Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(10): 7274–7298
DOI: 10.3934/era.2026315
Received: 02 July 2026
Revised: 10 August 2026
Accepted: 13 August 2026
Published: 25 August 2026

Research article

RDAS: Reliability-driven adaptive supervision for noisy-label facial expression recognition

Xuefeng Zhao* **and Yixuan Dong**

School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, China

* **Correspondence:** Email: zhaoxf@jou.edu.cn.

Abstract: In-the-wild facial expression recognition (FER) is inherently affected by ambiguous expressions and subjective annotations, making the observed labels only partially reliable. Conventional one-hot supervision nevertheless treats these labels as ground truth and may drive deep networks to memorize corrupted annotations. This paper proposes RDAS, a reliability-driven adaptive supervision framework for noisy-label FER. The central idea was to use label reliability not merely as a sample weight, but as a variable that determines the supervision policy of each training sample. RDAS enhances expression-sensitive representations by combining RGB appearance with structural high-frequency cues, estimates label reliability with an exponential moving average (EMA) teacher through the margin between the annotated class and its strongest competing class, and assigns reliability-specific objectives to high-, medium-, and low-reliability samples. Reliable samples retained hard-label learning, ambiguous samples received teacher-guided soft correction, and unreliable samples were prevented from dominating training with misleading gradients. Experiments on RAF-DB, AffectNet, and FERPlus datasets showed that RDAS improved both standard recognition performance and robustness under synthetic label noise. On RAF-DB, RDAS achieved an average accuracy of 89.70% on the original dataset and 86.15% under 30% label noise. Ablation studies provided quantitative support for reliability modeling and adaptive supervision, while the visual analyses provided complementary qualitative evidence consistent with the proposed mechanism.

Keywords: facial expression recognition; noisy-label learning; label reliability; adaptive supervision; complex-enhanced representation

1. Introduction

Facial expression recognition (FER) supports affective computing, human–computer interaction, and behavior analysis [1–3]. Large in-the-wild datasets, including FER2013, FERPlus, RAF-DB, and AffectNet, have advanced deep FER models [4–7], yet variations in identity, pose, illumination,

occlusion, resolution, and background still obscure subtle expression cues.

Label unreliability is a more intrinsic obstacle because expressions are continuous, mixed, and visually overlapping. Fear and surprise may share widened eyes and an opened mouth, while sadness and neutral expressions may differ only through weak muscle changes. Low-intensity, occluded, or mixed expressions also admit subjective annotations [5, 6]; thus, in-the-wild labels are observed annotations rather than error-free ground truth.

Standard cross-entropy nevertheless treats every observed label as a trustworthy one-hot target, although deep networks may eventually memorize corrupted annotations [8]. With already uncertain FER boundaries, this behavior can bias class separation and encourage unstable non-expression cues, especially as the noise rate increases.

Existing noisy-label FER studies have explored several complementary directions. Self-Cure Network (SCN) and Relative Uncertainty Learning (RUL) reduce the influence of unreliable annotations through sample weighting and relative uncertainty modeling [9, 10]. Erasing Attention Consistency (EAC) and subsequent consistency-based methods regularize attention or predictions under perturbations to mitigate noise memorization [11–13]. Other methods refine supervision through label correction, adversarial learning, dual-branch noise suppression, or collaborative adaptation [14–18]. However, instantaneous student estimates may fluctuate, binary clean/noisy separation is too coarse for ambiguous FER samples, and scalar weighting retains the same loss form for every reliability level. Robust FER should therefore determine both how much a sample is trusted and how it supervises the model.

We therefore propose RDAS, a reliability-driven adaptive supervision framework that treats reliability as a supervision-policy variable. Its representation–reliability–supervision pipeline first combines RGB appearance and structural high-frequency cues, and then uses an exponential moving average (EMA) teacher to compare support for the annotated class with its strongest competitor. The resulting reliability assigns high-, medium-, and low-reliability samples to base supervision, teacher-guided soft correction, or weak consistency without misleading label-dependent fitting, respectively.

On RAF-DB with 30% label noise, RDAS improves the baseline from $82.93 \pm 0.15\%$ to $86.15 \pm 0.19\%$. Ablations, reliability analysis, and visualizations further support the proposed mechanism.

The main contributions of this work are summarized as follows:

- We propose a reliability-driven formulation for noisy-label FER, where sample reliability determines both the strength and form of supervision rather than serving only as a scalar loss weight.
- We develop RDAS as a unified representation–reliability–supervision framework, combining complex-enhanced expression representation, EMA-stabilized label reliability modeling, and adaptive supervision for high-, medium-, and low-reliability samples.
- Experiments on RAF-DB, AffectNet, and FERPlus demonstrate that RDAS improves both original-dataset recognition and robustness under 10%, 20%, and 30% synthetic label noise, with ablation and visualization results supporting the proposed design.

2. Related work

The most relevant literature spans expression representation, noisy-label FER, and reliability-aware supervision. These threads motivate the representation–reliability–supervision structure of RDAS.

2.1. Facial expression representation learning

Early FER studies analyzed handcrafted texture and shape cues, while later surveys summarized the transition toward data-driven affect recognition [2, 3]. Large in-the-wild datasets, including FER2013, FERPlus, RAF-DB, and AffectNet, further enabled convolutional neural network (CNN)-based hierarchical representation learning [4–7]. Deep features are substantially more robust to appearance variation, but subtle local deformations remain difficult to preserve under pose, illumination, and occlusion changes.

Attention-based designs such as squeeze-and-excitation (SE), convolutional block attention module (CBAM), region attention network (RAN), and deeply-supervised attention network (DSAN) enhance expression-sensitive channel, spatial, or regional responses [19–22]. Transformer and cross-attention variants further model nonlocal facial relations and feature interactions [23–26]. Recent FER models also incorporate online label correction, landmark-aware learning, or multi-branch cross-fusion to strengthen region-level evidence [27–29]. These designs highlight the importance of the eyes, eyebrows, nose, and mouth, where local muscle movements carry much of the class-discriminative evidence.

