This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Robust Multimodal AU Detection: STN-Enhanced Visual Encoding and Audio-Visual Spatial-Temporal Alignment

Jun Yu[1] , Yunxiang Zhang[1,†] , Fengzhao Sun[1] , Leilei Wang[1] , Renjie Lu[1] , Lingsi Zhu[1] , Xilong Lu[1] , Yang Zheng[1] , Yongqi Wang[1]
[1] University of Science and Technolog of China
harryjun@ustc.edu.cn
{mesa,sunfz,sa22218164,renjielu,ls-zhu24,luxilong,zhengyang,wangyongqi}@mail.ustc.edu.cn

Abstract

Facial Action Unit (AU) detection is crucial for understanding human emotional expressions and enhancing the perceptual capabilities of intelligent systems, particularly in in-the-wild environments. The key challenge lies in accurately extracting subtle motion features and effectively integrating complex facial dynamics. However, existing methods are limited by the complexity of multimodal data and the varying environmental conditions, leading to suboptimal performance. To address these shortcomings, we propose a novel multimodal domain alignment strategy. Specifically, we design a multimodal audio-visual feature extraction and fusion framework that analyzes pre-processed audio and video frames at both the global and local levels. For video frames, we employ ConvNeXt integrated with a Spatial Transformer Network (STN) as the image encoder, while for audio, we extract Mel spectrogram features using Whisper. Additionally, we introduce a Temporal Convolutional Network (TCN) for temporal modeling of multimodal features. Experimental results demonstrate that our approach achieves superior performance on the Aff-Wild2 dataset, significantly improving AU detection accuracy compared to existing methods and showcasing state-of-the-art detection capabilities. Notably, our method secured first place in the AU Detection track of the 8th Competition on Affective & Behavior Analysis in-the-Wild (ABAW8), further highlighting its effectiveness and robustness in real-world scenarios.

1. Introduction

In recent years, as digital humans and human-computer interaction technologies have become focal points in the field of artificial intelligence[4], facial expression recognition has garnered increasing attention. These technologies hold the promise of enabling a more accurate understanding of human emotions[33], thereby facilitating the design of more lifelike digital humans and more user-friendly interaction systems. Derived from the Facial Action Coding System (FACS)[34], Action Units (AU) are regarded as the fundamental building blocks of facial expressions and play a crucial role in emotion recognition. The 8th Affective Behavior Analysis in-the-wild (ABAW8)[8, 9, 18, 20–22] Competition aims to address challenges related to human affective behavior analysis. The competition has built a multimodal facial expression dataset, Aff-Wild2[10–17, 19, 45], and introduced the Action Unit (AU) Detection Challenge as one of its tracks. In this challenge, participants are required to detect 12 types of Action Units in every frame of the provided videos.

The Action Unit (AU) recognition challenge faces numerous difficulties[42]. Firstly, the complexity of multimodal data significantly increases the difficulty of the detection task. Second, the uncertain lighting conditions and noise inherent in in-the-wild environments further exacerbate the challenge, placing even greater demands on the robustness of detection methods[41, 47, 49]. Furthermore, the changes in facial actions in videos constitute a dynamic process that requires modeling and analysis over the temporal dimension[25], rather than relying solely on independent processing of static frames.

To address these issues, this study designs and optimizes solutions from the perspectives of feature extraction and enhancement, as well as multimodal fusion. These efforts have led to improved recognition performance and achieved state-of-the-art excellent results. Our contributions are summarized below:

- We apply the pre-trained Whisper model and the ConvNeXt model integrated with a Spatial Transformer Network (STN)[7] to process audio and video streams, respectively. Additionally, we construct both global and local views for video and audio. The local views, obtained

†Corresponding author

by extracting enhanced spatiotemporal clips from video and enhanced time-frequency segments from audio, effectively capture fine-grained key features.
- Furthermore, we integrate an advanced cross-modal alignment and fusion module to synchronize audio and visual features[43]. This module leverages multi-scale self-attention and an adaptive sliding window mechanism to capture temporal context and fuse complementary information, thereby enhancing the accuracy and robustness of AU recognition[38, 48].

