Micro-Expression Spotting in Conversations via Noise-Disentangling and Boundary Aware Learning

Yigui Feng[a], Qinglin Wang[a,*], Yang Liu[b], Ke Liu[a], Haotian Mo[a], Gencheng Liu[a], Jie Liu[a]

[a]*College of Computer Science, National University of Defense Technology, Deya Road 109, Changsha, 410073, Hunan, China*
[b]*Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, xingye Road 777, Guangzhou, 511442, Guangdong, China*

Abstract

Accurately analyzing spontaneous micro-expressions in "in-the-wild" conversational scenes is a critical challenge, as performance of lab-trained models degrades dramatically due to speech-related noise. To address this, we introduce the Wild-Dialogue Micro-expression Dataset(WDMD), the first meticulously collected and annotated dataset focused on in-the-wild conversational scenarios. We propose a Micro-Expression Localization and Detection with Enhancer Framework (MELDAE), a novel end-to-end framework featuring a MicroExpression Enhancer module that uses learnable tokens to disentangle subtle ME signals from significant speech noise. We design a novel Boundary-Aware Loss function that dramatically improves temporal localization accuracy by explicitly penalizing onset and offset boundary errors. We conduct extensive experiments on WDMD and public benchmarks, demonstrating state-of-the-art results. On the challenging WDMD dataset, MELDAE achieves a 50.5% performance improvement over the SOTA model.

Keywords: Micro-expression Spotting, Boundary-Aware Loss, Temporal Localization, Affective Computing, Deep Learning, Pattern Recognition

*Corresponding author
Email address: wangqinglin@nudt.edu.cn (Qinglin Wang)

1. Introduction

Micro-expressions (MEs), as involuntary facial muscle movements characterized by their extremely brief duration (typically less than 500ms) and subtle amplitude, are widely regarded as reliable cues for revealing an individual's true emotions that they attempt to suppress or conceal [1]. The ability to automatically detect and analyze MEs holds immense potential across diverse fields, including clinical psychology for diagnostics, national security for deception detection, human resources for negotiation analysis, and advanced human-computer interaction [2].

However, despite considerable advancements in micro-expression(ME) analysis over the past decades [3], the field faces a critical "lab-to-reality" gap. The vast majority of research has heavily relied on data collected in strictly controlled laboratory environments [4]. These settings, while simplifying data acquisition, are ecologically invalid; they lack the complex variables of real-world interactions. "In-the-wild" settings introduce a series of formidable challenges, including unconstrained head poses, variable illumination conditions, occlusions, and, most notably, significant facial interference from other voluntary facial movementscite[5].

Among all "in-the-wild" scenarios, the analysis of ME within conversational contexts is particularly crucial and uniquely challenging. During a natural dialogue, the human face is in constant motion, producing strong, high-amplitude muscle movements associated with speech (e.g., lip articulation, jaw movement, cheek puffing, eyebrow raising for emphasis). These speech-related motions often temporally overlap with and spatially obscure the subtle, low-amplitude movements of a micro-expression. This creates a severe signal-to-noise ratio (SNR) problem where the "noise" (speech) is often stronger than the "signal" (ME). Existing models, trained on silent, non-interactive laboratory data, are not robust to this interference and exhibit a drastic performance degradation [6].

Furthermore, according to Truth-Default Theory[7], humans tend to operate on a default state of belief during communication. The authentic "emotional leakage," manifesting as MEs, can become a critical key to discerning true intentions precisely within these interactions. Therefore, developing methods that can robustly spot MEs

despite speech-related noise is a holy grail for practical affective computing. Nevertheless, a significant data-level deficiency impedes progress: nearly all publicly available micro-expression datasets are collected in non-conversational, non-interactive settings [8]. This lack of representative data has prevented the development of robust models for this critical application[9].

To bridge these gaps, this paper presents an end-to-end architecture, Micro-Expression Localization and Detection with Enhancer Framework(MELDAE). MELDAE learns a rich spatiotemporal representation of facial dynamics. A Vision Transformer (ViT) encodes each frame's spatial features, whose [CLS] tokens form a global sequence for a Bi-LSTM to capture long-range context. Parallel "region pooling" summarizes fine-grained patch information per frame. To amplify the sparse micro-expression signals amid noisy articulations, we introduce learnable query tokens that attend (via cross-attention) to the most relevant spatiotemporal regions, producing enhanced ME-specific features. A multitask prediction head then outputs (a) clip-level probability of a micro-expression (global ME classifier), (b) conversational state (speaking vs listening), and (c) framewise detection scores. Crucially, we propose a Boundary-Aware Loss (BAL) for the localization branch: in addition to a standard overlap loss (Focal-Tversky) for aligning predicted segments with ground truth, we add a weighted binary-cross-entropy term that up-weights the annotated start/end frames. This compels the model to focus on the exact temporal boundaries, addressing the notoriously ambiguous onset/offset of ME.

The key contributions of this work are summarized as follows:

- Wild-Dialogue Micro-expression dataset(WDMD). We build the first in-the-wild conversational micro-expression dataset, with fine onset/offset and dialogue-role annotations, to enable realistic benchmarking of ME spotting methods.
- MELDAE framework. We propose a novel end-to-end architecture combining ViT, LSTM, and an attention-enhancement module, trained with multitask supervision to robustly spot micro-expressions in noisy videos.
- Boundary-Aware Loss (BAL). We design a temporally aware loss that explicitly penalizes errors at micro-expression boundaries, significantly improving

Dataset	Environment	Elicitation Method	Interaction Scenario	Samples (MEs)	Annotation	Key Challenges
SMIC	Laboratory	Induced	Non-interactive	164	Apex	Controlled, Low resolution
CAS(ME)^2	Laboratory	Induced	Non-interactive	357 (57 MEs)	Onset, Offset, Apex	Controlled, Emotion types
SAMM	Laboratory	Spontaneous	Non-interactive	159	Onset, Offset, Apex	High resolution, Spontaneous
MMEW	Laboratory	Spontaneous	Non-interactive	300	Onset, Offset, Apex	High resolution, Diverse poses
WDMD (Ours)	**In-the-wild**	**Spontaneous**	**Conversational scenario**	**2253 (502 MEs)**	**Onset, Offset, Speaking state**	**Speech interference, Head poses, Illumination,Diverse poses, Emotion types**

Table 1: Comparison of major micro-expression datasets (only major representatives are selected, not all are included).

localization accuracy.

