From Uncertainty to Certainty: A Robust Dynamic Facial Expression Recognition Framework via Dynamic Uncertainty Constraint Modeling

Feng-Qi Cui[1], Jie Zhang[1], Jinyang Huang*, Anyang Tong, Ziyu Jia, Zhi Liu, Dan Guo*, Meng Wang, *Fellow, IEEE*

***Abstract*—Various uncertainties stem from unpredictable expression evolution, data uncertainty, and unrobust models, significantly downgrading the performance of dynamic facial expression recognition (DFER). While deep learning-based methods have made remarkable progress, they often struggle to effectively handle these uncertainties, inevitably limiting their performance in complex real-world scenarios. To deal with these uncertainties, we propose an efficient and robust DFER framework called Uncertainty Suppression Flow (*UnSFlow*), which integrates three uncertainty suppression modules, namely: Uncertainty-Constrained Augmentation (UCA), Dynamic Uncertainty-Guided Network (DUGN), and Uncertainty Sharpness-Aware Minimization (USAM). Firstly, to mitigate expression evolution uncertainty, by introducing appropriate randomness, UCA is introduced to enhance data diversity while maintaining augmentation consistency. Subsequently, by integrating the uncertainty estimation branch for quantitative uncertainty modeling and uncertainty-aware routing attention for adaptive feature refinement, DUGN is designed to suppress data uncertainty through hybrid spatiotemporal modeling. Finally, to tackle model uncertainty, USAM regulates gradient perturbations via predictive uncertainty to stabilize parameter updates and enhance robustness for minority-class and ambiguous samples, thereby achieving uncertainty-adaptive optimization. Extensive experiments on two widely used in-the-wild datasets, DFEW and FERV39k, demonstrate that *UnSFlow* achieves competitive or superior performance compared with representative DFER methods in terms of recognition accuracy, computational efficiency, and robustness, providing a practical and effective solution for real-world dynamic facial expression recognition. Our code is available at: https://github.com/QIcita/Uncertainty2Certainty.**

***Index Terms*—Dynamic facial expression recognition, uncertainty learning, emotion ambiguity.**

I. Introduction

SINCE facial expressions play a key role in various domains, *e.g.*, social interactions [1], [2], psychological assessment [3], and human-computer interaction (HCI) [4], [5], it becomes a research focus in both academia and industry. In general, facial expression recognition can be categorized into two main types: Static Facial Expression Recognition (SFER) [6] and Dynamic Facial Expression Recognition (DFER) [7]. SFER performs emotion recognition based solely on single-frame images [8], making it difficult to capture the temporal evolution of facial expressions and contextual relationships between frames [9], thus limiting its applicability in real-world scenarios. In contrast, by capturing the temporal evolution of expressions, DFER can more accurately reflect emotional changes, demonstrating greater adaptability and robustness in dynamic and complex real-world environments [7].

Fig. 1: Causes of Temporal Evolution Uncertainty in Dynamic Facial Expressions. Variations in expression trajectories, where an initially Neutral state transitions into distinct emotional states such as Happy, Neutral, and Angry, highlight the uncertainty challenges in accurately modeling spatiotemporal dynamics in dynamic facial expression recognition.

However, despite the many advantages of current DFER methods, the lack of in-depth analysis of uncertainties limits their application. Specifically, 1) **The insufficient analysis of expression evolution uncertainty limits the modeling of temporal dynamics.** As shown in Fig. 1, dynamic facial expressions exhibit complex evolution characteristics due to the diversity of expression transition paths and uncertainties in temporal scales. Existing methods typically rely on fixed time-step feature extraction strategies, making it difficult to effectively capture long-term temporal dependencies with high uncertainty, thereby restricting their ability to extract fine-grained dynamic features efficiently. **2) Ignoring data uncertainty exacerbates feature confusion during learning.** Variations in collection devices, environmental conditions, and labeling biases cause data uncertainty, inevitably leading to noise, ambiguity, or feature confusion. These factors further increase the

Feng-Qi Cui is with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China.

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

Jinyang Huang, Anyang Tong, Dan Guo, and Meng Wang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China.

Feng-Qi Cui and Jinyang Huang are with the Hefei Joy Embodied Intelligence Technology Co., Ltd., Hefei, China.

Ziyu Jia is with the Beijing Key Laboratory of Brainnetome and Brain-Computer Interface, and the Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Zhi Liu is with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo, Japan.

[1]Equal contribution, *Corresponding author.

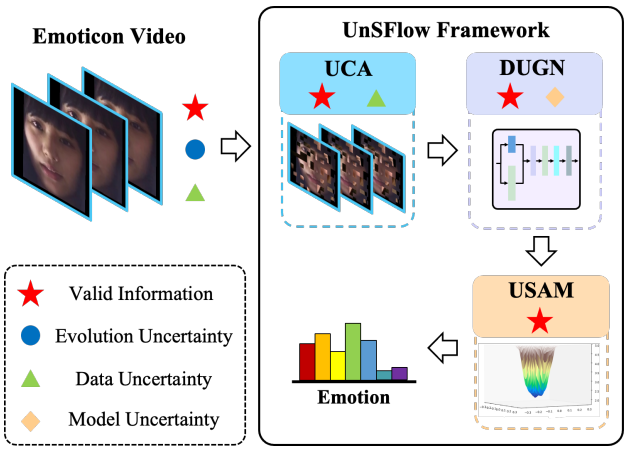


Fig. 2: Uncertainty reduction workflow of the proposed *UnSFlow* framework for dynamic facial expression recognition. UCA, DUGN, and USAM progressively suppress expression evolution uncertainty, data uncertainty, and model uncertainty from data augmentation, feature modeling, and optimization perspectives, respectively.

uncertainty in data distribution, making it difficult for models to stably learn discriminative features. Such uncertainty not only affects feature extraction but also results in unstable decision boundaries, finally reducing classification accuracy. **3) The lack of constraints on model uncertainty weakens expression discrimination capability.** DFER datasets often face the challenge of model uncertainty, where high feature similarity between different expressions increases optimization difficulty. This model's uncertainty results in a downgrade of recognition performance for minority expressions and the ambiguity of categories. Additionally, gradient fluctuations during training are inevitably amplified, making it difficult for optimization to converge to a smoother loss landscape, severely impacting classification reliability.

To address the aforementioned challenges, we propose an efficient and robust dynamic facial expression recognition framework, **Uncertainty Suppression Flow (*UnSFlow*)**, which consists of three key modules to deal with three different uncertainties, *i.e.*, Uncertainty-Constrained Augmentation (UCA), Dynamic Uncertainty-Guided Network (DUGN), and Uncertainty Sharpness-Aware Minimization (USAM). As illustrated in Fig. 2, these modules effectively suppress expression evolution uncertainty, data uncertainty, and model uncertainty, respectively, thereby enhancing the model's ability to learn complex dynamic expressions while improving both robustness and generalization. Firstly, to ensure that augmentation introduces sufficient diversity while avoiding excessive randomness that may disrupt feature distributions, UCA incorporates temporal consistency, mixing ratio, and mixing granularity constraints to perform spatial-level block mixing between samples of the same emotional category but with different representations. This approach effectively mitigates expression evolution uncertainty while preventing feature distortion caused by overly random augmentations, thus enabling the model to stably learn spatiotemporal features. Next, DUGN combines short-term and long-term dynamic modeling and integrates an Uncertainty Estimation Branch along with an Uncertainty-Aware Routed Attention to cope with data uncertainty. In particular, the Uncertainty Estimation Branch quantifies feature uncertainty to guide the model in dynamically adjusting feature weights, highlighting discriminative regions while suppressing ambiguous features, thereby enhancing the model's capability to model complex expression dynamics. The Uncertainty-Aware Routed Attention further employs an uncertainty-guided multi-expert routing strategy to adaptively adjust attention allocation, allowing the model to accurately capture dynamic facial expression features and reducing the interference of data uncertainty in feature learning. Finally, by using the predictive entropy to adaptively adjust the gradient perturbation magnitude, USAM incorporates an uncertainty-driven gradient perturbation optimization strategy to deal with model uncertainty, which enables the optimization process to adapt to varying levels of uncertainty in the data distribution. During training, USAM focuses more effectively on high-uncertainty samples while reducing gradient oscillations in the optimization trajectory, which ensures a smoother loss landscape, further alleviates model uncertainty, and improves model generalization. Overall, *UnSFlow* models and mitigates uncertainty at three levels, namely expression evolution, data distribution, and model optimization, thereby improving the robustness, generalization, and stability of DFER in complex real-world scenarios.