2. Related Work

Due to the complexity of Facial Action Unit (AU) composition, the field faces significant challenges, including variations in identity, facial structure, and expression intensity across widely used datasets [31, 39]. These variations make it difficult to develop robust models that generalize well across diverse populations. Furthermore, accurately extracting the most relevant local features for each AU remains a major obstacle, as different individuals may express the same AU with subtle differences. Given the intricate and often ambiguous nature of AU annotations, traditional approaches [2, 6]—particularly those that rely on manually defined feature specifications—exhibit substantial limitations. Such methods struggle to generalize beyond the predefined rules used during training and often fail to capture nuanced expressions effectively.

To overcome these challenges, recent studies have incorporated additional facial landmark annotations to extract meaningful local features and have leveraged multi-task learning frameworks to enhance the performance of AU detection models. For instance, SEV-Net [44] generates local region attention maps guided by textual descriptions, allowing the model to focus selectively on salient facial areas. This method enhances the interpretability of the network and improves its ability to identify subtle AU activations. Additionally, Tang et al. [40] proposed a three-stage training strategy that explicitly integrates facial landmark information within a multitask learning framework. This approach strategically guides the model's attention toward key facial regions, ensuring a more structured learning process. However, while these methods improve AU detection performance, they rely heavily on supplementary landmark annotations, which may not always be available or consistently labeled across different datasets. Furthermore, they do not fully capture the intricate dependencies and co-occurrence patterns among AUs.

To address this issue, Luo et al. [28] introduced a two-stage training method based on a graph neural network (GNN) to model the interdependencies among AUs. By treating AUs as nodes in a graph, their approach aims to leverage the inherent relationships between different facial movements. However, their method represents AU nodes using simple fully connected layers and does not incorporate additional landmark annotations, which limits its capacity to fully exploit spatial dependencies. Additionally, this approach necessitates an initial training phase to effectively learn meaningful node representations before performing AU classification, adding complexity to the overall pipeline. Although these studies have made significant progress in AU detection, they remain constrained by their reliance on additional annotations, specific training paradigms, or limited model expressiveness. These challenges highlight the need for more flexible and generalizable frameworks that can effectively capture both local facial features and high-level semantic relationships among AUs without excessive dependence on external annotations.

3. Method

In this section,we introduce each component that we have attempted in our approach. Our complete system integrates robust audio-visual feature extraction, advanced cross-modal alignment, classification and auxiliary optimization. The pipeline consists of three key modules, described in detail below.

3.1. Data Preprocessing & Feature Extraction

In this subsection, we describe our processing methods from both the audio and visual perspectives separately.

Audio: To extract rich feature representations, we generate both "global" and "local" views from the audio input[36, 46]. The global view is constructed using the full spectrogram, which captures broad temporal and spectral information. Specifically, we use the entire log-Mel spectrogram[30] as input for the global view, preserving the full context of the audio.

For the local view, we focus on extracting enhanced time-frequency crops from the spectrogram[29]. These crops are small, overlapping sections of the Mel spectrogram that capture fine-grained time-frequency patterns. We randomly sample a segment of the spectrogram and apply frequency masking as augmentations. This helps capture localized variations in the audio signal[37], such as specific speech or sound event patterns, which are crucial for distinguishing finer details in the audio stream.

In the audio preprocessing stage, we adopt the efficient approach validated in the Whisper model [35]to convert raw audio signals into high-dimensional feature representations. All raw audio is first resampled to 16,000 Hz and globally normalized, mapping the signal amplitude to the range $[-1, 1]$in order to minimizes the effects of varying recording conditions. The normalized audio signal $x(t)$ is segmented using a 25-millisecond window with a 10-millisecond stride to perform the Short-Time Fourier Trans-

