LUT University

AU-TTT: Vision Test-Time Training model for Facial Action Unit Detection

Xing Bohao, Yuan Kaishen, Yu Zitong, Liu Xin, Kälviäinen Heikki

This is a Author's accepted manuscript (AAM) version of a publication

published by IEEE

in 2025 IEEE International Conference on Multimedia and Expo (ICME)

DOI: 10.1109/ICME59968.2025.11209184

Copyright of the original publication:

© IEEE 2025

Please cite the publication as follows:

Xing, B., Yuan, K., Yu, Z., Liu, X., Kälviäinen, H. (2025). AU-TTT: Vision Test-Time Training model for Facial Action Unit Detection. In: 2025 IEEE International Conference on Multimedia and Expo (ICME), Nantes, France, 2025. pp. 1-6. DOI: 10.1109/ICME59968.2025.11209184

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This is a parallel published version of an original publication. This version can differ from the original published article.

AU-TTT: Vision Test-Time Training model for Facial Action Unit Detection

Bohao Xing[1,*], Kaishen Yuan[2,*], Zitong Yu[3], Xin Liu[1,†], Heikki Kälviäinen[1,4,5]
[1]Lappeenranta-Lahti University of Technology LUT, Finland
[2]The Hong Kong University of Science and Technology (Guangzhou), China [3]Great Bay University, China
[4]Rensselaer Polytechnic Institute, USA [5]Brno University of Technology, Czech Republic
{bohao.xing, xin.liu, heikki.kalviainen}@lut.fi,
yuankaishen01@gmail.com, zitong.yu@ieee.org

***Abstract*—Facial Action Units (AUs) detection is a cornerstone of objective facial expression analysis and a critical focus in affective computing. Despite its importance, AU detection faces significant challenges, such as the high cost of AU annotation and the limited availability of datasets. These constraints often lead to overfitting in existing methods, resulting in substantial performance degradation when applied across diverse datasets. Addressing these issues is essential for improving the reliability and generalizability of AU detection methods. Moreover, many current approaches leverage Transformers for their effectiveness in long-context modeling, but they are hindered by the quadratic complexity of self-attention. Recently, Test-Time Training (TTT) layers have emerged as a promising solution for long-sequence modeling. Additionally, TTT applies self-supervised learning for iterative updates during both training and inference, offering a potential pathway to mitigate the generalization challenges inherent in AU detection tasks. In this paper, we propose a novel vision backbone tailored for AU detection, incorporating bidirectional TTT blocks, named AU-TTT. Our approach introduces TTT Linear to the AU detection task and optimizes image scanning mechanisms for enhanced performance. Additionally, we design an AU-specific Region of Interest (RoI) scanning mechanism to capture fine-grained facial features critical for AU detection. Experimental results demonstrate that our method achieves competitive performance in both within-domain and cross-domain scenarios.**

***Index Terms*—Facial AU detection, test-time training.**

I. Introduction

Facial action units, as defined by the Facial Action Coding System, represent specific movements or deformations of facial muscles, providing a framework for analyzing complex facial expressions [1]–[5]. Consequently, AU detection has become a highly active area of research in recent years, spurring the development of numerous advanced methods. Traditionally, most approaches have relied on Convolutional Neural Networks (CNNs) [6] or Graph Neural Networks (GNNs) [7], which limit the model's focus to localized facial regions. Recently, however, innovative methods [8]–[14] have introduced Transformers [15], leveraging their exceptional capability to model long-range dependencies and extract richer global context information from facial data.

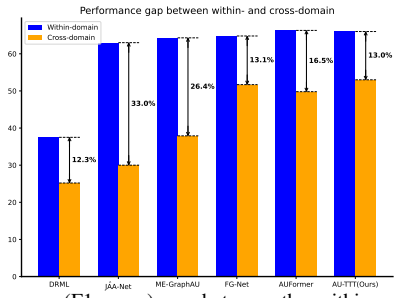

Fig. 1. Performance (F1 score) gap between the within- and cross-domain AU detection for DRML [16], JÂA-Net [17], MEGraphAU [10], FG-Net [18], AUFormer [13], and AU-TTT(Ours). The within-domain performance is averaged between DISFA and BP4D, while the cross-domain performance is averaged between BP4D to DISFA and DISFA to BP4D.