2. Related Work

This section reviews the landscape of micro-expression analysis, focusing on public datasets, spotting methodologies, and related techniques in temporal localization and loss function design.

2.1. Micro-expression Databases

The evolution of ME analysis has been intrinsically linked to the availability of datasets. As shown in Table 1, early datasets were foundational but limited. SMIC [10] and $CAS(ME)^2$[11] were collected in controlled laboratory settings using posed or induced paradigms (e.g., watching emotional videos). While valuable for initial algorithm development, they lack ecological validity, featuring static subjects and fixed-frontal poses. Subsequent datasets like SAMM [12] and MMEW [13]improved on this by capturing more spontaneous MEs from a larger, more diverse pool of subjects. They feature higher resolution and more variation in expression. However, they are still fundamentally non-interactive. Subjects are recorded in isolation, staring at a screen, without any conversational partner.This reveals a critical gap: no publicly available dataset before ours focuses specifically on spontaneous MEs occurring within an active, "in-the-wild" conversational dialogue. This is the unique contribution of our WDMD dataset, which introduces the core challenge of co-occurring speech articulation.

2.2. Micro-expression detection methods

ME detection aims to localize the temporal segments of MEs within long video streams. Early approaches relied on handcrafted features. Methods like LBP-TOP

[14] and LBP-SIP [15] extended Local Binary Patterns to the spatiotemporal domain to capture facial dynamics. While computationally efficient, these methods are highly sensitive to the noise, illumination, and pose variations prevalent in "in-the-wild" scenarios.

More recently, deep learning has dominated the field. These methods can be broadly categorized:CNN-based: CMNET [16] utilizes a contrastive magnification network to amplify subtle features. Others [17] use adaptive facial graphs. These methods are powerful spatial feature extractors but often require a separate temporal modeling component.RNN-based: LSTMs and GRUs are natural choices for modeling temporal dependencies. LTR3O [18] employs an RNN to learn onset-occurring-offset representations. Transformer-based: Transformers have shown promise due to their ability to capture long-range dependencies. u-BERT [19] and the more recent MOL [20] leverage Transformer architectures for recognition.

However, a common limitation persists: most existing deep methods [21, 22]are designed and validated on controlled, non-interactive datasets. Their ability to disentangle MEs from conversational speech articulations remains largely unaddressed and unproven.

2.3. Loss Functions for Localization

Precise temporal localization demands effective loss functions. For segmentation-based approaches, simple Binary Cross-Entropy (BCE) [23] is common but suffers from class imbalance. Losses like Dice or IoU [24] are better as they maximize overlap, but they treat all temporal points equally. Focal Loss[25] was proposed to focus on hard-to-classify samples, which is relevant given the rarity of MEs. However, none of these directly address the core challenge of MEs: ambiguous boundaries. Our work argues that for MEs, explicitly penalizing errors at the onset and offset frames is a critical, missing component for achieving high temporal precision.

3. WDMD Dataset

To spur research in realistic conversational ME analysis, we constructed the WDMD. This section details its collection, annotation, and statistical properties.

3.1. Data Collection and Curation

The WDMD is used to analyze sentiment in natural dialogue scenarios. Its main collection method is to capture a large number of publicly available movies, TV shows, and public interview dialogues, etc. We selected these sources to capture realistic, emotionally charged dialogues "in-the-wild". The focus on micro-expression clips from open-source movies is intended to lay the groundwork for our future micro-expression video generation model; the data we are currently using will be used to guide our micro-expression video generation model. We then conducted a rigorous screening process (combining manual review with automated filtering, such as keyword detection and Toxic-BERT detection[26]) to eliminate instances that were harmful, personal, or emotionally irrelevant. The selection criteria included: (1) High-definition footage (1080p or higher, 60fps) to ensure subtle movements are captured; (2) Clear, frontal, or near-frontal views of the subjects' faces; (3) Unscripted interactions (e.g., interviews, documentaries) to promote spontaneous emotional expression.

A total of over 1000 hours of footage was initially reviewed. Clips containing potential MEs were extracted by trained annotators, resulting in the final dataset of 2,253 clips (at 2560x1440 resolution).

3.2. Annotation Protocol and Quality Assurance

Given the subtlety of MEs, a rigorous annotation protocol is essential.

Annotators: The annotation was performed by 3 expert psychologists trained in the Facial Action Coding System (FACS).

Annotation Tool: We used ELAN for precise, frame-by-frame annotation.

Protocol: Annotators were instructed to identify and label the precise onset (start frame) and offset (end frame) of any involuntary facial movement matching the definition of a micro-expression. Crucially, they also annotated the subject's conversational context for each frame as either "Speaking" or "Listening/Silent".

Inter-Annotator Agreement (IAA): To validate the reliability of our annotations, we calculated the IAA on a randomly selected 20% subset of the data. For the discrete "speaking" state label, we achieved a Krippendorff's Alpha of 0.85, indicating high reliability. For the temporal segments (onset/offset), we used the Intersection over

Union (IoU) metric, achieving an average IoU of 0.78 between annotator pairs. These scores indicate a high level of agreement, confirming the quality of the labels despite the inherent difficulty of the task.

3.3. Dataset Statistics

The WDMD dataset employed in this study comprises ME samples captured in various contexts. To comprehensively understand the intrinsic characteristics of the dataset, this section provides a detailed statistical analysis of the data distribution from two key dimensions: speaking interaction state and expression duration.

Distribution of Speaking States (Speaking vs. Listening): We first analyzed the distribution of ME in two core speaking interaction states: "Speaking" and "Listening". According to the pie chart shown in Figure 1, we observe that in the "Speaking" state (blue region), the samples generated during the "Speaking" state account for slightly more than half of all collected ME samples. Through visual estimation, this proportion is approximately between 55% and 60%. In the "Listening" state (orange region), the samples generated during the "Listening" state account for slightly less than half, estimated to be between 40% and 45%. This data indicates that the number of ME samples in the "Speaking" state is slightly higher than that in the "Listening" state in this dataset. This may suggest that the speaking process (involving cognitive processing, language organization, and possible emotional suppression) is an important context for eliciting micro-expressions. Despite this slight skew, the sample sizes of the two states remain relatively balanced overall, avoiding a severe imbalance in the dimension of speaking state, which provides a solid data foundation for subsequent comparative studies (e.g., comparing the types or intensity of MEs in the two states).

