# Towards Trustworthy Dynamic Facial Expression Recognition via Information Bottleneck Modeling

Feng-Qi Cui, Anyang Tong, Jinyang Huang*, Jie Zhang, Meng Li, Xin Yan, Linsheng Huang, Dan Guo, and Meng Wang, *Fellow, IEEE*

***Abstract*—Due to the presence of semantic ambiguity among similar expression categories and the inherent imbalance in spatio-temporal feature intensities, dynamic facial expression recognition (DFER) in the wild poses significant challenges for building trustworthy and robust systems. These factors often lead to inconsistent feature representations and unreliable decision boundaries, which hinder the model's ability to perform stable and accurate recognition under uncertainty, and further pose a serious safety hazard, *e.g.*, misdiagnosis of depression. To tackle these challenges, we propose a novel adaptive framework, Semantic-Aware Facial Expression Recognition framework (*SAFE*), which is developed from an Information Bottleneck (IB)-inspired perspective to improve the robustness and prediction reliability of DFER in complex, unconstrained scenarios. Specifically, we first design a Temporal-aware Augmentation Module (TAM) to introduce structurally perturbed yet temporally coherent training samples, effectively mitigating spatio-temporal feature imbalance. Then, to ensure stable long-range modeling under temporal variation, we introduce the Spatio-temporal Modeling Module (STM) with a sparsity-aware state-space fusion gate. Furthermore, an Ambiguity-aware Calibration Loss (ACL) is formulated to dynamically refine decision boundaries by focusing on confusing and underrepresented categories, improving the model's resilience to distributional skew and semantic uncertainty. Extensive experiments on two large-scale in-the-wild DFER benchmarks, DFEW and FERV39k, demonstrate that *SAFE* consistently outperforms state-of-the-art methods across multiple metrics, particularly under ambiguous and imbalanced conditions. These results validate the effectiveness of our approach in promoting more robust and stable expression recognition, which is important for trustworthy DFER in real-world environments. Codes are released at https://github.com/QIcita/SAFE_DFER.**

***Index Terms*—Trustworthy facial expression recognition, information bottleneck, emotion ambiguity, category imbalance.**

## I. Introduction

FACIAL expression recognition plays a crucial role in emotional communication and human-computer interaction and has wide applications in fields such as psychological diagnosis [1]–[3] and intelligent security systems [4]–[7]. Although static facial expression recognition (SFER) has achieved impressive progress in recent years [8]–[10], its inability to model temporal dynamics limits its effectiveness in capturing contextual emotional cues, making it less suitable for real-world scenarios where expressions evolve continuously over time. This motivates growing interest in dynamic facial expression recognition (DFER), which seeks to model temporal transitions of emotions for more accurate and reliable facial behavior analysis in real-world scenarios. Recent works in trustworthy vision systems [11]–[13] have highlighted the importance of building privacy-preserving and interpretable facial behavior models for secure human-centered applications. However, these studies mainly focus on static identity-level tasks, leaving temporal expression modeling under ambiguity insufficiently addressed.

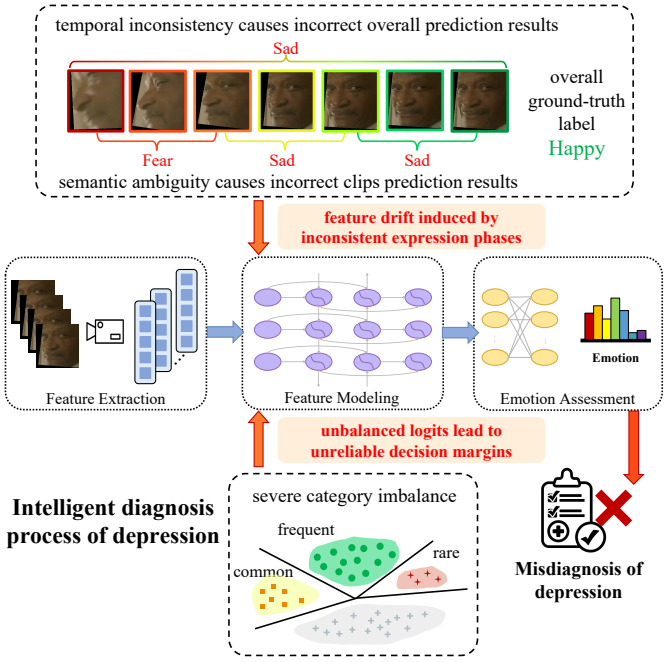



Fig. 1: The imbalance issues in wild dynamic facial expressions. (a) shows contextual feature intensity imbalance leading to error segment classification. (b) shows the category imbalance in current dynamic facial expression datasets. They collaboratively exacerbate the model's fuzzy classification and inductive bias.

Early DFER approaches relied on handcrafted features, such as FACS+ [14] and HOG-3D [15], to encode facial motion. These methods inevitably struggled to generalize under complex variations and noise, which further causes misidentification and system security hazards. With the advancement of deep learning, data-driven DFER methods have gained increasing popularity. However, significant challenges still remain, *e.g.*, particularly in the context of building robust and reliable DFER systems for security-critical applications. A key underlying cause is the expressive deviation problem, which encompasses both semantic inconsistency and distributional irregularities in facial behavior. This problem manifests in multiple intertwined forms: 1)

Feng-Qi Cui is with the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition and the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China.

Feng-Qi Cui, Anyang Tong, Jinyang Huang, Meng Li, Dan Guo, Meng Wang are with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine and the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China.

Linsheng Huang is with the National Engineering Research Center for Agro-Ecological Big Data Analysis and Application, Hefei 230601, China.

Jie Zhang is with Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore.

Xin Yan is with Cylingo Group, Beijing 100086, China.

Corresponding author*: Jinyang Huang (Email: hjy@hfut.edu.cn).

Semantic ambiguity and spatio-temporal inconsistency reduce decision reliability and weaken model robustness. In the absence of fine-grained annotations, existing DFER datasets (*e.g.*, DFEW [16] and FERV39k [17]) typically assign a single emotion label to an entire video, ignoring intra-video transitions and variations in feature intensity. As shown in Fig. 1 (a), this coarse labeling leads to semantic ambiguity and blurred category boundaries, posing a major obstacle to stable and consistent classification. Furthermore, conventional data augmentation or multi-scale modeling strategies often disrupt temporal coherence [18], which can degrade the system's ability to make stable and reliable predictions under ambiguous or evolving input conditions. For many important DFER-driven applications, *e.g.*, intelligent monitoring or emotion-aware decision systems, instability may compromise serious safety issues, *e.g.*, inmates misidentified or depression misdiagnosed. 2) Severe category imbalance introduces inductive bias and undermines robustness in minority or high-risk categories. As illustrated in Fig. 1 (b), many real-world DFER datasets suffer from long-tailed category distributions. Models trained under such an imbalance often overfit dominant categories while neglecting rare or ambiguous expressions. This not only reduces generalization but also creates vulnerabilities in scenarios where accurate detection of rare emotional cues is essential for safety or human-centered responses, *e.g.*, the faint smile of a depressed person. Without explicit mechanisms to handle imbalance, models tend to develop biased decision boundaries and exhibit downgraded performance in uncertain or high-risk prediction cases [19], [20]. 3) High computational complexity limits scalability and secure deployment. While recent Transformer-based DFER architectures [21] demonstrate strong modeling capacity, their computational overhead makes them unsuitable for deployment in latency-sensitive or resource-constrained environments. In safety-critical systems, *e.g.*, on-device emotion monitoring or real-time human–machine interaction, efficient and robust models are essential to ensure responsiveness, interpretability, and secure integration.

To address the aforementioned challenges stemming from expressive bias, we propose a lightweight and robust DFER framework, *SAFE*. This framework is developed from an Information Bottleneck (IB)-inspired perspective, aiming to retain essential emotional cues while effectively filtering out redundant or disruptive representations. First, to alleviate semantic ambiguity and temporal inconsistency, we propose the Temporal-aware Augmentation Module (TAM). This module performs structure-preserving spatial mixing while maintaining frame-level alignment, which effectively guides the model to suppress unstable or redundant facial patterns and focus instead on temporally consistent, task-relevant cues. In addition, to handle category imbalance and promote stable decision boundary formation, the Ambiguity-aware Calibration Loss (ACL) is specifically designed. This loss integrates confidence-aware dynamic weighting with supervised contrastive learning, enhancing category separability in the feature space and improving robustness in recognizing ambiguous, minority, or high-risk emotional categories. Furthermore, to improve temporal modeling under expression drift, we introduce the Spatio-temporal Modeling Module (STM), which organically incorporates gated sparse activation and residual compression mechanisms. These designs act as computational bottlenecks, effectively filtering out irrelevant temporal noise while preserving emotionally salient dynamics, thereby reducing inference cost and supporting computationally efficient DFER in practical scenarios. Moreover, while prior research has explored robustness against adversarial spoofing [22]–[24] and security vulnerabilities in face authentication systems [25], [26], they primarily address static facial identity verification rather than dynamic expression understanding under temporal ambiguity and category imbalance. Our work complements this line of research by introducing a principled, dynamic modeling framework for trustworthiness-oriented facial expression recognition. Extensive experiments on two large-scale in-the-wild DFER datasets demonstrate that *SAFE* not only outperforms existing state-of-the-art methods in terms of recognition accuracy and decision stability but also shows favorable robustness and deployment efficiency, making it a promising solution for secure facial behavior understanding in real-world applications.

Overall, the main contributions of this work are as follows:

- We are among the first to investigate the IB perspective in the DFER task. Guided by the IB perspective, we construct a modular framework named *SAFE*, which explicitly targets expressive bias by compressing task-irrelevant features while preserving emotion-relevant information, laying an interpretable basis for reliable and efficient DFER modeling.
- To address semantic ambiguity and temporal inconsistency, by injecting spatial diversity with consistent temporal alignment, we propose a temporal-structure-preserving augmentation module, named TAM. TAM not only enables context-invariant feature learning but also guides the model to focus on emotionally stable cues, which enhances generalization under uncertain dynamics.
- To mitigate category imbalance and confusion, we design a dual-level optimization mechanism composed of STM and ACL. STM incorporates gated sparse activation and residual compression for robust long-range reasoning, while ACL employs confidence-adaptive contrastive supervision to refine decision boundaries and enhance discrimination for minority and ambiguous expressions. Together, these designs improve temporal representation learning and boundary calibration, leading to more reliable recognition under category imbalance and semantic confusion.
- Extensive experiments on DFEW and FERV39k benchmarks demonstrate that *SAFE* achieves state-of-the-art performance in terms of recognition accuracy, boundary stability, robustness across diverse scenarios, and computational complexity. These results further highlight its practicality and suitability for deployment in real-world facial expression understanding-based decision-making systems.

