2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

SAT-Net: Self-Attention and Temporal Fusion for Facial Action Unit Detection

Zhihua Li, Zheng Zhang, Lijun Yin
Department of Computer Science
State University of New York at Binghamton, USA
zli191, zzhang27@binghamton.com; lijun@cs.binghamton.edu

***Abstract*—Research on facial action unit detection has shown remarkable performances by using deep spatial learning models in recent years, however, it is far from reaching its full capacity in learning due to the lack of use of temporal information of AUs across time. Since the AU occurrence in one frame is highly likely related to previous frames in a temporal sequence, exploring temporal correlation of AUs across frames becomes a key motivation of this work. In this paper, we propose a novel temporal fusion and AU-supervised self-attention network (a so-called SAT-Net) to address the AU detection problem. First of all, we input the deep features of a sequence into a convolutional LSTM network and fuse the previous temporal information into the feature map of the last frame, and continue to learn the AU occurrence. Second, considering the AU detection problem is a multi-label classification problem that individual label depends only on certain facial areas, we propose a new self-learned attention mask by focusing the detection of each AU on parts of facial areas through the learning of individual attention mask for each AU, thus increasing the AU independence without the loss of any spatial relations. Our extensive experiments show that the proposed framework achieves better results of AU detection over the state-of-the-arts on two benchmark databases (BP4D and DISFA).**

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413260

I. Introduction

Facial action units (AUs) are facial muscle actions at certain facial locations defined by Facial Action Coding System (FACS) [1]. AU detection has been an essential task for understanding facial behavior and mental activities. As deep learning methods have shown strong ability in image classification and many deep learning based models have been proposed recently [2]. Thus, AU detection research has been focused from traditional handcraft features to deep features. More and more deep convolutional models have been adapted to AU detection problems [3]. However, there are still several drawbacks existing in the current approaches:

1) Because different facial action units appear in different facial areas, one single AU's occurrence just depends on a region of interest, and most of the previous works have focused on splitting feature maps into patches and used separate convolution kernels for individual patch. One drawback of such methods is the ignorance of the global spatial relations or requirement of fusion of different ROIs features at the end of the networks, thus greatly increasing the complexity.
2) There are also some works that utilized attention maps to enhance some AU center guided or expert knowledge based regions. Although those methods have shown a certain performance improvement, the handcraft attention map needs the facial landmarks to be detected and cannot be easily extended to other unregistered AU detection problems. Therefore, it will be more efficient to apply a self-learned attention technique with the supervision of AU labels for learning the AU regions automatically.
3) Most existing approaches have not fully utilized the temporal dependency of consecutive frames. Although for temporal relation modeling, traditional fully connected recurrent neural nets (FC-LSTM) can model the temporal relation well in natural language processing and image caption [4], [5]FC-LSTMs are not able to learn the spatial relations simultaneously in image classification. 3D convolution [6] could be a remedy to alleviate the problem. 3D convolution extends one more kernel dimension to the temporal domain and can extract the temporal features as well as spatial features. However, it requires one more dimension for the kernel, thus greatly increasing the training parameters and network complexity. More importantly, unlike the sequence based classification problem that one video has only one label, AU detection has to be based on consecutive frames with every single frame being labeled, hence extending one more dimension for kernels in the temporal domain is not reasonable for such a task as the training process is significantly slow.

To tackle the AU detection problems as above mentioned, we propose to develop a new self-attention and temporal fusion network (SAT-Net) to model the spatial-temporal relations. The self-attention module is learned through the AU label supervision and it does not require facial landmarks or prior knowledge of AU locations. Our redesign of the input-output of Convolutional LSTM [7] enables the model to learn the temporal relations without losing any location information. The synergy of the two modules makes a significant performance boost when tested on two widely used databases.

The contribution of this paper is manifested in following tow-fold:

- This is the first work to introduce the self-attention into the AU detection task, which can focus on more important facial areas with the supervision of AU labels automatically. Our new design of the Conv-LSTM based temporal learning scheme fits better to the temporal representations of facial action unit.
- The light weighted self-attention and temporal learning framework (SAT-Net) does not increase the network complexity as compared to the baseline ResNet [8]. Impor-

tantly, this new network is end to end trainable.

II. Related Work

AU occurrence detection is a binary classification task by using static facial images or image sequences to detect the appearance of AUs. We summarize the previous works which are highly related to this paper in two subsections, spatial based detection and spatial-temporal based detection.

