similar between two locations of the feature map. As the learning process of feature maps, it is more reasonable to make A data dependent. A is derived from the most current feature map and the size is a quadratic order of feature map resolution. When the resolution of feature map is high which preserves the most information of input data, direct computation of A is practically infeasible. [14] learns eigenvalues, eigenvectors first and comes up with an mathematically equivalent expression by multiplying them with input step-by-step. They can circumvent the high demand of computing complexity. In light of this work, we take one more step to instantiate the eigenvector as global context w.r.t. each AU. In this way, we can model the connections between each AU and all other spatial locations.

In this paper, we propose a novel AU detection framework which consists of ROI map regression and deep feature refinement with graph convolution. By encoding all AUs into the ROI maps, the network is able to learn their locations which improve the recognition rate. Deep features from backbone network are fed into graph convolution to explore the relationship among all spatial locations. We regularize the learning of eigenvector matrix by AU classification loss to make similarity matrix decomposed in AU semantic space. As a result, refined features are more discriminative achieving improved performance on two benchmark datasets.

The main contribution of this work lies in three-fold:

(1) We formulate AU detection into ROI map regression problem with positive and negative occurred AU localized and classified at the same time. To some extent, it moderates overfitting and data imbalance issues caused by the larger number of negative samples.
(2) To actively utilize AU co-occurrence or mutual exclusive pattern, we employ graph convolution to refine the deep features. The dynamic modeling of spatial dependency facilitates the exploration of AU relationships and makes the entire framework end-to-end trainable.
(3) We propose to supervise the learning of eigenvector matrix by AU labels. In this case, the decomposition procedure of similarity matrix would concentrate on relationships between AUs and their spatial contexts.

2 RELATED WORK

In recent years, as the deep features show impressive generalization power over hand-crafted ones, AU researchers widely adopt deep networks for AU recognition task. [8] is one of the pioneer works in the FERA [26] challenge, which demonstrated impressive performance on both AU occurrence detection and intensity classification tasks. We refer readers for survey papers of deep features on AU analysis [37] or broader category on FER [5, 11]. Roughly speaking, the development of AU detection is summarized into the following categories, which are also the active trends that researchers can dive into for better performance.

Region learning. Different from other classification tasks where holistic features may be nearly efficient, AU recognition desires a finer grained analysis of human faces. The standard convolution operation on the inputs shares weights within the layer. But in human faces, different regions could have different statistics. Following the spirit of locally connected layers, Zhao [36] proposed region layers which uniformly slice the first feature map and apply independent filters on each local face patch. Landmarks represent the salient regions of human faces. They are well localized hence become reliable tools to define representative regions of AUs. These regions can either be emphasized [12, 22], as advised by attention mechanisms, or cropped out [4, 10, 12] for further feature engineering independent of each other.

Relationship learning AUs relationship has already been attempted [29, 35] from the age of hand-crafted features. The goal is to utilize AUs' dependency to improve the overall performance of a multi-label recognition problem. The study by Zhao et al. [35] concluded that AU has two kinds of relations, positive correlation or negative competition. Li et al. [10] extended this concept into a graph, treating AUs as nodes, and refines AU features with gated graph neural network (GGNN). However, the graph is statically defined according to the statistics of datasets hence lose the adaptation power during the feature learning process. Instead of refining AU features, Corneanu et al. [4] pass the patch predictions of CNN

into Conditional Random Field (CRF) and apply structure inference to obtain a final AU predictions. Shao et al. [23] also propose to refine attention weights with CRF at pixel-level. L-Net in [19] regards each cell of feature maps as a representation of local region and feeds them into LSTM for relationship learning.

Due to the close relationship between AU recognition and landmark detection tasks, researchers propose to jointly optimize these two problems. Therefore, they can be a good supervision of each other. Wu et al. [30] used Restricted Boltzmann Machine (RBM) to learn the joint probabilities of landmark coordinates and AU occurrence labels and update both tasks iteratively. Shao et al. [22] followed by a deep framework which detects AU and regress landmark coordinates in a multi-task learning fashion. In contrast, our framework is not learning landmarks. We explore the idea of the heatmap and adapt it to AU study. Another related work is [7] which integrates semantic correspondence module with heatmap regression for AU intensity estimation. The graph convolution in [7] is a spatial based method treating feature channels as nodes. Instead, our work is spectral-based and explores relationships between spatial locations.

3 METHODOLOGY

In this section, we will elaborate the details of our proposed framework. We first illustrate how to generate ROI maps from landmarks and give an overview of our framework. Then we will discuss how the refinement works with a graph convolution and extra supervisions we introduced. Lastly, we conclude the optimization objectives and the difference between training and testing phases.