Specifically, early AU detection methods relied heavily on hand-crafted features, which struggled to capture the complexity and subtlety of facial muscle activations, resulting in limited performance. The advent of deep learning, particularly CNNs, marked a significant breakthrough by enabling automatic feature extraction. Methods such as JÂA-Net [17] employed multi-scale hierarchical designs to capture features at varying resolutions, thereby improving the detection of AUs across diverse scales of muscle activation.

To address the intricate relationships among AUs, researchers explored GNNs [10], [19], which effectively model the inherent dependencies between AUs using graph-based structures. For example, ME-GraphAU [10] utilized a graph with multi-dimensional edge features and Gated Graph Convolutional Networks [20] to capture correlations between AU pairs, leading to enhanced detection accuracy. However, despite their effectiveness, GNN-based methods primarily emphasize local features and fail to fully exploit the global dependencies present in the data.

The recent integration of Transformers into AU detection has significantly enhanced the field by enabling the modeling of long-range dependencies [8], [9], [11]–[14]. For instance, methods like FAUDT [8] leveraged Transformer blocks to

This work is supported in part by the China Scholarship Council (Grant No. 202406250043), and the Finnish IT Center for Science is gratefully acknowledged.

*Equal contributions.

†Corresponding author.

capture global correlations between AUs, thereby providing richer contextual information for more accurate detection. However, despite these advancements, the substantial number of learnable parameters in Transformer-based models poses a challenge, increasing the risk of overfitting, particularly when training on datasets with limited AU annotations. Furthermore, the quadratic complexity of the self-attention mechanism significantly increases computational costs.

In addition, manual annotations for AUs are cumbersome and costly, as they require trained coders to label each frame individually. Common AU datasets, i.e., DISFA [22] and BP4D [23], only contain a limited number of subjects (27 and 41 subjects respectively). Current AU detection methods are predominantly evaluated using within-domain validation, where both training and testing data are sourced from the same dataset. While these methods often demonstrate strong within-domain performance, such results can obscure overfitting as shown in Figure 1, as their ability to generalize to unseen domains remains insufficiently explored. Cross-domain evaluations frequently reveal significant performance degradation when models are applied to datasets with variations in camera settings, environmental conditions, and subject demographics [24], [25].

Although some studies [13], [18], [24], [25] have made efforts to improve cross-domain generalization, their scope remains limited. Existing approaches to enhance generalization include identity-aware training on diverse datasets, as demonstrated by IdenNet [26], and domain adaptation techniques [27], but these methods typically require access to the target domain. DeepFN [28] addresses this issue by normalizing facial expressions onto a common template, but significant challenges persist in achieving robust cross-domain performance. AUFormer [13] leverages large-scale pretraining [29] and parameter-efficient transfer learning to improve cross-domain robustness but does not incorporate targeted strategies to address domain shifts. Similarly, FG-Net [18] utilizes generative model features to mitigate cross-domain challenges. Moreover, both methods rely on pretrained weights obtained from extensive additional data.

Apart from recent architectures such as Transformer and Mamba [30], [31], the concept of learning at test time has a long history in machine learning [32], [33]. More recently, Test-Time Training (TTT) extended this idea to Recurrent Neural Networks (RNNs) [34] and has emerged as an effective strategy for addressing out-of-distribution (OOD) problem [21], [35], [36]. The core idea of TTT is that each test instance defines its own learning problem, where the test instance itself serves as the target for generalization. In traditional settings, a predictor f is trained on all training instances and used to directly predict $f(x)$ for a given test instance x . In contrast, TTT formulates a specific learning problem centered on x. It updates a model f_x using x (with f as its initialization) and then predicts $f_x(x)$. Since the test instance lacks labels, the learning problem is defined using a self-supervised task. Previous studies demonstrated that TTT with reconstruction tasks significantly enhances performance, particularly for outlier instances [35].

In this paper, we introduce AU-TTT, a novel facial action unit detection method designed to enhance cross-domain generalization. AU-TTT leverages Test-Time Training to improve the model's generalization ability. Additionally, the integration of bidirectional scanning and AU RoI scanning enables AU-TTT to effectively capture both general image information and fine-grained AU-related facial features, further enhancing its robustness and performance.

Our main contributions can be summarized as follows:

- We propose AU-TTT, a novel framework that integrates Test-Time Training into facial AU detection to address the challenge of out-of-distribution generalization.
- We optimize the mini-batch strategy of the original TTT and propose a bidirectional scanning mechanism for image tasks.
- We design AU RoI scanning mechanisms within AU-TTT tailored to the characteristics of AUs.
- Our method demonstrates competitive performance in both within-domain and cross-domain scenarios.