Most representation methods optimize recognition directly. RDAS uses representation enhancement for the related but distinct purpose of making label assessment more dependable: a compact complex-enhanced stem introduces structural high-frequency responses, and CBAM refines them into facial evidence for label reliability modeling (LRM). This design requires neither external landmarks nor multiple local branches and remains compatible with a standard backbone.

2.2. Robust FER under label noise

Noisy-label FER methods either suppress unreliable samples or refine their supervision. SCN learns sample weights, and RUL models relative uncertainty to reduce mislabeled-sample influence [9, 10]. Distribution Mining and Uncertainty Estimation (DMUE) estimates latent label distributions through multiple branches, whereas reliable label noise suppression (ReSup) exchanges weights between networks according to prediction–label similarity [14, 30]. Recent work further explores uncertainty-aware distributions, online correction, and ambiguity navigation [15, 27, 31].

Consistency-based approaches regularize attention, predictions, or features under perturbations [11–13]. In particular, EAC uses erasing-attention consistency to discourage overfitting and preserve expression-related attention. More recent three-way adaptive uncertainty-suppressing model (3WAUS), dual-branch suppression, and collaborative noisy label adaptive learning (CNLA) investigate three-way uncertainty, explicit noise extraction, and collaborative adaptation [17, 18, 32].

Beyond FER, related studies use adversarial transformers for weakly supervised crack segmentation [33], cross-domain stylization and dual adversarial learning for domain adaptation [34], and region-wise distillation with prototype-balanced contrastive learning for class-incremental segmentation [35]. Although these tasks differ from noisy-label FER, they provide a broader perspective on extracting dependable signals under imperfect supervision; RDAS specializes this concern to the compatibility between facial evidence and observed expression labels.

EMA teachers stabilize prediction targets through temporal ensembling [36]. Building on the standard EMA update used in the mean teacher method, RDAS employs the teacher for a purpose distinct from conventional consistency learning. Specifically, the teacher compares the observed FER label with its strongest competing class, constructs reliability-dependent targets for Medium-reliability samples, and guides the selection of RAS supervision policies. The temporally smoothed predictions therefore serve

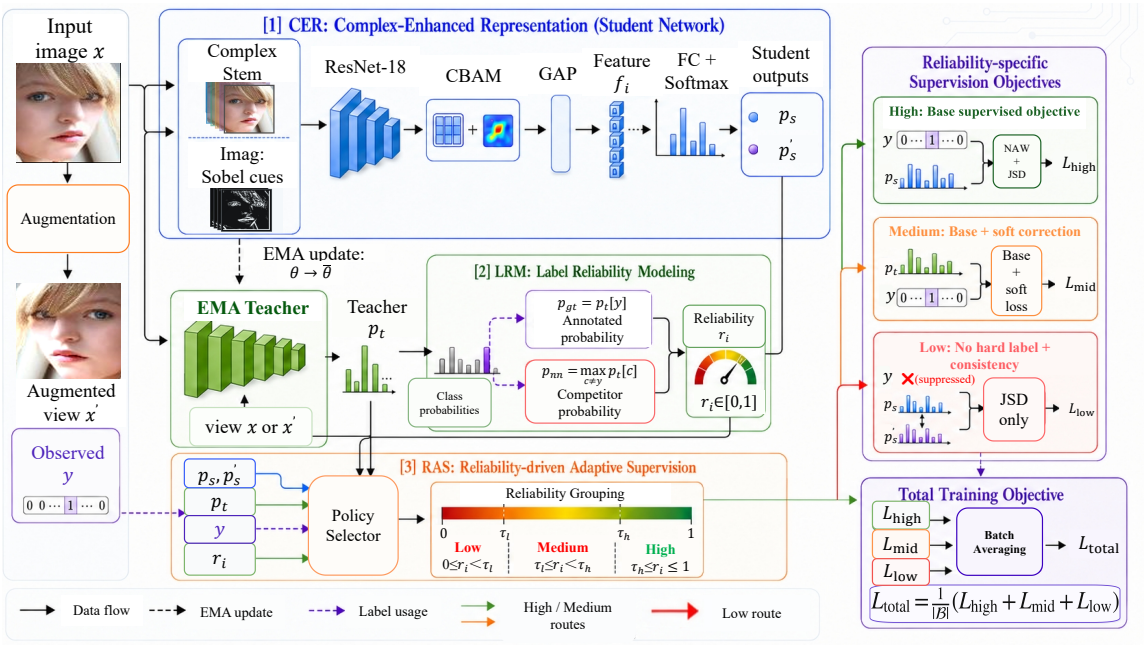


Figure 1. Overall framework of RDAS. The CER student network extracts expression representations from the original and augmented views. The EMA teacher provides stable predictive distributions for label reliability modeling. RAS then assigns reliability-specific supervision objectives according to the estimated reliability.

to assess annotation compatibility rather than merely providing generic consistency targets.

2.3. Reliability estimation and adaptive supervision

Prediction uncertainty and label reliability are distinct: the former measures confidence in an output, whereas the latter measures compatibility between the observed annotation and facial evidence. An ambiguous but correctly labeled sample may remain useful despite low confidence; conversely, a mislabeled sample may receive high confidence for another class. RDAS therefore uses the teacher margin between the annotated class and its strongest competitor, so strong support for an alternative directly reduces reliability rather than being mistaken for trustworthy confidence.

Binary clean/noisy separation may discard ambiguous but learnable samples, whereas continuous weighting leaves the loss form unchanged. RDAS instead assigns High samples base supervision, Medium samples an auxiliary teacher correction, and Low samples weak consistency without label-dependent fitting. Unlike landmark-assisted LA-Net [28], RDAS does not require external landmark detection. It also differs from collaborative CNLA [18] and uncertainty-suppression methods [9, 10, 32] by explicitly mapping a label-consistency margin to reliability-specific supervision objectives. RDAS thereby integrates expression-sensitive representation, EMA-stabilized reliability estimation, and adaptive supervision within a unified framework.