3.1 ROI maps

We first define region of interest (ROI) for each AU. Similar to

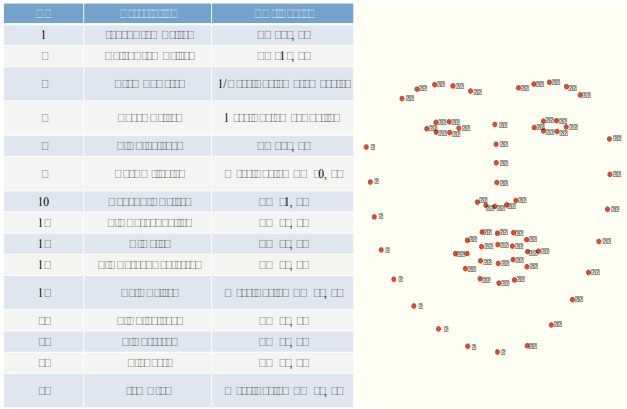


Figure 2: Left: ROI centers defined for AUs, 'scale' is measured by inner-ocular distance. Right: Landmark indices.

the rules in [12, 22], we sample two points symmetrically on the face based upon the most representative landmarks, see details in Fig. 2. Therefore, we will have two ROIs for each AU. They are small regions centered around those two points. By looking at facial expressions inside ROIs, we are able to infer the occurrence of specific AU. Note that some differences from rules in [12, 22] are the locations of AU1 and AU2. Instead of shifting a distance from inner and outer brow, we directly take their landmark positions which have more visual context and can be clearly identified.

After definition of ROIs, we are able to generate ground-truth ROI maps. Fig 3 gives two examples. Each map R_c^{gt} consists of one Gaussian window with maximum value at the center p_c^{gt} of ROI.

$$R_c^{gt}(p) = \mathbb{1} \exp\left(-\frac{\left\|p - p_c^{gt}\right\|_2^2}{2\sigma^2}\right), \quad p \in \Omega \tag{1}$$

σ controls the standard derivation, Ω is the set of all pixel locations in ROI. $\mathbb{1}$ is the indicator function $\mathbb{1} :\longrightarrow \{1, -1\}$ according to the ground-truth AU label. For p not in Ω, it represents a background pixel which intensity is initialized as 0.

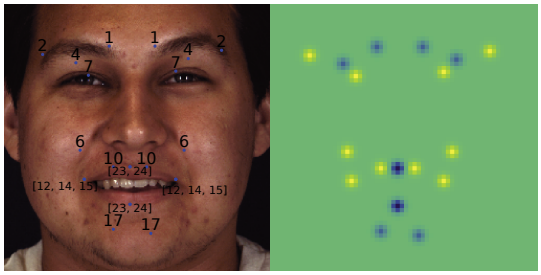
ROI maps have C channels where C corresponds to the total number of ROIs. Different from the maps in [7, 12, 22], which only encode positively occurred AUs, we also encode *non-occurred AUs* into the maps. In this case, the peak value is adapted to be either 1 (AU occurred) or -1 (AU absent). We re-formulate AU occurrence detection problem into two stages. The first one is to spot the possible locations of AUs and the second step is to classify if the target AU is active or not. Another advantage of this formulation is that negative samples would contribute to AU localization as well as the classification. Given that we usually have much more negative samples than the positive ones due to the nature of facial expressions, such additional 'task' would help the network better understand the AU positions across different facial structures.

3.2 Graph Convolution

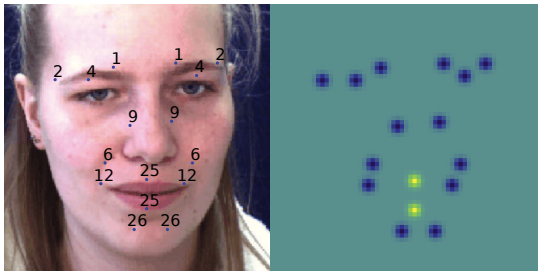
With shared features among different AUs, deep models are trying to find a generalized representation which can be easily classified and achieve the lowest average loss. There is a lack of consideration on AU relationships. The graph convolution grants us more flexibility to model the relationship and can be integrated into a common recognition pipeline which is end-to-end trainable.

