# Action unit detection by exploiting spatial-temporal and label-wise attention with transformer

Lingfeng Wang, Jin Qi, Jian Cheng
University of Electronic Science and Technology of China
Chengdu, China
202021011417@std.uestc.edu.cn, jqi@uestc.edu.cn, chengjian@uestc.edu.cn

Kenji Suzuki
Tokyo Institute of Technology
Tokyo, Japan
suzuki.k.di@m.titech.ac.jp

## Abstract

*The facial action units (FAU) defined by the Facial Action Coding System (FACS) has become an important approach of facial expression analysis. Most work on FAU detection only considers the spatial-temporal feature and ignores the label-wise AU correlation. In practice, the strong relationships between facial AUs can help AU detection. We proposed a transformer based FAU detection model by leverage both the local spatial-temporal features and label-wise FAU correlation. To be specific, we firstly designed a visual spatial-temporal transformer based model and a convolution based audio model to extract action unit specific features. Secondly, inspired by the relationship between FAUs, we proposed a transformer based correlation module to learn correlation between AUs. The action unit specific features from aural and visual models are further aggregated in the correlation modules to produce per-frame prediction of 12 AUs. Our model was trained on Aff-Wild2 dataset of the ABAW3 challenge and achieved state of art performance in the FAU task, which verified that the effectiveness of the proposed network.*

## 1. Introduction

Facial affective behavior analysis plays an important role in human-computer interaction [15]. Facial Action Coding System (FACS) is one of the most important approach of face analysis. FACS deconstruct facial expressions into individual components of basic muscle movement, called Action Units (AUs).It allows computer systems to understand human feelings and behaviors, which makes human computer interaction more applicable.

Since each facial action are defined at specific region within a short duration. Local spatial and temporal attention could help model to extract facial features. We therefore employ attention mechanism to focus on regions of interest with spatial and temporal transformer. Moreover, The task of FAU detection can be formulated as a multi-label binary classification problem. Most existing studies for FAU detection usually treat each AU label as an one-vs-all binary classification problem, which fail to exploit dependencies among AUs. However, the AU labels are highly dependant to each another. Exploiting the correlation between labels could help to boost multi-label classification performance.

In the challenge for Affective Behavior Analysis in-the-wild (ABAW3) Competition [8–15, 20], the organizers collect a large scale in-the-wild database Aff-Wild2 to provide a benchmark for valence-arousal (VA) estimation, expression (Expr) classification, action unit (AU) detection, multi-task-learning (MTL) tasks respectively.

In this paper, we describe our approach for Action Unit Detection challenge in the ABAW 2022 competition. we firstly trained a spatial-temporal transformer based video model to extract visual feature as well as a convolution based aural model for audio feature. Features from both the aural and visual model are then pass through 12 independent fully connected layer to extract 12 discriminative action unit specific features. After that, we proposed a action units correlation module to learn relationships between the action unit specific features and refine FAU prediction result.

## 2. RELATED WORKS

There are three stream in FAU detection research: spatial attention, temporal modeling, and AU correlation [1].

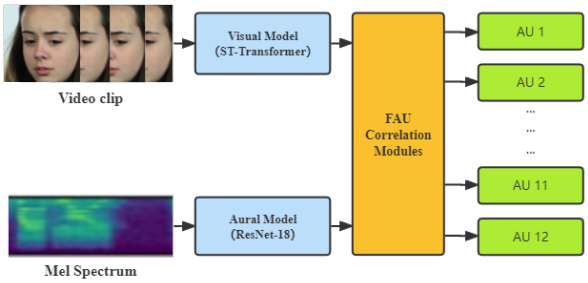



Figure 1. Overview of proposed architecture

Previous studies have proposed some effective facial action unit detection models by exploiting spatial attention. Zhao *et al*. [22] jointly learn image patches around detected landmarks and multi-label classification. JAA-Net [18] jointly estimates the location of landmarks and predict action unit.

Since transformer proposed by Vaswani *et al*. [19] has achieved the state of the art performance in many tasks, more and more researchers utilize transfomer for temporal modeling.Yue Jin *et al*. [7] proposed a transformer based model to merge feature from visual and audio model in sequence. zhao *et al*. [23] proposed a transformer with spatial and temporal attention for facial expression analysis.

Recently, several works also take the relationships of AUs into consideration. Chu *et al*. [1] designed a hybrid network that jointly learns Spatial representation, temporal modeling, and AU correlation for multi-label AU detection. Jacob *et al*. [5] proposed an attentin branch network for spatial attention learning and a transformer correlation module to learn relationship between action units.