Classification of Micro-expression Duration: For fine-grained temporal analysis, we performed binning statistics on the duration of all ME in the dataset. According to Figure 2, the time dimension is divided into the following seven intervals: 0.0s - 0.5s (covering the typical duration defined for traditional MEs), 0.5s - 1.0s, 1.0s - 1.5s, 1.5s - 2.0s, 2.0s - 2.5s, 2.5s - 3.0s, greater than 3.0s (> 3.0s).Although MEs are typically short (<0.5s), we include longer expressions (>3.0s) in our dataset statistics to represent macro-expressions or complex emotional compounds present in conversational settings,

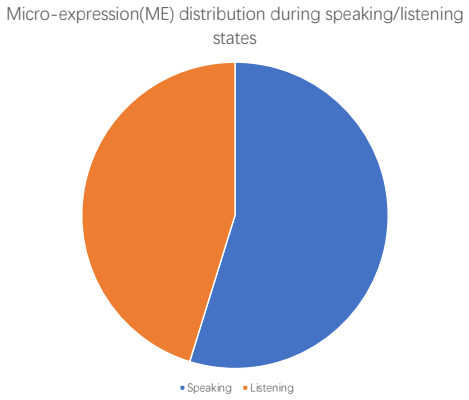

Figure 1: Distribution of Speaking States

which serve as important contrastive samples for the spotting task.

By combining the statistics from the two dimensions, we can outline a complete profile of this dataset: it is a micro-expression library that has sufficient sampling in both "Speaking" (approximately 55%-60%) and "Listening" (approximately 40%-45%) states, and possesses fine-grained time segmentation. The dataset is relatively balanced in the dimension of speaking state and has established clear time binning criteria.

4. Methodology: The MELDAE Framework

To address the challenges of ME spotting in conversational videos, we propose MELDAE, an end-to-end framework following an "encoder-enhancer-decoder" paradigm. The overall architecture is depicted in Figure 3.

4.1. Temporal Implicit Feature Extractor

This module extracts robust spatiotemporal features from the noisy video input ($X \in \mathbb{R}^{T\times C\times H\times W}$).

ViT Encoder: We employ a pre-trained Vision Transformer[27] as the frame-level spatial encoder. Unlike CNNs, which rely on local receptive fields, ViT's self-attention

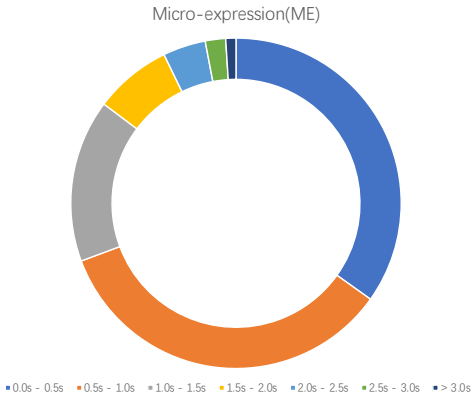

Figure 2: Classification of Micro-expression Duration

mechanism captures long-range dependencies between image patches. We hypothesize this is crucial for MEs, which often involve coordinated, subtle movements across distant facial regions (e.g., eyes and mouth). The ViT processes each frame, yielding a global classification token and a set of local patch tokens.

Temporal Modeling: The sequence of classification tokens is fed into a Bi-LSTM network. This explicit temporal modeling is vital for understanding the dynamics of facial movements, capturing context from both past and future frames to generate a comprehensive global temporal feature, $F_{\text{global}} \in \mathbb{R}^{T \times D}$. Concurrently, the patch tokens are processed via Region Pooling to form F_{regional}, retaining local spatial information.

4.2. MicroExpression Enhancer

This module is the core innovation of MELDAE, designed to isolate subtle ME signals from high-magnitude noise, especially speech articulations.

Learnable Query Tokens: We introduce a set of N learnable micro-expression query tokens (we found $N = 16$ to be optimal). These tokens are initialized randomly and act as abstract "prototypes" or "probes" for MEs. During training, they learn to represent common spatiotemporal patterns associated with micro-expressions (e.g., a token for "eyebrow twitch," another for "lip corner pull").

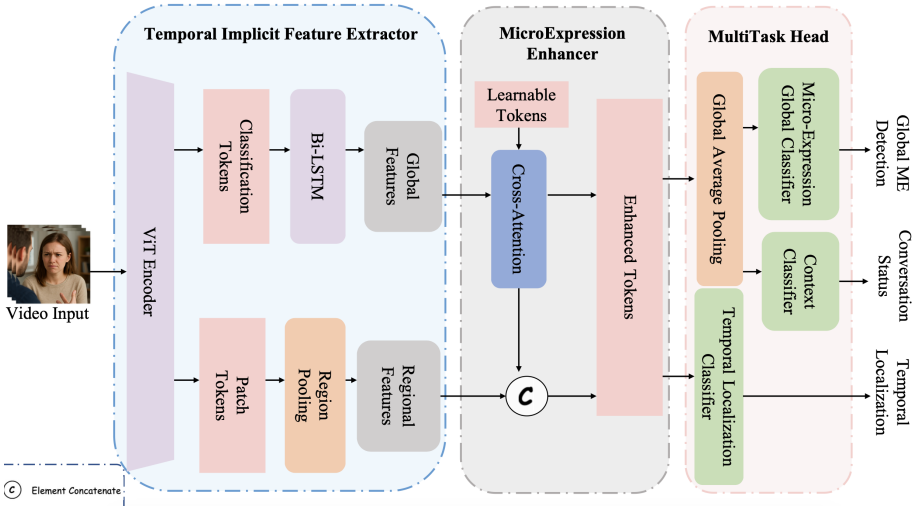

Figure 3: Overall framework of the proposed MELDAE.

Cross-Attention Mechanism: We leverage a cross-attention mechanism where the learnable tokens serve as the Query, and the regional video features F_{regional} serve as both Key and Value. This allows the query tokens to "interrogate" the video features, adaptively aggregating the most relevant information from all spatial regions across all frames that match their learned ME patterns. This process effectively filters out irrelevant noise (like jaw movements from speech) and forms a set of enhanced micro-expression representations, F_{enhanced}.

Feature Fusion: Finally, we concatenate F_{enhanced} with the global temporal features F_{global} to provide a highly informative and discriminative feature representation for the downstream tasks.