## II. Related Work

### A. Dynamic Facial Expression Recognition

DFER has made significant progress with the advent of deep learning and has found broad application in domains such as driver monitoring, intelligent surveillance, and emotion-aware human–machine interaction. Early DFER approaches mainly relied on 3D convolutional neural networks, such as C3D [27] and I3D [28], to jointly model spatio-temporal features. These models were later extended by combining CNN-based spatial encoders with sequential architectures such as RNNs or LSTMs [16] to enhance temporal reasoning. Although such methods improved

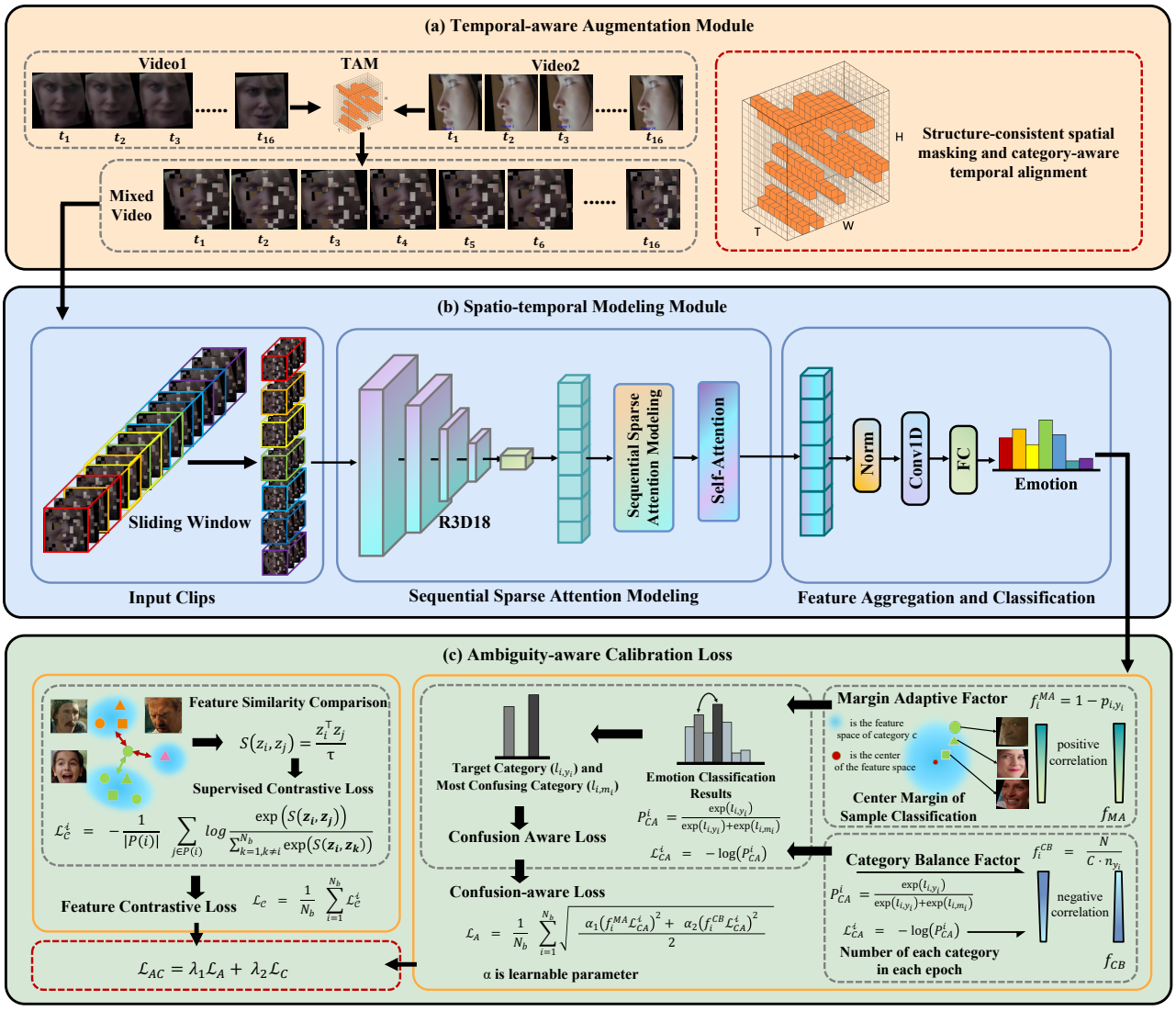



Fig. 2: An overview of the proposed *SAFE* framework. (a) The flow of the proposed Temporal-aware Augmentation Module (TAM). (b) The pipeline of Spatio-temporal Modeling Module (STM) in *SAFE*. (c) The sketch of Ambiguity-aware Calibration Loss (ACL).

short-term temporal modeling, they remained limited by fixed-length inputs, coarse temporal granularity, and weak clip-level supervision. In datasets such as DFEW [16] and FERV39k [17], a single emotion label is typically assigned to an entire video, ignoring intra-video variation in expression intensity. As a result, these early paradigms are still vulnerable to semantic ambiguity, subtle expression drift, and spatio-temporal misalignment, which undermines temporal stability and predictive confidence in real-world scenarios.

More recently, Transformer-based models [29] have shown strong long-range dependency modeling. For example, FormerDFER [21] uses self-attention to capture global context and alleviate occlusion effects. However, these models often over-attend to dominant regions and incur high computational cost, which limits their deployment in real-time or resource-constrained settings. To improve generalization under ambiguity, imbalance, and noisy supervision, recent DFER studies have explored attention regularization [30], semi-supervised pretraining [31], [32], multimodal fusion [33], robust learning [34], and cross-task transfer from static FER [35]. Methods such as IAL [19] and M3DFEL [36] alleviate label confusion and feature redundancy, while ExpLLM [37] and FEALLM [38] enhance semantic reasoning for FER. Nevertheless, these methods do not explicitly address semantic ambiguity, intra-video imbalance, and decision-boundary calibration within a unified DFER framework.

Our prior work HDF [11] improved DFER from the perspective of distributional robustness by introducing adversarial perturbations and Wasserstein-regularized attention. However, HDF mainly focuses on sample-level distribution shifts and does not provide dedicated mechanisms for suppressing fine-grained semantic ambiguity, correcting intra-video imbalance, or stabilizing decision boundaries. By contrast, *SAFE* addresses these issues through temporally consistent augmentation, lightweight spatio-temporal modeling, and ambiguity-aware calibration.

### *B. Data Augmentation in DFER*

In security-critical and uncertainty-prone scenarios, data augmentation serves not only as a regularization technique but also as a crucial means to enhance the robustness and reliability of DFER systems. Spatial-only augmentation strategies, *e.g.*, Mixup [39],

CutMix [40], and FaceMixup [41], have been shown to improve performance in SFER tasks by increasing semantic diversity. However, these methods are not directly transferable to DFER due to their disregard for temporal consistency, which is essential for modeling expression transitions across video frames. When directly extended to video sequences, such spatial perturbations may disrupt temporal coherence and induce semantic drift, making the evolving expression trajectory less compatible with the clip-level label. For instance, frame-wise mixing strategies like 3D-DSwin [18] introduced uncontrolled spatial perturbations that disregarded emotional continuity, while curriculum-based slicing in TACL [42] failed to adaptively capture transient or ambiguous states, both leading to unstable predictions and reduced reliability under real-world deployment conditions.

Recent advances in self-supervised learning have shown that structure-aware masking, *e.g.*, MAE [43], can guide models to focus on contextually relevant information. In particular, extensions of this concept to the spatio-temporal domain [44], [45] have offered a promising direction for designing augmentations that preserve temporal semantics while suppressing redundant or misleading cues. From an IB perspective, such temporally consistent perturbations can be viewed as a way to weaken task-irrelevant spatial redundancy while preserving emotion-relevant temporal dynamics. Following this intuition, our TAM adopts temporally aligned spatial mixing to introduce controlled structural perturbations during training, encouraging the model to suppress nuisance spatial variations while retaining temporally consistent and task-relevant emotional cues. Specifically, TAM introduces controlled cross-sample perturbations that promote context-invariant and emotion-relevant pattern learning, while reducing overreliance on dominant but potentially spurious regions. By explicitly addressing spatio-temporal feature imbalance during training, TAM not only enhances generalization but also strengthens the model's capacity to handle ambiguous or underrepresented expressions, contributing to the overall robustness and computational efficiency of DFER systems in unconstrained environments.

### C. Information Bottleneck in Visual Representation Learning

IB principle [46] offers a theoretical foundation for learning compressed yet task-relevant representations by maximizing mutual information with the target while minimizing that with the input. This dual objective enables models to suppress irrelevant or noisy content while preserving discriminative features. In DFER, where facial sequences often contain semantic redundancy, contextual noise, and ambiguous labels, the IB principle is particularly well-suited. It provides a principled way to filter spurious variations while retaining emotionally salient cues, thereby mitigating expressive bias and improving decision stability.