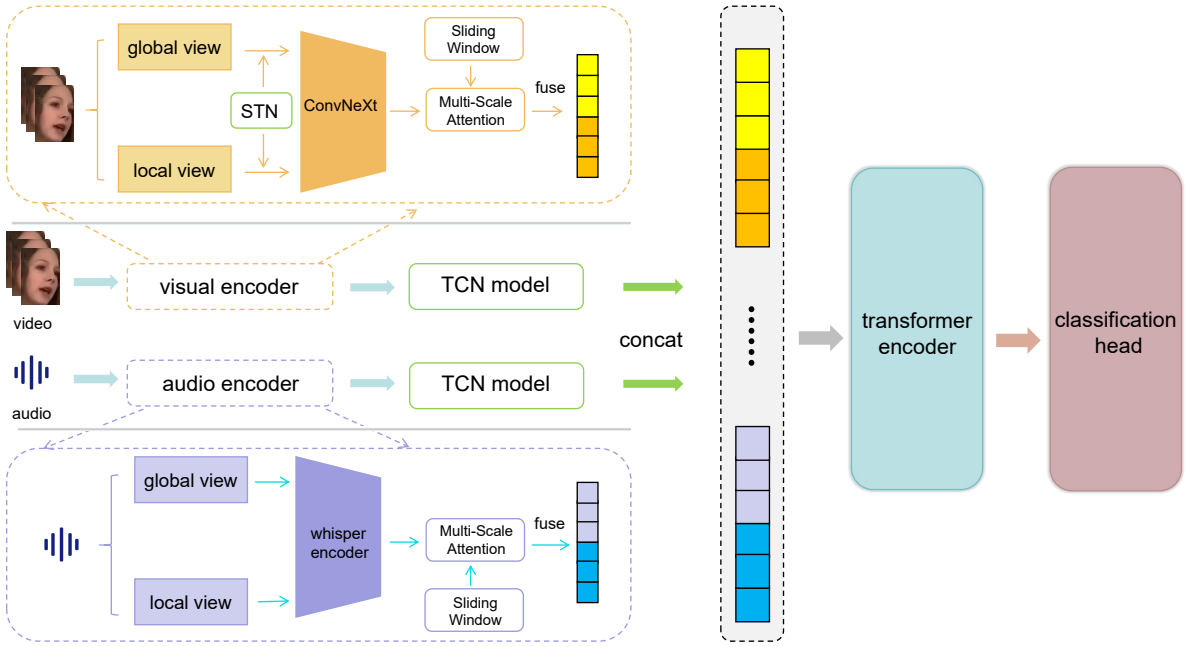

Figure 1. This is our pipeline for AU detection. Integrating audio and video modalities.For each modality, we perform feature extraction and fusion on both global and local views separately. Both modalities features pass through a Temporal Convolutional Network (TCN) to capture temporal dependencies. Their representations are concatenated and fed into a transformer encoder for enhanced feature learning. Finally, a classification head predicts affective states based on the refined multimodal representations.

form (STFT), yielding the spectral representation $S(t,k)$:

$$S(t,k) = \sum_{n=0}^{N-1} x(n) \cdot w(n-t) \cdot e^{-j\frac{2\pi kn}{N}}, \qquad (1)$$

where $w(n)$ denotes the window function and N is the window length. An 80-channel Mel filter bank $\mathbf{M} \in \mathbb{R}^{80\times N}$ is then applied to project the spectrum onto the Mel frequency scale. The log-energy of the projected spectrum is computed as:

$$X(t,f) = \log\left(\sum_k M_{f,k}\,|S(t,k)|^2\right), \quad f = 1,\ldots,80. \qquad (2)$$

This process produces an 80-dimensional log-Mel spectrogram that effectively captures the time-frequency characteristics of speech.

Visual: For the visual stream, we first employ Spatial Transformer Networks (STN)[7] for preprocessing. STN extends the concept of differentiable attention to encompass various spatial transformations. By integrating a differentiable geometric transformation module into the network architecture[27], STN enables neural networks to learn spatial transformations and apply them to input data. This capability is particularly valuable for handling the diverse facial poses and angles present in the Aff-Wild2 dataset[26].

Once preprocessed by STN, the facial image $I \in \mathbb{R}^{H\times W\times 3}$ is fed into the pre-trained ConvNeXt network.