As for Aff-Wild2 dataset, previous solution in the ABAW competition have reveal the effectiveness of aural-visual multi modal method. Kuhnke *et al*. [16] proposed a two-stream aural-visual network to combine vision and audio information for emotion recognition and achieves superior performance. In the ABAW3 Competition this year, the winner team [21] proposed a transformer network to fuse multi-modal information including spoken words, speech prosody, and visual expression in videos for AU detection. The runner up team SituTech [6] use a convolutional model pretrained on private AU dataset and achieved comparable performance. Team RPL [17] use two branches of transformer layers to merge temporal features and ranked the third.

# 3. METHODOLOGY

## 3.1. Framework

Figure 1 shows the framework of our multi-task affective behavior analysis model. All the video in the dataset are splitting into image clip and audio mel spectrogramm during training and inference stage. The two streams are pre-processed and fed into the aural-visual model synchronously.

For the Visual stream, the input frames are cropped facial region images. These facial crops are all aligned according to 5 point template (eye centers, nose tip, outer mouth corners).Each input clip contains l frames and the frames are sampled with dilation d. Here we choose clip length l = 16 and dilation d = 3. As for audio stream, we compute a mel spectrogramm for all audio stream extracted from the video using TorchAudio package. For each clip, spectrogram is cut into a smaller sub-spectrogram with the center of sub-spectrogram aligning with the current frame at time t.

The two stream are pass through Aural and Visual model respectively. We employ spatial-temporal transformer backbone described in [23] to extract spatio-temporal information from visual stream as well as a ResNet-18 [4] model for mel spectrogram feature learning. Finally, the output features of both models are merged into FAU Correlation module and give the joint prediction of action units.

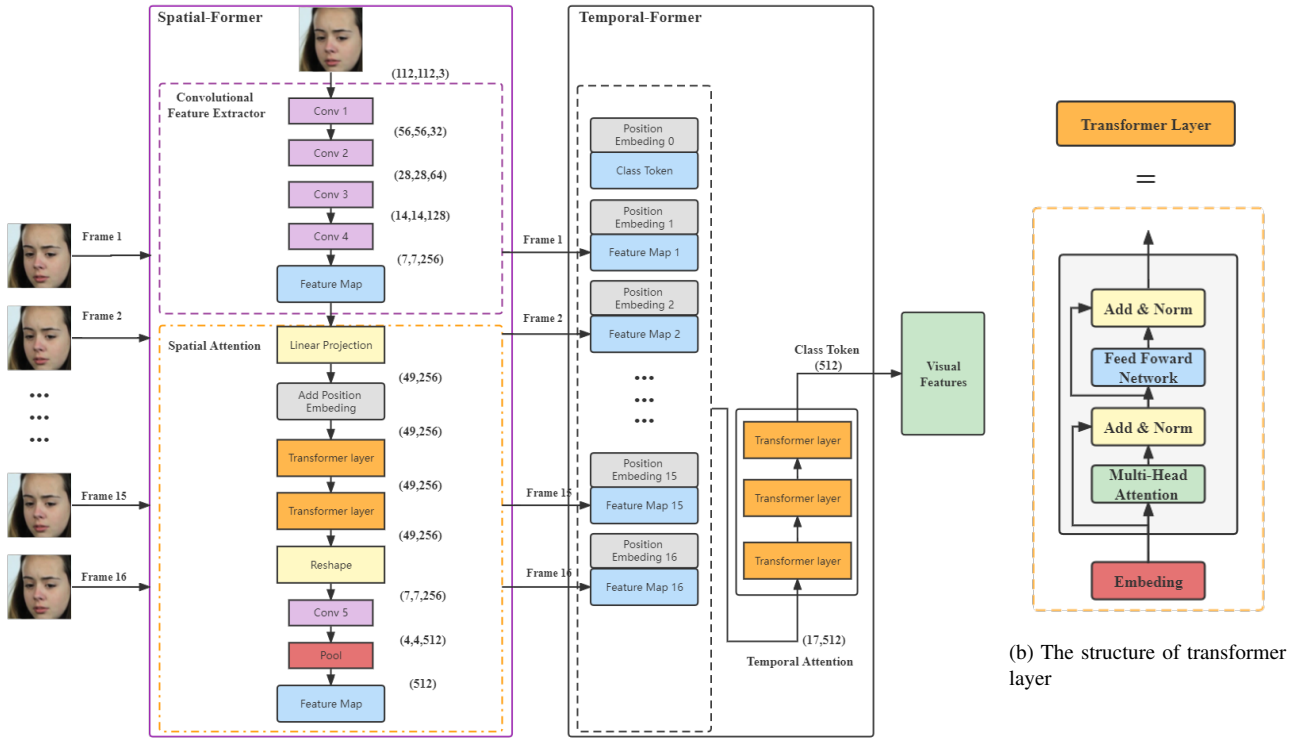
## 3.2. Transformer encoder

The transformer encoder has two main components: Multi Head Attention and Feed Forward Networks [19]. Layer normalization is also applied to accelerate model convergence. The structure of transformer encoder is shown in Figure 2b.The input of transformers includes a variety of tokens. The input tokens are fed into multi head self-attention module, where information aggregating globally. In this way transformer can model long-distance feature dependencies effectively.