IB has been increasingly used to explain why compact intermediate representations can improve robustness and generalization in deep visual learning [47]–[49]. In particular, recent studies have shown that IB-style compression is closely related to improved generalization in deep networks [50], [51]. However, its role in DFER remains largely underexplored, where semantic ambiguity, temporal inconsistency, and category imbalance make the selective preservation of emotion-relevant information especially important. In *SAFE*, this perspective is reflected in a shared design principle across TAM, STM, and ACL, where

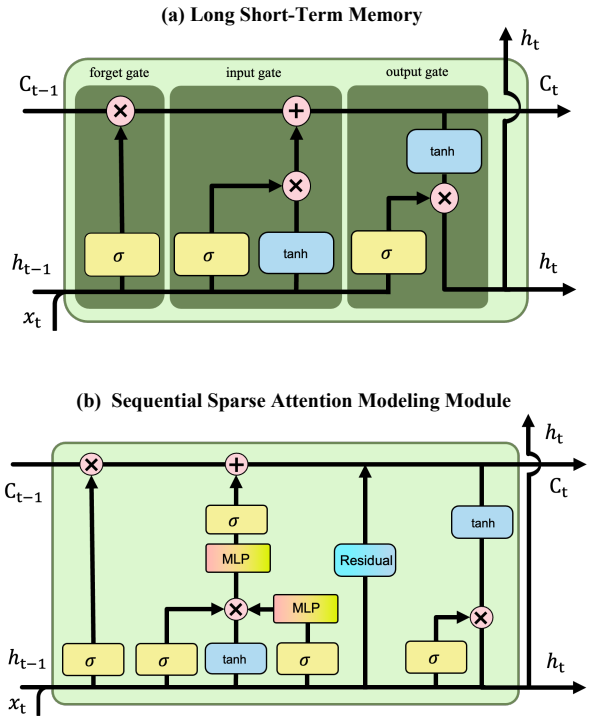



Fig. 3: Comparison between (a) LSTM and (b) Sequential Sparse Attention Modeling Module structures.

nuisance information is progressively suppressed while emotion-relevant cues are retained for recognition. More specifically, TAM reduces unstable spatial redundancy through temporally aligned perturbation, STM filters irrelevant temporal dynamics while preserving salient emotional trajectories, and ACL promotes more discriminative feature organization and clearer decision boundaries by compressing nuisance variation and preserving label-relevant distinctions. This formulation does not impose a separate variational IB objective, but instead incorporates the IB perspective into the functional design of the three modules, providing a more coherent explanation of how *SAFE* improves robustness and reliability under real-world uncertainty.

## III. Method

### A. Overview

The overall architecture of the proposed framework, *SAFE*, is illustrated in Fig. 2. Guided by the IB perspective, *SAFE* is designed as a modular and reliability-oriented solution to systematically mitigate expressive bias in DFER, which is mainly reflected in semantic ambiguity, temporal inconsistency, and category imbalance. It consists of three synergistic components, the augmentation module TAM, the modeling module STM, and the calibration module ACL. Specifically, TAM performs structure-preserving spatial mixing with temporal alignment to inject controlled perturbations that encourage the model to learn context-invariant and emotion-relevant features, thereby acting as an implicit information compressor. STM further refines temporal reasoning by employing gated sparse activation and residual compression, filtering out redundant dynamics while retaining salient expression trajectories, thus supporting efficient and robust long-term modeling under evolving affective states. Finally, ACL integrates confidence-guided dynamic weighting with supervised

**Algorithm 1** Workflow of TAM

**Require**: Training batch $\mathcal{B} = \{(I_i, y_i)\}_{i=1}^{N_b}$, dimensions $(C, T, H, W)$
**Require**: application ratio $\lambda_{TAM}$, mask ratio $\rho$, block size $B$
**Output**: Augmented batch $\mathcal{B}_{\text{mix}}$

Select $N_{TAM} = \lfloor \lambda_{TAM} N_b \rfloor$ clips from $\mathcal{B}$ for augmentation
**for** each selected clip $(I_u, y)$ **do**
  Sample a category-consistent paired clip $I_s$ satisfying $\text{label}(I_s) = y$
  $n_p = \left(\frac{H}{B}\right) \times \left(\frac{W}{B}\right)$ {*number of patches per frame*}
  $n_m = \text{int}(\rho \times n_p)$ {*number of masked patches*}
  $m = [0] \times (n_p - n_m) + [1] \times n_m$ {*initialize the patch-level mask*}
  $m \leftarrow \text{Shuffle}(m)$
  $\hat{m} = \text{reshape}\left(m, \frac{H}{B}, \frac{W}{B}\right)$
  $M = \text{repeat}(\hat{m}, B, B)$ {*expand each patch to a $B \times B$ spatial block*}
  $\hat{M} = \text{tile}(M, T)$ {*replicate the same spatial mask across all frames*}
  $I_{\text{mix}} = I_u \odot (1 - \hat{M}) + I_s \odot \hat{M}$ {*temporally aligned spatial mixing*}
  Replace $I_u$ with $I_{\text{mix}}$ in $\mathcal{B}_{\text{mix}}$
**end for**
Keep the remaining clips unchanged
**return** $\mathcal{B}_{\text{mix}}$

contrastive learning to adaptively calibrate decision boundaries, enhancing discrimination for ambiguous or minority categories and promoting more stable and reliable predictions in uncertain scenarios.

Taken together, these components suppress irrelevant variability and preserve task-critical emotional cues, forming a coherent and deployable framework for robust DFER in safety-sensitive environments.

### *B. Temporal-aware Augmentation Module*

In DFER tasks, facial expressions unfold over time, yet spatial feature distributions often exhibit localized redundancy and intensity imbalance. This spatio-temporal inconsistency, especially under real-world noise and semantic ambiguity, undermines the extraction of stable emotional cues and may cause the model to overfit dominant but unreliable facial regions. To alleviate this issue, we propose TAM, a structure-aware augmentation strategy that introduces controlled perturbations while preserving temporal coherence.

TAM performs spatial mixing under strict temporal alignment, allowing informative variations to be introduced without breaking the overall trajectory of expression evolution. As illustrated in Algorithm 1, given a training batch, TAM first selects a proportion $\lambda_{TAM}$ of clips for augmentation and keeps the remaining clips unchanged. For each selected clip, TAM samples a category-consistent paired clip from the same emotion category and replaces masked spatial regions with the corresponding regions from the paired clip. The same spatial mask is shared across all frames, thereby preserving temporal continuity and structural integrity. By applying category-consistent perturbations with a shared temporal mask, TAM reduces the model's reliance on unstable local regions and encourages it to focus on temporally coherent emotional cues. This behavior aligns with the IB perspective, where task-irrelevant variations are suppressed while label-relevant information is preserved. Compared with conventional static augmentations that may disrupt temporal dynamics, TAM preserves the overall temporal structure of facial expression evolution while enriching local spatial variation. By discouraging the encoder from relying on spatially redundant details and spurious local responses, this information-filtering process helps disentangle emotion-relevant cues from nuisance information, thereby improving representation robustness.

By restricting the paired clip to the same category as the selected clip, TAM avoids additional label inconsistency while expanding intra-class variation, thereby reducing sensitivity to spurious correlations in ambiguous or underrepresented categories. In summary, TAM improves generalization, reduces overfitting, and serves as the first step in suppressing expressive bias within our framework.

### *C. Spatio-temporal Modeling Module*

Reliable temporal modeling is essential for robust and reliable DFER, particularly in real-world settings where facial expression trajectories are often ambiguous, noisy, or unevenly distributed over time. To address this issue, we propose the STM, which aims to extract temporally discriminative and context-aware features while suppressing redundant or spurious fluctuations. STM is built upon a unified IB-guided perspective, where emotionally relevant dynamics are retained while nuisance variability is filtered out via structure-aware sparsity and fusion.

*1) Sliding Window-based Temporal Fragmentation:* To maintain local motion continuity and improve modeling of transient emotional signals, we first segment the input sequence $[B, C, T, H, W]$ into a set of overlapping clips using a sliding window strategy. Given window size $T_w$ and stride $T_s$, the number of windows can be expressed as:

$$N_w = \left\lfloor \frac{T - T_w}{T_s} \right\rfloor + 1. \tag{1}$$

This segmentation preserves the short-term temporal context within each clip while introducing cross-window redundancy to capture inter-segment consistency. It also mitigates sensitivity to abrupt expression transitions or frame-level noise, which are common under unconstrained conditions.

*2) Sequential Sparse Attention Modeling:* To model temporal dependencies more robustly and reduce expressive drift, we propose the Sequential Sparse Attention Modeling Module (S2AM2), as illustrated in Fig. 3 (b). This module integrates three core mechanisms inspired by information-theoretic principles of sparsity and redundancy reduction.

Firstly, sparse gating is employed to dynamically learn a temporal mask $\boldsymbol{M}_t$ that selectively filters out non-informative temporal states. The gating function can be defined as follows:

$$\boldsymbol{M}_t = \sigma(\text{MLP}(\boldsymbol{h}_{t-1})), \quad \tilde{\boldsymbol{c}}_t = \tilde{\boldsymbol{c}}_t \odot \boldsymbol{M}_t, \tag{2}$$

which encourages the model to attend to emotionally salient transitions while discarding irrelevant fluctuations.

Secondly, state-aware feature scaling then modulates the contributions of temporal features by applying learnable, context-sensitive weights. The update rule is given by:

$$\boldsymbol{c}_t = (\boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t) \odot \boldsymbol{w}_a, \tag{3}$$

where the adaptive scaling weight $\boldsymbol{w}_a$ enhances informative dimensions while suppressing ambiguous or redundant ones, thereby improving the model's robustness to feature imbalance.

Thirdly, residual state fusion is introduced to inject historical memory via residual pathways, mitigating gradient vanishing and stabilizing long-range temporal learning, and is implemented as:

$$\boldsymbol{c}_t = \boldsymbol{c}_t + \mathrm{Res}(\boldsymbol{h}_{t-1}), \tag{4}$$

$$\mathrm{Res}(\boldsymbol{h}_{t-1}) = \sigma(\boldsymbol{W}_r \boldsymbol{h}_{t-1}) \odot \mathrm{ReLU}(\boldsymbol{U}_r \boldsymbol{h}_{t-1}). \tag{5}$$

This operation effectively ensures smoother memory evolution across fluctuating expression patterns and significantly contributes to more coherent temporal representation.

To further enhance temporal modeling granularity, a lightweight self-attention layer is integrated on top of S2AM2, which allows the network to reweight each time step based on semantic importance. This mechanism dynamically emphasizes emotionally informative segments while effectively suppressing noise-induced uncertainty. Through the combination of temporal fragmentation, selective attention, and context-aware fusion, STM constructs compact and resilient temporal features. It provides the foundation for stable and interpretable dynamic reasoning in DFER, particularly under safety-critical and unconstrained real-world conditions.

### *D. Ambiguity-aware Calibration Loss*

In dynamic and safety-critical scenarios, DFER models must remain reliable when confronted with ambiguous expressions, severe category imbalance, or uncertain feature distributions. However, existing loss functions often fail to account for inter-class semantic confusion or underrepresented instances, resulting in biased optimization, blurry decision boundaries, and degraded performance on rare or high-risk categories. To address these issues, we introduce the ACL, a dual-level objective that operates at both the logit and feature levels to enhance discriminative power while suppressing expressive bias.