II. AU-TTT

In this section, we first outline the fundamentals of the TTT Layer [21]. We then present AU-TTT, our proposed adaptation of the Test-Time Training framework, specifically designed for facial AU detection.

A. Preliminaries

We briefly introduce the original TTT Block and TTT Layer proposed in [21], as illustrated in Figure 2. The TTT Block replaces the self-attention in the basic building block of Transformers with the TTT Layer. The TTT Layer leverages self-supervised reconstruction to encode information into the model's weights. This process is demonstrated using input sequences $X = [x_t], t \in \{1, \ldots, N\}$. It begins with an initial hidden state W_0, which is iteratively refined through self-supervised learning. For each input token x_t, the model computes an output as follows

$$z_t = f(x_t; W_t), \tag{1}$$

where W_t represents the current hidden state, and z_t is the updated embedding. The hidden state W_t is refined using a gradient descent step on a self-supervised loss ℓ, which minimizes the difference between the model's predicted embedding and the actual embedding. The update rule, illustrated in Figure 2 (C), can also be expressed as

$$W_t = W_{t-1} - \eta \nabla \ell(W_{t-1}; x_t), \tag{2}$$

where η denotes the learning rate. The TTT Layer further incorporates a multi-view reconstruction mechanism to enhance the optimization of the reconstruction process. The updated self-supervised loss is described as

$$\ell(W; x_t) = \left\| f(\theta_K x_t; W) - \theta_V x_t \right\|^2, \tag{3}$$

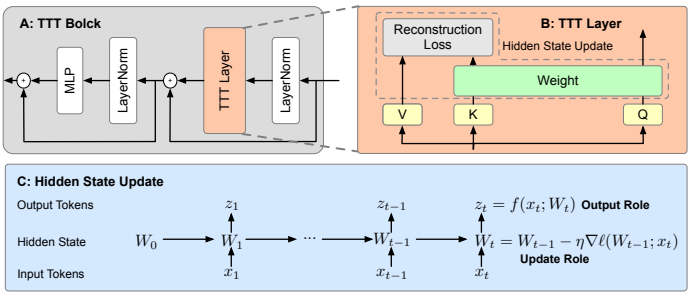

Fig. 2. **A**: Original TTT Block. The basic building block of Transformers, originally based on self-attention, is replaced with the TTT Layer in the Transformer backbone. **B**: Original TTT Layer, driven by self-supervised loss, updates the weights adaptively. **C**: Hidden state update rule. The key idea of TTT is to make the hidden state itself a model f with weights W , and the update rule a gradient step on the self-supervised loss ℓ. Therefore, updating the hidden state on a test sequence is equivalent to training the model f at test time [21].

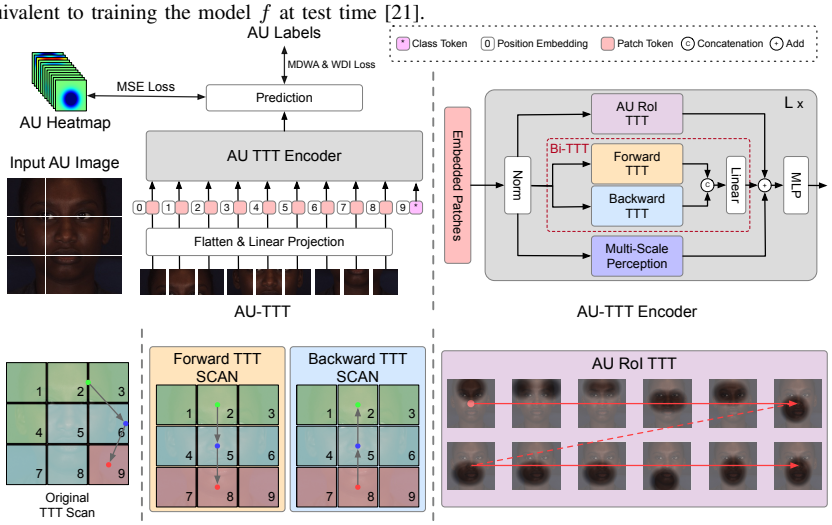

Fig. 3. **Top Left**: The overall AU-TTT framework. **Top Right**: The architecture of the AU-TTT Block. **Bottom Left**: The original TTT Layer scanning method. **Bottom Center**: The bidirectional TTT scanning method. **Bottom Right**: The AU RoI TTT scanning method.

where θ_K and θ_V are the projections for the training view and label view, respectively. The corresponding output rule is as follows

$$z_t = f\left(\theta_Q x_t; W_t\right), \tag{4}$$

where θ_Q is the projection for the test view.