Fig. 4 illustrates our proposed framework. We choose HRNetV2 [28] as the backbone feature extractor, which has a similar or smaller number of parameters and computation cost to ResNet-50 and Hourglass [18]. Due to its high resolution characteristics and capable of fusing multi-scale information, HRNetV2 [28] achieved a superior performance on tasks such as semantic segmentation and facial landmark detection. While AUs detection task studies the finer details of facial muscle movement, it fits well to the strength of HRNetV2 [28] which does not lose much information during the encoding process. Assume that input images have a size of $H * W * 3$, X_0 from the backbone network is downscaled to $H/4 * W/4 * C$. In the baseline model, a $1 * 1$ convolution is applied to reduce C to C_{out} features, where C_{out} corresponds to the number of target AUs. We employ the graph convolution in [14] to refine X_0, enhancing the spatial wise relationship. As shown in Eq. 2, X_1 represents the feature after refinement. L is the Laplacian matrix which has a symmetric normalized form of $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, Θ is a trainable weight matrix, and σ is the *ReLU* activation function.

$$X_1 = \sigma(L X_0 \Theta) \tag{2}$$



(a)



(b)

Figure 3: (a) Left: ROI centers defined for 12 AUs in BP4D, two for each AU. Blue dots stand for the locations and corresponding AU indices are labeled above. Right: Ground truth ROI maps. Each map is designed for each AU which has two ROIs. We sum all maps into one for display purpose (better viewed in color, pixel intensities range from -1 (blue) to 1 (green)). (b) Similar to (a) except for 8 AUs in DISFA.

Similarity matrix $A(X_0)$ is derived from input feature X_0 and the degree matrix D equals to diag $(d_1, d_2, \ldots, d_{H'W'})$ where $d_i = \sum_j A_{ij}$. Here we use data-dependent $A(X_0)$ due to its adaptation capability as the change of input. It is not restricted to any database statistics or empirical definition, hence it is able to capture the spatial structure better. Because of the symmetric property of similarity matrix, we can decompose it into an eigenvalue matrix Λ and an eigenvector matrix Φ. We divide the network into two branches for learning in parallel as a function of input X_0. Since we are targeting AUs and interested in C_{out} number of ROIs, we expect A to be decomposed based on those regions. Therefore, eigenvector matrix $\Phi(X_0) \in \mathbb{R}^{H'W' \times C_{out}}$ and eigenvalue matrix $\Lambda(X_0) \in \mathbb{R}^{C_{out} \times C_{out}}$. For simplicity, here we note $H/4, W/4$ as H', W'. For the branch $\Lambda(X_0)$, X_0 is first average-pooled along spatial domain and then goes through a linear layer to reduce channels from C to C_{out}. Then it is sigmoid activated and transformed to a diagonal matrix. On the other branch, 1×1 convolution is applied on X_0. We reshape the output into $H'W' \times C_{out}$ to generate $\Phi(X_0)$. Consequently, A is learnt by multiplying $\Phi(X_0)$ and $\Lambda(X_0)$ together, as shown in Eq. 3.

$$A(X_0) = \Phi(X_0)\Lambda(X_0)\Phi^T(X_0) \tag{3}$$

For each location of feature X_0, we are able to model its similarity to the other features.

3.3 Supervised Decomposition

$\Lambda(X)$ proposed in [14] learns a better distance metric for the similarity matrix. Inherit from this property, we further extend to draw connections with our AU detection task. As aforementioned, we expect $A(X_0)$ to be decomposed along the most discriminative areas and model the relationship between these attended areas and their neighbors. To achieve this goal, we introduce an auxiliary loss (Loss1 in Fig. 4) to regularize the decomposition process. For the branch computing $\Phi(X_0)$, each eigenvector models the spatial context with respect to each AU. The assumption behind the ROI map regression is that the designated ROIs are the most representative areas for recognizing AUs. Inevitably, it could narrow down the receptive field of AU detection task and could potentially overlook some helpful spatial context. To remedy this problem, we supervise the learning of $\Phi(X_0)$ by minimizing its discrepancy with ground-truth labels. For each AU, we first adopt a linear layer f to learn the weights θ_f of all $H' \times W'$ locations and compute the cross entropy (CE), as shown in Eq. 4, between weighted features and labels.

$$Loss1 = \frac{1}{C_{out}} \sum_{j=1}^{C_{out}} CE_j(\sigma(f(\Phi(x); \theta_f)), y_j) \tag{4}$$

σ stands for the sigmoid activation, y_j is the binary occurrence label of jth AU.

3.4 Complexity and Objective

Given that $A \in \mathbb{R}^{H'W' \times H'W'}$, a large amount of memory would be consumed if we compute A explicitly. Therefore, we adopt the same workaround as in [14]. LX_0 in Eq. 2 can be rewritten as the following form:

$$\begin{aligned} LX_0 &= X_0 - D^{-\frac{1}{2}}\Phi\Lambda\Phi^T D^{-\frac{1}{2}} X_0 \\ &= X_0 - P\left(\Lambda\left(P^T X_0\right)\right) \end{aligned} \tag{5}$$

where $P = D^{-\frac{1}{2}}\Phi$, we omit the data dependent annotation for simplicity and,

$$D = diag(A \cdot \overrightarrow{1}) = diag\left(\Phi\left(\Lambda\left(\Phi^T \cdot \overrightarrow{1}\right)\right)\right) \tag{6}$$

$\overrightarrow{1}$ stands for all-one vector in $\mathbb{R}^{H'W'}$. Parentheses indicate the order of operation. Every step in Eq. 6 is a multiplication with a vector. By computing from the inner parentheses of Eq. 5 to the outer one, we are able to avoid the quadratic order of complexity $O\left(H^2W^2\right)$.