*1) Logit-level Confusion-aware Calibration:* At the decision level, we first propose a Confusion-aware Loss (CAL) that explicitly penalizes the semantic overlap between the ground-truth class and its most confusing non-target class. For the $i$-th sample, let $y_i$ denote its ground-truth label and $\boldsymbol{\ell}_i \in \mathbb{R}^C$ denote the predicted logit vector over $C$ emotion categories. The most confusing non-target category is defined as $m_i = \arg\max_{c \neq y_i} \ell_{i,c}$, where $\ell_{i,c}$ denotes the logit of sample $i$ for class $c$. The sample-wise CAL is then formulated as:

$$\mathcal{L}_{CA}^{i} = -\log\left(\frac{\exp(\ell_{i,y_i})}{\exp(\ell_{i,y_i}) + \exp(\ell_{i,m_i})}\right), \tag{6}$$

where $\ell_{i,y_i}$ is the ground-truth-class logit and $\ell_{i,m_i}$ is the highest-scoring non-target logit. Unlike conventional softmax-based losses that consider all classes jointly, CAL focuses more directly on the local decision boundary where semantic confusion is most likely to occur, thereby encouraging sharper separation between highly entangled categories.

To further improve boundary robustness, we introduce two dynamic weighting factors inspired by uncertainty-aware and imbalance-sensitive learning:

$$f_i^{MA} = 1 - p_{i,y_i}, \quad f_i^{CB} = \frac{N}{C \cdot n_{y_i}}, \tag{7}$$

where $p_{i,y_i}$ denotes the predicted probability assigned to the ground-truth class of sample $i$, $N$ is the total number of training samples, $C$ is the number of categories, and $n_{y_i}$ denotes the number of training samples belonging to class $y_i$. Here, $f_i^{MA}$ acts as a margin-aware factor that emphasizes low-confidence samples, while $f_i^{CB}$ rebalances the gradient contribution toward minority categories in long-tailed distributions.

To produce the final adaptive logit-level loss, we combine these two factors through a learnable Root Mean Square fusion form:

$$\mathcal{L}_A = \frac{1}{N_b} \sum_{i=1}^{N_b} \sqrt{\frac{1}{2}\left[\alpha_1 \left(f_i^{MA} \mathcal{L}_{CA}^{i}\right)^2 + \alpha_2 \left(f_i^{CB} \mathcal{L}_{CA}^{i}\right)^2\right]}, \tag{8}$$

where $N_b$ denotes the batch size, and $\alpha_1$ and $\alpha_2$ are learnable scalar parameters. Based on sample confidence and category frequency, this formulation adaptively modulates the supervision intensity, improving boundary calibration under both semantic ambiguity and category imbalance.

*2) Feature-level Compactness via Contrastive Regularization:* While CAL helps distinguish semantically similar categories at the decision level, it may not fully regulate the feature distribution in the embedding space, especially for sparse or noisy categories. To address this issue, inspired by supervised contrastive learning, we further incorporate a Feature Contrastive Loss (FCL). Let $\boldsymbol{z}_i$ denote the normalized embedding of sample $i$, and let $P(i) = \{j \mid j \neq i, y_j = y_i\}$ denote the set of positive samples sharing the same label with sample $i$. FCL is formulated as:

$$\mathcal{L}_C = -\frac{1}{N_b} \sum_{i=1}^{N_b} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(S(\boldsymbol{z}_i, \boldsymbol{z}_j))}{\sum_{k=1, k \neq i}^{N_b} \exp(S(\boldsymbol{z}_i, \boldsymbol{z}_k))}, \tag{9}$$

where $S(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_j}{\tau}$ denotes the temperature-scaled cosine similarity, and $\tau$ is the temperature parameter. FCL pulls semantically similar features closer while pushing dissimilar ones apart, thereby improving intra-category compactness and inter-category separability. Such feature organization suppresses intra-category nuisance variation while preserving emotion-relevant distinctions that are critical for recognition, forming a compact and discriminative representation shaping process. As a result, the feature space supports clearer decision regions and reduces semantic drift in high-variance or minority categories.

*3) Final Objective:* We integrate the above terms with the standard cross-entropy loss to form the final objective:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_A + \lambda_2 \mathcal{L}_C, \tag{10}$$

where $\lambda_1$ and $\lambda_2$ are tunable weights balancing the two auxiliary components. The joint optimization simultaneously refines boundary sharpness, aligns feature structure, and mitigates bias, leading to more reliable predictions under real-world expressive uncertainty.

By integrating decision-level calibration and representation-level regularization, ACL acts as a reliability-oriented optimization objective. It improves boundary fidelity, controls overconfidence, and strengthens category discrimination under expressive bias, making it a key component of *SAFE* for ambiguity-aware and imbalance-aware DFER.

TABLE I: Comparison (%) of our *SAFE* with state-of-the-art methods on DFEW (Ha: happiness, Sa: sadness, Ne: neutral, An: anger, Su: surprise, Di: disgust and Fe: fear). **Bold**: Best, Underline: Second best.

| Method | Years | Accuracy of Each Emotion(%) | | | | | | | Metrics (%) | | FLOPs (G) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Hap. | Sad. | Neu. | Ang. | Sur. | Dis. | Fea. | WAR | UAR | |
| ResNet18+LSTM [16] | ACM MM'20 | 78.00 | 40.65 | 53.77 | 56.83 | 45.00 | 4.14 | 21.62 | 53.08 | 42.86 | 7.78 |
| EC-STFL [16] | ACM MM'20 | 79.18 | 49.05 | 57.85 | 60.98 | 46.15 | 2.76 | 21.51 | 56.51 | 45.35 | 8.32 |
| Former-DFER [21] | ACM MM'21 | 84.05 | 62.57 | 67.52 | 70.03 | 56.43 | 3.45 | 31.78 | 65.70 | 53.69 | 9.11 |
| Logo-Former [52] | ICASSP'23 | 85.39 | 66.52 | 68.94 | 71.33 | 54.59 | 0.00 | 32.71 | 66.98 | 54.21 | N/A |
| T-MEP [53] | T-CSVT'24 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 68.85 | 57.16 | 24.70 |
| CFAN-SDA [54] | T-CSVT'24 | **90.84** | 70.91 | 65.72 | 69.97 | 57.86 | 13.10 | 35.36 | 69.19 | 57.70 | N/A |
| GCA+IAL [19] | AAAI'23 | 87.95 | 67.21 | 70.10 | 76.06 | 62.22 | 0.00 | 26.44 | 69.24 | 55.71 | 9.63 |
| M3DFEL [36] | CVPR'23 | 89.59 | 68.38 | 67.88 | 74.24 | 59.69 | 0.00 | 31.63 | 69.25 | 56.10 | **1.65** |
| LG-DSTF [55] | T-MM'24 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 69.82 | 58.89 | N/A |
| CLIPER [56] | ICME'24 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 70.84 | 57.56 | N/A |
| KFE-SC [57] | Inf. Sci.'25 | 89.12 | 68.11 | **71.33** | **78.27** | 63.25 | 4.02 | 33.86 | 70.30 | 57.35 | N/A |
| RDFER [58] | T-BIOM'25 | 89.69 | 69.22 | 70.18 | 71.47 | 62.08 | 0.69 | 28.71 | 69.73 | 56.93 | N/A |
| HDF [11] | ACM MM'25 | 89.67 | 71.20 | 67.42 | 73.03 | **64.44** | 12.41 | 41.63 | 71.60 | 60.40 | N/A |
| *SAFE* (Ours) | - | 90.61 | **72.33** | 68.05 | 76.47 | 64.36 | **13.79** | **44.63** | **72.06** | **60.55** | 2.94 |

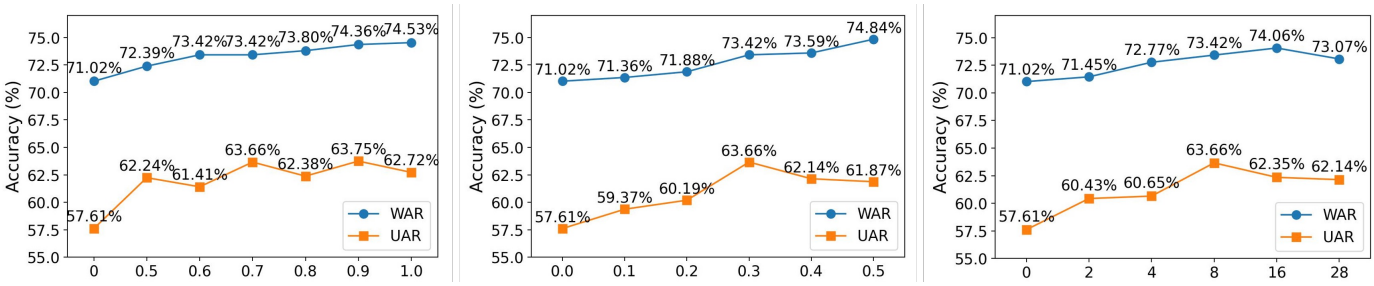



Fig. 4: Hyperparameters evaluation for TAM. left: application ratio, middle: mask ratio, right: mix pixel block size.

## IV. Experiment

### A. Experimental Setup

*1) Datasets:* We conduct extensive experiments on two popular in-the-wild DFER datasets, namely DFEW [16] and FERV39k [17]. DFEW is a large-scale dynamic expressions in-the-wild dataset introduced in 2020, containing over 16000 video clips. These clips are collected from more than 1500 films worldwide and include various challenging disturbances such as extreme lighting and pose changes. Each clip is annotated by ten trained annotators under professional guidance and assigned to one of seven basic expressions: happy, sad, neutral, angry, surprised, disgusted, and fearful. FERV39k is the largest available in-the-wild DFER dataset, containing 38935 video clips collected from 4 scenes, further subdivided into 22 fine-grained scenes. It is the first DFER dataset with large-scale 39K clips, scene-to-scene segmentation, and cross-domain support. Each video clip in FERV39k is annotated by 30 professional annotators to ensure high-quality labeling and assigned to one of the same seven main expressions as DFEW. We use the training and test sets provided by FERV39k for fair comparison.