B. AU-TTT

The original TTT Layer was designed for NLP tasks and is not directly applicable to image-based AU detection. To address this limitation, we made several improvements in this subsection. The first issue to consider is that the mini-batch size within the original TTT Layer is suboptimal for CV tasks. Through experimentation, we discovered that adjusting the mini-batch size based on the number of tokens corresponding to the image's feature rows significantly improves performance. We refer to this modified version of TTT as forward TTT. Next, we incorporate bidirectional scan TTT (Bi-TTT) into the AU-TTT block to mitigate the impact of sequential modeling on image tasks. Additionally, we introduce an AU RoI scanning mechanism (AU RoI TTT) specifically designed for AU detection tasks.

An overview of the proposed AU-TTT framework is shown in Figure 3. The standard TTT Layer is originally designed for 1D sequences. To adapt it for vision tasks, we first transform the input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$ into flattened patches $\mathbf{X} \in \mathbb{R}^{J\times(P^2\cdot C)}$, where (H, W) denotes the height and width of the image, C represents the number of channels, and P is the size of each image patch. Next, we linearly project $\mathbf{X}$ into a vector of size D and add learnable positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(J+1)\times D}$, as follows

$$\mathbf{Z}_0 = \left[X^1\mathbf{W}; X^2\mathbf{W}; \cdots; X^J\mathbf{W}; \text{CLS}\right] + \mathbf{E}_{\text{pos}}, \tag{5}$$

where X^j represents the j-th patch of $\mathbf{X}$, $\mathbf{W} \in \mathbb{R}^{(P^2\cdot C)\times D}$ is the learnable projection. Inspired by ViT [15], we also incorporate a class token (CLS) to represent the entire patch sequence. The token sequence $\mathbf{Z}_{l-1}$ is then passed through the l-th layer of the AU-TTT encoder, producing the output $\mathbf{Z}_l$.

Specifically, the input token sequence Z_{l-1} is first normalized using a normalization layer. Then, the token sequence is processed in both forward and backward directions. The bidirectional TTT process is formulated as follows (for simplicity, the layer l notation is omitted in the equations below):

$$\begin{aligned}\text{Bi-TTT}(\mathbf{Z}) &= \text{Bi-TTT}([\mathbf{z}^1; \mathbf{z}^2; \ldots; \mathbf{z}^J; \text{CLS}]) \\ &= \text{Linear}(\text{Cat}[\text{TTT}([\mathbf{z}^1; \mathbf{z}^2; \ldots; \mathbf{z}^J; \text{CLS}]); \\ &\qquad \text{Reverse}(\text{TTT}([\mathbf{z}^J; \mathbf{z}^{J-1}; \ldots; \mathbf{z}^1; \text{CLS}]))]).\end{aligned} \tag{6}$$

where $\mathbf{Z}$ represents the input sequence, $\mathbf{z}^j$ denotes the j-th token, and CLS is the class token. The operation Reverse$(\cdot)$ reverses the sequence order, excluding the CLS token. TTT

represents the foward TTT Layer, Cat$[\cdot;\cdot]$ denotes the concatenation of results along the channel dimension, and Linear is a linear projection used to reduce the dimension.

Additionally, we leverage facial landmarks to generate a heatmap for extracting local features of AU RoI regions [18]. Specifically, we first perform an element-wise multiplication between the facial features and the generated mask:

$$\mathbf{Z}_{\text{AU}} = \mathbf{Z} \cdot \mathbf{Mask}, \tag{7}$$

where $\mathbf{Z} \in \mathbb{R}^{H' \times W' \times D}$ and $\mathbf{Mask} \in \mathbb{R}^{\text{N}_{\text{AU}} \times H' \times W'}$. Here, H' and W' represent the height and width of the features, respectively. N_{AU} represents the number of AUs in the dataset. Through broadcasting, the element-wise multiplication results in $\mathbf{Z}_{\text{AU}} \in \mathbb{R}^{\text{N}_{\text{AU}} \times H' \times W' \times D}$. Next, we compress $\mathbf{Z}_{\text{AU}}$ into tokens $\mathbf{Z}'_{\text{AU}} \in \mathbb{R}^{\text{N}_{\text{AU}} \times D}$ using mean pooling, which are then modeled using the TTT layer to capture the relationships between different AUs.