Detection with spatial information. For spatial features learning, previous AU detection techniques can be divided into two types, handcraft appearance and geometry based feature representation and self-learning feature representation. The former extracts the facial texture information which reflects the magnitude and direction of facial surface, or features that describe the facial deformation information. The latter can automatically learn the representations from the training data hierarchically utilizing modern deep learning techniques. Valstar et al. [9] extracts Gabor wavelet features and then classified by Adaboot and SVM. Zhao et al. [10] designed a joint patch and multi-label learning (JPML) framework that identified important facial patches' SIFT features and utilized AU occurrence relations to constrain the AU detection. Later Zhao et al. [11] employed deep learning for AU detection by extracting deep features with CNNs, dividing the facial images into different regions, and training the kernels separately. Li et al. [3] proposed EAC-Net by inserting an enhancement and cropping layer to a VGG net, adding more weights to AU centered facial areas, dividing features into regions according to facial landmarks with separate convolution filters. Shao et al. [12] developed a end-to-end deep learning framework: JAA-Net, which utilized the correlations of AU detection and face alignment, and refined the AU attention map to optimize the local feature learning. Li et al. [13] proposed an AU semantic relationship embedded representation learning (SR-ERL) framework that utilized a multi-scale CNN for features representation and a GGNN (Gated Graph Neural Network) for AU semantic relationship learning. These methods have significantly improved the AU detection accuracy with the power of deep learning. However, these regional attention or prior knowledge based methods require extra landmark information; there is a lack of self-learned attention, and importantly, they ignore the sequential frame wise relationship in the temporal space.

Detection with spatial-temporal fusion. Valstar et al. [14] first considered temporal dynamics in AU detection, and they extracted geometrical features and introduced a hybrid SVM/HMM classifier to model temporal information, and they found that modeling temporal dynamics can significantly increase the performance. As recurrent neural networks (RNN) show powerful ability for temporal dependency modeling. Chu et al. [15] have first introduced Long Short-Term Memory (LSTM) to learn AU temporal relations, and combined CNNs and LSTMs to model both spatial and temporal cues, with adding both cues to generate frame-based AU detection. Ma et al. [16] redefined AU partition rules with expert prior knowledge and divided the facial areas into related groups, different AU groups are separately learned to reduce the interactions. Then they fused each AU group's spatial-temporal features with Convolutional LSTMs. The key issue is that they ignored the global relationship of different AUs and fused the temporal relations near the end of the network, where the features maps' location information is not preserved, as a result, there is no performance gain being achieved. Additionally this framework also requires extra landmark information and cannot be easily extend to other unregistered AUs detection once the groups division is fixed. Romero et al. [17] leveraged temporal relations by introducing an optical flow and textures based two-stream network. Although it is effective in video classification, such a network is heavy, hard to train, highly dependent on the quality of optical flow images, and not end-to-end trainable. Yang et al. [18] have applied a 3D convolution into AU detection, in which two layers of 2D convolutional net and 3D convolutional net have been used to extract spatial-temporal features and spatial features, respectively, followed by a fully connected layer. However, the 3D convolution has much more parameters to train and can poorly model temporal information for frame-based tasks such as AU detection.

III. Spatial-Temporal Learning

A. Overview

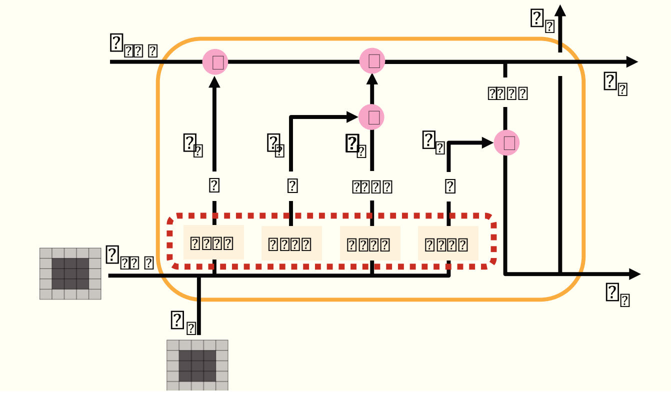


Fig. 1: Structure of ConvLSTM, in ConvLSTM, all fully connected layers are replaced with convolutional layers.

Convolutional neural networks (CNN) like ResNet [8] has been proven effective on feature learning and classification. Our SAT-net (Fig. 2) is adapted from ResNet and hence inherits its spatial feature representation power. SAT-net mainly consists of two components: temporal fusion and self-attention module (Fig. 2 and Fig. 3). We employ Convolutional LSTM (Conv-LSTM) [7] which extends the idea of LSTM and models the spatial-temporal relations that have both input to state and state-to-state transitions. We also develop a self-learned attention module to improve the spatial learning process.

SAT-Net also has four blocks. Given an input x, T images with resolution of $H \times W \times 3$, the spatial resolution of feature maps going through each block is decreased by 2 times and become more and more correlated to the tasks semantic space. Noted that output M_1 after the first two blocks f_{res1} has c_1 number of feature maps in $h_1 \times w_1$.