*2) Metrics:* We conduct experiments on two in-the-wild DFER benchmarks, DFEW [16] and FERV39k [17]. To ensure fair comparison with previous methods, we adopt weighted average recall (WAR) and unweighted average recall (UAR) as the primary recognition metrics, which are defined as follows:

$$\mathrm{WAR} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} N_c}, \quad \mathrm{UAR} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{N_c}, \tag{11}$$

where $C$ is the number of categories, $TP_c$ denotes the number of correctly classified samples in class $c$, and $N_c$ denotes the total number of samples in class $c$. WAR reflects the overall recognition performance under the original data distribution, while UAR better measures balanced performance across categories under class imbalance. In addition, to evaluate reliability-related properties, we further report Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier Score.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{12}$$

where $n$ is the number of samples, $M$ is the number of confidence bins, $B_m$ denotes the $m$-th bin, and $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the empirical accuracy and average confidence of bin $B_m$, respectively.

$$\mathrm{NLL} = -\frac{1}{n}\sum_{i=1}^{n} \log p_{i,y_i}, \tag{13}$$

where $p_{i,y_i}$ is the predicted probability assigned to the ground-truth label $y_i$ of sample $i$.

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\left(p_{i,c} - \mathbb{I}(y_i = c)\right)^2, \tag{14}$$

where $p_{i,c}$ is the predicted probability of sample $i$ belonging to class $c$, and $\mathbb{I}(\cdot)$ is the indicator function. Since ECE, NLL, and

TABLE II: Comparison (%) of our *SAFE* with the state-of-the-art methods on FERV39k.

| Method | Accuracy of Each Emotion(%) | | | | | | | Metrics (%) | |
|---|---|---|---|---|---|---|---|---|---|
| | Hap. | Sad. | Neu. | Ang. | Sur. | Dis. | Fea. | WAR | UAR |
| 2C3D [17] | 54.85 | **52.91** | 60.67 | 31.34 | 5.96 | 2.36 | 6.96 | 41.77 | 30.72 |
| ResNet18+LSTM [17] | 61.91 | 31.95 | 61.70 | 45.93 | 14.26 | 0.00 | 0.70 | 42.59 | 30.92 |
| 2ResNet18+LSTM [17] | 59.00 | 45.87 | 61.90 | 40.15 | 9.87 | 1.71 | 0.46 | 43.20 | 31.28 |
| VGG13+LSTM [17] | 66.26 | 51.26 | 53.22 | 37.93 | 13.64 | 0.43 | 4.18 | 43.37 | 32.42 |
| 2VGG13+LSTM [17] | 69.65 | 47.31 | 52.55 | 47.88 | 7.68 | 1.93 | 2.55 | 44.54 | 32.79 |
| Former-DFER [21] | 65.65 | 51.33 | 56.74 | 43.64 | 21.94 | 8.57 | 12.53 | 46.85 | 37.20 |
| M3DFEL [36] | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 47.67 | 35.94 |
| Logo-Former [52] | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 48.13 | 38.22 |
| LG-DSTF [59] | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 48.19 | 39.84 |
| GCA+IAL [19] | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 48.54 | 35.82 |
| RDFER [58] | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 48.60 | 36.47 |
| KFE-SC [57] | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 49.37 | 37.56 |
| CFAN-SDA [54] | 69.65 | 49.46 | **62.21** | 46.20 | 22.26 | 11.56 | **15.55** | 49.48 | 39.56 |
| *SAFE* (Ours) | **72.39** | 52.34 | 56.26 | **50.27** | **28.78** | **18.52** | 12.41 | **50.15** | **40.48** |

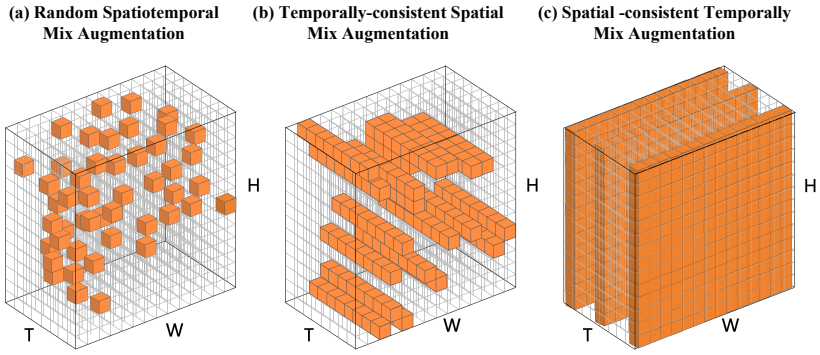



Fig. 5: Three spatio-temporal mix augmentation strategies in TAM.

TABLE III: Comparison of different spatio-temporal mix augmentation strategies in TAM.

| Setting | Method | WAR | UAR |
|---|---|---|---|
| a | None | 71.02 | 59.61 |
| b | Random | 71.79 | 59.99 |
| c | Frame | 68.88 | 60.04 |
| d | VideoMix-ST [60] | 72.35 | 62.43 |
| e | TAM | **73.42** | **63.66** |

Brier Score are all lower-is-better metrics, smaller values indicate better confidence calibration and more reliable probabilistic predictions.

*3) Implementation Details:* All experiments are conducted on a single NVIDIA RTX A6000 GPU, and all face images are resized to $112 \times 112$. We use augmentation techniques such as random cropping and horizontal flipping. For each video, 16 frames are extracted as samples. The feature extraction network is the standard R3D model with Torchvision pre-trained weights. The model is trained using the AdamW optimizer and cosine scheduler for 200 epochs. The learning rate is set to 5e-4, with a minimum learning rate of 5e-6, and weight decay is set to 0.05. For Sliding Window, the window size is set to 4 frames, and the sliding step size is 2. Empirically, the loss coefficients $\lambda_1$ and $\lambda_2$ for $L_A$ and $L_C$ are set to 0.05 and 0.001, respectively.

### *B. Comparison with State-of-the-art Methods*

*1) Results on DFEW:* As shown in Tab. I, our approach achieves the best performance on both WAR and UAR under the 5-fold cross-validation protocol, consistently outperforming previous methods. Notably, the gain is more evident in minority categories, which usually suffer from sparse supervision and semantic ambiguity. These results suggest that the proposed training strategy, centered on TAM, STM, and ACL, is effective in alleviating inter-category confusion and mitigating the impact of category imbalance. Such improvements indicate that *SAFE* helps preserve task-relevant emotional dynamics while suppressing redundant or noisy variations. This contributes to stronger recognition performance, especially in high-uncertainty or sample-scarce scenarios where conventional methods often produce unstable predictions.

*2) Results on FERV39k:* Tab. II reports the results on the more challenging FERV39k dataset. Despite its larger scale and more diverse real-world variations, *SAFE* achieves competitive performance on both WAR and UAR. In particular, the improvement in UAR indicates that our method is more effective in handling long-tailed distributions and reducing confusion among minority or ambiguous categories. These results further demonstrate the adaptability of *SAFE* to complex in-the-wild scenarios and highlight its practical advantages in terms of balanced recognition performance.

*3) Calibration Analysis:* We compare *SAFE* with the baseline in terms of ECE, NLL, and Brier Score during training. As shown in Fig. 6, we plot the epoch-wise ratio curves of Ours/Baseline for all three metrics, where a value below 1 means that our method yields lower calibration error and better probabilistic calibration than the baseline. After a brief early-stage fluctuation, *SAFE* achieves ratios below 1 for most epochs on all three metrics, with particularly clear improvements on ECE and NLL. The Brier ratio also remains consistently below 1 during the middle and late training stages, indicating better agreement between predicted confidence and actual outcomes. These observations suggest that *SAFE* produces better calibrated confidence estimates than the baseline and provides complementary evidence for the reliability-

oriented design of the proposed framework.

TABLE IV: Ablation study of TAM, S2AM2 of STM, and ACL in the proposed *SAFE* framework on DFEW fd5.

| Setting | Method | | | Metric (%) | |
|---|---|---|---|---|---|
| | TAM | S2AM2 | ACL | WAR | UAR |
| a | ✗ | ✗ | ✗ | 68.32 | 58.00 |
| b | ✓ | ✗ | ✗ | 70.85 | 63.18 |
| c | ✗ | ✓ | ✗ | 68.81 | 58.64 |
| d | ✗ | ✗ | ✓ | 70.16 | 58.73 |
| e | ✓ | ✓ | ✗ | 71.10 | 61.74 |
| f | ✓ | ✗ | ✓ | 73.42 | 62.01 |
| g | ✗ | ✓ | ✓ | 71.02 | 59.61 |
| h | ✓ | ✓ | ✓ | **73.42** | **63.66** |

### C. Ablation Studies

To validate the contribution of each module in our framework *SAFE*, we conduct ablation studies on Fold 5 of the DFEW dataset. The baseline system employs a conventional backbone combining Sliding Window segmentation, R3D encoder, BiLSTM temporal modeling, and Conv1D classifier, trained with standard cross-entropy loss.

*1) Effectiveness of Individual Modules:* Tab. IV presents the performance of various configurations by progressively integrating our proposed components. Introducing the TAM notably improves performance by mitigating spatial redundancy and overfitting to dominant facial regions. Specifically, by enforcing temporal consistency across frames while introducing controlled spatial perturbations, TAM enhances the model's robustness to contextual variation without disrupting the overall structure of expression evolution. This process acts as an implicit information-filtering mechanism, reducing the model's reliance on unstable or spurious spatial patterns while preserving task-relevant emotional cues. Besides, by implementing S2AM2 to replace the conventional BiLSTM with our STM, the system achieves consistent gains in temporal reasoning. Through sparse gating, residual fusion, and bi-directional attention mechanisms, S2AM2 enables the model to adaptively emphasize emotionally salient dynamics while suppressing temporal noise, thus improving coherence under ambiguous or evolving emotional trajectories. The ACL further boosts performance by explicitly optimizing the decision boundary under semantic confusion and category imbalance. It combines a confusion-aware logit loss with supervised contrastive regularization in the embedding space, guided by dynamically learned confidence and frequency-aware weights. This significantly enables the model to better distinguish between semantically overlapping categories and enhances generalization in low-frequency or high-risk emotion categories.

When these modules are integrated, their complementary nature becomes evident. By coupling augmentation-driven diversity with adaptive temporal filtering, the joint application of TAM and STM promotes stronger spatio-temporal representation. Combining TAM with ACL enables the model to refine classification boundaries while maintaining context-aware augmentation. Notably, even without spatial augmentation, the fusion of STM and ACL still achieves substantial improvements, demonstrating that adaptive temporal modeling and loss-level regularization alone can significantly enhance robustness under expressive drift.