After getting feature X_1, another 1×1 convolution is applied to reduce channels to C_{out}. By comparison with ground truth maps, we have the following loss function:

$$Loss0 = \frac{1}{N} \sum_{i=1}^{N} \left\| Conv_{1\times1}(X_1)_i - R_i^{gt} \right\|_2^2 ; where\ N = H' * W' * C_{out} \tag{7}$$

Final objective function is the minimization of the summed two losses:

$$L = Loss0 + \alpha_1 Loss1 \tag{8}$$

with α_1 is a hyperparameter that controls the importance of auxiliary loss.

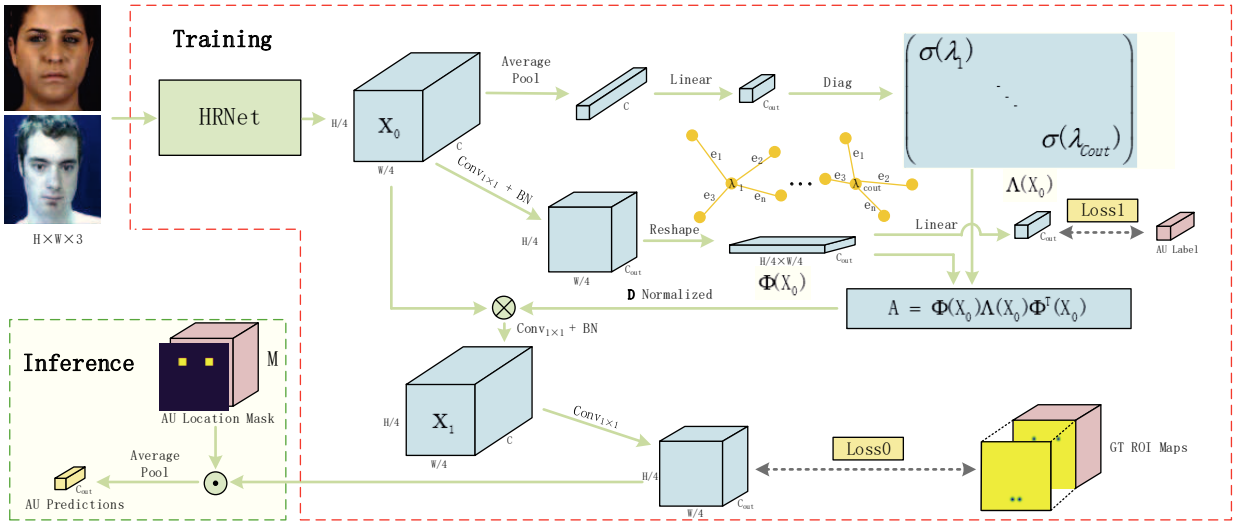


Figure 4: Overview of proposed framework for AU detection. We first extract features X_0 by a backbone network. Then X_0 is spatially refined into X_1 with graph convolution. The decomposition of A is supervised by AU labels. Predicted ROI maps are compared with ground truth (GT) maps for training. It is multiplied with the location mask M and then decoded for inference. $\odot$ and $\otimes$ stand for dot product and matrix multiplication respectively.

3.5 Inference

At the inference stage, conventional heatmap regression works decode the spatial location from the cth channel of predicted heatmaps (R_c^{pred}) which has the largest activation $\tilde{P}_c = argmax\ R_c^{pred}$. For the AU occurrence detection, it is not needed to know the exact $\tilde{P}_c$. Thus, we can consider specific AU to be active if $max/mean\ (R_c^{pred})$ is larger than a certain threshold. However, such a decoding protocol may relax the spatial constraint of AUs too much and cause many false positives. For instance, it is not optimal to determine eyebrow related AUs by activation of regions far from connected. To remedy this problem, we propose to use the mask M. Similar to Eq. 1, M is defined as $M(p) = 1,\ p \in \Omega$ and the positive criteria of each AU becomes to check $mean(M \cdot R_c^{pred}) > 0$ from either of two ROIs. In other words, only the activations inside the ROIs can count toward final predictions.

4 EXPERIMENTS

We evaluate our new framework on two datasets BP4D [34] and DISFA [17]. They are publicly available benchmarks for the AU occurrence detection task.

BP4D: It contains 41 subjects captured under laboratory environments. 8 tasks are designed to elicit a range of spontaneous emotions. There are $328 (= 41 \times 8)$ sequences in total with a frame rate of 25. Expert coders selected the most expressive 20s of each sequence for AU coding. Around 140,000 labeled frames are included in our experiments and split into subject-exclusive 3 folds for a fair comparison with state-of-the-art algorithms.