## 3.3. Spatial-Temporal attention

We use Spatial-Temporal transformer from [23] as visual backbone. The structure of Spatial-Temporal transformer is shown in Figure 2a. It mainly consists of a convolutional spatial transformer (S-Former) and a temporal transformer (T-Former). The S-Former takes still frame as input and extract spatial facial feature. Following that, features at each frames are pass through T-Former in sequence and generate temporal feature representation.

The convolutional spatial transformer (S-Former) is a CNN-transformer hybrid architecture which consists of five ResNet-18 [4] basic blocks and two spatial transformer encoders. We didn't use pure ViT architecture in [3] due to the fact ViT can capture long-distance feature dependencies effectively but fail to extract local feature details. As




(a) The structure of visual model with spatial and temporal attention.

(b) The structure of transformer layer

Figure 2. The structure of visual model(a) and transformer layer(b)

for CNN, traditional CNN architecture cannot capture rich global contextual information due to the limit of CNN receptive field. Proposed CNN-transformer hybrid design can leverage both global and local information.

We utilize temporal transformer to to modeling the relationship between frames. The temporal transformer (T-Former) consists of three temporal encoders. Features extracted by S-Former at each frames are pass through T-Former in sequence. An extra class token is concatenate at the top of sequence. And then the sequence is add by positional embedding as input for T-Former. After three cascaded transformer encoder, the embedding of class token is output as spatial-temporal visual feature.

### 3.4. FAU attention

Figure 3 shows the framework of proposed AU Correlation architecture. We use 12 independent fully connected layers to extract features for 12 AU labels. The discriminative features from the of 12 AU branches are provided as input to the AU correlation module. We use two layer of transformer encoders as correlation module. The discriminative features is then add by positional embedding as input embeddings for transformer. Relationships between FAU features are learnt in the transformer encoder. The output of transformer encoder are passed to a predict head with 12 fully connected layer to predict the labels.

Our visual model and audio model are both connected to the AU Correlation architecture in 3. When training visual and audio model, their losses are computed by groundtruth label and predict result of predict head. When training the joint model, we discard the predict heads of visual and audio model and output correlated features directly. The correlated features from visual branch $F_v \in \mathbb{R}^{12\times256}$ and features from aural branch $F_a \in \mathbb{R}^{12\times256}$ are concatenate into $F_{av} \in \mathbb{R}^{12\times512}$. The merged features are then fed into a final correlation module where the multi-modal features are aggregated. Following that, we use a final prediction head to produce per-frame prediction of 12 AUs.

### 3.5. Loss Function

Facial AU detection can be regarded as a multi-label binary classification problem. However, action unit samples in Aff-Wild2 dataset suffer from class imbalance problem. We use binary cross entropy loss (BCE) loss with position weight to tackle the challenge. The position weights here is proportional to the ratio of positives in the total number for each AU class in training set. Weighted BCE allows model to achieve trade-off between recall and precision.

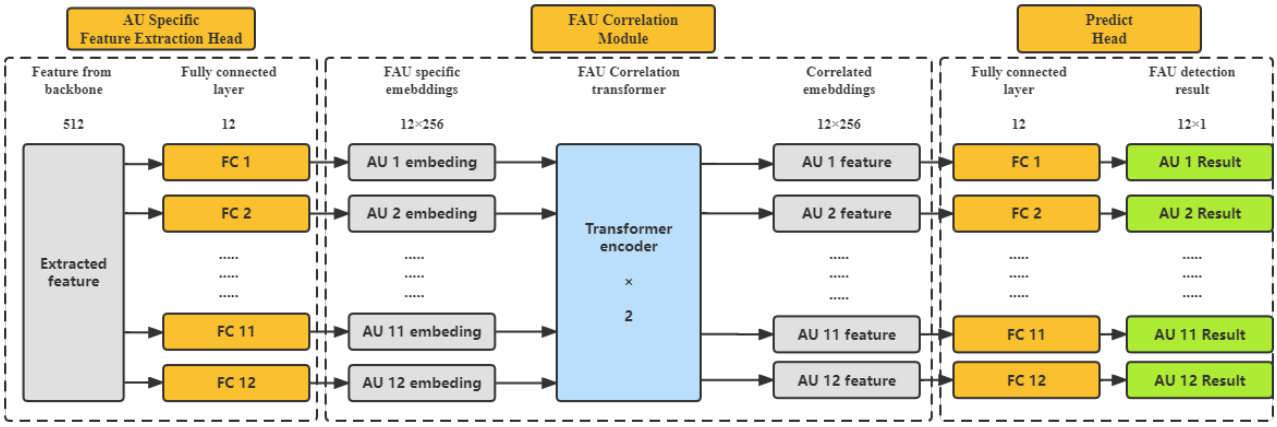



Figure 3. Framework of proposed FAU correlation architecture

$$L_{BCE} = \mathbb{E}[-\sum(w_i t_i \cdot log p_i + (1-t_i)\cdot log(1-p_i))] \quad (1)$$