In summary, our main contributions are as follows:

- We present a unified uncertainty suppression framework for DFER that explicitly considers expression evolution uncertainty, data uncertainty, and model uncertainty. Correspondingly, *UnSFlow* integrates three uncertainty-aware modules and constructs a robust learning pipeline from augmentation, representation modeling, to optimization for DFER tasks.
- Following a progressive uncertainty suppression strategy, we design three modules to address different sources of uncertainty. UCA balances randomness and consistency through constrained augmentation to enhance generalization under diverse expression evolutions. Then, DUGN dynamically models feature-level uncertainty and adaptively refines attention to boost discriminability under ambiguity. Finally, USAM adjusts sharpness-aware perturbations based on predictive uncertainty, smoothing optimization trajectories, and improving robustness for minority and hard-to-learn expressions.
- Extensive experiments conducted on two widely used in-the-wild DFER datasets, DFEW and FERV39k, validate the effectiveness of our approach. The results demonstrate that *UnSFlow* achieves competitive or superior performance compared with representative DFER methods in recognition accuracy, spatiotemporal feature modeling, and adaptability to complex data distributions, showing its practical value for real-world DFER applications.

II. Related Work

A. Dynamic Facial Expression Recognition

DFER aims to model the temporal evolution of facial expressions and capture finer-grained emotional variations than

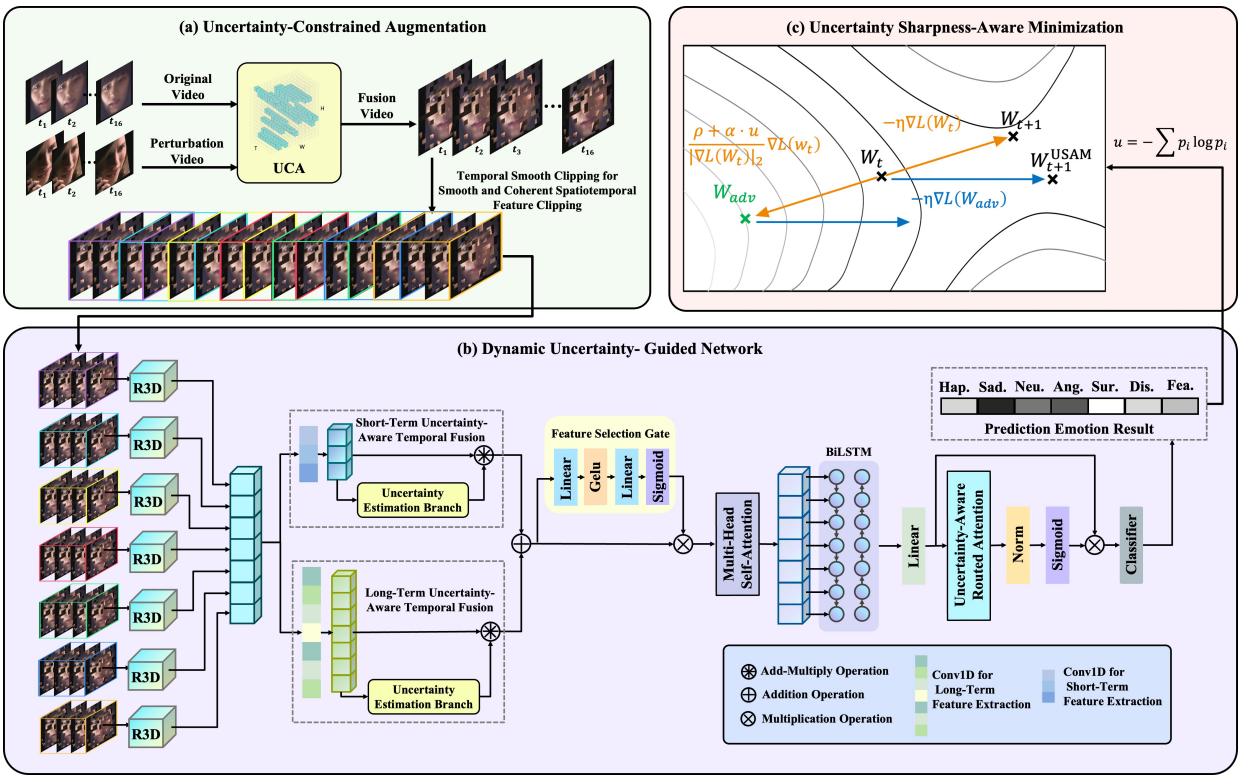


Fig. 3: Overview of the proposed UnSFlow framework. (a) UCA introduces constrained intra-class video mixing to increase expression diversity while preserving temporal consistency. (b) DUGN estimates feature reliability and performs uncertainty-aware temporal modeling to suppress ambiguous spatiotemporal responses. (c) USAM adaptively scales sharpness-aware perturbations according to predictive uncertainty to stabilize optimization.

SFER. Early methods mainly relied on handcrafted spatiotemporal descriptors, such as LBP-TOP [10] and HOG-TOP [11], but their dependence on domain priors limits their adaptability to complex in-the-wild scenarios. With the development of deep learning, supervised visual DFER methods have become a widely used paradigm. CNN-RNN based methods extract spatial features with CNNs and model temporal dependencies using RNNs, LSTMs, or GRUs [12]–[16], while 3D CNN-based methods jointly learn spatiotemporal representations [17]. For example, M3DFEL [18] adopts a multi-instance 3D CNN framework to reduce the influence of noisy frames, and IAL [19] improves class separation with intensity-aware supervision. However, these supervised methods usually focus on specific issues such as temporal modeling, noisy frames, or class separation, while the joint modeling of multiple uncertainty sources remains insufficiently explored.

Recently, Transformer- and graph-based methods have been introduced to capture long-range dependencies and structured facial dynamics [20] . These methods improve spatiotemporal representation learning, but they may not explicitly model data uncertainty and often introduce additional computational cost. Beyond supervised visual DFER, recent studies also explore self-supervised pre-training, vision-language learning, and multimodal representation learning. For instance, MAE-DFER [21] learns transferable representations through masked autoencoding, DK-CLIP [22] incorporates domain knowledge into vision-language pre-training, and AVF-MAE++ [23] exploits audio-visual masked pre-training. Although these methods have advanced DFER from the perspectives of representation pre-training and cross-modal knowledge transfer, their construction and deployment usually require additional resources, such as large-scale pre-training data, external textual knowledge, or synchronized audio-visual inputs. In practical in-the-wild scenarios, such requirements may increase data collection cost, training complexity, and dependence on modality availability. Therefore, it remains necessary to improve the robustness of visual DFER models under standard supervised settings, especially when only facial video inputs are available.

Overall, existing DFER studies have improved dynamic expression recognition from different perspectives. However, many of them mainly enhance representation capacity through stronger architectures, external knowledge, or additional modalities, while the uncertainties inherent in visual DFER data are still not systematically addressed. In contrast, UnSFlow focuses on the uncertainty sources within the visual DFER pipeline itself and progressively suppresses expression evolution uncertainty, data uncertainty, and model uncertainty through constrained augmentation, uncertainty-aware spatiotemporal modeling, and uncertainty-guided optimization.

B. Uncertainty Learning

Uncertainty learning aims to improve model reliability under unreliable predictions, ambiguous supervision, and distributional variation [24], [25]. Representative studies include Bayesian deep learning [26] and Monte Carlo Dropout [27], which estimate predictive uncertainty through probabilistic

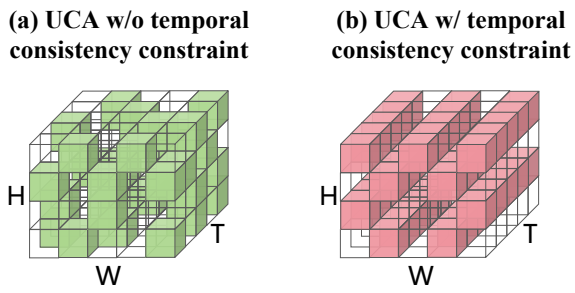


Fig. 4: Comparison of different temporal constraints in UCA.

modeling or stochastic inference. Recent FER studies also address uncertainty caused by ambiguous expressions and noisy labels through label distribution learning or landmark-aware reliability modeling [28], [29]. However, most of these methods mainly focus on static images, prediction confidence, or label-level ambiguity. For DFER, uncertainty also arises from temporal expression evolution, low-intensity or noisy visual observations, and unstable learning from hard or mislabeled samples [19], [30]. Our recent work HDF [31] explored distributional heterogeneity in DFER through robust optimization, but it does not explicitly coordinate uncertainty suppression across augmentation, spatiotemporal modeling, and optimization.

From the data perspective, augmentation can be regarded as a practical way to regulate training distribution uncertainty by enlarging the observed sample space [32]–[35]. Mixup [36] and CutMix [37] are widely used in static recognition tasks, while AutoAugment [38] and UDA [39] show that augmentation strength should be carefully controlled to avoid excessive distribution shift or insufficient regularization. For video-based expression recognition, independent frame-wise perturbations may break temporal continuity, and unconstrained region mixing may introduce semantic drift. Therefore, DFER augmentation should increase data diversity while preserving temporal coherence and expression semantics.