3. Method

Figure 1 presents RDAS as reliability-conditioned empirical risk minimization with three connected components: complex-enhanced representation (CER) extracts expression-sensitive features, label

reliability modeling (LRM) evaluates the observed annotation with an EMA teacher, and reliability-driven adaptive supervision (RAS) converts that estimate into different objectives.

3.1. Problem formulation and framework overview

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote an in-the-wild FER training set, where $x_i \in \mathbb{R}^{3\times H\times W}$ is a facial image and $y_i \in \{1, \ldots, K\}$ is the observed expression label. Because annotations may be affected by subjectivity and ambiguity, y_i may differ from the latent true label y_i^*. Conventional empirical risk minimization optimizes

$$\theta_{\text{ce}}^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{ce}}(F(x_i; \theta), y_i), \tag{1}$$

which implicitly treats every observed label as reliable. This assumption is overly strong for FER, where visually ambiguous samples and subjective annotations make $P(y_i = y_i^* \mid x_i) < 1$ for a non-negligible portion of the training set.

RDAS relaxes this assumption by introducing a sample-wise reliability score $r_i \in [0, 1]$. The learning objective becomes

$$\theta_{\text{rdas}}^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(x_i, y_i; r_i, \theta), \tag{2}$$

where r_i determines the form of supervision. A high reliability score indicates that the annotation is sufficiently supported and can be used as a hard target. A medium score indicates that the annotation is informative but uncertain and should be softened. A low score suggests that the annotation may be misleading and should not dominate optimization. Thus, reliability is used to select the supervision policy rather than only to rescale the loss.

RDAS contains a student model $F_s(\cdot; \theta)$ and an EMA teacher $F_t(\cdot; \bar{\theta})$. Given an image x_i and an augmented view x_i', the student produces p_i^s and $p_i^{s\prime}$, while the teacher produces a temporally smoothed distribution p_i^t. The observed label y_i selects the annotated-class probability from p_i^t, and the largest probability among the remaining classes defines the strongest competing class. The relationship between these two probabilities forms the reliability estimate that drives the subsequent adaptive supervision.

3.2. Complex-enhanced expression representation

Accurate reliability estimation depends on features that preserve expression-discriminative evidence. In FER, important cues often appear as subtle high-frequency structures, such as eyebrow deformation, eye contours, nasolabial folds, and mouth corners. A simple concatenation of handcrafted structural maps and RGB channels treats appearance and structure as independent inputs. RDAS instead organizes RGB appearance and structural responses as the real and imaginary components of a unified complex representation.

For image x_i, the complex input is defined as

$$z_i = x_i + \mathrm{i}\,\psi(x_i), \tag{3}$$

where $\psi(\cdot)$ denotes a Sobel or high-pass operator that extracts structural responses. A complex convolution maps z_i into

$$U_i = \text{Conv}^{\mathbb{C}}(z_i) = A_i + \mathrm{i}B_i, \tag{4}$$

where $A_i = \mathrm{Re}(U_i)$ and $B_i = \mathrm{Im}(U_i)$. The complex response is projected to a real feature map by

$$\tilde{U}_i = A_i + \tanh(\gamma) B_i, \tag{5}$$

where γ is a learnable fusion coefficient. Unlike fixed structural fusion, this projection allows the contribution of high-frequency responses to be adjusted according to the training objective. For an objective $\mathcal{L}$, the gradient with respect to γ is

$$\frac{\partial \mathcal{L}}{\partial \gamma} = \left\langle \frac{\partial \mathcal{L}}{\partial \tilde{U}_i}, \mathrm{sech}^2(\gamma) B_i \right\rangle, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes summation over spatial and channel dimensions. The structural component can therefore be strengthened when it supports expression discrimination and weakened when it introduces noise.

The projected feature is passed through a ResNet-18 backbone $\Phi(\cdot)$:

$$G_i = \Phi(\tilde{U}_i). \tag{7}$$

CBAM further recalibrates the representation through channel attention M_c and spatial attention M_s:

$$\hat{G}_i = M_s(M_c(G_i) \otimes G_i) \otimes (M_c(G_i) \otimes G_i), \tag{8}$$

where $\otimes$ denotes element-wise multiplication. The prediction of the student network is then obtained by

$$f_i = \mathrm{GAP}(\hat{G}_i), \quad o_i^s = W f_i + b, \quad p_i^s = \mathrm{softmax}(o_i^s). \tag{9}$$

The augmented view x_i' is processed by the same student network to obtain $p_i^{s\prime}$, which supports consistency learning without introducing additional inference branches.

3.3. EMA-stabilized label reliability modeling

Reliability estimation should measure whether the observed label agrees with the semantic evidence in the image. Directly estimating this quantity from the current student may be unstable because the student is updated by mini-batches containing noisy labels. RDAS therefore maintains an EMA teacher:

$$\bar{\theta} \leftarrow \alpha \bar{\theta} + (1 - \alpha)\theta, \tag{10}$$

where α is the EMA decay. The teacher is not optimized by back-propagation; it acts as a temporal ensemble of historical students and provides a smoother distribution

$$p_i^t = F_t(x_i; \bar{\theta}). \tag{11}$$

Following the temporal-ensembling principle of the mean teacher method [36], α controls the trade-off between responsiveness and temporal smoothing. A smaller value tracks the current student more rapidly but produces a less-stable reliability signal, whereas a value closer to one suppresses short-term fluctuations but may lag behind the student. We use $\alpha = 0.995$ after a five-epoch warm-up and keep it fixed across all datasets and noise settings, avoiding dataset-specific tuning of the teacher dynamics.