TABLE VI: Ablation study of CAL and FCL in ACL on DFEW fd5.

| Setting | ACL | | Metric (%) | |
|---|---|---|---|---|
| | CAL | FCL | WAR | UAR |
| a | ✗ | ✗ | 71.10 | 61.74 |
| b | ✓ | ✗ | 73.12 | 61.31 |
| c | ✗ | ✓ | 73.07 | 61.78 |
| d | ✓ | ✓ | **73.42** | **63.66** |

Ultimately, the full integration of TAM, STM, and ACL achieves the best overall performance. These results indicate that our unified framework can effectively alleviate spatio-temporal inconsistency, improve boundary calibration under ambiguity, and mitigate the adverse impact of category imbalance in DFER.

*2) The Impact of Hyperparameters for TAM:* To analyze the robustness and adaptability of the proposed TAM, we conduct detailed experiments on three key hyperparameters, as illustrated in Fig. 4 (left to right). These include application ratio, which determines the proportion of training samples subjected to TAM-based augmentation, mask-ratio, which specifies the percentage of masked regions in the source video during spatial mixing, and mix-block size, which controls the granularity of spatial replacement patches.

The results indicate that appropriate levels of temporal-aware augmentation play a crucial role in balancing representation diversity and structural coherence. Specifically, a moderate application ratio improves generalization by introducing sufficient perturbation to discourage overfitting, while excessive perturbation can erode temporal integrity, which is especially detrimental in tasks where emotional continuity is semantically meaningful. Similarly, tuning the mask ratio to an intermediate range ensures that the spatial structure is sufficiently altered to enhance robustness, without introducing semantic noise that may compromise class boundaries. In terms of spatial granularity, medium-sized mixing blocks consistently yield superior results, providing a desirable trade-off between spatial diversity and local semantic preservation. These findings suggest that, when properly configured, TAM serves as an effective structured perturbation strategy that suppresses unstable patterns while preserving useful emotional dynamics.

Beyond hyperparameter tuning, we further evaluate the effectiveness of our augmentation design by comparing TAM with three alternative spatio-temporal mixing strategies, including random spatial mixing, frame-wise mixing, and a temporally consistent VideoMix [60] baseline. Here, the Random strategy is conceptually close to applying CutMix-style [61] spatial replacement independently to each frame, whereas VideoMix-ST uses a shared tube-shaped replacement region across frames. As shown in Fig. 5 and Tab. III, TAM (setting e) consistently outperforms Random (setting b), Frame (setting c), and VideoMix-ST (setting d) in terms of both WAR and UAR. These results suggest that, although temporal consistency is important, large contiguous tube replacement may excessively corrupt discriminative facial regions and weaken subtle local expression cues. By contrast, TAM introduces finer-grained and structure-preserving perturbations,

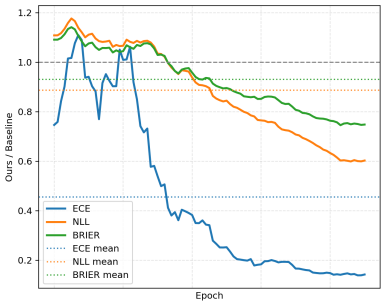



Fig. 6: Epoch-wise Ours/Baseline ratio curves of ECE, NLL, and Brier Score.

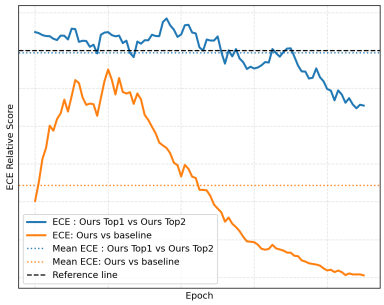



Fig. 7: Epoch-wise relative ECE comparison during training.

TABLE V: Sensitivity analysis of confusing-class selection strategies in ACL on DFEW fd5.

| Strategy | WAR | UAR |
|---|---|---|
| Top-1 (ours) | **73.42** | **63.66** |
| Top-2 | 71.78 | 62.41 |
| Top-3 | 71.62 | 62.15 |

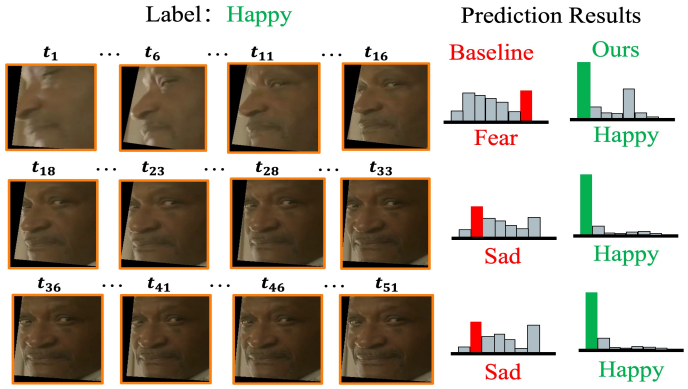



Fig. 8: Visualization of the effects of *SAFE* versus baseline.

which better maintain useful facial dynamics under ambiguous or imbalanced conditions.

*3) The Effectiveness of Two Proposed Factors in ACL:* We conduct extensive ablation studies on CAL and FCL in the proposed ACL to evaluate their individual contributions and synergistic effects in improving model performance. As shown in Table VI, introducing CAL alone (setting b) significantly improves WAR, indicating its effectiveness in dynamically perceiving confusing categories and optimizing classification boundaries. However, its limited improvement in UAR suggests that it is less effective in addressing category imbalance. In contrast, employing FCL alone (setting c) improves UAR without significantly boosting WAR, demonstrating the advantage of contrastive learning in enhancing the discrimination of minority categories and structuring the feature space. Finally, combining CAL and FCL (setting d) yields the best performance in both WAR and UAR, validating their complementary roles in dynamic confusion modeling and feature space optimization.

*4) Sensitivity to Confusing-Class Selection:* We further analyze the sensitivity of ACL to different confusing-class selection strategies. Specifically, we compare the default Top-1 strategy with Top-2 and Top-3 variants, where the top-$k$ non-target logits are aggregated when computing the confusion-aware term. As shown in Tab. V, the default Top-1 strategy achieves the best WAR and UAR among all compared settings, indicating that focusing on the most confusing class is already effective for boundary optimization in practice. To further examine the calibration behavior during training, Fig. 7 presents the epoch-wise relative ECE comparison between Top-1 and Top-2, together with the Ours-versus-baseline reference curve. Although Top-2 introduces a broader aggregation over confusing classes, Top-1 remains more competitive throughout training and yields better final recognition performance. These results suggest that the simple Top-1 design provides a favorable trade-off between effectiveness and stability, and is therefore adopted as the default setting in ACL.

TABLE VII: Robustness comparison of *SAFE* w/o TAM and *SAFE* under common perturbations on DFEW fd5. Results are reported as mean±std over five random seeds.

| Perturbation | *SAFE* w/o TAM | *SAFE* |
|---|---|---|
| | WAR / UAR | WAR / UAR |
| Clean | 70.84±0.55 / 59.28±0.73 | **72.08±0.43 / 62.69±0.58** |
| Gaussian | 67.21±0.78 / 55.03±0.91 | **69.63±0.57 / 59.34±0.70** |
| Frame drop | 67.57±0.69 / 56.49±0.84 | **69.92±0.51 / 59.47±0.66** |

*5) Robustness under Common Perturbations:* To further evaluate robustness-related properties under input uncertainty, we compare *SAFE* with *SAFE* w/o TAM under two common perturbation settings, Gaussian noise and Frame drop, and report the mean and standard deviation over five random seeds. Gaussian noise is added to all frames to simulate appearance-level corruption, while Frame drop randomly removes a portion of frames to test the sensitivity of temporal modeling to incomplete observations. As shown in Tab. VII, *SAFE* consistently outperforms *SAFE* w/o TAM under both perturbations in terms of mean WAR and UAR, while maintaining relatively small standard deviations across different runs. These results suggest that the temporally aligned structured perturbation introduced by TAM not only improves clean-set recognition, but also enhances robustness and stability under realistic noisy conditions.

## D. Visualization Analysis

*1) Visualization of SAFE Cases:* We first analyze visually ambiguous samples where emotion expressions are subtle, low-intensity, or context-dependent. As shown in Fig. 8, compared with the baseline, which yields low confidence and scattered predictions across categories, our method consistently assigns high-confidence predictions to the correct categories. This reflects *SAFE* 's ability to capture contextual emotional dependencies and suppress spurious or unstable cues. The improved confidence concentration and semantic alignment provide qualitative support for the effectiveness of our modules, particularly TAM and STM, under fuzzy conditions.

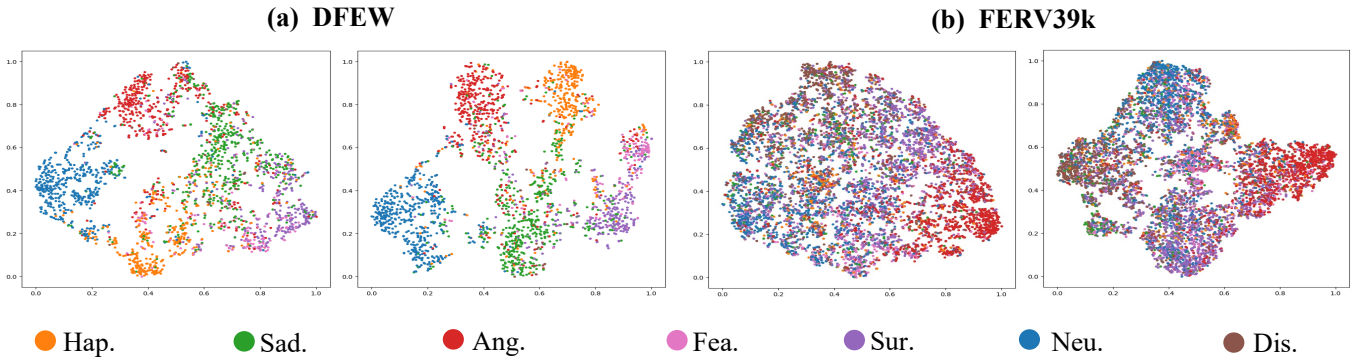



Fig. 9: Visualization of the feature distribution learned by the baseline (left) and our method (right) on two datasets.