DISFA: There are 27 subjects in the DISFA database, 12 females and 15 males. Videos were captured when subjects were watching emotive videos. There are $\sim$ 130,000 frames in the database and for every frame, DISFA provides intensity code with $0-5$ different scales. Following the evaluations in [22], the frames with intensities equal or larger than 2 are regarded as AU occurred while the rest are absent. Also, we divide the dataset into subject-exclusive 3 folds and report the performance through the cross validation.

4.1 Implementation details

The bounding box of faces in raw images are obtained using the publicly available library OpenFace [2]. Given the high resolution of raw input containing a large portion of background, we first preprocess the image by cropping out facial areas, and resize the image into 256∗256∗3 to fit the network. So H, W are 256 and H', W' are 64 in our experiments. Each input image is randomly rotated $-30°$ to $30°$ (BP4D), $-10°$ to $10°$ (DISFA) and flipped horizontally for data augmentation. We use the landmarks provided by the datasets to find ROI centers, then generate the ground-truth maps R^{gt} and AU location masks M. Each ROI has a size of 7×7. We also standardize each image by subtracting mean values and by dividing the variances of input channels. The backbone HRNetV2-W18 [28] is pre-trained on ImageNet [21]. We use the Adam optimizer with an initial learning rate of $1e-4$. It decreases by 0.1 every 10 epochs. For both datasets, the weight decay is set to $5e-4$ and the batch size is 32. α_1 is set as 0.001 for the following experiments.

4.2 Metrics

F-score (F_1) is the most popular evaluation metric used for AU detection, which is computed as harmonic mean of precision and recall $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

4.3 Results and Discussions

We compare our proposed algorithms against state-of-the-art algorithms JPML [35], DRML [36], EAC [12], DSIN [4], JAA [22], ARL [23], SRERL [10], LP [19] on BP4D [34] and DRML [36], EAC [12], DSIN [4], JAA [22], ARL [23], SRERL [10], LP [19] on DISFA [17] database. The results are reported in Table 1 and Table 2 regarding each individual AU and their average (Noted as Avg.). On BP4D, our model outperforms all state-of-the-art algorithms and achieves an average F_1 of 63.5. And the major improvement comes from the less likely occurred AUs (AU1, AU2, AU4, AU23) in the dataset. These AUs have a lower occurrence rate ($\sim$ 20%) as discussed in previous work [10, 12, 23], therefore they lack optimization when jointly trained with other labels.

On DISFA, we observe a significant F1 increment from 58.7 to 62.0. Consistent with findings in BP4D, main improvement is from AU1, AU2 and AU4. Our method employs the graph convolution to explore the relations among all spatial locations. Therefore, it promotes the co-occurrence or mutual exclusive patterns of AUs. Moreoever, graph networks show great potential on certain AUs. Together with SRERL [10] which utilizes GGNN, we outperform other works in terms of AU2, AU7, AU12, AU15 in BP4D and AU1, AU2 in DISFA.

4.3.1 Ablation Study. We further provide ablation study on the BP4D dataset to investigate the effectiveness of each proposed module. Table 3 shows the performance comparisons in terms of both individual AUs and the average. All experiments are conducted in a three-fold cross validation with the same settings. Heatmap stands for the baseline model that encodes the positive AUs occurrence only. The decoding threshold is 0.5. It achieves an average F_1 of 62.0, which is inferior to the state-of-the-art algorithms. This is as expected since modeling positive AUs only will render a large number of non-activated heatmap channels, given the sparsity nature of ground truth AU labels. We argue that the negative samples would contribute not only to classification but localization of ROIs, which is in return of benefit to AU recognition. As shown in the third column of Table 3, many AUs (such as AU2, AU10, AU14, AU17) have dramatic improvement over their counterparts in the second column. The overall performance 62.5 is on par with the peer models. We further applied the graph convolution [14] which learns the similarity matrix A based on X_0 without supervision of the decomposition process by AU labels. Although the F_1 is increased from 62.5 to 62.8, it does not exercise the full power of spatial refinement for AU detection. If we consider the refinement as the message passing scheme between one spatial location and all others in X_0, a similarity matrix is the bridge that controls the amount of information to communicate. The extra supervision of eigenvectors essentially contributes to two aspects: (1) It encourages the ROIs to anchor the decomposition of similarity matrix; (2) It aggregates the spatial context w.r.t. each AU, which complements the region learning for AU recognition. Note that the second aspect is particularly crucial given the potential limitation of ROI definitions and mis-tracked landmarks in some training samples. Therefore, we observe a significant performance increment on half of the AUs (e.g., AU4, 7, 15, 17, 23, 24), as shown in the last column of Table 3.