At the same time, we employ a set of dilated convolutions with varying dilation rates as a Multi-Scale Perception (MSP) module to extract multi-scale facial features. Finally, the outputs of the AU RoI TTT, bidirectional TTT, and MSP modules are combined through addition and passed through a MLP to produce the output token sequence $\mathbf{Z}_l$.

C. Loss Function

AU detection is typically formulated as a multi-label binary classification problem. To generate the final predictions, we transform the CLS token and the heatmap obtained from the last block of AU-TTT. For the classification loss, we adopt the Margin-Truncated Difficulty-Aware Weighted Asymmetric Loss (MDWA-Loss) proposed by AUFormer [13]. This loss function emphasizes activated AUs, adaptively evaluates the difficulty of unactivated AUs, and discards potentially mislabeled samples. The MDWA-Loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{\text{MDWA}} = -\frac{1}{N}\sum_{i=1}^{N} \omega_i [y_i \log(p_i) \\ + (1-y_i) p_{m,i}^{\gamma_i} \log(1-p_{m,i})],\end{aligned} \tag{8}$$

where y_i and p_i represent the ground truth and predictions for the i-th AU, respectively. ω_i is the weight for addressing class imbalance, and it can be formulated as $\omega_i = N(1/r_i)/\sum_{j=1}^{N}(1/r_j)$, where r_i is the occurrence rate of the i-th AU. $p_{m,i}$ is used to discard potentially mislabeled samples and can be formulated as $p_{m,i} = \max(p_i - m, 0)$, where $m \in [0, 1]$ is the truncation margin. γ_i is used to distinguish the different difficulty levels of unactivated AUs and can be formulated as $\gamma_i = B_L + (B_R - B_L) \times r_i$, where B_L and B_R are left and right boundaries of γ_i respectively.

Additionally, referring to [17], we also introduce a weighted multi-label dice loss (WDI-Loss) to alleviate the issue of AU prediction bias towards non-occurrence, as follow:

$$\mathcal{L}_{\text{WDI}} = \frac{1}{N}\sum_{i=1}^{N} \omega_i \left(1 - \frac{2 y_i p_i + \epsilon}{y_i^2 + p_i^2 + \epsilon}\right), \tag{9}$$

TABLE I
AVERAGE F1-SCORE (IN %) RESULTS FOR WITHIN-DOMAIN EVALUATIONS ON BP4D AND DISFA. THE BEST AND SECOND-BEST RESULTS FOR EACH COLUMN ARE BOLDED AND UNDERLINED, RESPECTIVELY.

Method	Venue	BP4D	DISFA
DRML [16]	CVPR 16	48.3	26.7
LP-Net [19]	CVPR 19	61.0	56.9
SRERL [37]	AAAI 19	62.9	55.9
JÂA-Net [17]	IJCV 21	62.4	63.5
HMP-PS [38]	CVPR 21	63.4	61.0
SEV-Net [39]	CVPR 21	63.9	58.8
FAUDT [8]	CVPR 21	64.2	61.5
PIAP [40]	ICCV 21	64.1	63.8
KDSRL [41]	CVPR 22	64.5	64.5
ME-GraphAU [10]	IJCAI 22	65.5	62.4
BG-AU [42]	CVPR 23	64.1	58.2
KS [43]	ICCV 23	64.7	62.8
FG-Net [18]	WACV 24	64.3	65.4
AUFormer [13]	ECCV 24	**66.2**	**66.4**
AU-TTT (Ours)	-	<u>65.6</u>	**66.4**

where ϵ is a smooth term and the meanings of y_i, p_i, and ω_i remain consistent with those defined in Equation (8).

In addition, we utilize Mean Squared Error (MSE) loss to enhance the network's capability in capturing AU RoI location-related information which can be expressed as:

$$\mathcal{L}_{MSE} = \|m - \hat{m}\|_2^2, \tag{10}$$

where m represents the ground-truth heatmap and $\hat{m}$ represents the predicted heatmap.

The overall loss of AUFormer can be expressed as:

$$\mathcal{L} = \lambda_{\text{MDWA}} \cdot \mathcal{L}_{\text{MDWA}} + \lambda_{\text{WDI}} \cdot \mathcal{L}_{\text{WDI}} + \lambda_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}}, \tag{11}$$

where λ_{MDWA}, λ_{WDI}, and λ_{MSE} are hyperparameters.