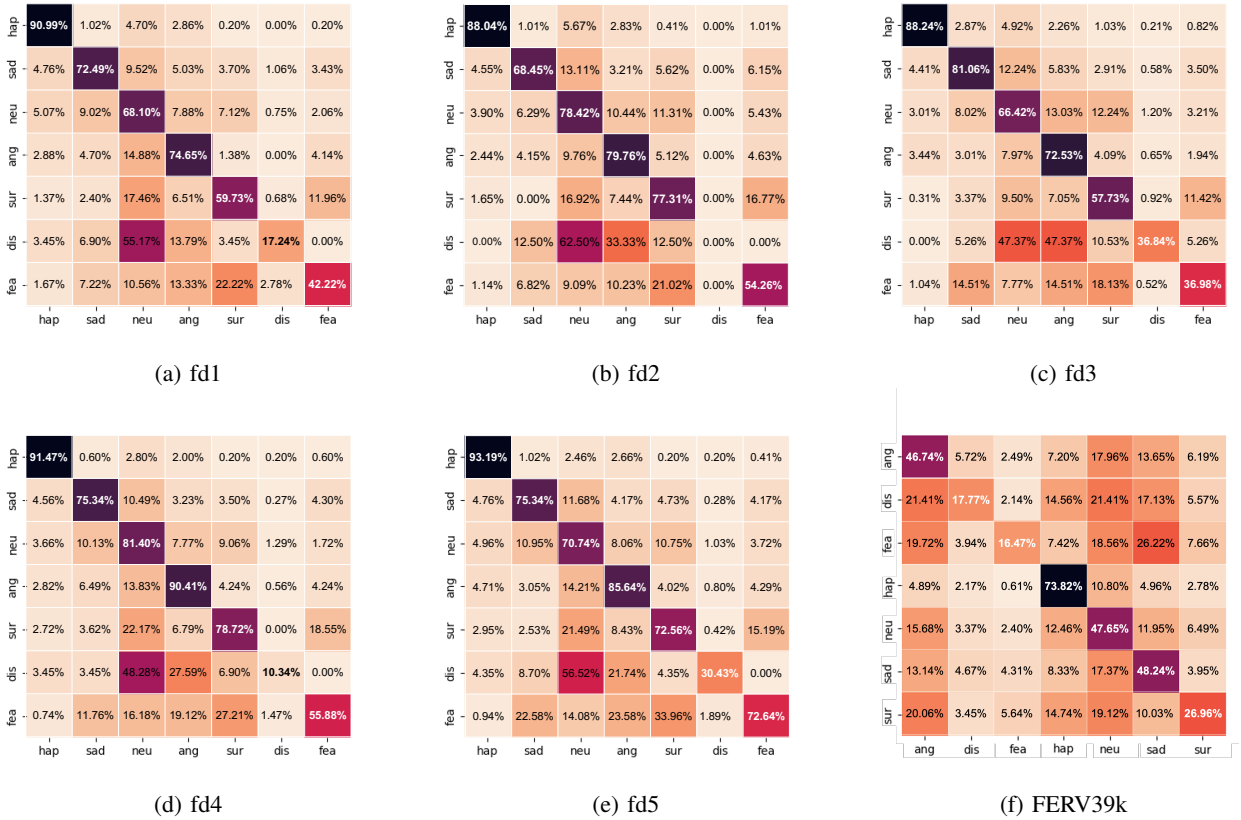



Fig. 10: The confusion matrix of our proposed *SAFE* evaluated on DFEW Fold 1-5 and FERV39k.

*2) t-SNE Feature Distribution:* To illustrate the effect of our method on feature separability and intra-category compactness, we apply the t-SNE method [62] to visualize the learned embeddings. As shown in Fig. 9, *SAFE* yields tighter clustering for intra-category samples and clearer boundaries across different emotion categories, especially for minority or overlapping categories. Compared with the baseline, which exhibits entangled and poorly separable clusters, our approach significantly improves category-specific coherence. This observation supports the role of ACL in enhancing feature-level discrimination through contrastive regularization. It is also consistent with our information-filtering motivation, where nuisance variations are reduced and task-relevant cues are organized into more compact and separable representations.

*3) Confusion Matrix Analysis:* We further visualize the aggregated confusion matrix across 5-fold cross-validation on DFEW in Fig. 10. The results indicate that *SAFE* reduces the overall confusion for minority categories such as Fear and Sad, demonstrating improved robustness to category imbalance. However, extreme underrepresented categories like Disgust still present challenges due to their sparse and ambiguous nature. Notably, we observe a slight drop in Neutral category accuracy, which we attribute to the model's rebalancing effect. Since Neutral often dominates in imbalanced datasets, conventional models tend to over-predict it for uncertain inputs. Our approach, in contrast, redistributes decision confidence across underrepresented or fuzzy categories, which may slightly reduce dominant-category accuracy while enhancing fairness and overall robustness. These

results provide further evidence that *SAFE* effectively mitigates expressive bias by improving feature compactness, boundary calibration, and contextual consistency in DFER.

## V. Conclusion

In this paper, we present *SAFE*, a reliability-oriented framework for DFER, designed to address spatio-temporal feature imbalance and semantic category ambiguity. Inspired by the IB perspective, *SAFE* aims to suppress nuisance variations while preserving emotion-relevant cues through three synergistic components, TAM, STM, and ACL. TAM introduces structure-consistent spatial perturbations while maintaining temporal alignment, enhancing feature diversity without disrupting emotional continuity. STM employs sparsity-aware attention and residual compression to extract context-sensitive emotional dynamics, enabling stable and efficient long-term modeling. ACL jointly calibrates decision boundaries and regularizes feature representations via uncertainty-aware reweighting and supervised contrastive learning, improving robustness under ambiguous or minority conditions. Extensive experiments show that *SAFE* achieves state-of-the-art performance in terms of recognition accuracy and balanced classification performance, while also providing improved calibration, robustness, and computational efficiency under real-world uncertainty. These results support the effectiveness of *SAFE* as a practical framework for more reliable dynamic facial expression recognition. In future work, we plan to explore frame-level uncertainty annotation, adaptive cross-domain generalization, and multi-modal fusion with audio or physiological signals to further improve the reliability and applicability of dynamic emotion understanding systems.

## VI. Acknowledgments

This work is supported by Fundamental Research Funds for the Central Universities (Grant No. JZ2025HGTB0225), Major Scientific and Technological Project of Anhui Provincial Science and Technology Innovation Platform (Grant No. 202305a12020012), National Natural Science Foundation of China (Grant No. 62302145).

## References

[1] R. Wang, J. Huang, J. Zhang, X. Liu, X. Zhang, Z. Liu, P. Zhao, S. Chen, and X. Sun, "Facialpulse: An efficient RNN-based depression detection via temporal facial landmarks," in *ACM Multimedia*, 2024.

[2] J. Kim, A. B. J. Teoh *et al.*, "Flexible secure biometrics: A protected modality-invariant face-periocular recognition system," *IEEE Transactions on Information Forensics and Security*, 2025.

[3] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "An enhanced adaptive confidence margin for semi-supervised facial expression recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[4] Y. Du, C. Lei, Z. Zhao, Y. Dong, and F. Su, "Video-based visible-infrared person re-identification with auxiliary samples," *IEEE Transactions on Information Forensics and Security*, 2024.

[5] S. Liu, Y. Zhang, T. Wang, Z. Zhan, and H. Jin, "Efficient privacy-preserving facial expression classification," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[6] L. Yuan, W. Chen, X. Pu, Y. Zhang, H. Li, Y. Zhang, X. Gao, and T. Ebrahimi, "Pro-face c: Privacy-preserving recognition of obfuscated face via feature compensation," *IEEE Transactions on Information Forensics and Security*, 2024.

[7] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services," *Information Sciences*, 2019.

[8] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, 2023.

[9] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Discriminant functional learning of color features for the recognition of facial action units and their intensities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] L. Liang, C. Lang, Y. Li, S. Feng, and J. Zhao, "Fine-grained facial expression recognition in the wild," *IEEE Transactions on Information Forensics and Security*, 2021.

[11] F.-Q. Cui, A. Tong, J. Huang, J. Zhang, D. Guo, Z. Liu, and M. Wang, "Learning from heterogeneity: Generalizing dynamic facial expression recognition via distributionally robust optimization," in *Proceedings of the 33nd ACM International Conference on Multimedia*, 2025.

[12] J. Gao and et al., "Trading-off privacy, utility and explainability in deep learning-based image data analysis," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[13] J. Huang, Y. Feng, F.-Q. Cui, X. Zhang, Z. Liu, X. Liu, J. Liu, F. Zhang, and M. Li, "Identifying who you are no matter what you write through abstracting handwriting style," *IEEE Transactions on Dependable and Secure Computing*, 2026.

[14] I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[15] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, 2008.

[16] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[17] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] R. Wang and X. Sun, "Dynamic facial expression recognition based on vision transformer with deformable module," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023.

[19] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, 2023.

[20] B. Lee, H. Shin, B. Ku, and H. Ko, "Frame level emotion guided dynamic facial expression recognition with emotion grouping," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[21] Z. Zhao and Q. Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[22] M. Luo, H. Wu, H. Huang, W. He, and R. He, "Memory-modulated transformer network for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, 2022.

[23] S. Mao and H. Li, "Hyperbolic metric learning for generalizable face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, 2025.

[24] X. Long, J. Zhang, and S. Shan, "Confidence aware learning for reliable face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, 2025.

[25] T. Wang, W. Wen, X. Xiao, Z. Hua, Y. Zhang, and Y. Fang, "Beyond privacy: Generating privacy-preserving faces supporting robust image authentication," *IEEE Transactions on Information Forensics and Security*, 2025.

[26] J. Wei, H. Huang, Y. Wang, R. He, and Z. Sun, "Towards more discriminative and robust iris recognition by learning uncertain factors," *IEEE Transactions on Information Forensics and Security*, 2022.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015.

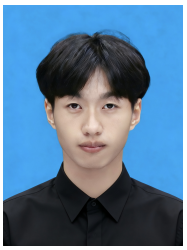
[28] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[30] C. Huang, F. Jiang, Z. Han, X. Huang, S. Wang, Y. Zhu, Y. Jiang, and B. Hu, "Modeling fine-grained relations in dynamic space-time graphs for video-based facial expression recognition," *IEEE Transactions on Affective Computing*, 2025.