Given the observed label y_i, RDAS defines

$$p_{\mathrm{gt},i} = p^t_{i,y_i}, \quad p_{\mathrm{nn},i} = \max_{k \neq y_i} p^t_{i,k}. \tag{12}$$

Here $p_{\mathrm{gt},i}$ denotes teacher support for the annotated class, while $p_{\mathrm{nn},i}$ denotes support for the strongest competing class. Their difference

$$m_i = p_{\mathrm{gt},i} - p_{\mathrm{nn},i} \tag{13}$$

is interpreted as a label-consistency margin. A positive margin means that the observed label is more plausible than any alternative category under the teacher distribution. A negative margin indicates that another expression class explains the image better than the annotation, which may imply label corruption or strong ambiguity. This margin is more informative than maximum confidence alone because high confidence for a class different from y_i should decrease, rather than increase, label reliability.

The final reliability score combines relative margin evidence and absolute annotation confidence:

$$r_i = \lambda_m \sigma(\eta m_i) + (1 - \lambda_m) p_{\mathrm{gt},i}, \tag{14}$$

where $\sigma(\cdot)$ is the sigmoid function, η controls the sharpness of the margin transformation, and λ_m balances relative and absolute evidence. The margin term evaluates whether the annotation defeats its strongest semantic competitor, while the confidence term prevents uniformly uncertain samples from being overestimated. The resulting r_i is a soft estimate of label reliability rather than a binary clean/noisy indicator.

3.4. Noise-aware supervised objective

Before applying reliability-specific supervision, RDAS defines a base objective that combines label-dependent classification and label-independent consistency. For the student prediction p^s_i, let

$$p^s_{\mathrm{gt},i} = p^s_{i,y_i}, \quad p^s_{\mathrm{nn},i} = \max_{k \neq y_i} p^s_{i,k}. \tag{15}$$

The vector $q_i = [p^s_{\mathrm{gt},i}, p^s_{\mathrm{nn},i}]^T$ represents the student-side support for the annotation and its strongest competitor. A noise-aware weight is defined as

$$w_i = \exp\left(-\frac{1}{2}(q_i - \mu)^T \Sigma^{-1} (q_i - \mu)\right), \tag{16}$$

where μ and Σ specify a reference region in the two-dimensional confidence space. Their implementation is fully specified as follows. When $p^s_{\mathrm{gt},i} > p^s_{\mathrm{nn},i}$, we use $\mu^+ = [0.5, 0.5]^T$ and

$$\Sigma^+(t) = \begin{bmatrix} 0.08 & 0.04c_t \\ 0.04c_t & 0.08 \end{bmatrix}, \qquad c_t = 1 - \exp\left(-\frac{10t}{T}\right), \tag{17}$$

where t and T denote the current and total numbers of training epochs, respectively. Otherwise, we use $\mu^- = [0.3, 0.15]^T$ and $\Sigma^- = \left[\begin{smallmatrix} 0.05 & 0.02 \\ 0.02 & 0.05 \end{smallmatrix}\right]$. These quantities are fixed by the training protocol and are not estimated from the test set. The weight therefore depends on the joint relation between the annotated class and its strongest competitor rather than on annotated-class confidence alone. The weighted classification loss is

$$\ell_{\mathrm{naw},i} = (1 + w_i)\ell_{\mathrm{ce}}(o^s_i, y_i). \tag{18}$$

This term increases the influence of samples whose confidence pattern is close to the expected reliable region.

For the original and augmented views, let $P_i = p_i^s$, $Q_i = p_i^{s\prime}$, and $M_i = (P_i + Q_i)/2$. Prediction consistency is imposed by Jensen-Shannon divergence:

$$\ell_{\mathrm{jsd},i} = \frac{1}{2} D_{\mathrm{KL}}(P_i \| M_i) + \frac{1}{2} D_{\mathrm{KL}}(Q_i \| M_i). \tag{19}$$

The base supervised objective is

$$\ell_{\mathrm{sup},i} = \lambda_s \ell_{\mathrm{naw},i} + (1 - \lambda_s)\ell_{\mathrm{jsd},i}, \tag{20}$$

where λ_s balances classification and consistency. This objective provides the foundation for the reliability-conditioned losses described below.

3.5. Reliability-driven adaptive supervision

RAS converts reliability into a concrete supervision policy. A binary clean/noisy split is too restrictive for FER because many samples are ambiguous but still useful. A purely continuous weighting scheme is also limited because it applies the same loss form to all samples. RDAS therefore partitions each mini-batch into three reliability regions:

$$\mathcal{B} = \mathcal{B}_h \cup \mathcal{B}_m \cup \mathcal{B}_l, \tag{21}$$

where $\mathcal{B}_h$, $\mathcal{B}_m$, and $\mathcal{B}_l$ denote high-, medium-, and low-reliability subsets. For a mini-batch containing B samples, the batch statistics are computed as $\bar{r} = B^{-1} \sum_{i=1}^{B} r_i$ and $s_r = \sqrt{B^{-1} \sum_{i=1}^{B} (r_i - \bar{r})^2}$. The thresholds are

$$\tau_h = \bar{r} + k s_r, \quad \tau_l = \bar{r} - k s_r. \tag{22}$$

Samples with $r_i \geq \tau_h$ are assigned to $\mathcal{B}_h$, samples with $r_i \leq \tau_l$ are assigned to $\mathcal{B}_l$, and the remaining samples are assigned to $\mathcal{B}_m$. To keep all three supervision branches nonempty in a finite mini-batch, if no sample naturally satisfies the upper or lower boundary, the sample with the highest or lowest reliability score is assigned to $\mathcal{B}_h$ or $\mathcal{B}_l$, respectively. The coefficient k controls the width of the medium-reliability interval relative to the current mini-batch distribution; it is not a fixed reliability threshold. We set $k = 1.0$ as a dataset-independent default and retain the same value under all noise rates. This design adapts the partition to the reliability scale of different batches without requiring the true noise ratio, while the batch-size and training-stage stability of the resulting groups, together with the frequency of the High-group fallback, is evaluated in the subsequent analysis.