Overall, existing studies improve robustness by estimating prediction uncertainty, handling ambiguous labels, enlarging training distributions, or stabilizing optimization. However, they usually address individual uncertainty factors, while the coupled uncertainty sources in DFER remain insufficiently coordinated. In contrast, *UnSFlow* treats uncertainty suppression as a progressive pipeline and mitigates expression evolution uncertainty, data uncertainty, and model uncertainty through constrained augmentation, uncertainty-aware spatiotemporal modeling, and uncertainty-guided optimization, respectively.

III. PROPOSED METHOD

Fig. 3 depicts the proposed *UnSFlow* framework, which consists of three key modules: UCA, DUGN, and USAM. Specifically, UCA introduces constrained intra-class video mixing to enrich expression variations while controlling augmentation-induced distributional deviation within a learnable range. Then, DUGN combines short-term and long-term spatiotemporal modeling, where the Uncertainty Estimation Branch estimates feature reliability and the Uncertainty-Aware Routed Attention adaptively selects reliable tokens for attention refinement. This approach suppresses noise and emphasizes key spatiotemporal features, thereby improving the model's ability to perceive dynamic expressions. Finally, USAM introduces uncertainty-driven perturbation adjustments during optimization to smooth gradient updates, improving the model's adaptability to class imbalance and noise. By integrating these three modules, *UnSFlow* forms a progressive uncertainty suppression pipeline that controls augmentation-induced uncertainty, suppresses feature-level uncertainty, and stabilizes optimization uncertainty from data, representation, and parameter perspectives, respectively.

Algorithm 1 The workflow of UCA

Require: Original Video $I_u \sim Q$, Perturbation Video $I_s \sim Q$, dimensions (T, H, W)
Require: Mixing Ratio Constraint λ, Mixing Granularity Constraint B
Output: Augmented Video $I_{mix} \sim P_A$, dimensions (T, H, W)

$n_p = \left(\frac{H}{B}\right) \times \left(\frac{W}{B}\right)$ {*Compute the number of spatial blocks per frame*}
$n_m = \text{int}(\lambda \times n_p)$ {*Compute the number of masked spatial blocks per frame*}
$\mu = [0] \times (n_p - n_m) + [1] \times n_m$ {*Initialize a binary spatial block mask*}
$\mu = \text{shuffle}(\mu)$ {*Shuffle the spatial mask array*}
$\mu = \mu.\text{reshape}\left(\frac{H}{B}, \frac{W}{B}\right)$ {*Reshape to spatial block layout*}
$\mu = \text{np.repeat}(\text{np.repeat}(\mu, B, \text{axis} = 0), B, \text{axis} = 1)$ {*Expand to frame size*}
$\hat{\mu} = \text{np.tile}(\mu, (T, 1, 1))$ {*Share the same spatial mask across frames to enforce temporal consistency*}
$\hat{\mu} = \hat{\mu}.\text{unsqueeze}(1)$ {*Add the channel dimension*}
$I_{mix} = I_u * (1 - \hat{\mu}) + I_s * \hat{\mu}$ {*Perform uncertainty-constrained intra-class augmentation*}
return I_{mix}

A. Uncertainty-Constrained Augmentation (UCA)

To alleviate classification confusion caused by diverse expression evolution, UCA introduces controlled intra-class perturbations during training. Specifically, UCA constructs mixed videos between samples from the same emotion category, increasing expression variation while reducing the risk of semantic drift. Since dynamic facial expressions contain continuous temporal evolution, the augmentation process should enlarge the training distribution while preserving temporal coherence and semantic stability.

Generally, following the KL-divergence-based view for measuring the discrepancy between two distributions [40], [41], the augmentation-induced distributional deviation can be intuitively described as:

$$\mathcal{D}_{KL}(P_A \| Q) = \sum_{x \in \mathcal{X}} P_A(x) \cdot \log \frac{P_A(x)}{Q(x)}, \tag{1}$$

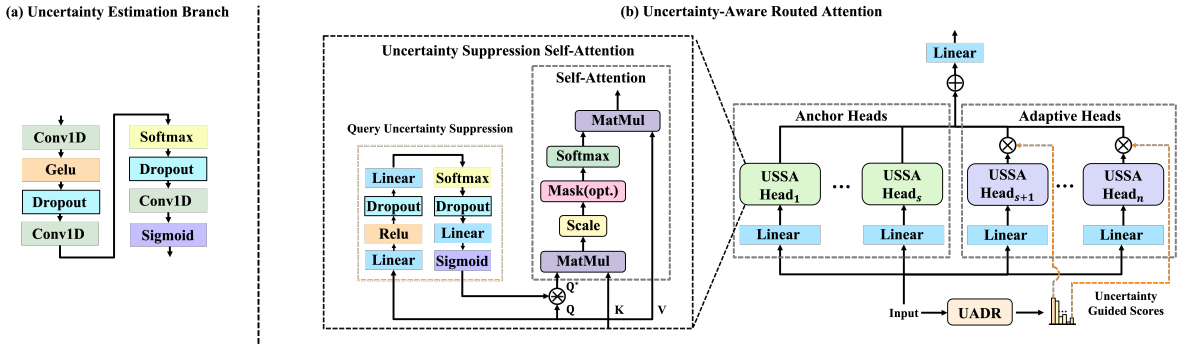


Fig. 5: Structures of (a) the Uncertainty Estimation Branch and (b) the Uncertainty-Aware Routed Attention. UEB estimates uncertainty-aware weights for short-term and long-term features through dropout-based uncertainty modeling, while UARA uses uncertainty-adjusted routing to allocate attention to stable and discriminative tokens.

where $P_A(x)$ and $Q(x)$ represent the distributions of the augmented and original data, respectively.

In this formulation, $P_A(x)$ reflects the augmented sample distribution generated by the mixing operation, while $\log \frac{P_A(x)}{Q(x)}$ characterizes its discrepancy from the original distribution. Accordingly, UCA introduces practical constraints to regulate both the deviation magnitude and the structural form of P_A. The mixing ratio λ controls the proportion of replaced regions, thereby adjusting the perturbation strength and the discrepancy term $\log \frac{P_A(x)}{Q(x)}$. The block size B determines the spatial granularity of the selected regions, affecting the local structural composition of $P_A(x)$ while limiting facial structure distortion. To further constrain the temporal structure of $P_A(x)$, UCA shares the selected spatial mixing regions across frames. Compared with independently sampling masks for different frames, this tube-wise mask-sharing strategy preserves the temporal continuity of expression evolution and avoids abrupt frame-wise perturbations. In this way, λ controls the perturbation magnitude, B regulates the spatial structure of the mixed regions, and temporal mask sharing constrains the temporal coherence of the augmented distribution. Together, these constraints balance augmentation diversity, facial structure preservation, and temporal consistency.

Alg. 1 illustrates the implementation of UCA. Specifically, UCA first generates a binary spatial block mask according to the mixing ratio λ and the block size B, and then tiles the same mask along the temporal dimension to construct tube-wise coherent mixing regions, as shown in Fig. 4(b). This design differs from frame-wise independent masking, where different masks are sampled for different frames and may disrupt the continuity of expression evolution. Thus, the mixing ratio controls the replaced region proportion, the block size determines the spatial granularity, and temporal mask sharing maintains coherent perturbation patterns across frames. These constraints jointly reduce excessive perturbations and help preserve the semantic stability of augmented videos.

B. Dynamic Uncertainty-Guided Network (DUGN)

1) Overview: DUGN aims to effectively capture the spatiotemporal dependencies of dynamic facial expressions by integrating short-term and long-term spatiotemporal modeling with uncertainty-guided mechanisms. The input video sequence is first divided into overlapping sliding windows to ensure temporal coherence and smooth contextual transitions between windows. Each window undergoes spatiotemporal feature extraction using R3D [42]. Then, through the Uncertainty Estimation Branch, these window-level features are used to generate short-term and long-term temporal representations, and the estimated uncertainty is further used to adjust their contribution weights. Next, bidirectional temporal modeling is applied to further aggregate frame-wise information through self-attention mechanisms.

Finally, the Uncertainty-Aware Routed Attention refines token-wise attention allocation by suppressing unreliable responses and emphasizing more discriminative features, thereby improving the spatiotemporal representation for downstream classification.

2) The Uncertainty Estimation Branch: As shown in Fig. 5 (a), the Uncertainty Estimation Branch is designed to tackle the high-entropy phenomenon in feature distributions caused by data uncertainty, which inevitably downgrades feature quality.