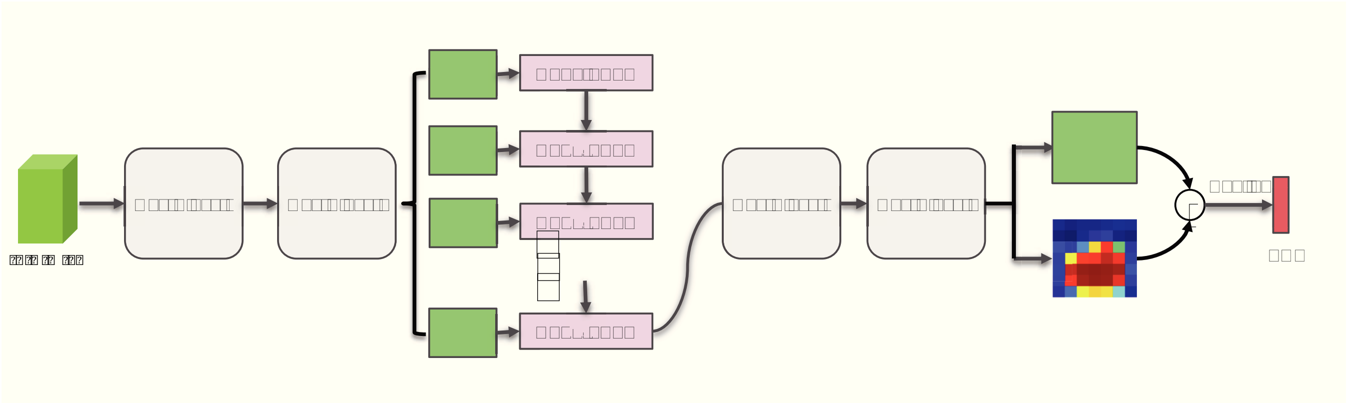


Fig. 2: Structure of our proposed framework. SAT-Net stems from ResNet with 4 residual blocks. It is characterized by two main components. **Conv-LSTM**: features generated from Res-Block-2 are sequentially fed into this module. The hidden state of current frame is processed by the rest of residual blocks. **Self-Attention**: learning feature maps and attention masks at the same time, after which attended features map are pooled for predictions.

B. Temporal Fusion

We would like to consider the temporal context among consecutive frames. A natural choice is the family of recurrent neural networks (RNN), LSTM for instance, which has strong ability to model temporal patterns. Regarding AUs recognition task, previous works [15], [19] employ fully connected long short-term memory (FC-LSTM) to learn the temporal dependency. One drawback of FC-LSTM is that we need to reduce the spatial dimensions for input at the risk of losing location information. The spatial context is important for AUs recognition, but is hard to be retrieved by LSTM. To remedy this issue, we opt to use Conv-LSTM which directly takes M_1 as input and models the temporal relationship without averaging out spatial context.

Detailed structure of Conv-LSTM [20] we employ is shown in Fig. 1. Compared to FC-LSTM, where the gates i_t, f_t, o_t are represented by activations of fully connected layers that take in one dimensional vectors, Conv-LSTM use convolution operations to deal with the inputs.

$$\begin{aligned} i_t &= \sigma(Conv(x_t; w_{xi}) + Conv(h_{t-1}; w_{hi}) + b_i) \\ f_t &= \sigma(Conv(x_t; w_{xf}) + Conv(h_{t-1}; w_{hf}) + b_f) \end{aligned} \tag{1}$$

Eq. 1 shows the input and forget gates where $x_0 = M_1$. Noted that there is no need to reduce the inputs x_t and hidden states h_{t-1} from last cell as we are using convolutions. The input gate i_t determines the amount of information accumulated to the current cell state c_t and the forget gate f_t decides how much information to forget about the last cell state c_{t-1}, as shown in Eq. 2.

$$\begin{aligned} c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \qquad where, \\ \tilde{c}_t &= \tanh(Conv(x_t; w_{xg}) + Conv(h_{t-1}; w_{hg}) + b_g) \end{aligned} \tag{2}$$

After update of the current cell state c_t, we pass it through $\tanh$ activation layer and further multiply with the an output gate o_t. Gate o_t controls how much the current cell state c_t to be propagated to the hidden state h_t.

$$\begin{aligned} o_t &= \sigma(Conv(x_t; w_{xo}) + Conv(h_{t-1}; w_{ho}) + b_o) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

In action recognition literature, Conv-LSTM is widely adopted due to its superior performance to model spatial and temporal relations [20], [21]. A typical task in action recognition is video-based classification. One label is given for each segment of frames. Therefore, every individual frame in this segment will contribute to the same label. Under this scenario, it makes sense to utilize hidden states of all frames and find a way (i.e. concatenation, summation) to aggregate them.