4.3. Multi-Task Head

The fused features are decoded by three parallel branches for multi-task learning. Global ME Classifier: Predicts the probability $p_{\text{ME}} \in [0, 1]$ that the entire clip contains an ME. Conversational Scene Classifier: Predicts the probability $p_{\text{State}} \in [0, 1]$ that the subject is in a "speaking" state. This task acts as an auxiliary regularizer. We hypothesize that by forcing the model to explicitly learn and disentangle speech-related features, we free the main enhancer module to focus on non-speech ME signals. Our ablation

studies confirm this. Temporal Locator: Generates a frame-by-frame confidence score $S_{\text{loc}} \in \mathbb{R}^T$ for the presence of an ME.

4.4. Boundary-Aware Multi-Task Learning Strategy

We use a weighted composite loss:

$$\mathcal{L}_{\text{total}} = w_1\mathcal{L}_{\text{ME}} + w_2\mathcal{L}_{\text{State}} + w_3\mathcal{L}_{\text{loc}} \tag{1}$$

For the classification losses ($\mathcal{L}_{\text{ME}}$, $\mathcal{L}_{\text{State}}$), we use Focal Loss to counteract the severe class imbalance. The core of our contribution is the temporal localization loss, $\mathcal{L}_{\text{loc}}$, for which we designed the **BAL:**

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{overlap}} + \lambda\mathcal{L}_{\text{boundary}} \tag{2}$$

Overlap Loss ($\mathcal{L}_{\text{overlap}}$): Implemented using Focal Tversky Loss [28]. This variant of the Tversky index is ideal for small, imbalanced targets (like MEs). It maximizes the overlap (IoU) between prediction S_{loc} and ground truth Y while focusing on hard samples.

Boundary Loss ($\mathcal{L}_{\text{boundary}}$): This component directly targets localization precision. It is a weighted binary cross-entropy (BCE) where the weight w_i for each frame i is significantly increased to a hyperparameter W_{boundary} (e.g., 10) for annotated onset and offset frames, and is 1 otherwise:

$$\mathcal{L}_{\text{boundary}} = -\frac{1}{T}\sum_{i=1}^{T} w_i \cdot [y_i \log(s_i) + (1 - y_i)\log(1 - s_i)] \tag{3}$$

This design compels the model to pay special attention to the transient boundaries, which are the most difficult part to localize, fundamentally enhancing localization precision. The hyperparameter λ (set to 1.0) balances the two loss components.

5. Experiments and Results

We conducted a comprehensive set of experiments to validate MELDAE, focusing on (1) performance on our challenging WDMD dataset, (2) generalization capabilities on public benchmarks, and (3) in-depth analysis of each model component.

5.1. Datasets

- **WDMD (Ours):** Our primary dataset for in-the-wild conversational ME analysis.
- $CAS(ME)^2$**:** A standard lab-controlled dataset for benchmarking.
- **SAMM:** A dataset of spontaneous MEs, providing a different "in-the-wild" (though non-conversational) challenge.
- **MMEW:** Another widely used spontaneous ME dataset.

5.2. Evaluation Metrics

To quantitatively evaluate our framework, we established a multi-dimensional metric system. For the binary classification tasks of global micro-expression detection and conversational state classification, we use standard Accuracy. For the more challenging core task of temporal localization, we employ an evaluation scheme based on Intersection over Union (IoU), where a predicted segment is deemed a True Positive (TP) if its IoU with a ground-truth segment exceeds a threshold of $\theta = 0.5$. Based on this, we compute Precision and Recall to derive the primary localization metric, the F1-score. To further account for the differing facial dynamics in conversation, we separately calculate localization F1-scores for speaking ($F1_{\text{speaking}}$) and listening ($F1_{\text{listening}}$) contexts and propose a single, comprehensive metric, the F1-score for Dialogue Roles ($F1_{DR}$), defined as their harmonic mean:

$$F1_{DR} = \frac{2 \cdot F1_{\text{listening}} \cdot F1_{\text{speaking}}}{F1_{\text{listening}} + F1_{\text{speaking}}} \tag{4}$$

This fused metric provides a fair and robust assessment of localization performance across distinct conversational states.For the three public benchmarks $CAS(ME)^2$, SAMM, and MMEW, we use the UF1 index to evaluate model performance.

$$UF1 = \frac{1}{K} \sum_{k=1}^{K} F_1^{(k)} \qquad \text{where} \quad F_1^{(k)} = \frac{2 \cdot Precision_k \cdot Recall_k}{Precision_k + Recall_k} \tag{5}$$

5.3. Implementation Details

Our model is implemented in PyTorch and trained on eight NVIDIA H100 80GB GPUs. We use the AdamW optimizer with a differential learning rate: 1×10^{-5} for

the ViT backbone and 5×10^{-4} for newly initialized parts. Batch size is 16. The loss weights are $w_1 = 0.5$, $w_2 = 0.2$, $w_3 = 1.0$. For BAL, $\lambda = 1.0$ and $W_{\text{boundary}} = 10$.

5.4. Baseline Models

Traditional Methods: LBP-TOP and LBP-SIP.

Deep Methods: LTR3O, CMNET, u-BERT, PLMaM-Net[29], SRMCL[30] and FFDIN[31].

Recent SOTA: SODA4MER[32] and MOL .

5.5. Comparison with State-of-the-Art

On WDMD: As shown in Table 2, MELDAE achieves an $F1_{\text{DR}}$ of 31.74, significantly outperforming all baselines. This represents a 50.5% relative improvement over the strongest recent baseline, SODA4MER (21.15), and more than doubles the performance of u-BERT (14.02). This highlights the critical failure of existing models in the face of conversational noise and the effectiveness of our Enhancer module. The high Acc_State (80.27%) also confirms the success of our multi-task learning.

On Public Benchmarks: As shown in Table 2, MELDAE's superiority is not limited to WDMD. On CAS(ME)^2, it achieves an UF1 of 37.93 (vs 25.04 from SODA4MER). On the spontaneous dataset SAMM, MELDAE achieved a score of 88.13, second only to the MOL model. On the MMEW dataset, MELDAE achieved a score of 73.04, once again setting a new state-of-the-art record. This strong generalization performance demonstrates that our framework learns the fundamental and robust features of ME, rather than overfitting to noisy patterns specific to a particular dataset.

5.6. Comparative Experiment of Loss Functions

To validate the effectiveness of the Boundary-Aware Loss (BAL) function proposed in this paper, we conducted a comparative analysis against five baseline loss functions: MAE (Mean Absolute Error)[33], MSE (Mean Squared Error)[34], IoU (Intersection over Union), Smooth L1 Loss)[35], and BCE (Binary Cross-Entropy) Loss.