For $i \in \mathcal{B}_h$, the annotation is sufficiently supported, and RDAS preserves strong supervision:

$$\ell_i = \ell_{\mathrm{sup},i}. \tag{23}$$

For $i \in \mathcal{B}_m$, the annotation is informative but uncertain. RDAS constructs a reliability-dependent soft target

$$\tilde{y}_i = (1 - a_i) y_i^{\mathrm{onehot}} + a_i p_i^t, \tag{24}$$

where

$$a_i = \beta(1 - r_i). \tag{25}$$

A lower reliability score leads to stronger teacher guidance. The soft target is used as an auxiliary correction term rather than a complete replacement of the base objective, allowing ambiguous samples to retain discriminative information while being pulled toward the teacher distribution. The medium-reliability objective is

$$\ell_i = \ell_{\text{sup},i} + \lambda_r(t)\ell_{\text{soft}}(o_i^s, \tilde{y}_i), \tag{26}$$

where ℓ_{soft} is the cross-entropy with the soft target and $\lambda_r(t)$ controls the strength of relabeling during training. For $i \in \mathcal{B}_l$, hard-label fitting is risky, so RDAS retains only weak consistency:

$$\ell_i = \lambda_l \ell_{\text{jsd},i}. \tag{27}$$

The final objective is

$$L_{\text{total}} = \frac{1}{|\mathcal{B}|}\left(\sum_{i\in\mathcal{B}_h}\ell_i + \sum_{i\in\mathcal{B}_m}\ell_i + \sum_{i\in\mathcal{B}_l}\ell_i\right). \tag{28}$$

Through this formulation, high-reliability samples preserve discriminative learning, medium-reliability samples reduce ambiguity through soft correction, and low-reliability samples reduce memorization by removing misleading hard-label gradients.

3.6. Training and inference

Training begins with a warm-up stage, during which all samples are optimized by the base objective while the EMA teacher is updated. After warm-up, the teacher distribution is used to estimate r_i, and each mini-batch is optimized with the reliability-specific objectives. Mixup is used as an additional regularizer:

$$\bar{x}_{ij} = \lambda x_i + (1-\lambda)x_j, \quad \lambda \sim \text{Beta}(\rho, \rho), \tag{29}$$

with the corresponding loss linearly combined according to λ. At inference time, reliability partitioning and adaptive supervision are no longer required. Consistent with the evaluation protocol, a test image is fed into a single CER network parameterized by the EMA-averaged weights, and the predicted category is

$$\hat{y} = \arg\max_k F_{\text{ema}}(x; \bar{\theta})_k. \tag{30}$$

The EMA network has the same architecture as the student and is retained as a single deployed model rather than an additional teacher branch. Thus, RDAS improves training robustness without introducing reliability estimation or extra forward passes at inference.

The additional cost of LRM and RAS is confined to training. Relative to the two student forward passes used by the original/augmented consistency objective, the EMA teacher requires one extra forward pass without gradient storage. Updating the teacher costs $O(P)$ operations for P model parameters, while reliability computation and thresholding cost $O(BC)$ and $O(B)$, respectively, for batch size B and C expression classes. After training, the instantaneous student weights, reliability estimator, and three-way loss branches are discarded, while the EMA-averaged CER weights are retained as the single deployed network. Therefore, adaptive supervision introduces no additional parameters or forward passes at inference.

As reported in Table 1, full RDAS requires 226.1 s per epoch, compared with 24.4 s for the plain baseline and 64.3 s for the CER student. Most of the additional training cost arises from processing

the augmented view and performing the EMA-teacher forward pass, whereas reliability estimation and three-way policy assignment are comparatively lightweight. On RAF-DB with 30% symmetric label noise, RDAS improves accuracy from 82.93 ± 0.15% to 86.15 ± 0.19%. Despite the higher offline training cost, RDAS retains the same single-network inference cost as the CER student.

4. Experiments

We evaluate clean FER performance, robustness under symmetric and class-dependent noise, component ablations, parameter and partition stability, reliability discrimination, feature structure, and attention localization.

Datasets. We use official splits. RAF-DB contains 12,271 training and 3068 test images from seven expressions [6]; AffectNet uses the common eight-class protocol including contempt [7]; and FERPlus uses its standard eight-class setting [4, 5].

Implementation details. All face images are resized to 224×224, and ResNet-18 initialized from MS-Celeb-1M pretraining is used as the backbone. The network is optimized with Adam using an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, a batch size of 128, and a cosine learning rate schedule for 60 epochs. Unless otherwise specified, the EMA decay is fixed at 0.995, the warm-up stage lasts 5 epochs, the dynamic threshold coefficient is set to $k = 1.0$, and Mixup uses $\rho = 0.2$. The remaining loss-related parameters are kept unchanged across datasets: $\lambda_m = 0.65$, $\eta = 8.0$, $\lambda_s = 0.5$, $\beta = 0.3$, $\lambda_l = 0.05$, teacher temperature is 0.7, and $\lambda_r(t)$ linearly increases from 0 to 0.2 after warm-up. For the LRM-without-RAS ablation, the reliability-weight floor is fixed at $\omega_{\min} = 0.1$. Unless otherwise noted, accuracy is reported using the EMA-averaged CER model, and the same EMA-averaged weights are retained as the single network for deployment.

Computational cost. Table 1 reports FP32 measurements for 224×224 inputs on one NVIDIA GeForce RTX 4060 Laptop GPU. Training time is averaged over all 60 optimization epochs with batch size 128 and excludes validation; memory is the peak allocation. Inference latency averages 200 single-image passes after 50 warm-up iterations and excludes data loading and host-to-device transfer. Deployable parameter counts express each complex coefficient as two real scalars, and each multiply–accumulate operation is counted as one operation.