To address this issue, it estimates the uncertainty of short-term and long-term temporal features and converts uncertainty cues into reliability-aware weights for adaptive feature fusion. Specifically, we estimate feature uncertainty using a Monte Carlo approach, where multiple forward passes under Monte Carlo Dropout generate different sampled representations $\{x_i^{(m)}\}_{m=1}^{M}$ for each feature x_i. The empirical variance

Algorithm 2 The operation flow of USAM

Requirements: Optimizer $\mathcal{O}$, model parameters W_t, base perturbation radius ρ, uncertainty weight α
Input: Loss function L, mini-batch $\mathcal{D}$
Output: Updated model parameters W_{t+1}^{USAM}

Compute gradient $\nabla L(W_t)$ on $\mathcal{D}$
Estimate entropy-based uncertainty: $u = -\sum p_i \log p_i$
Compute dynamic perturbation: $\rho_{\text{dyn}} = \rho + \alpha \cdot u$
Calculate perturbation: $\epsilon = \frac{\rho_{\text{dyn}}}{|\nabla L(W_t)|_2} \nabla L(W_t)$
Compute adversarial weight: $W_{adv} = W_t + \epsilon$
Compute adversarial gradient: $\nabla L(W_{adv})$
Update parameters: $W_{t+1}^{USAM} = W_t - \eta \nabla L(W_{adv})$

of each feature is computed as:

$$\mathrm{Var}(x_i) = \frac{1}{M-1}\sum_{m=1}^{M}\left(x_i^{(m)} - \overline{x}_i\right)^2, \tag{2}$$

$$\overline{x}_i = \frac{1}{M}\sum_{m=1}^{M} x_i^{(m)}, \tag{3}$$

where $\mathrm{Var}(x_i)$ represents the model's predictive uncertainty for that feature. The variance values are then converted into a normalized stability distribution:

$$p(x_i) = \frac{\exp\big(-\mathrm{Var}(x_i)\big)}{\sum_k \exp\big(-\mathrm{Var}(x_k)\big)}. \tag{4}$$

Due to the negative exponential mapping, lower-variance features obtain larger stability responses. Based on this distribution, the entropy of short-term and long-term features is computed as:

$$H_{\text{short}} = -\sum_i p(x_i)\,\log p(x_i), \tag{5}$$

$$H_{\text{long}} = -\sum_j p(x_j)\,\log p(x_j). \tag{6}$$

To dynamically adjust the contributions of different temporal features, we use entropy-based reliability weighting. The weight assigned to each temporal branch is inversely proportional to its entropy, which ensures high-quality low-entropy features are emphasized while noisy high-entropy features are relatively suppressed:

$$w_{\text{short}} = \frac{1}{H_{\text{short}} + \gamma}, \tag{7}$$

$$w_{\text{long}} = \frac{1}{H_{\text{long}} + \gamma}, \tag{8}$$

where γ is a smoothing constant that prevents numerical instability when entropy values are close to zero.

After obtaining channel-wise stability responses, we apply a pooling strategy to obtain branch-level reliability-aware modulation scores, denoted as $\mathbf{r}_{\text{short}}$ and $\mathbf{r}_{\text{long}}$. These scores represent the relative reliability of short-term and long-term features, rather than their raw uncertainty.

When fusing the short-term and long-term representations, we combine entropy-based weights with reliability-aware modulation scores:

$$\tilde{\mathbf{x}}_{\text{short}} = \mathbf{x}_{\text{short}} \cdot \big(1 + \beta \cdot w_{\text{short}} \cdot \mathbf{r}_{\text{short}}\big), \tag{9}$$

$$\tilde{\mathbf{x}}_{\text{long}} = \mathbf{x}_{\text{long}} \cdot \big(1 + \beta \cdot w_{\text{long}} \cdot \mathbf{r}_{\text{long}}\big), \tag{10}$$

and then combine them as follows:

$$\mathbf{x}_{\text{fused}} = \tilde{\mathbf{x}}_{\text{short}} \;+\; \tilde{\mathbf{x}}_{\text{long}}. \tag{11}$$

Through this entropy-guided reliability modulation, more stable temporal responses are assigned larger effective contributions, while high-entropy responses are prevented from dominating the fusion process.

Moreover, by leveraging uncertainty estimation and reliability-aware feature modulation, the Uncertainty Estimation Branch improves short-term and long-term spatiotemporal fusion for DFER.

3) Uncertainty-Aware Routed Attention: Fig. 5 (b) illustrates the structure of the proposed Uncertainty-Aware Routed Attention (UARA) module, which integrates uncertainty modeling with a routing strategy inspired by the Mixture-of-Experts mechanism [43]. Specifically, UARA employs Monte Carlo sampling to estimate token-wise predictive uncertainty and derives dynamic routing scores to guide attention allocation. By assigning lower weights to high-uncertainty features and amplifying attention on more reliable regions, UARA effectively suppresses noisy signals and enhances the model's robustness and discriminative capacity in DFER tasks.

First, given the input features $\mathbf{X} \in \mathbb{R}^{B\times N\times C}$, a Monte Carlo Dropout branch estimates the predictive variance $\mathrm{Var}(\mathbf{x}_i)$ for each token $\mathbf{x}_i$. This variance is then converted via a sigmoid activation into an uncertainty score $\mathbf{u}_i$, with higher $\mathbf{u}_i$ indicating a more unreliable feature. Next, these uncertainty scores dynamically adjust the routing weights of shared and routed heads. For the shared heads, we define them as follows:

$$g_i^{\text{shared}} = \alpha_1 \cdot \mathrm{Softmax}\big(\mathbf{W}_s\,\mathbf{x}_i\big), \tag{12}$$

For the routed heads, we focus on tokens whose uncertainty is relatively lower, following a top-k strategy:

$$g_i^{\text{routed}} = \begin{cases} \alpha_2 \cdot \mathrm{Softmax}\big(\mathbf{W}_r\,\mathbf{x}_i\big)\,(1-\mathbf{u}_i), & \text{if } i \in \text{top-}k, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Specifically, the top-k selection is performed according to the uncertainty-adjusted routing response $\mathrm{Softmax}(\mathbf{W}_r\mathbf{x}_i)(1-\mathbf{u}_i)$, so routing candidates with stronger activation and lower uncertainty are preferred.

We then combine them into a total routing weight: $g_i = g_i^{\text{shared}} + g_i^{\text{routed}}$, with α_1, α_2 being adaptively learned via $[\alpha_1, \alpha_2] = \mathrm{Softmax}\big(\mathbf{W}_h\,\mathbf{x}_i\big)$. During the attention stage, Uncertainty-Aware Routed Attention further modifies each query vector $\mathbf{q}_i$ based on $\mathbf{u}_i$ to emphasize low-uncertainty (*i.e.*, more reliable) features: $\mathbf{q}_i^{\text{adjusted}} = \mathbf{q}_i \cdot \big(1 - \mathbf{u}_i\big)$. The adjusted query vectors $\mathbf{q}_i^{\text{adjusted}}$ are then aggregated into a full query matrix $\mathbf{Q}^{\text{adjusted}}$ across all tokens, where $\mathbf{Q}^{\text{adjusted}} = \mathbf{Q}\cdot(1-\mathbf{U})$, which ensures a consistent uncertainty-aware modulation before applying scaled dot-product attention:

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\Big(\frac{\mathbf{Q}^{\text{adjusted}}\,\mathbf{K}^{\top}}{\sqrt{d}}\Big)\,\mathbf{V}. \tag{14}$$

Hence, high-uncertainty or noisy tokens are downweighted, which allows more attention to be devoted to regions of greater value.

Overall, by integrating Monte Carlo-based uncertainty estimation with uncertainty-adjusted routing and query modulation, UARA prioritizes reliable tokens without changing the original attention computation pipeline. This mechanism effectively strengthens the model's capacity for discrimination and generalization in dynamic facial expression recognition.

C. Uncertainty Sharpness-Aware Minimization (USAM)

USAM extends sharpness-aware minimization [51] by introducing uncertainty-guided dynamic perturbation scaling. This enhances the trade-off between local flatness and loss minimization, mitigating gradient oscillations and local overfitting

TABLE I: Comparison (%) of our *UnSFlow* with state-of-the-art methods on DFEW. (Hap.: happiness, Sad.: sadness, Neu.: neutral, Ang.: anger, Sur.: surprise, Dis.: disgust and Fea.: fear.) (**Bold**: Best, Underline: Second best.)