All models were trained under identical dataset and experimental settings for a total of 20 epochs. We adopted the $F1_{\text{DR}}$ score as the Key Performance Metric (KPM), where

Model	WDMD (Acc ME)	WDMD (Acc State)	WDMD ($F1_{\mathrm{DR}}$) ↑	$CAS(ME)^2$ (UF1) ↑	SAMM (UF1) ↑	MMEW (UF1) ↑
LBP-TOP	23.91	19.80	3.62	5.92	39.54	45.12
LBP-SIP	25.17	22.03	3.90	7.33	52.11	47.89
LTR3O	76.27	75.84	12.82	13.73	54.09	52.53
CMNET	74.80	72.97	10.16	10.89	73.94	68.34
u-BERT	77.15	76.02	14.02	15.41	77.89	72.17
PLMaM-Net	69.07	66.74	8.25	8.14	7.90	59.76
SRMCL	70.02	69.11	8.07	8.98	70.15	63.45
FFDIN	75.78	74.51	11.30	12.77	65.98	60.22
MOL	77.50	75.36	15.30	17.12	**89.72**	65.88
SODA4MER	79.01	77.25	21.15	25.04	78.93	70.46
MELDAE* (w/o Enhancer)	78.05	76.59	18.87	20.68	76.05	65.10
MELDAE (Full)	**81.76**	**80.27**	**31.74**	**37.93**	88.13	**73.04**

Table 2: Performance comparison with SOTA methods on WDMD, CAS(ME)^2, SAMM, and MMEW datasets.

a higher score indicates better overall model performance. The experimental results are illustrated in Figure 4.

As clearly observed in Figure 4, there are significant disparities in the $F1_{\mathrm{DR}}$ scores achieved by models trained with different loss functions. As the number of training epochs increases, the performance of all models shows varying degrees of improvement, tending to converge after approximately 14 epochs. The loss functions are distinctly grouped into two performance clusters. The traditional regression losses, MAE and MSE, performed the worst. Their $F1_{\mathrm{DR}}$ scores remained below 0.15 throughout the training process, ultimately converging at approximately 0.12. This suggests they may be unsuitable for the specific task addressed in this study. The IoU, Smooth L1, and BCE losses performed substantially better than MAE and MSE. At convergence, the IoU loss achieved an $F1_{\mathrm{DR}}$ score of approximately 0.27, Smooth L1 reached about 0.285, and BCE approached 0.30.

The BAL loss function proposed in this paper (labeled "BAL (Ours)" in the chart) demonstrated optimal performance at all stages of training. It took the lead from the 2nd epoch, and its advantage became more pronounced as training progressed. Ultimately, the BAL loss function enabled the model to achieve an $F1_{\mathrm{DR}}$ score of approximately

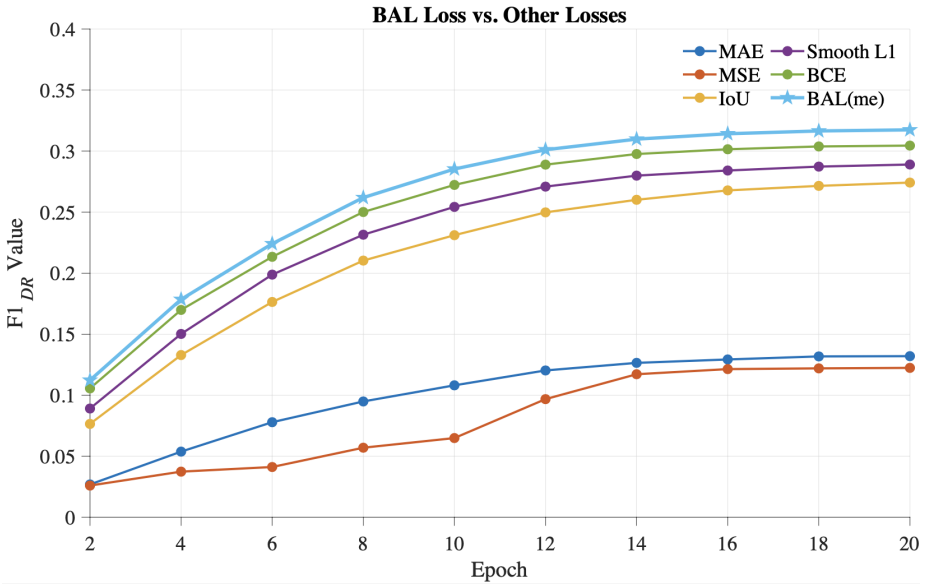

Figure 4: Experimental comparison results of BAL and other losses.

0.32, which is markedly higher than all other compared methods. The experimental results strongly demonstrate that, compared to existing loss functions, the proposed BAL loss can more effectively guide model training and achieve superior performance on the $F1_{\mathrm{DR}}$ metric.

5.7. Ablation Studies

We conducted in-depth ablation studies on the WDMD dataset to dissect the contribution of each component of MELDAE. Table 3 provides compelling evidence for our design choices.

Effect of Enhancer (Row 2): Removing the Micro-Expression Enhancer causes a catastrophic 40.6% relative drop in $F1_{\mathrm{DR}}$. This is the largest drop observed, confirming that this module is the most critical component for filtering conversational noise and isolating the ME signal.

Effect of Backbone (Row 3): Replacing the ViT backbone with a standard ResNet-50 (while keeping other components) results in a 17.7% drop. This suggests that ViT's ability to capture long-range spatial dependencies is superior for modeling the

coordinated muscle movements of the face.

Effect of Multi-Task Loss (Row 5): Removing the auxiliary "speaking state" loss ($\mathcal{L}_{\text{State}}$) causes a significant 20.6% drop. This strongly supports our hypothesis: forcing the model to explicitly identify and disentangle "speaking" features allows the main Enhancer to better focus on the non-speech ME signals.

Effect of BAL (Rows 6-8): The choice of loss function is critical. Replacing BAL with standard BCE or even IoU losses leads to major performance degradation. Notably, even using just the Focal Tversky component (Row 8), which is already a strong baseline, is still 10.7% worse than the full BAL. This proves that our novel Boundary Loss ($\mathcal{L}_{\text{boundary}}$) term is highly effective and necessary for achieving precise temporal localization.