But AUs detection problem is formulated differently in popular AU-related datasets. Every frames in the sequence are annotated individually, though considering the temporal context, with multi-labels w.r.t number of AUs. Therefore, the occurrence of AUs is detected in a frame-based pattern. How to successfully inject Conv-LSTM to the network for AU recognition is open to discussion. In particular, what are the inputs and how to deal with the outputs of Conv-LSTM module are our main contributions.

We propose to feed M_1 after the second residual block into Conv-LSTM and use the hidden state of the current frame h_t for the rest of the residual blocks. Since the residual block closer to data layer preserves more spatial information while the one at the top of network is highly abstracted and aligned to task space, if we consider temporal context at the very end of spatial learning, as in [15], [19], the location information of each frame and the state transition in pixel level between frames will be lost. h_t not only takes in the feature maps of current frame x_t which is a direct clue of the facial appearance, but accounts $h_0 \dots h_{t-1}$ from the previous frames with cell state transition. With the accumulation of state information, every previous frame has impacts to the final Conv-LSTM output.

C. Spatial Learning with Self-Attention

In vanilla ResNet, global average pooling (GAP) followed by one fully connected (fc) layer is an effective strategy to derive categorical labels. Instead of flatting feature maps and using multiple fc layers, GAP simply averages out spatial information, hence proved more robust to spatial translations and parameters efficient. But AUs detection is a multi-label problem. One linear layer after shared feature maps is inadequate to represent different AUs. A prevalent solution is to increase model capacities in account for each AU. [12], [13] fall into this category. After a shared network as global context, they tailed with parallel branches so that each AU has one set of parameters that are updated independently. Although proved effective, the number of parameters will be increased dramatically w.r.t the number of AUs.

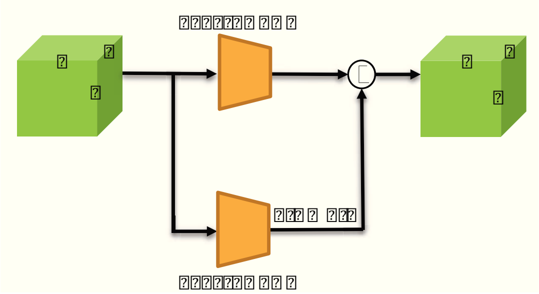


Fig. 3: Self-Attention Module in SA-Net/SAT-Net, C, N stand for the number of input and output channels, h and w are the spatial resolution of feature maps.

Inspired by the attention mechanism, we come up with a workaround to learn attention maps while learning the features of each AU. As elaborated in Fig. 3, self-attention module has two streams. The upper one is to learn features for each of the AUs with 1×1 convolutions, and the lower one use one more set of 1×1 convolutions followed by sigmoid activation for attention maps. Then we merge these two streams by performing element wise multiplication. In terms of the overall recognition pipeline, self-attention module is appended after Res-Block-4, where attention maps are learnt in parallel to highlight facial areas which best describe different AUs. Therefore the attended feature maps are average pooled spatial-wisely to give AU occurrence predictions.

One major advantage of our network is the number of trainable parameters. Extra parameters accounting for the temporal context are greatly reduced. Assuming a kernel size of $K \times K$, input channel I, and output channel O, our proposed network has $4 \times K^2 \times (O+I) \times I$ parameters for 4 gates (omitting the bias), which is $4 \times (128+128) \times 128 \times 9 \approx 1.2\ million$ in the experiments. Such temporal fusion module has much less parameters than backbone ResNet-18 ($11\ million$), as well as peer works [16], which also models the temporal dependency with Conv-LSTM: $4 \times (2048+256) \times 256 \times 9 \approx 21.2\ million$, and [19] which uses FC-LSTM: $4 \times (2048+512) \times 512 \approx 5.2\ million$. Thus our proposed model is efficient on training parameters.

D. Imbalanced Classification

Some of the AUs are likely to occur while some are under representative. To balance the positive sampling rate of each AUs, during the training stage, we add a weight w_c to the cross entropy loss [22] as shown in Eq. 4.

$$\begin{aligned} Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} & w_c y_{n,c} \cdot log\sigma\left(x_{n,c}\right) \\ & +(1-y_{n,c}) \cdot log(1-\sigma(x_{n,c})) \end{aligned} \tag{4}$$

where C is the number of labels (AUs) considered, N is the number of samples in mini-batch. $x_{n,c}$ is the output of the network in the forward pass which is going through σ activation and $y_{n,c}$ is the ground truth labels. w_c is obtained by computing the occurrence rate of each AU in the training set.