Method	Venue	Accuracy of Each Emotion(%)							Metrics (%)		FLOPs (G)
		Hap.	Sad.	Neu.	Ang.	Sur.	Dis.	Fea.	WAR	UAR	
ResNet18+LSTM [44]	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 [44]	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 [20]	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 [45]	ICASSP'23	85.39	66.52	68.94	71.33	54.59	0.00	32.71	66.98	54.21	N/A
T-MEP [46]	T-CSVT'24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	68.85	57.16	24.70
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 [18]	CVPR'23	89.59	68.38	67.88	74.24	59.69	0.00	31.63	69.25	56.10	**1.65**
LG-DSTF [7]	T-MM'24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	69.82	58.89	N/A
RDFER [47]	T-BIOM'25	89.69	69.22	70.18	71.47	62.08	0.69	28.71	69.73	56.93	N/A
CLIPER [48]	ICME'24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	70.84	57.56	N/A
KFE-SC [49]	Inf. Sci.'25	89.12	68.11	**71.33**	78.27	63.25	4.02	33.86	70.30	57.35	N/A
ST-RDGCN [50]	T-AFFC'25	89.57	69.92	62.17	**79.31**	62.24	20.69	30.39	69.37	59.18	30.38
HDF [31]	ACM MM'25	89.67	**71.20**	67.42	73.03	**64.44**	12.41	**41.63**	71.60	60.40	N/A
UnSFlow (Ours)	-	**90.27**	70.58	69.04	77.31	60.88	**26.89**	37.63	**72.29**	**62.64**	2.90

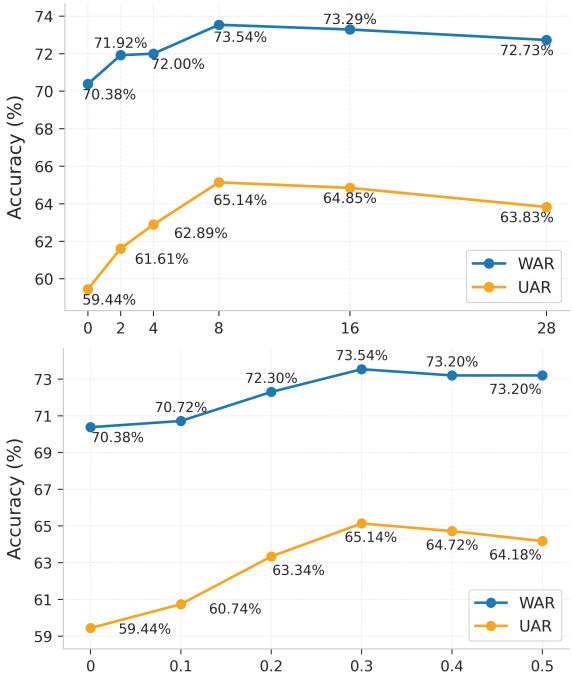


Fig. 6: Over- and under-augmentation constraints in UCA and their effects on DFEW (fd5). The top and bottom plots show the effects of the mixing granularity constraint and the mixing ratio constraint, respectively.

caused by model uncertainty. Sharpness-aware minimization applies a perturbation ϵ to the parameters W_t, jointly minimizing both the sharpness and the value of the loss function:

$$L_{\text{SAM}}(W_t) = \max_{|\epsilon|_2 \leq \rho} L(W_t + \epsilon), \tag{15}$$

where ρ is a fixed perturbation radius controlling local flatness. However, for imbalanced or ambiguous training samples, a fixed perturbation radius may not adapt to different uncertainty levels. To address this, USAM dynamically adjusts ρ_{dyn} according to entropy-based predictive uncertainty: $\rho_{\text{dyn}} = \rho + \alpha \cdot u$, where $u = -\sum p_i \log p_i$. High-uncertainty samples usually lie near ambiguous or noisy decision regions, and a larger local perturbation radius encourages the optimizer to search for flatter neighborhoods around these unstable samples. In the first optimization step, USAM computes an adaptive perturbation ϵ:

$$\epsilon = \frac{\rho_{\text{dyn}}}{|\nabla L(W_t)|_2} \nabla L(W_t), \tag{16}$$

which is further employed to generate an adversarial weight perturbation:

$$W_{adv} = W_t + \epsilon. \tag{17}$$

Instead of directly updating W_t, USAM first computes the gradient at the perturbed weights W_{adv}, i.e., $\nabla L(W_{adv})$. Finally, the parameters are updated using the adversarial gradient:

$$W_{t+1}^{USAM} = W_t - \eta \nabla L(W_{adv}). \tag{18}$$

As depicted in Alg. 2, by incorporating uncertainty-driven perturbation scaling and adversarial gradient evaluation, USAM applies stronger smoothing to high-uncertainty samples, achieving robust optimization for dynamic facial expression recognition.

IV. Experiments

A. Experimental Setup

1) Datasets: We conduct experiments on two widely used in-the-wild DFER datasets, namely DFEW [44] and FERV39k [52]. DFEW contains over 16,000 video clips collected from more than 1,500 films worldwide, covering challenging factors such as illumination changes, pose variations, and complex backgrounds. Each clip is annotated by ten trained annotators and assigned to one of seven basic expression categories: happiness, sadness, neutral, anger, surprise, disgust, and fear. FERV39k contains 38,935 video clips from 4 coarse scenes and 22 fine-grained scenes, providing a challenging benchmark for evaluating DFER models under diverse real-world and cross-scene conditions. Each clip is annotated by 30 professional annotators using the same seven expression categories

TABLE II: Comparison (%) of *UnSFlow* with the state-of-the-art methods on FERV39k. (**Bold**: Best, Underline: Second best.)

Method	Accuracy of Each Emotion(%)							Metrics (%)	
	Hap.	Sad.	Neu.	Ang.	Sur.	Dis.	Fea.	WAR	UAR
ResNet18+LSTM [52]	61.91	31.95	61.70	45.93	14.26	0.00	0.70	42.59	30.92
VGG13+LSTM [52]	66.26	51.26	53.22	37.93	13.64	0.43	4.18	43.37	32.42
2C3D [52]	54.85	52.91	60.67	31.34	5.96	2.36	6.96	41.77	30.72
2ResNet18+LSTM [52]	59.00	45.87	61.90	40.15	9.87	1.71	0.46	43.20	31.28
2VGG13+LSTM [52]	69.65	47.31	52.55	**47.88**	7.68	1.93	2.55	44.54	32.79
Former-DFER [20]	65.65	51.33	56.74	43.64	21.94	8.57	12.53	46.85	37.20
M3DFEL [18]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	47.67	35.94
GCA+IAL [19]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	48.54	35.82
Logo-Former [45]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	48.13	38.22
LG-DSTF [7]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	48.19	39.84
MICACL [53]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	48.57	40.25
RDFER [47]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	48.60	36.47
CFAN-SDA [54]	69.65	49.46	**62.21**	46.20	22.26	11.56	15.55	49.48	39.56
UnSFlow (Ours)	**73.18**	**55.30**	47.86	41.75	**34.64**	**19.27**	**18.79**	**49.69**	**41.54**

TABLE III: The importance of UCA, DUGN, and USAM in the proposed *UnSFlow* framework on DFEW (fd5).

Setting	Method			Metric	
	UCA	DUGN	USAM	WAR	UAR
a	✗	✗	✗	68.58	58.14
b	✔	✗	✗	71.62	62.92
c	✗	✔	✗	69.61	58.95
d	✗	✗	✔	71.23	59.46
e	✔	✔	✗	71.79	61.16
f	✔	✗	✔	72.90	62.83
g	✗	✔	✔	70.38	59.44
h	✔	✔	✔	**73.54**	**65.14**

TABLE IV: Ablation study of UEB and UARA in DUGN on DFEW (fd5).

Setting	DUGN		Metric	
	UEB	UARA	WAR	UAR
a	✗	✗	72.90	62.83
b	✔	✗	73.29	62.76
c	✗	✔	73.24	64.22
d	✔	✔	**73.54**	**65.14**

as DFEW. We use the official training and test splits of FERV39k for fair comparison.

2) Metrics: Following previous DFER methods, we adopt weighted average recall (WAR) and unweighted average recall (UAR) as the main evaluation metrics. WAR reflects overall recognition performance under the original class distribution, while UAR measures the average recall across all categories and better evaluates class-balanced recognition under imbalanced data distributions.

3) Implementation Details: In our experiments, all images are resized to 112×112, and 16 frames are evenly sampled from each video as input. We adopt random cropping, horizontal flipping, and color jitter with a factor of 0.4 for data augmentation. The backbone is the standard R3D model initialized with Torchvision pre-trained weights. The model is trained for 300 epochs with a cosine learning rate scheduler

TABLE V: Ablation study of the temporal consistency constraint in UCA on DFEW (fd5).

Setting	Method	WAR	UAR
a	w/o UCA	70.38	59.44
b	UCA w/o tcc	71.15	61.38
c	UCA w/ tcc	**73.54**	**65.14**

and a 20-epoch warm-up period, and the batch size is set to 64. The initial learning rate is set to 7×10^{-4}, the minimum learning rate is 5×10^{-6}, and the weight decay is 0.05. AdamW is used as the base optimizer, and USAM is applied on top of it with the base perturbation radius $\rho = 0.05$ and uncertainty weight $\alpha = 0.1$. For DUGN, the sliding window size and stride are set to 4 and 2, respectively, producing 7 temporal windows for each video. The dropout probability in the uncertainty branches is set to 0.1, and the number of MC Dropout samples is set to $M = 5$ during training.

B. Experimental Results

In this section, we compare the proposed *UnSFlow* with representative DFER methods on the DFEW and FERV39k benchmarks. Following the common evaluation setting in DFER, we mainly focus on visual video-based methods that use facial expression clips as input, so as to provide a fair comparison under the same recognition paradigm.

1) Comparison on the DFEW Dataset: Consistent with previous works, we employ the 5-fold cross-validation protocol on DFEW, and the results are reported in Tab. I.