First, the input is processed through the patchify module of ConvNeXt. This module applies non-overlapping convolutions with a kernel size of 4 to downsample the input image. The downsampling operation helps reduce input dimensionality while effectively capturing relevant features. Next, the downsampled inputs pass through the ConvNeXt stem stage. Each stem stage is followed by a Squeeze-and-Excitation (SE) block, which recalibrates the feature maps before proceeding to the next stage. This recalibration enhances the model's ability to extract discriminative facial features, enabling more accurate emotion recognition. Compared to traditional neural networks such as ResNet[5] , this operation effectively reduces the spatial dimensions while preserving local structure.

Following the stem, let the output be denoted as $F \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times C}$, where s is the overall downsampling factor and C is the number of channels.

To obtain a compact, global representation of the face, we apply global average pooling over the spatial dimensions:

$$f = \frac{1}{\frac{H}{s}\cdot\frac{W}{s}} \sum_{i=1}^{\frac{H}{s}} \sum_{j=1}^{\frac{W}{s}} F(i,j). \qquad (3)$$

resulting in a feature vector $f \in \mathbb{R}^C$. This formulation

not only preserves the essential texture and structural information but also provides translation invariance—critical for capturing subtle facial expressions.

In addition to the global view, we generate local views by applying spatio-temporal crops to the video frames. These crops focus on specific temporal segments of the video, allowing us to capture more detailed facial movements or expressions that occur over time. The local views are generated by selecting 1-second temporal segments of the video and performing random augmentations, such as cropping, flipping, or rotating. These enhanced spatio-temporal crops allow the model to focus on particular regions of interest within the video, capturing short-term variations in facial action units that are critical for accurate recognition[1].

Both the global and local views for audio and video are then projected into a shared embedding space. For the audio, the local views are reshaped into smaller time-frequency patches, while the video local views are reshaped into smaller spatio-temporal cuboids. These views are processed through their respective encoders, and the outputs are flattened into vectors, which are then projected onto the embedding space. These embeddings are subsequently fed into the cross-modal alignment and fusion stage, where complementary information from the audio and visual streams will be integrated for facial action unit recognition.

3.2. Advanced Cross-Modal Fusion

In this work, we introduce a cross-modal alignment and fusion module that effectively synchronizes audio and visual features to enhance the accuracy of AU recognition[24]. This module leverages multi-scale self-attention mechanisms and adaptive sliding windows to capture temporal dependencies and fuse complementary information across modalities.

The multi-scale self-attention mechanism operates at different temporal resolutions, enabling the model to focus on both fine-grained and long-term patterns in the audio and video streams. By employing adaptive sliding windows, we account for varying temporal contexts, ensuring that the fusion process dynamically adapts to different events in the sequence. This approach allows the model to capture subtle variations in facial expressions while maintaining a coherent understanding of the overall context, thus improving robustness.

The fusion process can be mathematically expressed as follows:

$$F_{\text{fusion}}(t) = \sum_{i=1}^{N} \left(\alpha_i \cdot A_i(t) + \beta_i \cdot V_i(t)\right). \tag{4}$$

Where $F_{\text{fusion}}(t)$ represents the fused feature at time t. $A_i(t)$ and $V_i(t)$ are the audio and visual features at scale i. α_i and β_i are the learned weights that control the importance of each modality at scale i. N is the number of scales considered in the multi-scale attention process.

3.3. Temporal Modeling & Classification Optimization

Temporal Modeling is implemented using a Temporal Convolutional Network (TCN)[23] to capture sequential patterns in our audio-visual features. We input the high-dimensional embeddings from the cross-modal fusion module into a TCN that employs several layers of 1D convolutions with increasing dilation factors. This design captures both local and long-range dependencies while ensuring causality, as each convolution uses only current and past inputs. By directly leveraging TCN for temporal modeling, we bypass the limitations of traditional recurrent networks and simplify the overall architecture. Residual connections further help maintain information flow and ease training.

Following temporal feature extraction, the output is fed into a multilayer perceptron (MLP) [32]classifier. The MLP consists of two fully connected layers: the first with 512 units and ReLU activation, followed by dropout to mitigate overfitting, and the final layer applies softmax activation to produce action unit predictions. We optimize the entire network using cross-entropy loss on the Aff-Wild2[10–17, 19, 45] dataset. This approach provides a streamlined yet effective framework for modeling temporal dynamics and enhancing the accuracy and robustness of AU recognition in complex video data.