IV. Experiments

A. Datasets and Settings

We evaluate our approach on two benchmark datasets BP4D [23] and DISFA [24]. They contain sequences with spontaneous emotions and labels that are manually annotated by FACS coders. In align with peer works, we only consider AUs which have occurrence rate greater than 5% in our experiments.

BP4D. There are 23 female and 18 male subjects with diverse ethnicities and backgrounds in the BP4D data collection. Each subjects are required to perform 8 tasks for a total of 328 emotional sequences. In each sequence, both 2D texture images and 3D geometric models are provided. $\sim 140,000$ frames have valid AU occurrence codes.

DISFA. It contains 12 female and 15 male subjects. The subjects are asked to watch videos while spontaneous facial expressions are captured. One video is provided for each subject. The video frames are labeled by $0-5$ intensity values with 0 means the absent of AU and 5 the most expressive. In our experiments, we define the AU intensity greater or equal than 2 as occurred and less than 2 as not. More than $100,000$ annotated images can be extracted from the videos. The data imbalance issue is more severe than that of BP4D. Some AUs have much more negative samples than positive ones.

Implementation Details. The whole framework is implemented in PyTorch [25]. For pre-processing the data, we use face detection tools to locate and crop the face which greatly removes the backgrounds and then resize the input frames into 224×224 before feeding the networks. We augment input frames by random horizontal flipping. Due to the dramatic increasing of parameters introduced by extra dimension on temporal, we choose ResNet-18 as our backbone which achieves a good trade-off between accuracy and computing complexity. We conduct subject-exclusive 3-fold cross-validation following the same splitting protocol as in [15]–[17], [23]. We employ stochastic gradient descent (SGD) as our optimizer with a mini-batch size of 20, a Nesterov momentum of 0.9, and a weight decay of 0.0002. The initial learning rate is 0.01 and it decays 0.1 every 8 epochs until convergence.

Evaluation Metrics. We report the performance of proposed framework in terms of F_1-score. F_1-score is formulated as $F_1 = 2PR/(P+R)$, where P and R denote precision and recall, respectively.

AU	JPML	DRML	CNN-LSTM	EAC	JAA	LP	$AR_{ConvLSTM}$	SRERL	ResNet	T-Net	SA-Net	SAT-Net
1	32.6	36.4	31.4	39.0	47.2	43.4	48.0	46.9	50.8	45.9	52.0	[**54.1**]
2	25.6	41.8	31.1	35.2	44.0	38.0	43.2	45.3	45.4	43.9	45.1	[**49.5**]
4	37.4	43.0	71.4	48.6	54.9	54.2	53.1	55.6	56.2	55.8	[**60.0**]	58.3
6	42.3	55.0	63.3	76.1	77.5	77.1	76.9	77.1	77.1	76.5	[**78.0**]	77.7
7	50.5	67.0	77.1	72.9	74.6	76.7	78.4	[**78.4**]	76.6	76.8	76.9	77.7
10	72.2	66.3	45.0	81.9	[**84.0**]	83.8	82.8	83.5	82.3	82.6	83.8	83.6
12	74.1	65.8	82.6	86.2	86.9	87.2	[**87.9**]	87.6	86.7	86.8	87.3	86.5
14	65.7	54.1	[**72.9**]	58.8	61.9	63.3	67.7	60.6	57.2	59.5	61.0	63.2
15	38.1	36.7	33.2	37.5	43.6	45.3	45.6	52.2	49.3	[**53.0**]	49.5	49.1
17	40.0	48.0	53.9	59.1	60.3	60.5	63.4	[**63.9**]	60.5	62.8	61.0	61.8
23	30.4	31.7	38.6	35.9	42.7	48.1	47.9	47.1	48.1	[**50.0**]	47.5	48.7
24	42.3	30.0	37.0	35.8	41.9	54.2	[**56.4**]	53.3	50.0	48.5	47.6	49.3
Avg	45.9	48.3	53.2	55.9	60.0	61.0	62.6	62.9	61.7	61.8	62.5	[**63.3**]