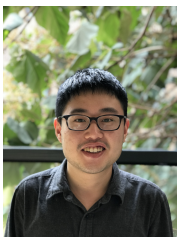
[31] J. Yu, Z. Wei, Z. Cai, G. Zhao, Z. Zhang, Y. Wang, G. Xie, J. Zhu, W. Zhu, Q. Liu, and J. Liang, "Exploring facial expression recognition through semi-supervised pre-training and temporal modeling," in *2024 IEEE/CVF Con-*

*ference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.
[32] F.-Q. Cui, Z. Lin, X. Rao, A. Tong, S. Li, F. Wang, C. Chen, and B. Liu, "Micacl: Multi-instance category-aware contrastive learning for long-tailed dynamic facial expression recognition," in *IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 2025.
[33] E. Ryumina, M. Markitantov, D. Ryumin, H. Kaya, and A. Karpov, "Zero-shot audio-visual compound expression recognition method based on emotion probability fusion," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.
[34] F. Liu, H. Wang, and S. Shen, "Robust dynamic facial expression recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.
[35] Y. Chen, J. Li, Y. Zhang, Z. Hu, S. Shan, M. Wang, and R. Hong, "Static for dynamic: Towards a deeper understanding of dynamic facial expressions using static expression data," *IEEE Transactions on Affective Computing*, 2025.
[36] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
[37] X. Lan, J. Xue, J. Qi, D. Jiang, K. Lu, and T.-S. Chua, "Expllm: Towards chain of thought for facial expression recognition," *IEEE Transactions on Multimedia*, 2025.
[38] Z. Hu, K. Yuan, X. Liu, Z. Yu, Y. Zong, J. Shi, H. Yue, and J. Yang, "Feallm: Advancing facial emotion analysis in multimodal large language models with emotional synergy and reasoning," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.
[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
[40] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
[41] F. A. Faria, M. M. Souza, R. F. da S. Teixeira, and M. P. Segundo, "Facemixup: Enhancing facial expression recognition through mixed face regularization," 2024. [Online]. Available: https://arxiv.org/abs/2405.20259
[42] A. Tong, C. Tang, and W. Wang, "Semi-supervised action recognition from temporal augmentation using curriculum learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
[43] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[44] C. Feichtenhofer, H. Fan, and Y. L. K. He, "Masked autoencoders as spatiotemporal learners," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2022.
[45] J. Huang, J.-X. Bai, X. Zhang, Z. Liu, Y. Feng, J. Liu, X. Sun, M. Dong, and M. Li, "Keystrokesniffer: An off-the-shelf smartphone can eavesdrop on your privacy from anywhere," *IEEE Transactions on Information Forensics and Security*, 2024.
[46] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 2000.
[47] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.
[48] M. Sefidgaran, A. Zaidi, and P. Krasnowski, "Generalization guarantees for representation learning via data-dependent gaussian mixture priors," in *The Thirteenth International Conference on Learning Representations*, 2025.
[49] Z. Wang, S.-L. Huang, E. E. Kuruoglu, J. Sun, X. Chen, and Y. Zheng, "PAC-bayes information bottleneck," in *International Conference on Learning Representations*, 2022.
[50] K. Kawaguchi, Z. Deng, X. Ji, and J. Huang, "How does information bottleneck help deep learning?" in *Proceedings of the 40th International Conference on Machine Learning*, 2023.
[51] S. Wang and S. E. Palmer, "Towards understanding neural collapse in supervised contrastive learning with the information bottleneck method," *arXiv preprint arXiv:2305.11957*, 2023.
[52] F. Ma, B. Sun, and S. Li, "Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition," *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
[53] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[54] D. Chen, G. Wen, P. Yang, H. Li, C. Chen, and B. Wang, "Cfan-sda: Coarse-fine aware network with static-dynamic adaptation for facial expression recognition in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[55] Z. Zhang, X. Tian, Y. Zhang, K. Guo, and X. Xu, "Label-guided dynamic spatial-temporal fusion for video-based facial expression recognition," *IEEE Transactions on Multimedia*, 2024.
[56] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Cliper: A unified vision-language framework for in-the-wild facial expression recognition," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024.
[57] S. Yan, Y. Wang, X. Mai, Q. Zhao, W. Song, J. Huang, Z. Tao, H. Wang, S. Gao, and W. Zhang, "Observe finer to select better: Learning key frame extraction via semantic coherence for dynamic facial expression recognition in the wild," *Information Sciences*, 2025.
[58] F. Liu, H. Wang, and S. Shen, "Robust dynamic facial expression recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.
[59] Z. Zhang, X. Tian, Y. Zhang, K. Guo, and X. Xu, "Label-guided dynamic spatial-temporal fusion for video-based facial expression recognition," *IEEE Transactions on Multimedia*, 2024.
[60] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, "Videomix: Rethinking data augmentation for video classification," *arXiv preprint arXiv:2012.03457*, 2020.
[61] J. Oh and C. Yun, "Provable benefit of cutout and cutmix for feature learning," *Advances in Neural Information Processing Systems*, 2024.
[62] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, 2008.



**Feng-Qi Cui** (Student Member, IEEE) is currently pursuing M.E. degree in Computer Science at University of Science and Technology of China (USTC), Hefei, China. He is also affiliated with Anhui Provincial Key Laboratory of Affective Computing and Advanced Intelligent Machines, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. His research interests include Computer Vision, Multimedia Information Processing and Affective Computing.

**Anyang Tong** (Student Member, IEEE) was born in 1998. He is currently pursuing a Ph.D. degree with the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). During his studies, he interned at the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, where he worked on affective computing. His research interests include semi-supervised learning, reliable computing, and multimodal retrieval-augmented generation.

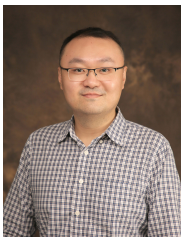


**Jinyang Huang** is an Associate Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). His research interests include Multimodal Perception, Human-computer Interaction, Wireless Security, and Signal Processing. In this area, he has published 60 papers in international peer-reviewed journals and conferences, including ToN, TMC, TIFS, TDSC, MobiCom, IEEE S&P, USENIX Security, Ubicomp, NeurIPS, INFOCOM, and ACM MM. He has served as a TPC member for conferences, including ACM MM, IEEE ICME, and Globecom, and has the honor of becoming ACM MM 2024 Outstanding Reviewers. He is a Guest Editor for the Technical Committee of ICME 25 Special Session. He is the recipient of the Young Scientist of Anhui Computer Federation and IEEE HITC Distinguished PhD Dissertation Award.

**Jie Zhang** (Member, IEEE) received his B.S. degree in 2017 from China University of Geosciences, Beijing and received his Ph.D.degree in 2022 from the University of Science and Technology of China (USTC). Currently, he is a Research Scientist of Center for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore. Before that, he was a research fellow at Nanyang Technological University. His primary research interests include IP protection for AI, Trustworthy generative AI, and AI Regulation.

**Xin Yan** received his bachelor's degree in Automation from East China University of Science and Technology. He is currently a Senior algorithm engineer at Cylingo Group. His research interests include automatic multimodal large models, human-computer interaction (HCI), and affective computing.

**Linsheng Huang** received the Ph.D. degree in Circuits and Systems from Anhui University, China in 2013. He is currently with the National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, China. His current research interests include remote sensing information processing, technology and applications of internet of things.



**Meng Li** (Senior Member, IEEE) is a Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China. He was a Post-Doc Researcher at the Department of Mathematics and HIT Center, University of Padua, Italy, where he is with the Security and PRIvacy Through Zeal (SPRITZ) research group led by Prof. Mauro Conti (IEEE Fellow). He obtained his Ph.D. in Computer Science and Technology from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2019. He was sponsored by ERCIM 'Alain Bensoussan' Fellowship Programme (from 2020.10.1 to 2021.3.31) to conduct Post-Doc research supervised by Prof. Fabio Martinelli at CNR, Italy. He was sponsored by China Scholarship Council (CSC) as a Joint Ph.D. student (from 2017.9.1 to 2018.8.31) supervised by Prof. Xiaodong Lin (IEEE Fellow) in the Broadband Communications Research (BBCR) Lab at University of Waterloo and Wilfrid Laurier University, Canada. He is supported by CSC as a Visiting Scholar (from 2025.3.1 to 2025.6.30) collaborating with Prof. Mauro Conti (IEEE Fellow) at the HIT Center, University of Padua, Italy. He was promoted to a Professor from December, 2024. His research interests include data sharing, IoV, security, privacy, applied cryptography, blockchain, TEE. In this area, he has published 156 papers in topmost journals and conferences, including T-IFS, TDSC, ToN, TMC, JSAC, TKDE, TODS, TPDS, TSE, TSC, COMST, IEEE S&P, USENIX Security, MobiCom, and INFOCOM. He is a Senior Member of IEEE, CCF, CACR, CIE, and CIC. He is a Senior Area Editor for T-IFS and an Associate Editor for TDSC. He has served as a TPC member for conferences, including ICDCS, Inscrypt, ACISP, ICICS, and TrustCom. He is the recipient of 2024 IEEE HITC Award for Excellence (Early Career Researcher) and 2025 IEEE TCSVC Rising Star Award. He was selected into the IEEE Computer Society Computing's Top 30 Early Career Professionals for 2025.

**Dan Guo** (Senior Member, IEEE) received the B.E. degree in computer science and technology from Yangtze University, Wuhan, China, in 2004, and the Ph.D. degree in system analysis and integration from Huazhong University of Science and Technology, Wuhan, China, in 2010. She is currently a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis. Dr. Guo serves as a Program Committee Member for top-tier conferences and prestigious journals in multimedia and artificial intelligence, such as ACM Multimedia, IJCAI, AAAI, CVPR, and ECCV. She also serves as an Senior Program Committee Member for IJCAI 2021. She is an Associate Editor of the IEEE Transactions on Multimedia (TMM).

**Meng Wang** (Fellow, IEEE) received the B.E. and Ph.D. degrees from the Special Class for the Gifted Young, Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, Hefei. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored or coauthored more than 200 book chapters and journal articles and conference papers in these areas. Dr. Wang was a recipient of the ACM SIGMM Rising Star Award in 2014. He is an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), the IEEE Transactions on Knowledge and Data Engineering (TKDE), the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), the IEEE Transactions on Multimedia (TMM), and the IEEE Transactions on Neural Networks and Learning Systems (TNNLS).