4. Experiments

4.1. Dataset

In this study, we use the Aff-Wild2[10–17, 19, 45] dataset, a large-scale extension of the original Aff-Wild dataset designed for emotion recognition and action unit (AU) detection in spontaneous facial expressions. Aff-Wild2 consists of 564 videos, totaling approximately 2.8 million frames, with annotations for seven basic expressions, twelve AUs, and continuous values for valence and arousal. The dataset includes 554 participants from diverse demographic backgrounds, including variations in age, ethnicity, and nationality, representing a wide range of emotional and environmental contexts. The annotations are frame-level, making it unique in that it provides both audio and video annotations, enabling cross-modal emotion recognition. Aff-Wild2 is also notable for its real-world complexity, capturing spontaneous emotional expressions in natural settings, which is crucial for training models that generalize well in practical applications.

Overall, the Aff-Wild2 dataset provides a comprehensive resource for emotion recognition and AU detection tasks, making it an invaluable asset for advancing research in these

Experimental Combination	STN	TCN	Sliding Window	F1 Score (%)
offical method				36.50
Baseline(ConvNeXt+Whisper)				46.05
+STN	✓			51.81
+TCN		✓		49.19
+Sliding Window			✓	49.70
+STN + TCN	✓	✓		54.94
+STN + Sliding Window	✓		✓	55.03
+TCN + Sliding Window		✓	✓	53.88
+STN +TCN + Sliding Window	✓	✓	✓	56.11

Table 1. Ablation study results on AU detection performance (F1 score %). The baseline model (ConvNeXt + Whisper) is incrementally enhanced with STN, TCN, and Sliding Window. The full model achieves the highest F1 score of 56.11% on validation set, demonstrating the effectiveness of each module.

fields.

4.2. Implementation details

Evaluation Metrics The F1 score serves as the primary evaluation metric for action unit (AU) recognition[3], balancing precision and recall to provide a comprehensive assessment of model performance. The overall performance measure P is defined as the average F1 score across all 12 AU categories:

$$P = \frac{1}{12}\sum_{i=1}^{12} F1_i \tag{5}$$

where $F1_i$ ($i \in \{1, 2, \cdots, 12\}$) denotes the F1 score for each action unit, computed as:

$$F1_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{6}$$

Here, Precision measures the proportion of correctly predicted positive samples among all predicted positives, while Recall represents the proportion of correctly predicted positive samples among all actual positives. The F1 score provides a harmonic mean of these two metrics, ensuring a balanced evaluation of both false positives and false negatives. By averaging the F1 scores across all 12 AU categories, the performance measure P offers a holistic assessment of the model's AU recognition capability.

4.3. Training Strategy Optimization

Sharpness-Aware Minimization (SAM): To enhance the generalization ability of our AU detection model, we incorporate Sharpness-Aware Minimization (SAM), which optimizes both the loss function and the loss landscape smoothness to mitigate overfitting. SAM achieves this by solving:

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} L(\theta + \epsilon) \tag{7}$$

where ρ controls the perturbation magnitude. The optimization consists of two steps: (1) an ascent step to find ϵ^* that perturbs the parameters to maximize the loss, and (2) a descent step that updates θ to minimize the worst-case perturbed loss. This method effectively prevents convergence to sharp minima, leading to improved generalization.

Stochastic Weight Averaging (SWA): To further stabilize the training process, we employ Stochastic Weight Averaging (SWA) during the later stages of model optimization. Instead of relying on a single model checkpoint, SWA maintains a running average of model weights across multiple iterations. By averaging multiple weight updates, SWA effectively smooths the loss landscape and improves the generalization ability of the model. In our framework, SWA is applied to the Transformer Encoder, ensuring that the learned representations remain robust and less sensitive to fluctuations during training. This strategy helps mitigate overfitting and enhances the model's performance on AU recognition tasks.

$$\theta_{\text{SWA}} = \frac{1}{K}\sum_{i=1}^{K} \theta_i \tag{8}$$

where K denotes the number of model checkpoints. SWA encourages convergence to flatter minima in the loss landscape, reducing test error and improving robustness. This approach is particularly beneficial when combined with adaptive learning rate schedules and plays a crucial role in optimizing deep neural networks for AU detection.