As shown in Tab. I, *UnSFlow* achieves competitive overall performance under the visual video-based DFER setting, obtaining the best results in both WAR and UAR among the compared methods. Compared with CNN-RNN based methods, Transformer-based methods, and recent uncertainty- or graph-enhanced DFER models, *UnSFlow* shows a consistent advantage in balanced recognition performance. This indicates that jointly modeling expression evolution uncertainty, data uncertainty, and model uncertainty can improve the robustness of dynamic expression representation learning.

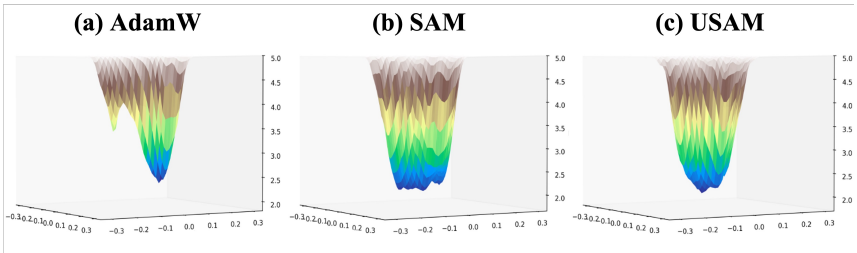


Fig. 7: The visualization of separate loss landscape on DFEW (fd5), optimized by AdamW, SAM and our USAM respectively.

TABLE VI: Comparison of results when using AdamW, SAM and USAM respectively on DFEW (fd5).

Setting	Method	WAR	UAR
a	AdamW	71.79	61.16
b	SAM	72.17	64.22
c	USAM	**73.54**	**65.14**

TABLE VII: Computational cost analysis of DUGN components. Since UCA and USAM are training-stage components, the inference-time cost mainly comes from DUGN. All results are measured under the same input setting.

DUGN		FLOPs (G / Δ)	Mem. (MB)	Time (ms/iter)	FPS
UEB	UARA				
✗	✗	2.81 / –	273	6.37	157
✓	✗	2.89 / +2.9%	333	7.08	141
✓	✓	2.90 / +3.3%	340	7.25	138

At the category level, *UnSFlow* also performs favorably on several small-sample or visually ambiguous expressions, such as disgust, surprise, and fear. These categories are usually more sensitive to class imbalance, weak expression intensity, and inter-class feature confusion. The improved recognition on these challenging categories suggests that the proposed uncertainty suppression pipeline helps reduce ambiguous feature responses and enhances the model's ability to distinguish difficult expressions in the wild. Overall, the results on DFEW validate the effectiveness of *UnSFlow* for robust visual DFER.

2) Comparison on the FERV39k Database: As shown in Tab. II, *UnSFlow* achieves competitive WAR and UAR on the large-scale FERV39k dataset, which contains diverse scenes and complex in-the-wild variations. This result further indicates the robustness and scalability of the proposed uncertainty suppression framework.

We also observe a class-wise trade-off. While several difficult categories, such as disgust and fear, are improved, the accuracy of Neutral slightly decreases. This phenomenon can be interpreted as a class-wise trade-off under ambiguous video-level annotations. In imbalanced and ambiguous in-the-wild DFER datasets, Neutral often acts as a broad and low-risk category, since low-intensity, transitional, or uncertain emotional states may be annotated as Neutral. By introducing uncertainty-aware augmentation, feature refinement, and optimization, *UnSFlow* becomes more sensitive to difficult categories and reduces over-reliance on broad ambiguous-category patterns. Meanwhile, the relatively lower accuracy on Anger may be related to its visual overlap with high-arousal expressions such as Fear, Surprise, and Disgust, which may share upper-face tension, eye-region activation, and mouth-region changes under complex scene variations. Overall, although some class-wise trade-offs remain, *UnSFlow* achieves better balanced recognition performance, as reflected by the improvement in UAR, and improves robust DFER performance under large-scale in-the-wild conditions.

C. Ablation Studies

We conduct ablation studies to validate the effectiveness of the main components in *UnSFlow*, including UCA, DUGN, and USAM. In the baseline setting, we use R3D, BiLSTM, and conventional self-attention for spatiotemporal feature learning, and adopt cross-entropy loss for optimization. We then progressively introduce UCA, DUGN, and USAM to analyze their individual and combined contributions.

1) The Importance of Our Designed Modules: As shown in Tab. III, each module contributes positively to the overall performance of *UnSFlow*.

Adding UCA improves the baseline by introducing constrained intra-class variations while preserving temporal consistency, indicating the benefit of controlling expression evolution uncertainty at the data level. Introducing DUGN further enhances spatiotemporal representation learning by estimating feature reliability and refining uncertain temporal responses. When USAM is incorporated, the model obtains additional gains, suggesting that uncertainty-aware perturbation scaling helps stabilize optimization for ambiguous and hard-to-learn samples.

The results with different module combinations further show that UCA, DUGN, and USAM are complementary rather than isolated improvements. UCA acts on the augmentation process, DUGN refines feature representation, and USAM regularizes the optimization trajectory. Their joint integration achieves the strongest overall performance, validating the effectiveness of progressively suppressing uncertainty from data, representation, and optimization perspectives.

We further analyze the two key components in DUGN, namely the UEB and UARA, as shown in Tab. IV. Using either UEB or UARA improves the baseline, indicating that temporal reliability estimation and uncertainty-aware routed attention are both useful for reducing feature-level uncertainty. Their combination achieves the best performance among the compared DUGN variants, showing that reliability modeling and routed attention refinement provide complementary benefits for robust spatiotemporal feature learning.

2) Hyperparameter Analysis of UCA: We analyze two key hyperparameters in UCA: the mixing ratio λ, and the block size B. Here, λ controls the proportion of replaced regions, B determines the spatial granularity of block-wise mixing.

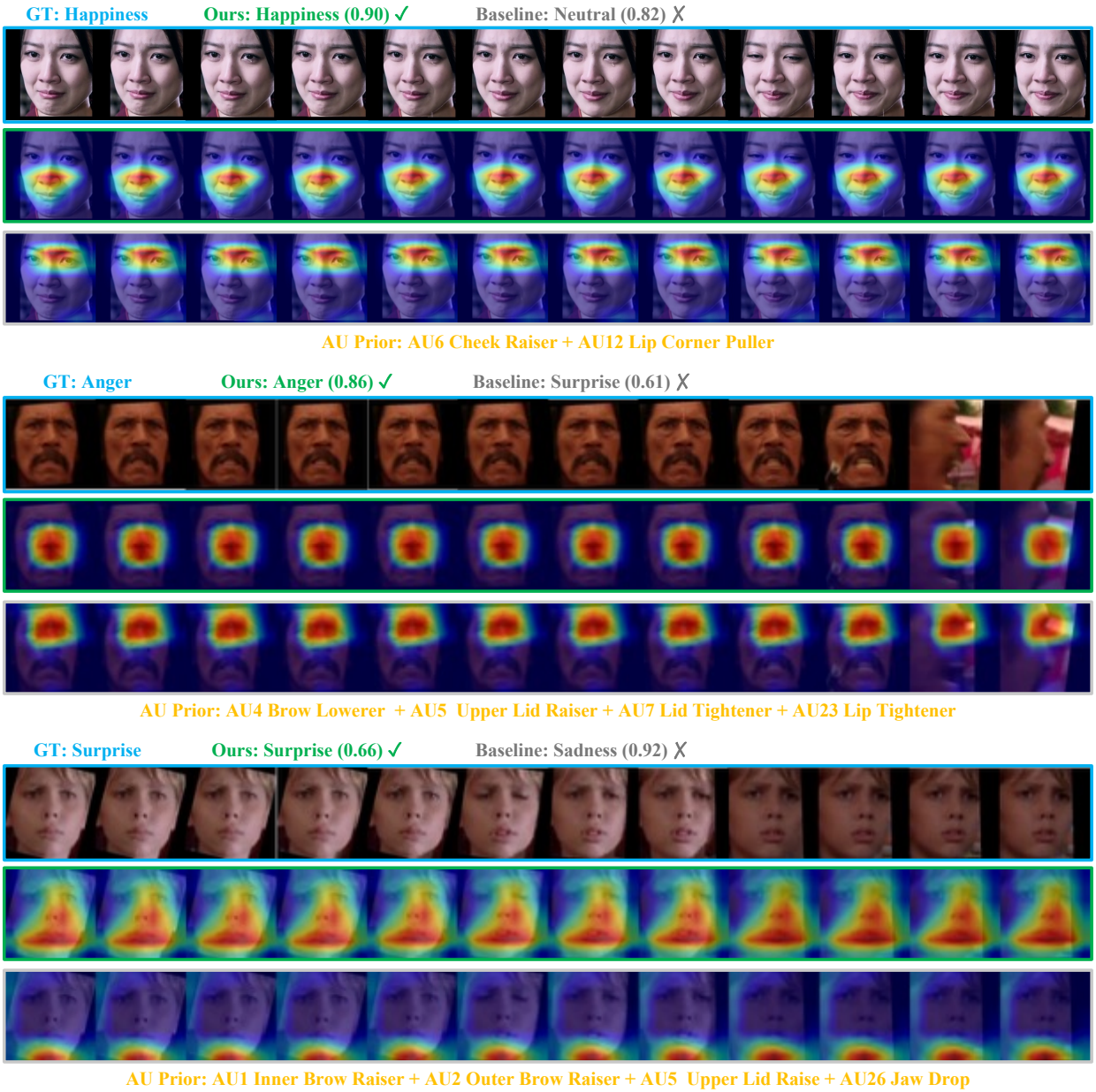


Fig. 8: Grad-CAM visualization with AU-based interpretation on representative DFEW samples. The listed common AU cues are used only for post-hoc interpretation and are not used during training.