Table 1. Computational cost of the plain baseline, CER student, and full RDAS. Deployable cost is measured using the single EMA-averaged CER network retained for inference.

Configuration	Params (M)	GFLOPs	Train (s/epoch)	Memory (GB)	Inference (ms/image)
Plain baseline	11.180	1.814	24.4	2.956	3.168
CER student	11.222	2.170	64.3	4.813	3.958
Full RDAS	11.222	2.170	226.1	5.485	3.958

CER adds 0.042M parameters. Full RDAS has the same deployment cost as CER because LRM and RAS operate only during training.

Statistical analysis. Baseline–RDAS differences are assessed by a two-sided paired Student's t-test over five matched seeds at $\alpha = 0.05$; 95% confidence intervals (CIs) are reported for mean paired improvements.

Symmetric synthetic noisy labels. At rates of 10%, 20%, and 30%, the corresponding proportion of training samples is randomly selected and each label is replaced uniformly by one of the other expression categories. Test labels remain unchanged. Because no transition is favored, this setting

provides a controlled measure of robustness to class-independent annotation errors.

Class-dependent asymmetric noisy labels. In contrast to uniform random replacement, asymmetric noise follows predefined transitions between visually or semantically confusable expressions. We construct this structured corruption on the RAF-DB training set using Surprise→Fear, Fear→Surprise, Disgust→Anger, Anger→Disgust, Happiness→Neutral, Sadness→Neutral, and Neutral→Sadness. For each source class, approximately 10%, 20%, or 30% of its samples are selected using a fixed corruption seed and changed according to the corresponding transition. The corrupted subsets are nested across increasing noise rates. Within each setting, all methods use identical corrupted training labels and are evaluated on the original clean test set.

Table 2. Comparison with existing methods on RAF-DB using a pre-trained ResNet-18 backbone. Baseline and RDAS results are reported as mean ± standard deviation over five matched random-seed runs, each evaluated at the final epoch. The results for CB, BBN, RUL, EAC, and MEK are literature-reported final-epoch point estimates from the unified ResNet-18 evaluation of MEK [37]; run-level variances are unavailable. Bold and underlined entries denote the highest and second-highest reported numerical values only and do not imply statistically significant superiority over literature point estimates.

Method	Overall	Mean	Happiness	Neutral	Sadness	Surprise	Disgust	Anger	Fear
Baseline	86.98 ± 0.10	78.21 ± 0.15	94.68 ± 0.18	88.03 ± 0.17	84.85 ± 0.11	84.86 ± 0.14	58.13 ± 0.44	78.02 ± 0.34	58.92 ± 0.74
CB [38]	88.04	79.26	95.11	<u>90.74</u>	84.73	86.93	64.38	73.46	59.46
BBN [39]	87.39	78.19	93.59	**91.62**	84.94	84.80	61.88	77.78	52.70
RUL [10]	88.66	81.66	<u>95.78</u>	87.06	86.19	**89.36**	65.00	<u>83.33</u>	64.86
EAC [11]	89.05	81.09	95.27	88.97	**90.17**	<u>87.84</u>	61.25	<u>83.33</u>	60.81
MEK [37]	**89.77**	<u>82.44</u>	**96.37**	89.56	89.33	<u>87.84</u>	<u>66.89</u>	80.86	<u>66.22</u>
RDAS (Ours)	<u>89.70 ± 0.13</u>	**83.35 ± 0.08**	95.71 ± 0.17	88.68 ± 0.15	<u>90.13 ± 0.09</u>	87.17 ± 0.14	**68.75 ± 0.44**	**84.07 ± 0.28**	**68.92 ± 0.96**

4.1. Comparison with existing methods on the original RAF-DB

We first evaluate RDAS on the original RAF-DB dataset using a pre-trained ResNet-18 backbone. Table 2 reports the overall, mean-class, and per-category accuracies. Mean-class accuracy assigns equal importance to all categories and is therefore more sensitive to performance on minority classes.

RDAS improves overall accuracy from 86.98±0.10% to 89.70±0.13% and mean-class accuracy from 78.21 ± 0.15% to 83.35 ± 0.08%. The paired overall gain is significant ($t(4) = 52.25$, $p < 0.001$; 95% CI [2.58, 2.87]), and the larger mean-class gain reflects stronger minority-class performance. RDAS is comparable to the 89.77% MEK point estimate and records the highest reported Disgust, Anger, and Fear values. Cross-paper differences remain descriptive because competitor variances are unavailable; larger Disgust and Fear deviations also reflect their small test sets of 160 and 74 images.

4.2. Robustness under synthetic label noise

Table 3 compares performance under 10%–30% label corruption on three datasets. Competing results are literature-reported point estimates, whereas RDAS results are averaged over five runs. Because preprocessing and training protocols cannot be fully controlled across studies, these comparisons should be interpreted descriptively rather than as matched statistical tests.

Accuracy decreases as the noise rate increases, while RDAS remains competitive across all three datasets. LA-Net provides the highest RAF-DB point estimate at the 10% noise rate, whereas RDAS obtains the highest numerical means at the 20% and 30% noise rates and across all AffectNet settings. On FERPlus, numerical differences of 0.02–0.06 percentage points are too small to support a claim of

Table 3. Comparison under synthetic label noise on RAF-DB, AffectNet, and FERPlus. Competing results are literature point estimates; RDAS reports mean ± standard deviation over five runs. Nominal noise rates are matched when available, but cross-paper preprocessing and training differences remain uncontrolled. Boldface identifies the highest reported numerical value and should not be interpreted as evidence of statistically significant superiority.