III. EXPERIMENTS

A. Settings

Datasets. We select two publicly available datasets, DISFA [22] and BP4D [23], which are commonly used for AU detection. These datasets were collected from different subjects under varying backgrounds and lighting conditions. Follow [6], [10], [13], [17], [18], we employ subject-exclusive 3-fold cross-validation.

Implementation Details. We pretrain AU-TTT on ImageNet-1K [44], following the configuration of ViT-S/16 [15]. Specifically, we set the patch size to 16, the embedding dimension to 384, the depth to 12, and the number of heads to 6. The dilated rates r_1, r_2, and r_3 in Multi-Scale Perception module are set to 1, 3, and 5. The left and right boundaries B_L and B_R of γ are set to 1 and 2. The truncation margin m for BP4D and DISFA are set to 0.1 and 0.15. The smooth term ϵ is set to 1. The number of AUs N_{AU} for DISFA and BP4D is set to 8 and 12, following [16].

Evaluation Metric. Following prior works [13], [16]–[18], [43], we use the F1-score as the evaluation metric for AU detection performance. The F1-score represents the harmonic mean of precision and recall.

TABLE II

F1-SCORES (IN %) FOR CROSS-DOMAIN EVALUATIONS BETWEEN BP4D AND DISFA. THE BEST AND SECOND BEST RESULTS FOR EACH COLUMN ARE BOLDED AND UNDERLINED, RESPECTIVELY. *THE NUMBERS ARE DERIVED FROM A REPLICATION OF [18] BASED ON OPEN-SOURCE CODE.

Direction	BP4D → DISFA						DISFA → BP4D					
AU	1	2	4	6	12	**Avg.**	1	2	4	6	12	**Avg.**
DRML* [16]	10.4	7.0	16.9	14.4	22.0	14.1	19.4	16.9	22.4	58.0	64.5	36.3
JÂA-Net* [17]	12.5	13.2	27.6	19.2	46.7	23.8	10.9	6.7	42.4	52.9	68.3	36.2
ME-GraphAU* [10]	43.3	22.5	41.7	23.0	34.9	33.1	36.5	30.3	35.8	48.8	62.2	42.7
Patch-MCD [27]	34.3	16.6	<u>52.1</u>	33.5	50.4	37.4	-	-	-	-	-	-
BG-AU [42]	-	-	-	-	-	38.8	-	-	-	-	-	-
IdenNet [26]	20.1	25.5	37.3	**49.6**	**66.1**	39.7	-	-	-	-	-	-
FG-Net [18]	**61.3**	**70.5**	36.3	42.2	<u>61.5</u>	**54.4**	**51.4**	**46.0**	36.0	49.6	61.8	49.0
AUFormer [13]	<u>49.7</u>	<u>49.9</u>	43.2	31.5	43.5	43.6	<u>49.8</u>	<u>43.8</u>	**45.7**	<u>66.6</u>	<u>74.5</u>	<u>56.1</u>
AU-TTT (Ours)	46.0	44.6	**56.7**	<u>44.4</u>	52.2	<u>48.7</u>	49.4	39.5	<u>46.0</u>	**73.2**	**77.9**	**57.2**

B. Experimental Results

The models are evaluated for both within-domain and cross-domain performance. Cross-domain evaluation is particularly critical for assessing the generalization capability of our AU detection approach.

Within-domain Evaluation. We conduct within-domain evaluations on the widely used DISFA and BP4D datasets. Table I reports the average F1-score results of various methods, demonstrating that AU-TTT achieves highly competitive performance on both BP4D and DISFA. Compared to previous CNN-based approaches and most Transformer-based methods, our approach not only delivers superior performance but also features linear complexity, effectively mitigating the quadratic complexity issue inherent to self-attention.

Cross-domain Evaluation. To evaluate the generalization capabilities of AU-TTT, we conduct bidirectional cross-domain evaluations, specifically from BP4D to DISFA and from DISFA to BP4D. Following the approach of FG-Net [18], we use two folds of data from the source domain as the training set, while the entire dataset from the target domain serves as the testing set. As shown in Table II, AU-TTT outperforms all methods in the DISFA-to-BP4D direction and ranks second only to FG-Net in the BP4D-to-DISFA direction, demonstrating highly competitive generalization capabilities. It is noteworthy that FG-Net leverages features from StyleGAN2 [45], pre-trained on FFHQ [46], a large-scale dataset with high-quality and diverse facial data, which helps mitigate overfitting on BP4D. In contrast, our method relies only on pre-trained weights from ImageNet-1K, showcasing strong generalization with minimal dependence on additional data sources.