# 4. EXPERIMENTAL

## 4.1. Dataset

Proposed model is trained on the large-scale in-the-wild Aff-Wild2 dataset only. This dataset contains 564 videos with frame-level annotations for valence-arousal estimation, facial action unit detection, and expression classification tasks. As for action unit detection task, Aff-Wild2 dataset provide 305 training and 105 validation samples. We use the official provided cropped and aligned images in the Aff-wild2 dataset directly.

## 4.2. Training Setup

Model is trained with official train split dataset only. As for visual branch, we trained spatial transformer model firstly. After that, we freeze the parameters of the spatial transformer and train the temporal transformer. At the same time, audio model is trained independently. Finally, we fuse the feature from visual branch and audio branch together and train joint model. Models are optimized using Adam optimizer and a learning rate of 0.0005. AutoAugment strategy for ImageNet described in [2] is applied for each input clip. The mini-batch size is set to 64.

## 4.3. Ablation Analysis

In order to analyze the effects of proposed framework design, we conduct ablation studies to compare performance with or without proposed components. The results can be seen in Table 1. Comparing to ResNet-18 model, proposed CNN-transformer hybrid architecture obtain improvement of 5.7%. For aural branch and Visual model, the use of correlation module can boost performance by 2.1% and 2.2%,

| Method | Score (F1 in %) |
|---|---|
| Competition Baseline [8] | 39.0 |
| Audio | 32.3 |
| Audio(with CM) | 34.4 |
| Visual ResNet only model | 39.8 |
| Visual spatial model | 45.5 |
| Visual spatial model(with CM) | 47.9 |
| Visual spatial temporal model(with CM) | 50.1 |
| Joint Aural and Visual | 52.3 |

Table 1. performance of models on official validation set, CM is short for correlation module

| Method | Score (F1 in %) | Pretrain |
|---|---|---|
| Netease Fuxi Virtual Human [21] | 49.89 | True |
| SituTech [6] | 49.82 | True |
| PRL [17] | 49.04 | False |
| Competition Baseline [8] | 36.50 | False |
| Ours | 48.83 | False |

Table 2. Results on the test set of the Aff-Wild2 dataset.

respectively, which indicates that the usage of correlation module allows the model to learn multi-label relationships and refine classification result. And the use of the temporal transformer can improve the F1 score by 2.1%. Moreover, if the aural branch and visual branch are both employed, the F1 score can reach 52.3%.

## 4.4. Comparison with State-of-the-Arts

We also evaluated our model on the official test set. The results on the test set can be seen in Table 2. Our model obtains F1 score of 48.83 % and outperforms the baseline model of [8] a lot. Our solution rank at 4th on the com-

petition leader board and achieve comparable performance comparing to the Top 3 teams. Our score is lower than team Netease and SituTech by 1 %. This could be due to that they use pre-trained expression model on additional dataset, whereas our model is training from scratch using the Aff-Wild2 database only. Moreover, team PRL [17] also did not use pretrained model. Our F1 score is quite close to their with gap of only 0.21 %, which verifies that the our method have obtained State-of-the-Art performance.

## 5. CONCLUSION

This paper describe an effective FAU detection transformer based model by exploiting spatial-temporal attention and FAU label-wise correlation. Our key idea is to firstly develop a visual spatial-temporal transformer based model and a convolution based audio model. Then we fuse the two branch together and employ correlation module to learn relationship between Action units for further refine FAU detection. Experimental results on test dataset show that our model have achieved State-of-the-Art performance, which verifies the effectiveness of proposed method.

## Acknowledgment

This work is supported by the NNSFC&CAAC under Grant U2133211.

## References

[1] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey 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)*, pages 25–32, 2017. 1, 2

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[6] Wenqiang Jiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. Facial action unit recognition with multi-models ensembling, 2022. 2, 4

[7] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition, 2021. 2

[8] Dimitrios Kollias. ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges, 2022. 1, 4

[9] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, 2020. 1

[10] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network, 2020. 1

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study, 2021. 1

[12] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, Feb 2019. 1

[13] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface, 2019. 1

[14] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework, 2021. 1

[15] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[16] Felix Kuhnke, Lars Rumberg, and Jorn Ostermann. Two-stream aural-visual affect analysis in the wild. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Nov 2020. 2

[17] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video, 2022. 2, 4, 5

[18] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021. 2

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[20] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild' challenge. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987, 2017. 1

[21] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis, 2022. 2, 4

[22] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[23] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 2