TABLE I: Performance on BP4D

AU	LSVM	APL	DRML	EAC	JAA	$AR_{ConvLSTM}$	SRERL	ResNet	SA-Net	T-Net	SAT-Net
1	10.8	11.4	17.3	41.5	43.7	26.9	[**45.7**]	29.8	32.3	36.6	41.2
2	10.0	12.0	17.7	26.4	46.2	24.4	[**47.8**]	29.3	33.1	32.5	33.1
4	21.8	30.1	37.4	[**66.4**]	56.0	58.6	59.6	56.6	62.3	64.9	63.0
6	15.7	12.4	29.0	50.7	41.4	49.7	47.1	[**57.3**]	52.2	53.1	56.4
9	11.5	10.1	10.7	[**80.5**]	44.7	34.2	[**45.6**]	35.5	33.3	35.8	43.0
12	70.4	65.9	37.7	[**89.3**]	69.6	71.3	73.5	71.8	71.2	74	73.1
25	12.0	21.4	38.5	[**88.9**]	88.3	83.4	84.3	84.6	84.0	82.2	82.9
26	22.1	26.9	20.1	15.6	58.4	51.4	43.6	55.2	59.6	55.7	[**60.6**]
Avg	21.8	23.8	26.7	48.5	56.0	50.0	55.9	52.5	53.5	54.3	[**56.7**]

TABLE II: Performance on DISFA

B. Comparison with state-of-art Methods

We compare our methods (T-Net, SA-Net, and SAT-Net) with state-of-art image based (JPML [26], DRML [10], EAC [3], JAA [12], LP [27], SRERL [13]) and sequence based (CNN-LSTM [15], $AR_{ConvLSTM}$ [16]) AU detection works under the same 3-fold cross validation setting.

Results on BP4D. Table I shows performance comparison on BP4D dataset reported in percentage, our baseline network is ResNet-18 pretrained with ImageNet, and the fully connected layers is adapted to have 512 input channels and 12 output channel which is the number of AUs to detect. We found that pretrained parameters are critical to performance because of the low variance in AU detection datasets. Images are very similar for each subject and sequences contain a large portion of neutral faces, tending to overfit the training set. For the following experiments, we always initialize the network with ImageNet pretrained parameters. By fine-tuning and balancing the positive negative loss, the baseline ResNet can achieve an average F_1 score of 61.7 across all AUs on BP4D. And the score further improves from 61.7 to 62.5 with the introduction of self-attention module (SA-Net). SA-Net outperforms EAC [3] and LP [27] by 11.8%, 4.2% respectively, in terms of the F_1 score. Both of these two works use handcraft landmark-based attention maps to attend to AU centered areas. Such performance improvement demonstrates the benefit of self-attention module. Compared to baseline ResNet, the T-Net (with temporal fusion module only) does not show much performance gain due to the lack of self-attention module. However, the combination of temporal fusion and self-attention module (SAT-Net) can greatly increase the F_1 score and achieves the best: 63.3% among the proposed networks in this paper, 1.3% more than the SA-Net. Comparing to [15] and $AR_{ConvLSTM}$ [16] which also address the temporal dependency. Our SAT-Net obtains the best performance and increases by 19.0% and 0.6% respectively. We found the redesign of input and output of temporal fusion model works better than $AR_{ConvLSTM}$ and FC-LSTM based CNN-LSTM [15] as well as showing the performance boost by combining Conv-LSTM and self-attention modules.

Results on DISFA. In DISFA dataset, AUs are annotated with $0 - 5$ intensity levels, we treat frames with intensities equal or larger than 2 as positive samples in our experiment in line with compared works. Noticed that DISFA is more imbalanced in terms of AUs occurrence rate than that of BP4D. One AU could have around 10 times positive samples than the other, leading to heavier performance fluctuation during the training phase. From Table II, we can see that our proposed SA-Net has increased F_1 score from 52.5 to 53.5. Some of the less representative AUs, such as AU1 and AU2, have remarkable improvement with the help of self attention module. While T-Net, taking advantage of Conv-LSTM, has larger performance boost from 52.5 to 54.3. And the combination of self-attention and temporal fusion further increases this value to 56.7. It is a 13% improvement over $AR_{ConvLSTM}$ [16] which also considers the temporal dependency, and 1.3% over JAA [12], achieving the state-of-the-art result. It is worth noting that the reason why temporal fusion has obvious performance increment in DISFA than BP4D is that DISFA's video of each subject shows a more consistent affection status for each task, which is different from videos in BP4D, where each subject may experience varied emotions during each individual task. Thus the videos' continuous property with expressions' consistency and regularity across subjects makes the DISFA data easier to model for the temporal dependency.

C. Discussion

We first visualize the feature maps (the first 32 channels) between SA-Net and SAT-Net in Fig. 4, which demonstrate the impact of introduced Conv-LSTM module. Fig. 4a are the activations after Res-Block-2 in SA-Net (same as baseline ResNet) while Fig. 4b are the ones after we fuse the temporal dependency with Conv-LSTM module. Comparing these two figures, we can clearly see the differences. For feature maps in Fig. 4a, most facial areas are dark. One can easily recognize the edges and contours of the face, meaning that a smaller percentage of pixels in each channel is considered important by the neural network and thus activated for AU predictions. Without any context from temporal domain, SA-Net can only focus on various interpretations of spatial information for the current frame therefore the activation maps are very sparse. In contrast, temporal fused features (Fig. 4b) have evenly spread values across all channels. Through the temporal learning process, some general information like face shapes, contours are averaged out. Those are considered less important on revealing muscle movement. And the network could pay more attention on the facial details. Fig. 4 indicates that the temporal information is an effective supplement for the facial action unit detection tasks.