4.3.2 Does loss balancing help? We noticed an extremely imbalanced data distribution in the DISFA database. Most likely occurred AU25 in Table 4 has an $\sim$ 6.6 times of positive samples than the least likely one (AU9). With 3-Fold partition of 27 subjects, this ratio in the training set of one fold may be even larger. Previous works [10, 19, 22, 23] propose to use weighted loss functions to prevent data imbalance issues from skewing the training process. We investigate if this strategy can also help in our method. Similar to those state-of-the-art algorithms, we use weighted versions (noted as 'weighted' in Table 4) of Loss0 and Loss1 for comparison. During the training phase, for both Eq. 4 and Eq. 7, we multiply with inverse occurrence rates before they average over AUs. Though the recognition rates of some minority AUs (like AU6, AU9, AU26) are increased, we don't observe a further increment on average F_1. We attribute it to the feature refinement by graph convolution which models the relationship among AUs.

4.4 Visualization

After we sample an image from the test set and feed it into the network after training, we visualize the activation maps of X_0 and X_1 in Fig. 5 to illustrate the effectiveness of proposed spatial refinement. We rearrange the channels (270 each) into grid structure for comparison purpose. Each cell represents one channel with a dimension of 64×64. As we can see, X_0 contains much more activations inside the face than X_1 has. With the help of graph convolution, the most unrelated activations in X_0 (Fig. 5a) have been smoothed out, making X_1 more uniformly distributed prior to entering the following Conv1 $\times$ 1 layer. Therefore, the channels can be easily weighted and classified by a linear layer. We also noticed that, in X_1, the boundary is likely to be activated in many channels. It is an artifact caused by the face cropping process but can also be tackled easily by the channel weighting. When inspecting both X_0 and X_1, there exist small dots inside the face. Those are ROIs defined in the map to be regressed. With reduced distractions of unrelated activations in X_0, X_1 pays more attention to learning those ROIs. More importantly, X_1 has a much diverse and sparse combination of ROIs, indicating the exploration of possible relations among them. This, in return, will reflect the co-occurrence or mutual exclusive AU patterns.

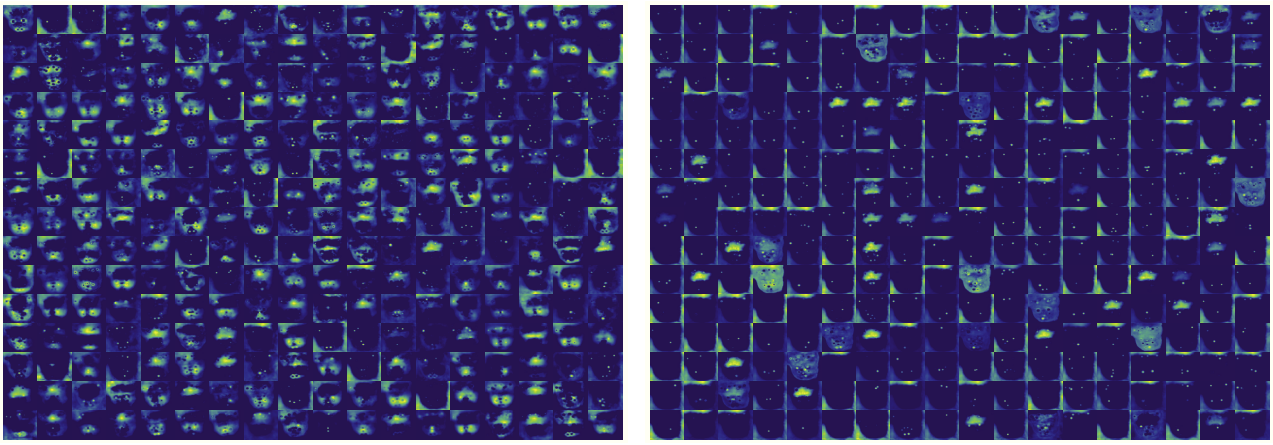
We also visualize the eigenvectors $\Phi(x)$ in Fig. 6 in order to examine how they aggregate the context w.r.t. each AU. The first column shows sampled images from the test set. *It is worth noting that one advantage of our approach is that given the input images, it can localize AU related regions by appearance on the face no matter whether this AU occurs or not.* It is very different from the traditional attention mechanisms which only highlight those active areas. Moreover, facial appearance is the reflection of facial muscle movements. Facial muscles are the "engines" that initiate the facial action and drive the motion of each AU region. For example, AU6 (cheek raiser) is triggered by the orbital part of orbicularis oculi muscle. AU10 (upper lip raiser) is caused by the lower end of the

Table 1: F1 of 12 AUs on BP4D database. Bold numbers indicate the best performance.