4.4. Experimental Results

For the evaluation of AU recognition, we measure model performance using the F1 score for each action unit.To evaluate the performance of our method for AU recognition, we measure model effectiveness using the F1 score for each action unit. The ablation study results, shown in Table 1, demonstrate the contribution of different compo-

nents, including Spatial Transformer Network (STN), Temporal Convolutional Network (TCN), and Sliding Window Mechanism, to overall performance. These results confirm that each component plays a crucial role in improving AU detection, the combination of all three modules leads to a robust and high-performing AU detection framework in the wild.

To further validate the effectiveness of our approach, we participated in the 8th Competition on Affective & Behavior Analysis in-the-Wild (ABAW8), specifically in the AU Detection track. The competition provides a challenging benchmark for AU detection, with the Aff-Wild2 dataset serving as the evaluation platform. Our method achieved an F1 score of 0.515 on the official test set and 0.561 on the validation set, significantly surpassing existing approaches and demonstrating strong robustness in real-world scenarios. This outstanding performance earned us first place in the AU Detection track, highlighting the advantages of our proposed multi-modal feature extraction, enhancement, and fusion strategies. The results confirm that incorporating both local and global representations, along with advanced cross-modal alignment techniques, effectively improves AU recognition accuracy. This achievement underscores the potential of our approach for broader applications in affective computing and intelligent perception systems.

5. Conclusion

In this study, we proposed a novel multimodal audio-visual feature extraction and fusion framework to enhance Facial Action Unit (AU) detection performance in complex environments. We design a multimodal domain alignment strategy that performs dual-modal feature extraction on both global and local views while leveraging a multiscale self-attention mechanism and an adaptive sliding window to capture contextual relationships. Additionally, we introduce a Temporal Convolutional Network (TCN) for temporal modeling, which ensures modality alignment while enhancing spatio-temporal understanding.

Experimental results on the Aff-Wild2 dataset demonstrate that the proposed method significantly outperforms existing approaches in AU detection, validating its effectiveness in emotion recognition tasks. Our method achieved first place in the AU detection track of the 8th Competition on Affective & Behavior Analysis in-the-wild, attaining a score of 0.515 on the official test set. In the future, we will further explore more advanced cross-modal fusion mechanisms and improve the model's generalization ability in more complex environments, promoting the widespread application of AU detection technology in intelligent perception systems.

Acknowledgement

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), Dreams Foundation of Jianghuai Advance Technology Center (2023-ZM01Z001).

References

[1] Mouath Aouayeb, Catherine Soladie, Wassim Hamidouche, Kidiyo Kpalma, and Renaud Seguier. Spatiotemporal features fusion from local facial regions for micro-expressions recognition. Frontiers in Signal Processing, 2:861469, 2022. 4

[2] Carlos Fabian Benitez-Quiroz, Yan Wang, Aleix M Martinez, et al. Recognition of action units in the wild with deep nets and a new global-local loss. In ICCV, pages 3990–3999, 2017. 2

[3] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics, 21:1–13, 2020. 5

[4] H. Onan Demirel, Salman Ahmed, and Vincent G. Duffy. Digital human modeling: A review and reappraisal of origins, present, and expected future methods for representing humans computationally. International Journal of Human–Computer Interaction, 38(10):897–937, 2022. 1

[5] 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. 3

[6] 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

[7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. Advances in neural information processing systems, 28, 2015. 1, 3

[8] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2328–2336, 2022. 1

[9] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In European Conference on Computer Vision, pages 157–172. Springer, 2023. 1

[10] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855, 2019. 1, 4

[11] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792, 2021.

[12] 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.

[13] 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, .

[14] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Irene Kotsia, UK Cogitat, Eric Granger, Marco Pedersoli, Simon Bacon, Alice Baird, Chunchang Shao, et al. Advancements in affective and behavior analysis: The 8th abaw workshop and competition. .

[15] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. arXiv preprint arXiv:1910.11111, 2019.

[16] 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, pages 1–23, 2019.