	Model Configuration	$F1_{\text{DR}} \uparrow$	$\Delta F1_{\text{DR}}$ (vs Full)
1	MELDAE (Full Model)	31.74	–
Component Analysis			
2	w/o Micro-Expression Enhancer (MELDAE*)	18.87	-12.87 (40.6% ↓)
Backbone Analysis			
3	w/ ResNet-50 (replace ViT)	26.13	-5.61 (17.7% ↓)
4	w/o Bi-LSTM (remove temporal modeling)	28.05	-3.69 (11.6% ↓)
Multi-Task Analysis			
5	w/o $\mathcal{L}_{\text{State}}$ (remove speaking state loss)	25.20	-6.54 (20.6% ↓)
Loss Function Analysis			
6	w/ BCE Loss (replace BAL)	22.04	-9.70 (30.6% ↓)
7	w/ IoU Loss (replace BAL)	27.19	-4.55 (14.3% ↓)
8	w/ Focal Tversky (i.e., BAL w/o $\mathcal{L}_{\text{boundary}}$)	28.33	-3.41 (10.7% ↓)

Table 3: Ablation studies on the WDMD dataset.

5.8. Hyperparameter Sensitivity

We analyzed the sensitivity of MELDAE to its two key new hyperparameters: the BAL weight λ and the number of learnable tokens N.

As shown in Table 4, the model performance is robust to a reasonable range of λ values, peaking at $\lambda = 1.0$, which indicates a balanced contribution from both the overlap and boundary components is optimal. For the number of query tokens N, performance improves up to $N = 16$ and then saturates, suggesting that 16 tokens are sufficient to capture the main "prototypes" of MEs without overfitting.

Parameter	Value	$F1_{\text{DR}}$
BAL λ		
	0.5	30.15
	1.0 (Selected)	**31.74**
	1.5	31.09
	2.0	29.88
Tokens N		
	4	28.30
	8	31.02
	16 (Selected)	**31.74**
	32	31.71

Table 4: Hyperparameter sensitivity analysis for BAL weight λ and number of learnable tokens N on WDMD.

5.9. Qualitative Analysis and Visualization

In this section, we provide a qualitative analysis to intuitively demonstrate the effectiveness and robustness of our proposed MELDAE model, particularly its ability to handle speech-related noise. We first present a direct case comparison and then visualize the internal attention mechanism of our Enhancer module.

Figure 5 presents a challenging case study where a true ME event (frames 30-45) significantly overlaps with speech noise (frames 20-60). The baseline model,

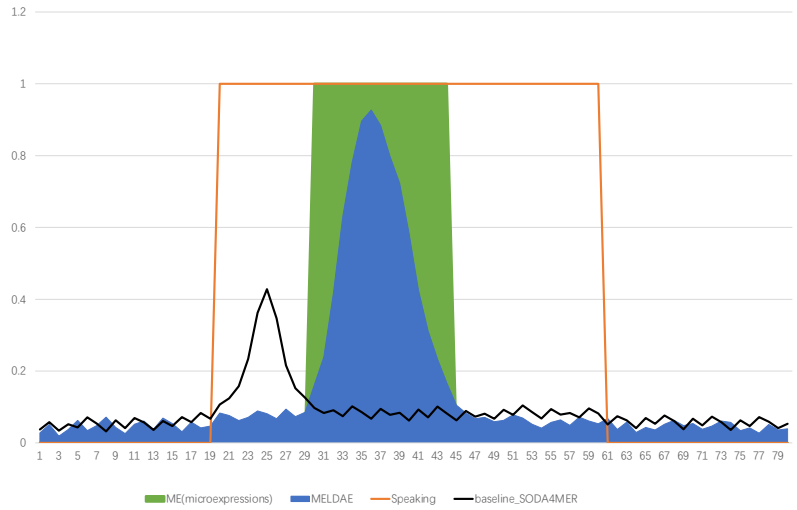

Figure 5: Case analysis. ME occurs in frames 30-45. 'Speaking' (speech noise) occurs in frames 20-60. MELDAE successfully detects ME in the noise, while Baseline-SODA4MER is interfered with by speech noise, producing false activations in frames 22-28 and missing the real ME.

SODA4MER, is clearly confused by the irrelevant motion from speaking. It not only produces false activations (frames 22-28) by misinterpreting speech as an ME, but it also completely misses the real ME event. In sharp contrast, our MELDAE model successfully suppresses the interference from the speech noise and accurately localizes the true ME in the correct temporal window.

To understand *how* MELDAE achieves this robustness, we visualize the attention weights of the Enhancer module. We represent the facial area as a 10x10 grid, with different row-groups corresponding to distinct facial regions (e.g., rows 3-4 for Eyes & Brows, rows 7-8 for Mouth).

Figure 6 shows the attention distribution of the model on frames containing only pure speech. As shown, the attention weights are generally low and dispersed, with a slight concentration in the mouth region (rows 7-8). This pattern is consistent with the characteristics of speech noise, indicating that the model can correctly identify motion sources but does not treat them as high-confidence ME signals. Figure 8 shows the attention distribution on non-ME, non-speech frames. As shown, the attention weights are very low and uniformly distributed across all regions, indicating that the model is in

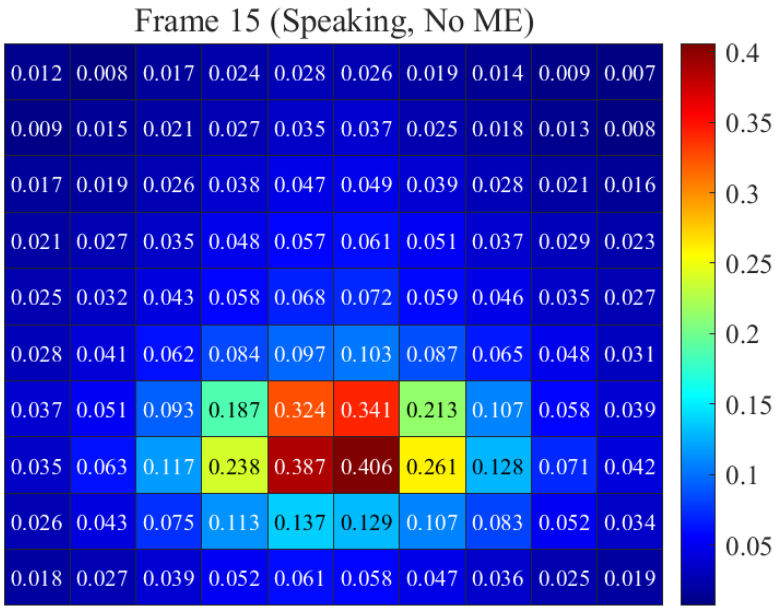

Figure 6: Attention visualization of pure speech frames. A 10x10 grid represents the facial region: rows 1-2: Forehead, rows 3-4: Eyes & Brows, rows 5-6: Nose & Cheeks, rows 7-8: Mouth, rows 9-10: Chin & Jaw. Attention weights are low and dispersed, mainly concentrated in the mouth region, consistent with the characteristics of speech noise.

an "idle" state and does not detect any significant facial activity.