AU	JPML [35]	DRML [36]	EAC [12]	DSIN [4]	JAA [22]	ARL [23]	SRERL [10]	LP [19]	Ours
1	32.6	36.4	39.0	51.7	47.2	45.8	46.9	43.4	**52.6**
2	25.6	41.8	35.2	40.4	44.0	39.8	45.3	38.0	**47.0**
4	37.4	43.0	48.6	56.0	54.9	55.1	55.6	54.2	**61.4**
6	42.3	55.0	76.1	76.1	**77.5**	75.7	77.1	77.1	76.8
7	50.5	67.0	72.9	73.5	74.6	77.2	78.4	76.7	**79.2**
10	72.2	66.3	81.9	79.9	**84.0**	82.3	83.5	83.8	83.5
12	74.1	65.8	86.2	85.4	86.9	86.6	87.6	87.2	**88.6**
14	**65.7**	54.1	58.8	62.7	61.9	58.8	63.9	63.3	60.4
15	38.1	33.2	37.5	37.3	43.6	47.6	**52.2**	45.3	49.3
17	40.0	48.0	59.1	62.9	60.3	62.1	**63.9**	60.5	62.6
23	30.4	31.7	35.9	38.8	42.7	47.4	47.1	48.1	**50.8**
24	42.3	30.0	35.8	41.6	41.9	**55.4**	52.3	54.2	49.6
Avg.	45.9	48.3	55.9	58.9	60.0	61.1	62.9	61.0	**63.5**

Table 2: F1 of 8 AUs on DISFA database. Bolded numbers indicate the best performance.

AU	DRML [36]	EAC [12]	DSIN [4]	JAA [22]	ARL [23]	SRERL [10]	LP [19]	Ours
1	17.3	41.5	42.4	43.7	43.9	45.7	29.9	**55.0**
2	17.7	26.4	39.0	46.2	42.1	47.8	24.7	**63.0**
4	37.4	66.4	68.4	56.0	63.6	59.6	72.7	**74.6**
6	29.0	**50.7**	28.6	41.4	41.8	47.1	46.8	45.3
9	10.7	**80.5**	46.8	44.7	40.0	45.6	49.6	35.2
12	37.7	**89.3**	70.8	69.6	76.2	73.5	72.9	75.3
25	38.5	88.9	90.4	88.3	95.2	84.3	**93.8**	93.5
26	20.1	15.6	42.2	58.4	**66.8**	43.6	65.0	54.4
Avg.	26.7	48.5	53.6	56.0	58.7	55.9	56.9	**62.0**



(a) **(b)**

Figure 5: Activation maps of (a) X_0 and (b) X_1. We re-arrange a total of 270 channels into grid structure to have a direct feature comparison before and after graph convolution. Small dots in each map stand for ROIs. In (b), non-related activations are removed but diverse patterns of ROIs are preserved. Since 270 channels are weighted summed into C_{out} ROI maps, each map is regarded as a combination of those patterns.

linked muscle (levator labii superioris). Learned eigenvector maps reflect the motions of these linked muscles. Therefore, our work is able to better analyze the structure of underlying muscle movements. Except for some artifacts caused by the image boundary, we

Table 3: Ablation study on the BP4D dataset

AU	Heatmap	+ Neg. AUs	+ Neg. AUs + GCN	+ Neg AUs + GCN + Supervison
1	**55.2**	52.3	50.9	52.6
2	47.8	**49.4**	48.1	47.0
4	59.0	56.9	60.6	**61.4**
6	77.0	**77.5**	**77.5**	76.8
7	77.6	77.7	77.4	**79.2**
10	82.6	**83.6**	83.1	83.5
12	88.5	88.5	88.5	**88.6**
14	58.3	**61.9**	61.5	60.4
15	47.5	46.6	46.7	**49.3**
17	55.6	61.2	61.6	**62.6**
23	47.1	46.1	48.6	**50.8**
24	48.1	48.5	48.8	**49.6**
Avg.	62.0	62.5	62.8	**63.5**

Table 4: Positive samples ratio of DISFA databset and performance comparison with/without loss balancing.

AU	1	2	4	6	9	12	25	26	Avg.
Occ. Rate	5.0	4.3	15.2	7.9	4.2	12.9	27.7	8.8	-
Weighted	52.7	61.0	72.5	**46.7**	**37.8**	73.4	**93.6**	**55.0**	61.6
Ours	**55.0**	**63.0**	**74.6**	45.3	35.2	**75.3**	93.5	54.4	**62.0**