C. Ablation Study

We conduct ablation studies on the BP4D dataset without ImageNet-1K pretraining to avoid extensive pretraining costs and employ subject-exclusive 3-fold cross-validation to investigate the effects of each component of AU-TTT. Table III presents the average F1-score results for various combinations of AU-TTT's components.

Original TTT. We use a linear projection without the Test-Time Training strategy as a baseline. The first two rows of Table III demonstrate the effectiveness of the original TTT Layer. **MSP.** The second and third rows of Table III verify the importance of multi-scale perception (MSP), highlighting its role in leveraging multi-scale knowledge effectively. **Foward TTT.** As shown in the third and fourth rows of Table III, adjusting the mini-batch size to better align with image tasks significantly enhances the performance of the TTT Layer on image tasks. **Bidirectional TTT.** We ablate the bidirectional design of AU-TTT, as demonstrated in the fourth and fifth rows of Table III. The results indicate that the bidirectional strategy improves performance by 0.8% compared to the unidirectional AU-TTT. **AU RoI TTT.** By comparing the last two rows of Table III, it is evident that incorporating additional AU RoI scanning enables the network to better extract AU-related local information, resulting in improved performance.

TABLE III

THE AVERAGE F1-SCORE (IN %) RESULTS OF VARIOUS COMBINATIONS OF AU-TTT COMPONENTS ON BP4D.

Baseline	Original TTT	MSP	Foward TTT	Backword TTT	AU RoI TTT	**Avg.**
✓						58.9
✓	✓					60.1
✓	✓	✓				60.4
✓		✓	✓			61.2
✓		✓	✓	✓		61.7
✓		✓	✓	✓	✓	62.1

IV. CONCLUSION

In this paper, we propose AU-TTT, an effective and innovative method for generalizable AU detection. AU-TTT leverages Test-Time Training to enhance the model's generalization ability and mitigate the overfitting caused by the limited availability of labeled AU datasets. By introducing TTT Layer and optimizing the scanning mechanism for images, AU-TTT effectively reduces the computational complexity associated with self-attention-based methods, particularly addressing the quadratic complexity challenge of Transformers. Additionally, the integration of an AU RoI scanning mechanism enables AU-TTT to better capture fine-grained, AU-related facial information, further improving its robustness and performance. These advancements make AU-TTT a promising approach for AU detection across diverse datasets and applications.

REFERENCES

[1] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic, "Automatic analysis of facial actions: A survey," *IEEE transactions on affective computing*, vol. 10, no. 3, pp. 325–347, 2017.

[2] Bohao Xing, Zitong Yu, et al., "Emo-llama: Enhancing facial emotion understanding with instruction tuning," *arXiv preprint arXiv:2408.11424*, 2024.

[3] Deng Li, Bohao Xing, and Xin Liu, "Enhancing micro gesture recognition for emotion understanding via context-aware visual-text contrastive learning," *IEEE Signal Processing Letters*, 2024.

[4] Yong Li, Menglin Liu, et al., "Counterfactual discriminative micro-expression recognition," *Visual Intelligence*, vol. 2, no. 1, pp. 29, 2024.

[5] Deng Li, Xin Liu, Bohao Xing, Baiqiang Xia, Yuan Zong, Bihan Wen, and Heikki Kälviäinen, "Eald-mllm: Emotion analysis in long-sequential and de-identity videos with multi-modal large language model," *arXiv preprint arXiv:2405.00574*, 2024.

[6] Kaiming He, Xiangyu Zhang, et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2016, pp. 770–778.

[7] Franco Scarselli, Marco Gori, et al., "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[8] Geethu Miriam Jacob and Bjorn Stenger, "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 7680–7689.

[9] Xin Liu, Kaishen Yuan, et al., "Multi-scale promoted self-adjusting correlation learning for facial action unit detection," *IEEE Transactions on Affective Computing*, 2024.

[10] Cheng Luo, Siyang Song, et al., "Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition," *arXiv preprint arXiv:2205.01782*, 2022.

[11] Xuehan Wang, CL Philip Chen, et al., "Semantic learning for facial action unit detection," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 3, pp. 1372–1380, 2022.

[12] Jing Yang, Jie Shen, et al., "Fan-trans: Online knowledge distillation for facial action unit detection," in *Proceedings of the IEEE/CVF WACV*, 2023, pp. 6019–6027.