Conversely, Figure 7 shows the attention map for a frame at the burst (apex) of an ME. The attention mechanism exhibits a dramatically different behavior. The weights become highly concentrated in the key facial regions critical for this specific ME, such as the eyes (rows 3-4) and the corners of the mouth (rows 7-8). Meanwhile, other regions like the jaw (rows 9-10) receive very low weights.

This comparison clearly indicates that the Enhancer module has learned to effectively distinguish task-relevant ME signals from task-irrelevant noise. It successfully captures and focuses on the subtle, localized muscle movements characteristic of micro-expressions, providing an interpretable basis for the model's superior and robust performance.

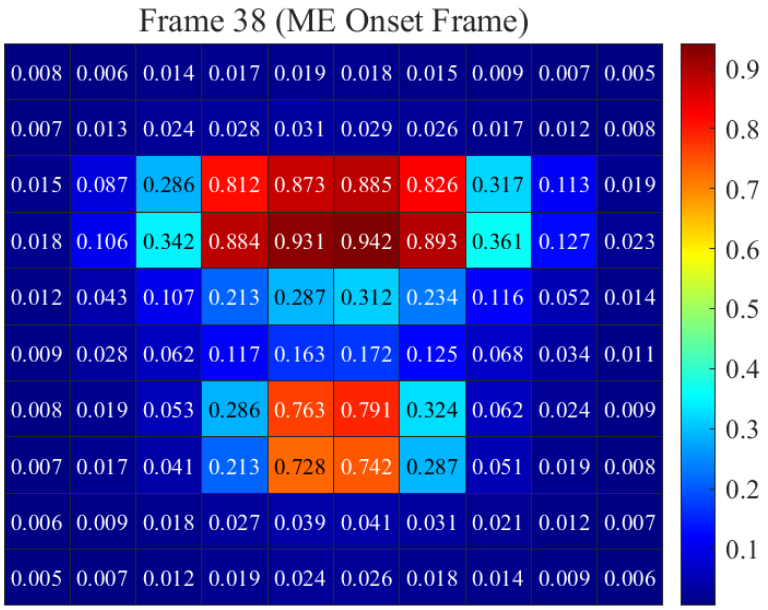

Figure 7: Attention visualization of ME burst frames. A 10x10 grid represents facial regions: rows 1-2: Forehead, rows 3-4: Eyes & Brows, rows 5-6: Nose & Cheeks, rows 7-8: Mouth, rows 9-10: Chin & Jaw. Attention weights are highly concentrated in key ME regions, such as the eyes (rows 3-4) and corners of the mouth (rows 7-8), while other regions (such as the jaw) have low weights. This indicates that the Enhancer module successfully captured the ME signal from the noise.

6. Discussion and Future Work

6.1. Limitations

Despite its strong performance, our work has limitations.

Dataset Source: While WDMD is the first conversational dataset, it is sourced from high-quality films and documentaries. These interactions, while unscripted, may not fully capture the nuance and variability of all real-life, "in-the-field" scenarios.

Computational Cost: The use of a ViT backbone and a Bi-LSTM, while effective, results in a computationally heavy model not suitable for real-time mobile applications without significant optimization.

Extreme Poses: As shown in our qualitative analysis, the model's performance

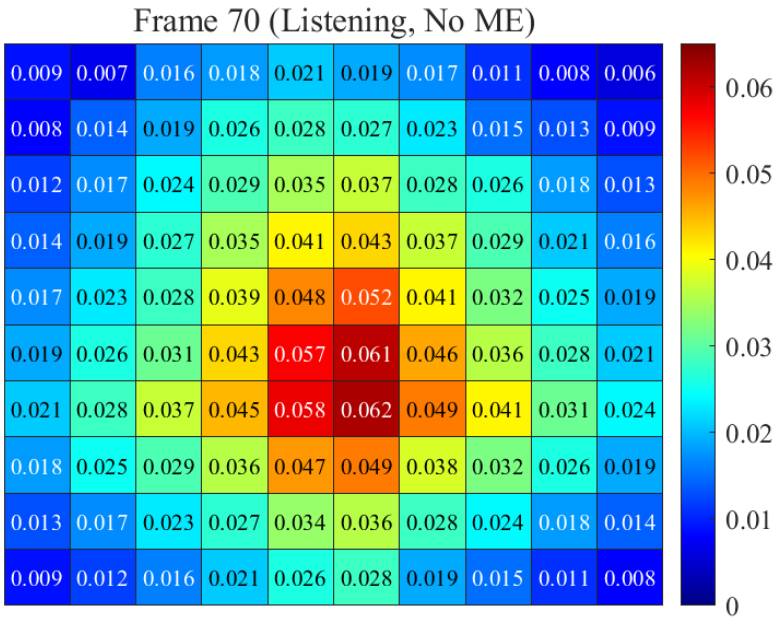

Figure 8: No-ME, no-speaking frames. A 10x10 grid represents the facial region: rows 1-2: Forehead, rows 3-4: Eyes & Brows, rows 5-6: Nose & Cheeks, rows 7-8: Mouth, rows 9-10: Chin & Jaw. Attention weights are very low and uniformly distributed across all regions, indicating that the model is in an "idle" state and no significant facial activity is detected.

degrades under extreme head poses (> 45 degrees) and severe occlusions (e.g., hand-over-mouth).

6.2. Future Work

This work opens several avenues for future research.

Dataset Expansion: We plan to expand dataset diversity by collecting data from more varied "in-the-wild" sources (e.g., video blogs, real-world interviews) to further challenge our models.

Lightweight Models: A critical next step is exploring lightweight models for real-time deployment. We are investigating knowledge distillation from our full MELDAE model into a smaller, efficient architecture (e.g., MobileNet + a compact Transformer) to bridge the gap between performance and efficiency.

Cross-Dataset Generalization: The domain shift between different ME datasets remains a significant challenge. Future work will explore advanced domain adaptation techniques to create a truly universal ME spotting system.

7. Conclusion

This paper presented a comprehensive framework for micro-expression spotting in challenging, "in-the-wild" conversational scenes. We made four primary contributions.