As shown in Fig. 6, λ and B jointly affect the trade-off between augmentation diversity and semantic consistency. A moderate λ introduces useful intra-class variations, whereas an overly large or small value may cause semantic drift or insufficient regularization. Similarly, B controls the granularity of spatial perturbations: a smaller B produces fine-grained local mixing, while a larger B leads to coarser regions and stronger structural continuity within each block. Thus, appropriate settings of λ and B are important for balancing local variation and facial structure preservation.

Tab. V further shows the effect of the temporal consistency constraint. Compared with the ablated variant without temporal consistency, where masks are independently sampled for different frames, the default UCA with temporal mask sharing achieves better performance. This indicates that coherent tube-wise mixing across frames helps preserve expression evolution and reduce augmentation-induced temporal disruption.

3) The Effectiveness of USAM Compared to AdamW and Sharpness-Aware Minimization: We compare USAM with AdamW and Sharpness-Aware Minimization (SAM) to evaluate the effect of uncertainty-aware optimization. As shown in Fig. 7 and Tab. VI, USAM achieves better WAR and UAR than AdamW and SAM under the same training setting. AdamW provides a standard optimization baseline but does not explicitly consider local sharpness or predictive uncertainty. SAM improves optimization robustness by searching for flatter local regions with a fixed perturbation radius. However, a fixed perturbation strategy may be less adaptive when training samples exhibit different uncertainty levels.

USAM extends SAM by dynamically adjusting the pertur-

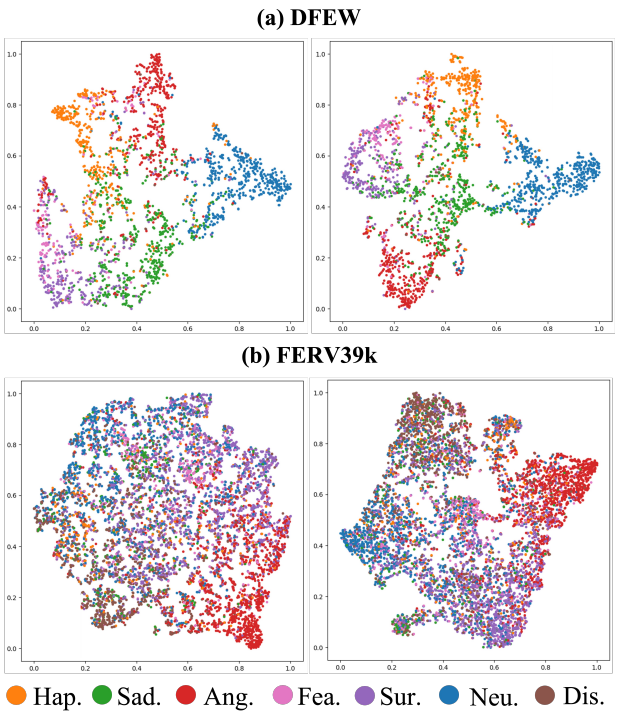


Fig. 9: Visualization of the feature distribution learned by the baseline (left) and our method (right) on DFEW (fd5) and FERV39k.

bation radius according to predictive uncertainty. This design allows the optimizer to apply uncertainty-aware smoothing during training, which is especially useful for ambiguous or hard-to-learn samples. As a result, USAM obtains more stable optimization behavior and better recognition performance in the later training stages. These results suggest that uncertainty-aware perturbation scaling can better adapt the optimization process to the uncertain and imbalanced nature of in-the-wild DFER data.

D. Computational Cost Analysis

To evaluate the efficiency of the proposed uncertainty-aware modeling, we report the inference cost under different DUGN configurations in Tab. VII, including FLOPs, GPU memory, time per iteration, and FPS. Since UCA and USAM are only used during training for data augmentation and optimization, respectively, they do not introduce additional inference FLOPs. Therefore, the inference-time overhead of *UnSFlow* mainly comes from DUGN. The results show that introducing UEB and UARA of DUGN only brings limited extra computational and runtime cost, while the model still maintains efficient inference speed. Moreover, the additional overhead introduced by UARA over UEB is marginal, indicating that the uncertainty-aware routed attention can refine features with little extra computation.

E. Visualization

1) Grad-CAM Visualization with AU-based Interpretation: To interpret the learned activation patterns, we visualize representative DFEW samples using Grad-CAM [55] and relate

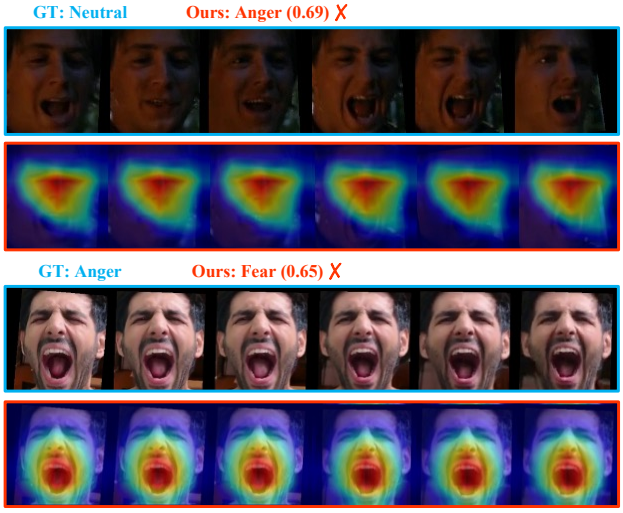


Fig. 10: Failure case analysis under ambiguous real-world expressions. These examples illustrate remaining challenges caused by ambiguous Neutral annotations and overlapping facial action patterns among high-arousal emotions.

the highlighted regions to commonly recognized emotion-AU associations defined in FACS/EMFACS [56], [57]. The listed AUs correspond to the ground-truth emotion categories and are used only as post-hoc interpretive cues, rather than as training supervision, automatic AU annotations, or additional model inputs.

As shown in Fig. 8, *UnSFlow* produces activation regions that are more consistent with emotion-related AU cues. For Happiness, it focuses on cheek and mouth-corner regions associated with AU6 and AU12; for Anger, it attends to brow, eye, and mouth tension related to AU4, AU5, AU7, and AU23; and for Surprise, it highlights brow, eye, and mouth regions associated with AU1, AU2, AU5, and AU26. In contrast, the baseline tends to produce broader or less specific activations, leading to confusion with Neutral, Surprise, or Sadness in these examples. These results suggest that the proposed uncertainty suppression pipeline helps reduce ambiguous responses and encourages the model to focus on more stable and emotion-discriminative facial dynamics.

2) Visualization of Loss Landscape: We adopt a visualization method similar to SAM [58] to analyze the loss landscape of the model, with results shown in Fig. 7. For a fair comparison, SAM and USAM use the same base perturbation radius as specified in the implementation details. Compared with AdamW and SAM, USAM produces a relatively flatter landscape by adapting the perturbation magnitude according to predictive uncertainty. This observation is consistent with the motivation of USAM, indicating that uncertainty-aware perturbation scaling can help reduce optimization sensitivity for ambiguous and high-uncertainty samples.

3) Visualization of t-SNE: We use t-SNE [59] to analyze the feature distributions learned by the baseline and our method, as shown in Fig. 9. For both DFEW and FERV39k, the proposed method produces more compact intra-class feature distributions and clearer inter-class boundaries compared with the baseline. This suggests that *UnSFlow* improves discriminative

representation learning and helps reduce feature entanglement among ambiguous or minority expression categories.

F. Limitations and Future Work

Although *UnSFlow* improves DFER performance by progressively mitigating uncertainty from augmentation, representation learning, and optimization perspectives, several challenges remain in complex in-the-wild scenarios. As shown in Fig. 10, some remaining errors are closely related to the intrinsic ambiguity of dataset annotations and the mismatch between discrete emotion labels and continuous facial expression evolution. Most DFER datasets assign a single categorical label to each video, whereas real facial expressions often evolve continuously and may contain mixed, transitional, or low-intensity emotional states. Such an annotation protocol inevitably compresses complex temporal affective dynamics into a fixed category, introducing label ambiguity and class-wise confusion.