[13] Kaishen Yuan, Zitong Yu, et al., "Auformer: Vision transformers are parameter-efficient facial action unit detectors," in *Proceedings of the European Conference on Computer Vision*. Springer, 2025, pp. 427–445.

[14] Zihan Wang, Siyang Song, et al., "Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2024, pp. 1270–1280.

[15] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] Kaili Zhao, Wen-Sheng Chu, et al., "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2016, pp. 3391–3399.

[17] Zhiwen Shao, Zhilei Liu, et al., "Jaa-net: joint facial action unit detection and face alignment via adaptive attention," *IJCV*, vol. 129, pp. 321–340, 2021.

[18] Yufeng Yin, Di Chang, et al., "Fg-net: Facial action unit detection with generalizable pyramidal features," in *Proceedings of the IEEE/CVF WACV*, 2024, pp. 6099–6108.

[19] Xuesong Niu, Hu Han, et al., "Local relationship learning with person-specific shape regularization for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2019, pp. 11917–11926.

[20] Xavier Bresson and Thomas Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[21] Yu Sun, Xinhao Li, et al., "Learning to (learn at test time): Rnns with expressive hidden states," *arXiv preprint arXiv:2407.04620*, 2024.

[22] S Mohammad Mavadati, Mohammad H Mahoor, et al., "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[23] Xing Zhang, Lijun Yin, et al., "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[24] Itir Onal Ertugrul, Jeffrey F Cohn, et al., "Cross-domain au detection: Domains, learning approaches, and measures," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–8.

[25] Itir Onal Ertugrul, Jeffrey F Cohn, et al., "Crossing domains for au coding: Perspectives, approaches, and measures," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 158–171, 2020.

[26] Cheng-Hao Tu, Chih-Yuan Yang, et al., "Idennet: Identity-aware facial action unit detection," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–8.

[27] Yufeng Yin, Liupei Lu, et al., "Self-supervised patch localization for cross-domain facial action unit detection," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2021, pp. 1–8.

[28] Javier Hernandez, Daniel McDuff, et al., "Deepfn: towards generalizable facial action unit recognition with deep face normalization," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2022, pp. 1–8.

[29] Alec Radford, Jong Wook Kim, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[30] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[31] Xinyu Xie, Yawen Cui, et al., "Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba," *Visual Intelligence*, vol. 2, no. 1, pp. 37, 2024.

[32] Léon Bottou and Vladimir Vapnik, "Local learning algorithms," *Neural computation*, vol. 4, no. 6, pp. 888–900, 1992.

[33] Alex Gammerman, Volodya Vovk, et al., "Learning by transduction," *arXiv preprint arXiv:1301.7375*, 2013.

[34] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[35] Yossi Gandelsman, Yu Sun, et al., "Test-time training with masked autoencoders," *Advances in NeurIPS*, vol. 35, pp. 29374–29385, 2022.

[36] Dian Meng, Bohao Xing, Xinlei Huang, Yanran Liu, Yijun Zhou, Zitong Yu, Xubin Zheng, et al., "scfusionttt: Single-cell transcriptomics and proteomics fusion with test-time training layers," *arXiv preprint arXiv:2410.13257*, 2024.

[37] Guanbin Li, Xin Zhu, et al., "Semantic relationships guided representation learning for facial action unit recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8594–8601.

[38] Tengfei Song, Zijun Cui, et al., "Hybrid message passing with performance-driven structures for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 6267–6276.

[39] Huiyuan Yang, Lijun Yin, Yi Zhou, et al., "Exploiting semantic embedding and visual feature for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 10482–10491.

[40] Yang Tang, Wangding Zeng, et al., "Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 12899–12908.

[41] Yanan Chang and Shangfei Wang, "Knowledge-driven self-supervised representation learning for facial action unit recognition," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2022, pp. 20417–20426.

[42] Zijun Cui, Chenyi Kuang, et al., "Biomechanics-guided facial action unit detection through force modeling," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2023, pp. 8694–8703.

[43] Xiaotian Li, Xiang Zhang, et al., "Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 20979–20989.

[44] Jia Deng, Wei Dong, et al., "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on CVPR*. Ieee, 2009, pp. 248–255.

[45] Tero Karras, Samuli Laine, et al., "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2020, pp. 8110–8119.

[46] Tero Karras, "A style-based generator architecture for generative adversarial networks," *arXiv preprint arXiv:1812.04948*, 2019.