The introduction and detailed analysis of WDMD, the first dataset focused on conversational MEs, which captures the critical challenge of speech-related noise. The MELDAE end-to-end framework, featuring a novel Micro-Expression Enhancer module that effectively isolates subtle ME signals from this noise using learnable tokens and cross-attention. The Boundary-Aware Loss (BAL), which significantly improves temporal localization precision by explicitly penalizing onset and offset boundary errors. Extensive validation on four datasets (WDMD, CAS(ME)^2, SAMM, MMEW) demonstrating that MELDAE not only sets a new state-of-the-art by a large margin on conversational data (50.5% relative $F1_{\mathrm{DR}}$ improvement) but also generalizes exceptionally well to traditional benchmarks.

Through detailed ablation studies and qualitative analysis, we validated the efficacy of each component. This work represents a significant step forward in moving micro-expression analysis from controlled laboratories into the complexity of real-world human interactions.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China: 2021YFBO300101.

References

[1] P. Ekman, W. V. Friesen, The repertoire of nonverbal behavior: Categories, origins, usage, and coding, Semiotica 1 (1) (1969) 49–98.

[2] Using machine learning to detect emotions and predict human psychology, IGI Global, 2024, [M].

[3] Y. H. Oh, J. See, A. C. Le Ngo, D. B. Jayagopi, R. C. W. Phan, A survey of automatic facial micro-expression analysis: databases, methods, and challenges, Frontiers in Psychology 9 (2018) 1128.

[4] L. Zhang, O. Arandjelović, Review of automatic microexpression recognition in the past decade, Machine Learning and Knowledge Extraction 3 (2) (2021) 414–434.

[5] Z. Shangguan, Y. Dong, S. Guo, et al., Facial expression analysis and its potentials in iot systems: A contemporary survey, ACM Computing Surveys 58 (2) (2025) 1–39.

[6] S. Chen, H. Huang, Y. Liu, Z. Wu, S. Wang, C. C. Loy, Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis, arXiv preprint arXiv:2508.13618 (2025). arXiv:2508.13618.

[7] T. R. Levine, Truth-default theory (tdt) a theory of human deception and deception detection, Journal of Language and Social Psychology 33 (4) (2014) 378–392.

[8] W. Merghani, A. K. Davison, M. H. Yap, A review on facial micro-expressions analysis: datasets, features and metrics, arXiv preprint arXiv:1805.02397 (2018). arXiv:1805.02397.

[9] S. Zhao, H. Tang, X. Mao, S. Wang, W. Yan, X. Liu, X. Fu, Dfme: A new benchmark for dynamic facial micro-expression recognition, IEEE Transactions on Affective Computing 15 (3) (2023) 1371–1386.

[10] X. Li, T. Pfister, X. Huang, et al., A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

[11] F. Qu, S. J. Wang, W. J. Yan, H. Li, S. Wu, X. Fu, Cas(me)2: a database for spontaneous macro-expression and micro-expression spotting and recognition, IEEE Transactions on Affective Computing 9 (4) (2017) 424–436.

[12] A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE Transactions on Affective Computing 9 (1) (2018) 116–129. doi:10.1109/TAFFC.2016.2573832.

[13] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). doi:10.1109/TPAMI.2021.3067464.

[14] A. C. Le Ngo, Y. H. Oh, R. C. W. Phan, J. See, Eulerian emotion magnification for subtle expression recognition, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 1243–1247.

[15] Y. Wang, J. See, R. C. W. Phan, Y. H. Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: Asian Conference on Computer Vision, Springer International Publishing, Cham, 2014, pp. 525–537.

[16] M. Wei, X. Jiang, W. Zheng, J. Liu, Cmnet: Contrastive magnification network for micro-expression recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 119–127.

[17] H. Sun, Z. Liu, S. Wang, et al., Adaptive attention-based graph representation learning to detect phishing accounts on the ethereum blockchain, IEEE Transactions on Network Science and Engineering 11 (3) (2024) 2963–2975.

[18] J. Zhu, Y. Zong, J. Shi, G. Zhao, Learning to rank onset-occurring-offset representations for micro-expression recognition, IEEE Transactions on Affective Computing (2025).

[19] X. B. Nguyen, C. N. Duong, X. Li, K. Luu, Micron-bert: Bert-based facial micro-expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1482–1492.

[20] Z. Shao, Y. Yuan, L. Ma, X. Zhu, Curvnet: Latent contour representation and iterative data engine for curvature angle estimation, Pattern Recognition (2025) 112546.

[21] J. Wei, J. Sun, G. Lu, et al., Multi-information hierarchical fusion transformer with local alignment and global correlation for micro-expression recognition, in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 5873–5882.

[22] C. Gan, J. Xiao, Q. Zhu, et al., Macro-expression-guided micro-expression recognition: A motion similarity perspective, Pattern Recognition (2025) 112237.

[23] Y. N. Wu, Cross entropy, in: Encyclopedia of Statistics in Quality and Reliability, Springer International Publishing, Cham, 2021, pp. 225–226.

[24] H. Rezatofighi, N. Tsoi, J. Y. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 658–666.

[25] T. Y. Lin, P. Goyal, R. Girshick, et al., Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[26] L. Hanu, Unitary team, Detoxify, Github: `https://github.com/unitaryai/detoxify`, accessed: [Your Access Date] (2020).

[27] K. Han, Y. Wang, H. Chen, et al., A survey on vision transformer, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (1) (2022) 87–110.

[28] N. Abraham, N. M. Khan, A novel focal tversky loss function with improved attention u-net for lesion segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), IEEE, 2019, pp. 683–687.

[29] F. Wang, Y. Zong, J. Zhu, W. Zheng, G. Zhao, Progressively learning from macro-expressions for micro- expression recognition, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 4390–4394.

[30] Y. Bao, C. Wu, P. Zhang, S. Fan, Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning, IEEE Transactions on Affective Computing 15 (4) (2024) 2083–2096.

[31] C. Li, R. Ba, X. Wang, J. Dong, W. Zheng, Structure representation with adaptive and compact facial graph for micro-expression recognition, IEEE Transactions on Biometrics, Behavior, and Identity Science (2024).

[32] B. Zhang, X. Wang, C. Wang, et al., Dynamic stereotype theory induced micro-expression recognition with oriented deformation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 10701–10711.

[33] C. M. Bishop, Pattern Recognition and Machine Learning, Springer Science+Business Media, LLC, Berlin, Germany, 2006.

[34] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.

[35] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.