AU 1 Inner Brow Raiser	AU 2 Outer Brow Raiser	AU 4 Brow Lowerer	AU 6 Cheek Raiser
AU 7 Lid Tightener	AU 10 Upper Lip Raiser	AU 12 Lip Corner Puller	AU 14 Dimpler
AU 15 Lip Corner Depressor	AU 17 Chin Raiser	AU 23 Lip Tightener	AU 24 Lip Pressor

TABLE III: Corresponding AU index in Fig. 5

Fig. 5 shows the the effectiveness of attention maps generated by SA-Net and SAT-Net on BP4D. Both figures show the attention module works well for two networks. Attention of AU1, AU2 and AU4 focus on the upper face near eyes which is related to the eyebrow muscle movement, and the followings focus on the middle and lower facial area which is consistent with the place where the corresponding AUs appear. From Fig. 5a, attention map generated without temporal fusion, we can see the high value weights in eye and brow area for AU1, AU2, and AU4, which are Inner Brow Raiser, Outer brow raiser and brow lower, and the values in the background and other facial areas are almost 0. The same for AU6 representing cheek raiser and the attention maps exclude the area nears eyes which is reasonable. For AU12, 14, 15, 17, 23, 24: AUs appears around lip, the attention maps are able to correctly show high weight values around lip and mouth area. As the self-learned attention has no ground truth bounding box of action unit areas, the attention does not appear exactly around AU centers and inevitably covers some unrelated pixels. However background noisy pixels are successfully excluded in both attention maps. And because our self-attention is placed following the last convolution layer, the feature map reception field are quite large therefore some AUs' attention are quite similar, however still focus on correct part of the face. Fig. 5b shows the attention maps generated with temporal fusion module, where we can observe that the attention map expands compared with Fig. 5a (attention maps from SA-Net). Which is also consistent with what we have observed from Fig. 4, where the feature maps show the temporal fusion makes feature maps expand due to the cell state information transfer of the sequence previous frames then information is fused into the last frame's feature maps. The last frame's features describe more information than a single frame thus cause the expansion.

To utilize the temporal information, we also consider the number of frames to use for each training step. Table IV and Table V show the AU occurrence frame length in two datasets, and we treat two continuous frames having 0 intensity as the AU end signal. The first three AUs are around 25 frames, and the following six AUs gather in 40-70 and the last three AUs are around 19 frames. In the SAT-Net, we use a frame sequence of 16 frames to train and test because 16 frames can cover the least median frames(18) that an AU will occur. Using more frames will cover some non-occurring AUs and thus introduce noise.

AU	1	2	4	6	7	10	12	14	15	17	23	24
Median	25	28	25	54	45	48	72	42	18	19	18	21

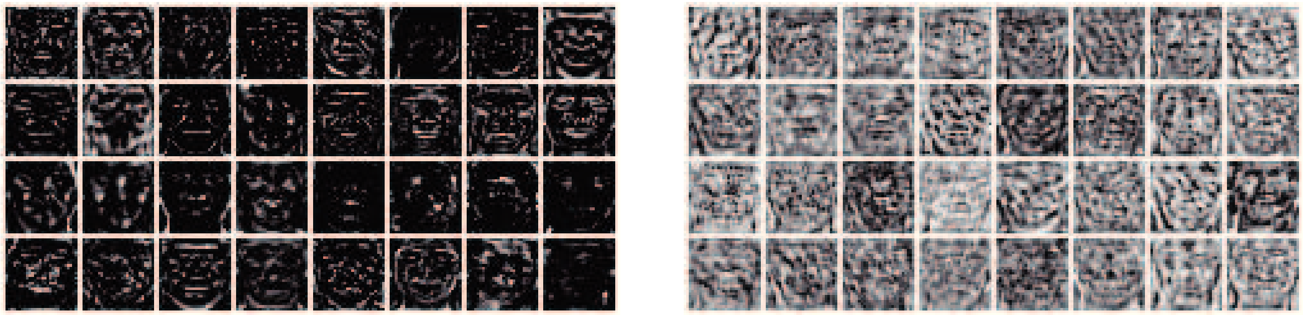
TABLE IV: AU occurrence median frames on BP4D