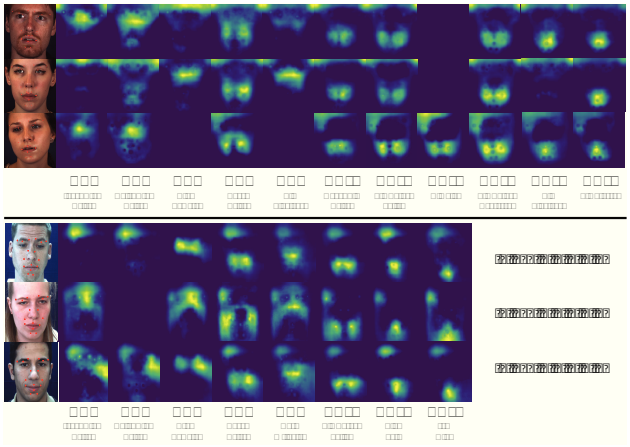


Figure 6: Visualization of eigenvectors $\Phi(x)$ w.r.t. each AU for BP4D (top) and DISFA (bottom) dataset. Leftmost column represents input images sampled from the test set. Failure cases are noted as blank maps.

can also see that the largest activation comes from the most related regions on the face, and does not have to be on the exact location of ROI centers. The intensity of activation decreases as the distance from ROI goes further. Please note that, even if some AUs share the same ROI centers (AU12, 14, 15 or AU 23, 24 in BP4D), their contexts are varied, which means their individual characteristics are still well exhibited. As an example, the third row of Fig. 6 illustrates that the activated regions of AU15 (lip corner depressor) is lower than that of AU12 (lip corner puller). When compared to AU12, AU15 has indeed a movement along the vertical direction. Although some eigenvectors may not be differentiable visually, their contexts are easily classified by the following linear layer.

4.5 Conclusion and future work

In this paper we have presented a new ROI map regression framework for action unit detection. Each ROI map consists of two ROIs defined for both positive and negative occurred AUs, thus allowing us to re-format the AU detection process into two separate stages: AU localization and AU classification. By leveraging the graph convolution with supervised spectral decomposition, features for producing the ROI maps are spatially refined. The results are better than the compared state-of-the-art algorithms when tested on two benchmark datasets. We have also conducted the ablation study to showcase the effectiveness of individual components. As well, we have visualized and scrutinized the eigenvectors and the activation maps before and after the graph convolution.

Our future work will take a more effective way to define ROIs in an attempt to make it more robust to significant landmark errors. We will also investigate how the number of ROIs (current number is two) defined for each AU would impact the recognition performance. To demonstrate the generality, we will apply our proposed framework to other applications with more databases included for cross-validation. Additionally, applying graph reasoning on multi-scale features would be another future direction.

5 ACKNOWLEDGEMENT

The material is based on the work supported in part by the National Science Foundation (NSF) under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

REFERENCES

[1] Timur Almaev, Brais Martinez, and Michel Valstar. 2015. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[3] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5177–5186.

[4] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. 2018. Deep Structure Inference Network for Facial Action Unit Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[5] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 38, 8 (2016), 1548–1568.

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

[7] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2020. Facial Action Unit Intensity Estimation via Semantic Correspondence Learning with Dynamic Graph Convolution. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

[8] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. 2015. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–5.

[9] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[10] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. 2019. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 8594–8601.

[11] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020).

[12] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. 2017. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 103–110.

[13] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. 2018. EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2018).

[14] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. 2020. Spatial Pyramid Based Graph Reasoning for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Yin Li and Abhinav Gupta. 2018. Beyond grids: Learning graph representations for visual recognition. In *Advances in Neural Information Processing Systems (NIPS)*. 9225–9235.

[16] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2016. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*. 4898–4906.

[17] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160.

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 483–499.

[19] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. 2019. Local Relationship Learning with Person-specific Shape Regularization for Facial Action Unit Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11917–11926.

[20] Guozhu Peng and Shangfei Wang. 2018. Weakly Supervised Facial Action Unit Recognition Through Adversarial Training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2188–2196.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision (IJCV)* 115, 3 (2015), 211–252.

[22] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2018. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[23] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. 2019. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing* (2019).

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5693–5703.

[25] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems (NIPS)*. 1799–1807.

[26] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. 2015. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–8.

[27] Can Wang and Shangfei Wang. 2018. Personalized multiple facial action unit recognition through generative adversarial recognition network. In *Proceedings of the 26th ACM international conference on Multimedia (ACMMM)*. 302–310.

[28] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence (PAMI)* (2020).

[29] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. 2013. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 3304–3311.

[30] Yue Wu and Qiang Ji. 2016. Constrained Joint Cascade Regression Framework for Simultaneous Facial Action Unit Recognition and Facial Landmark Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[32] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 466–481.

[33] Hang Xu, Chenhan Jiang, Xiaodan Liang, and Zhenguo Li. 2019. Spatial-aware graph relation network for large-scale object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9298–9307.

[34] Xing Zhang, Lijun Yin, Jeffrey Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing (IVC)* 32, 10 (2014), 692–706.

[35] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. 2015. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. 2016. Deep Region and Multi-Label Learning for Facial Action Unit Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[37] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. 2019. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer* (2019), 1–27.