Method	RAF-DB			AffectNet			FERPlus		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
SCN [9]	82.18	80.10	77.46	58.58	57.25	55.05	84.28	83.17	82.47
RUL [10]	86.17	84.32	82.06	60.54	59.01	56.93	86.93	85.05	83.90
EAC [11]	88.02	86.05	84.42	61.11	60.29	58.91	87.03	86.07	85.44
LA-Net [28]	**88.75**	87.12	85.33	62.85	61.72	60.82	88.02	86.85	86.01
NLFER [29]	87.93	86.59	85.04	62.29	60.32	58.17	87.62	87.10	85.79
3WAUS [32]	88.17	86.90	84.71	61.53	61.19	60.08	87.15	86.59	85.46
RDAS (Ours)	88.59 ± 0.12	**87.42** ± 0.15	**86.15** ± 0.19	**63.32** ± 0.07	**62.54** ± 0.05	**61.85** ± 0.18	**88.08** ± 0.09	**87.12** ± 0.13	**86.51** ± 0.15

Table 4. Comparison of the baseline and RDAS with fixed and dynamic reliability partitions under symmetric and class-dependent asymmetric label noise on RAF-DB. Results are reported as mean accuracy (%) ± standard deviation over five runs with matched training seeds. The symmetric-noise results of RDAS with dynamic partitioning are repeated from Table 3 for controlled comparison.

Noise type	Method	10%	20%	30%
Symmetric	Baseline	86.37 ± 0.16	84.51 ± 0.14	82.93 ± 0.15
	RDAS (fixed partition)	88.30 ± 0.18	87.14 ± 0.21	85.70 ± 0.12
	RDAS (dynamic partition)	**88.59 ± 0.12**	**87.42 ± 0.15**	**86.15 ± 0.19**
Asymmetric	Baseline	86.23 ± 0.14	84.35 ± 0.11	82.56 ± 0.18
	RDAS (fixed partition)	88.47 ± 0.15	86.99 ± 0.12	85.45 ± 0.17
	RDAS (dynamic partition)	**88.72 ± 0.10**	**87.28 ± 0.16**	**85.88 ± 0.13**

meaningful superiority. The small run-level standard deviations of RDAS nevertheless indicate stable performance across repeated runs.

4.3. Robustness under symmetric and class-dependent asymmetric label noise

To examine whether the dynamic reliability partition remains effective beyond uniform label corruption, we compare symmetric noise with structured class-dependent asymmetric noise in Table 4. The fixed variant assigns 60%/20%/20% of each mini-batch to the High-, Medium-, and Low-reliability groups, respectively, whereas the dynamic variant uses $\bar{r} \pm ks_r$ with $k = 1.0$ under otherwise matched conditions.

Across all evaluated noise types and corruption rates, RDAS with dynamic partitioning outperforms both the baseline and the fixed-partition variant. Relative to the baseline, the dynamic method improves accuracy by 2.22–3.22 percentage points under symmetric noise and by 2.49–3.32 percentage points under asymmetric noise. It also consistently exceeds the fixed-partition strategy and exhibits a smaller accuracy decline as the noise rate increases. At higher rates, structured flips are more harmful because they create repeated class-specific bias while remaining visually plausible. These results indicate that RDAS is not restricted to uniform corruption, while standard deviations of 0.10–0.21 indicate limited run-to-run variation.

Table 5. Module ablation on RAF-DB with 30% symmetric label noise. Accuracy (%) is reported as the mean ± standard deviation over five matched random-seed runs. CER, LRM, and RAS denote complex-enhanced representation, label reliability modeling, and reliability-driven adaptive supervision, respectively.

CER	LRM	RAS	RAF-DB
×	×	×	82.93 ± 0.15
✓	×	×	83.57 ± 0.18
×	✓	×	84.09 ± 0.17
✓	✓	×	84.81 ± 0.23
×	✓	✓	<u>85.47</u> ± 0.12
✓	✓	✓	**86.15** ± 0.19

4.4. Ablation studies

All ablations use RAF-DB with 30% noise and examine the main modules, internal components, supervision policies, and reliability parameters.

4.4.1. Ablation of main modules

We first examine the individual and combined contributions of CER, LRM, and RAS through progressive module ablation, as summarized in Table 5. CER comprises the complex stem and CBAM, LRM estimates label reliability, and RAS assigns reliability-specific supervision policies. Because RAS depends on the reliability estimates produced by LRM, configurations containing RAS without LRM are not considered.

For the two configurations with LRM enabled but RAS disabled, reliability affects optimization through continuous sample weighting rather than three-way policy selection. After the five-epoch warm-up, each sample is assigned the detached reliability weight

$$\omega_i^{\mathrm{rel}} = \omega_{\min} + (1 - \omega_{\min}) r_i, \qquad \omega_{\min} = 0.1,$$

and the mini-batch objective is

$$\mathcal{L}_{\text{LRM-only}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{sg}\left(\omega_i^{\mathrm{rel}}\right) \ell_{\mathrm{sup},i},$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient and $\ell_{\mathrm{sup},i}$ is the common base objective in Eq (20). During warm-up, all samples use the unweighted base objective. Thus, these configurations retain the same loss form for every sample and use r_i only to scale its contribution; they do not construct High/Medium/Low groups, add the Medium soft-target term, or remove hard-label supervision from Low samples.

When evaluated individually, CER and continuous LRM weighting improve the baseline by 0.64 and 1.16 percentage points, respectively. Combining the two components yields an accuracy of 84.81 ± 0.23%. Adding RAS to LRM in the absence of CER provides a further gain of 1.38 percentage points, indicating that reliability-specific supervision is more effective than continuous weighting alone. Full RDAS achieves 86.15 ± 0.19%, exceeding the baseline by 3.22 percentage points and the CER+LRM configuration by 1.34 percentage points. These results indicate that representation enhancement, reliability estimation, and adaptive supervision make complementary contributions.

Table 6. Fine-grained ablation on RAF-DB with 30% symmetric label noise. Panel (a) separates conventional regularization and representation components, where Concat. denotes real-valued channel concatenation of RGB and Sobel responses. Panel (b) isolates the EMA teacher and the supervision policies for medium- and low-reliability samples. A checkmark indicates that a component or policy is enabled, whereas a cross indicates that it is removed or replaced as defined in the preceding paragraph. Accuracy (%) is reported as the mean ± standard deviation over five matched random-seed runs, and Change denotes the mean-accuracy difference from the full RDAS.