AU	1	2	4	6	9	12	25	26
Median	22	25	23	84	45	82	85	33

TABLE V: AU occurrence median frames on DISFA

V. Conclusion

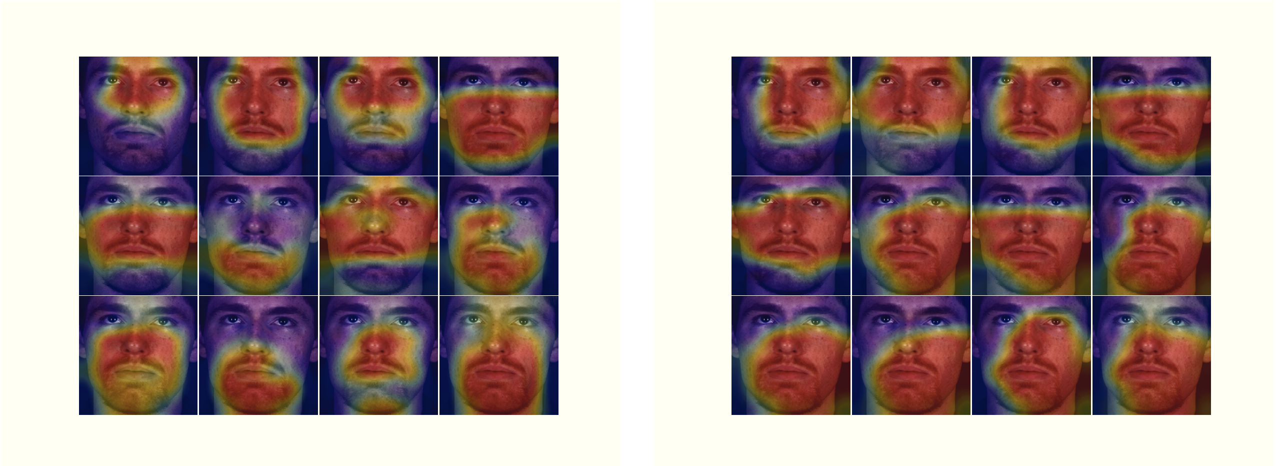
In this paper, we have proposed a novel light weighted spatial self-attention and temporal fusion network specifically for facial action unit detection. Our network utilized the least training parameters but achieves the state-of-the-art performance. Different from previous works that utilize prior knowledge based handcraft attention mechanism, we developed an AU label supervised self-learned attention module to enable the network to learn to pay more attention to different facial areas for the corresponding AUs. We have also proposed to use Conv-LSTM module to fuse the temporal information into AU detection problems and proved to be feasible with temporal information as a supplement in facial action unit detection. Our proposed network is evaluated on both BP4D database and DISFA database, and experimental results show that the proposed SAT-Net outperforms most of the handcraft attention based networks as well as temporal fusion networks, achieving



(a) SA-Net feature maps

(b) SAT-Net feature maps

Fig. 4: Visualizations of feature maps with and without Conv-LSTM module.



(a) SA-Net attention maps

(b) SAT-Net attention maps

Fig. 5: Self-attention maps with the supervision of AU labels

the state-of-the-art performance on BP4D and DISFA. Our experiments have also explained how attention and temporal fusion work inside the network.

Our future work will address the data unbalance issue from the datasets in order to improve the training of our network. And avoiding overfitting is also critical for stable training in AU detection, which will be explored in our future work. Moreover, to make the attention module more stable, we will further explore landmark information so as to supervise the attention training along with the self-learning based on the supervision of labeled AUs.

VI. Acknowledgement

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

References

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

[2] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[3] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "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 2017)*. IEEE, 2017, pp. 103–110.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[7] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, pp. 149–149.

[10] 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.

[11] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.

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

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

[14] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden markov models for modeling facial action temporal dynamics," in *International workshop on human-computer interaction*. Springer, 2007, pp. 118–127.

[15] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 25–32.

[16] 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.

[17] A. Romero, J. León, and P. Arbeláez, "Multi-view dynamic facial action unit detection," *Image and Vision Computing*, 2018.

[18] L. Yang, I. O. Ertugrul, J. F. Cohn, Z. Hammal, D. Jiang, and H. Sahli, "Facs3d-net: 3d convolution based spatiotemporal representation for action unit detection," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 538–544.

[19] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1841–1850.

[20] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[21] L. Zhang, G. Zhu, L. Mei, P. Shen, S. A. A. Shah, and M. Bennamoun, "Attention in convolutional lstm for gesture recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 1953–1962.

[22] PyTorch, "Bcewithlogitsloss," https://pytorch.org/docs/stable/index.html.

[23] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[24] 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.

[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[26] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2207–2216.

[27] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 917–11 926.