[17] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. arXiv preprint arXiv:2105.03790, 2021. 1, 4

[18] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5888–5897, 2023. 1

[19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 2813–2821, 2024. 1, 4

[20] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4587–4598, 2024. 1

[21] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Abhinav Dhall, Shreya Ghosh, Chunchang Shao, and Guanyu Hu. 7th abaw competition: Multi-task learning and compound expression recognition. arXiv preprint arXiv:2407.03835, 2024.

[22] Dimitrios Kollias, Panagiotis Tzirakis, Alan S. Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon L. Bacon, Alice Baird, Chris Gagne, Chunchang Shao, Guanyu Hu, Soufiane Belharbi, and Muhammad Haseeb Aslam. Advancements in Affective and Behavior Analysis: The 8th ABAW Workshop and Competition. 2025. 1

[23] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 156–165, 2017. 4

[24] Sumin Lee, Sangmin Woo, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Modality mixer for multimodal action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3298–3307, 2023. 4

[25] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6766–6775, 2017. 1

[26] Chang Liu, Kaoru Hirota, and Yaping Dai. Patch attention convolutional vision transformer for facial expression recognition with occlusion. Information Sciences, 619:781–794, 2023. 3

[27] Cristina Luna-Jiménez, Jorge Cristóbal-Martín, Ricardo Kleinlein, Manuel Gil-Martín, José M Moya, and Fernando Fernández-Martínez. Guided spatial transformers for facial expression recognition. Applied Sciences, 11(16):7217, 2021. 3

[28] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. arXiv preprint arXiv:2205.01782, 2022. 2

[29] Ian McLoughlin, Zhipeng Xie, Yan Song, Huy Phan, and Ramaswamy Palaniappan. Time–frequency feature fusion for noise robust audio event classification. Circuits, Systems, and Signal Processing, 39(3):1672–1687, 2020. 2

[30] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. IEEE access, 7:125868–125881, 2019. 2

[31] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pages 11917–11926, 2019. 2

[32] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, 8(7):579–588, 2009. 4

[33] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In 2016 IEEE 16th international conference on data mining (ICDM), pages 439–448. IEEE, 2016. 1

[34] Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. Facial action coding system. Environmental psychology & nonverbal behavior, 1, 2015. 1

[35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In International conference on machine learning, pages 28492–28518. PMLR, 2023. 2

[36] Pritam Sarkar and Ali Etemad. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 14875–14885, 2024. 2

[37] Christopher Schymura, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, and Dorothea Kolossa. Exploiting attention-based sequence-to-sequence architectures for sound event localization. In 2020 28th European Signal Processing Conference (EUSIPCO), pages 231–235. IEEE, 2021. 2

[38] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In Proceedings of the European conference on computer vision (ECCV), pages 705–720, 2018. 2

[39] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. IEEE transactions on affective computing, 13(3):1274–1289, 2019. 2

[40] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12899–12908, 2021. 2

[41] Ashish Tawari and Mohan Manubhai Trivedi. Face expression recognition by cross modal data association. IEEE Transactions on Multimedia, 15(7):1543–1552, 2013. 1

[42] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3405–3414, 2017. 1

[43] Zihan Wang, Siyang Song, Cheng Luo, Songhe Deng, Weicheng Xie, and Linlin Shen. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1270–1280, 2024. 2

[44] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10482–10491, 2021. 2

[45] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 1980–1987. IEEE, 2017. 1, 4

[46] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. IEEE transactions on pattern analysis and machine intelligence, 44(5):2402–2415, 2020. 2

[47] Ligang Zhang, Brijesh Verma, Dian Tjondronegoro, and Vinod Chandran. Facial expression analysis under partial occlusion: A survey. ACM Computing Surveys (CSUR), 51 (2):1–49, 2018. 1

[48] Xiang Zhang and Lijun Yin. Multi-modal learning for au detection based on multi-head fused transformers. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pages 1–8. IEEE, 2021. 2

[49] Hao Zheng, Ruili Wang, Wanting Ji, Ming Zong, Wai Keung Wong, Zhihui Lai, and Hexin Lv. Discriminative deep multi-task learning for facial expression recognition. Information Sciences, 533:60–71, 2020. 1