For example, a video annotated as Neutral may still contain subtle muscle changes, weak affective cues, or transitional expression patterns. Under low illumination, pose variation, or partial occlusion, these weak cues may become visually similar to emotion-related patterns such as Anger or Sadness. Similarly, high-intensity expressions may share overlapping AU-related facial patterns; for instance, Anger, Fear, and Surprise can all involve upper-face tension, eye-region activation, and strong mouth opening. These observations suggest that part of the remaining errors may originate from subjective annotation bias and the limited expressiveness of discrete labels, rather than simple model instability.

These limitations indicate several directions for future research. Finer-grained annotation protocols, such as frame-level emotion intensity, AU-level descriptions, or continuous valence-arousal labels, may help reduce video-level label ambiguity. Modeling dynamic emotional transitions and AU-guided temporal relations may provide a more faithful description of real facial expressions. Moreover, multimodal cues, such as audio, language, or physiological signals, may help resolve ambiguities caused by low-quality visual inputs, subjective annotations, and overlapping facial action patterns. In future work, we will investigate uncertainty-aware continuous emotion modeling and multimodal/self-supervised learning to further improve the reliability and generalization of DFER in real-world scenarios.

V. Conclusion

In this paper, we propose *UnSFlow*, an uncertainty suppression framework for dynamic facial expression recognition. *UnSFlow* addresses expression evolution uncertainty, data uncertainty, and model uncertainty through three components: UCA, DUGN, and USAM. Specifically, UCA introduces constrained intra-class augmentation to improve data diversity while preserving temporal consistency; DUGN performs uncertainty-aware spatiotemporal feature modeling to reduce feature confusion; and USAM stabilizes optimization by adapting perturbation strength according to predictive uncertainty. Experiments on DFEW and FERV39k demonstrate the effectiveness of *UnSFlow* in improving robust and balanced DFER performance under challenging in-the-wild conditions. In future work, we will explore finer-grained emotion modeling, multimodal fusion, and self-supervised learning to further improve the reliability and generalization of DFER systems.

Acknowledgments

This work is supported by the 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, 62272144, 72188101), the Fundamental Research Funds for the Central Universities (JZ2025HGTB0225, JZ2024AHST0337), National Key R&D Program of China (NO.2024YFB3311602), and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

[1] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," *IEEE Transactions on Multimedia*, 2024.

[2] F.-Q. Cui, J. Huang, S. Zhao, K. Li, Z. Liu, M. Li, Z. Jia, D. Guo, and M. Wang, "Persomoni: A comprehensive video-based benchmark dataset for fine-grained personality assessment with 15 trait dimensions," *IEEE Transactions on Affective Computing*, 2026.

[3] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," *IEEE Transactions on Affective Computing*, 2022.

[4] F.-Q. Cui, J. Huang, J.-C. Zhao, S. Zhao, Y. Yang, D. Guo, X. Zhou, Z. Zhong, F. Zhang, J. Lu, M. Li, and B.-k. Bao, "Sain: A hierarchical dynamics-aware eeg–fnirs framework for continuous affect estimation download," in *ACM International Conference on Multimedia*, 2026.

[5] 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.

[6] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[7] 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.

[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] E. Pei, M. C. Oveneke, Y. Zhao, D. Jiang, and H. Sahli, "Monocular 3d facial expression features for continuous affect recognition," *IEEE Transactions on Multimedia*, 2021.

[10] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[11] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, 2018.

[12] S. Wang, H. Shuai, and Q. Liu, "Phase space reconstruction driven spatio-temporal feature learning for dynamic facial expression recognition," *IEEE Transactions on Affective Computing*, 2022.

[13] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Transactions on Cybernetics*, 2019.

[14] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Transactions on Image Processing*, 2020.

[15] F.-Q. Cui, A. Tong, J. Huang, J. Zhang, M. Li, X. Yan, L. Huang, D. Guo, and M. Wang, "Toward trustworthy dynamic facial expression recognition via information bottleneck modeling," *IEEE Transactions on Information Forensics and Security*, vol. 21, pp. 6985–6999, 2026.

[16] 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.

[17] W. Chen, D. Zhang, M. Li, and D.-J. Lee, "Stcam: Spatial-temporal and channel attention module for dynamic facial expression recognition," *IEEE Transactions on Affective Computing*, 2023.

[18] 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.

[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*, 2023.

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

[21] L. Sun, Z. Lian, B. Liu, and J. Tao, "MAE-DFER: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[22] L. Li, Y. Zheng, S. Liu, X. Xu, and T. Li, "Domain knowledge enhanced vision-language pretrained model for dynamic facial expression recognition," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[23] X. Wu, H. Sun, Y. Wang, J. Nie, J. Zhang, Y. Wang, J. Xue, and L. He, "AVF-MAE++: Scaling affective video facial masked autoencoders via efficient audio-visual self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[24] Z. Qi, S. Wang, W. Zhang, and Q. Huang, "Uncertainty-boosted robust video activity anticipation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[25] F.-Q. Cui, J. Huang, S. Zhao, X. Li, X. Yan, Z. Jia, and X. Zhou, "Robust low-rank sparse framework for video-based affective computing," *IEEE Transactions on Consumer Electronics*, 2026.

[26] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[27] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016.

[28] N. Le, K. Nguyen, Q. D. Tran, E. Tjiputra, B. Le, and A. Nguyen, "Uncertainty-aware label distribution learning for facial expression recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[29] Z. Wu and J. Cui, "La-net: Landmark-aware learning for reliable facial expression recognition under label noise," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[30] H. Wang, X. Mai, Z. Tao, X. Tong, J. Lin, Y. Wang, J. Yu, S. Yan, Z. Zhou, and W. Zhang, "D2sp: Dynamic dual-stage purification framework for dual noise mitigation in vision-based affective recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[31] 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 *ACM Multimedia 2025*, 2025.

[32] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y.-G. Jiang, "Svformer: Semi-supervised video transformer for action recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] 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.

[34] ——, "Empc: Efficient multi-view parallel co-learning for semi-supervised action recognition," *Expert Syst. Appl.*, 2024.

[35] Q. Yang, X. Sun, X. Li, F. Cui, Y. Guo, S. Hu, P. Luo, and S. Li, "Prior knowledge distillation network for face super-resolution," in *Proceedings of the 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology*, 2025.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[37] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. J. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[38] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, 1951.

[41] C. Zheng, G. Wu, and C. LI, "Toward understanding generative data augmentation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[42] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[43] P. Jin, B. Zhu, L. Yuan, and S. YAN, "Moe++: Accelerating mixture-of-experts methods with zero-computation experts," in *The Thirteenth International Conference on Learning Representations*, 2025.

[44] 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.

[45] F. Ma, B. Sun, and S. Li, "Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[46] 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.

[47] F. Liu, H. Wang, and S. Shen, "Robust dynamic facial expression recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.

[48] 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.

[49] 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.

[50] 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.

[51] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2021.

[52] 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.

[53] 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 *2025 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 2025.

[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.

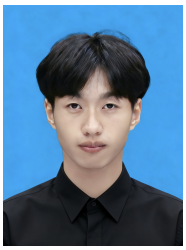
[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[56] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[57] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

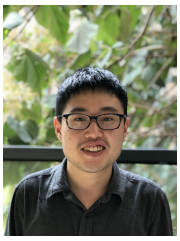
[58] M. Li, Y. Liu, Y. Lu, Y. Zhang, Y. ming Cheung, and H. Huang, "Improving visual prompt tuning by gaussian neighborhood minimization for long-tailed visual recognition," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[59] 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 the Ph.D. degree with the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, School of Information Science and Technology, University of Science and Technology of China, 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 and brain-inspired intelligence.

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.



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 74 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.

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. 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 uncertainty learning.

Ziyu Jia (Member, IEEE) is an Assistant Professor at the Institute of Automation, Chinese Academy of Sciences. His research focuses on time-series analysis methods and their applications in health and medicine, including multimodal affective computing, sleep stage classification, and brain-computer interfaces. He has published over 50 peer-reviewed papers in venues such as IEEE Transactions on Affective Computing, IEEE Transactions on Multimedia, IEEE Transactions on Neural Systems and Rehabilitation Engineering, KDD, and ICLR. Dr. Jia currently serves as an Associate Editor or Editorial Board Member for prestigious journals including IEEE Transactions on Affective Computing and Information Fusion, and he is an Area Chair for major AI and machine learning conferences such as IJCAI and IJCNN. In addition to his academic contributions, Dr. Jia has extensive industry experience, having successfully led multiple R&D projects and secured several patents. He has received numerous honors, including the MSRA StarTrack Award, and the CIE Young Talent Award.

Zhi Liu (S'11-M'14-SM'19) received the Ph.D. degree in informatics in National Institute of Informatics. He is currently an Associate Professor at The University of Electro-Communications. His research interest includes video network transmission and mobile edge computing. He is now an editorial board member of Springer wireless networks and IEEE Open Journal of the Computer Society. He is a senior member of IEEE.

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).