(a) Conventional baseline and representation components

Variant	Complex	Concat.	CBAM	JSD	Mixup	LRM/RAS	Accuracy (%)	Change
Plain CE	×	×	×	×	×	×	82.93 ± 0.15	−3.22
CE+Mixup+JSD+CBAM	×	×	✓	✓	✓	×	84.37 ± 0.22	−1.78
RGB+Sobel concatenation	×	✓	✓	✓	✓	✓	85.61 ± 0.27	−0.54
w/o CBAM	✓	×	×	✓	✓	✓	85.72 ± 0.28	−0.43
w/o JSD	✓	×	✓	×	✓	✓	85.34 ± 0.32	−0.81
w/o Mixup	✓	×	✓	✓	×	✓	85.89 ± 0.29	−0.26
Full RDAS	✓	×	✓	✓	✓	✓	**86.15** ± 0.19	–

(b) Reliability estimation and adaptive supervision policies

Variant	EMA	Medium correction	Low suppression	Accuracy (%)	Change
Student predictions	×	✓	✓	85.57 ± 0.35	−0.58
w/o medium soft correction	✓	×	✓	85.74 ± 0.26	−0.41
Hard labels retained for Low	✓	✓	×	85.07 ± 0.31	−1.08
Full RDAS	✓	✓	✓	**86.15** ± 0.19	–

4.4.2. Ablation of internal components

We next isolate the contributions of conventional regularization, representation design, and reliability-specific supervision. As summarized in Table 6, panel (a) compares full RDAS with a conventional configuration combining cross-entropy, Mixup, JSD, and CBAM, as well as with direct RGB–Sobel concatenation and variants that remove individual representation or regularization components. Panel (b) evaluates the effects of replacing EMA outputs with detached student predictions, removing the Medium-reliability correction, and retaining hard-label supervision for Low-reliability samples. All other experimental conditions remain unchanged.

The configuration combining cross-entropy with Mixup, JSD, and CBAM achieves 84.37 ± 0.22%, while full RDAS improves upon it by 1.78 percentage points. Direct RGB–Sobel concatenation achieves 85.61 ± 0.27%, trailing full RDAS by 0.54 percentage points. Removing CBAM, JSD, or Mixup reduces accuracy by 0.43, 0.81, and 0.26 percentage points, respectively. These results indicate that, among the evaluated alternatives, the complex representation provides an additional benefit over direct concatenation; however, they do not establish its superiority over all possible feature-fusion designs.

Replacing the EMA outputs with current-student predictions reduces accuracy by 0.58 percentage points. Removing the Medium-reliability correction and retaining hard-label supervision for Low-reliability samples lead to decreases of 0.41 and 1.08 percentage points, respectively. These results indicate the complementary benefits of temporal smoothing, auxiliary soft correction, and the suppression of potentially misleading hard-label gradients for Low-reliability samples.

4.4.3. Sensitivity to reliability-related parameters

Smaller k assigns more samples to the extreme groups, whereas larger k widens the Medium interval. We vary k, λ_m, and η one at a time, fixing the latter pair at 0.65 and 8 when not varied.

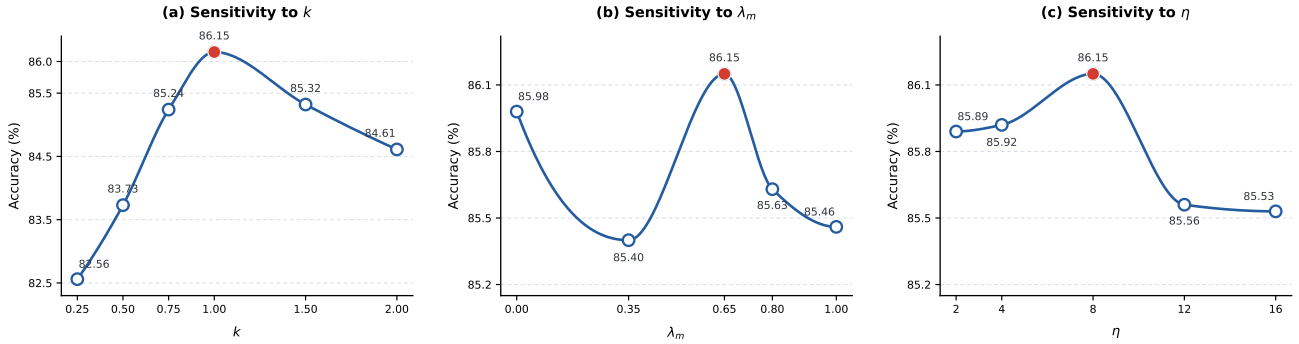


Figure 2. Sensitivity of k, λ_m, and η on RAF-DB with 30% symmetric label noise. Panels (a)–(c) vary one parameter at a time while keeping the others at their default values. Each marker reports the mean final-epoch accuracy over five matched random-seed runs. The smooth curves are used only to visualize the trends, whereas the markers denote the measured parameter settings. The default settings are highlighted in red.

The highest mean accuracy in Figure 2 is obtained at $k = 1.0$. Smaller values assign samples more aggressively to the High- and Low-reliability groups, whereas larger values reduce policy differentiation by retaining a greater proportion of samples in the Medium-reliability group. Among the evaluated settings, the default values $\lambda_m = 0.65$ and $\eta = 8$ also yield the highest mean accuracies.

4.5. Analysis and visualization

Beyond numerical accuracy, we analyze RDAS from three visual perspectives: sample reliability, feature distribution, and attention localization. The visual studies are conducted on RAF-DB because it is a representative in-the-wild FER benchmark, and its synthetic noisy-label setting provides known clean/noisy sample identities. These results are intended to complement the quantitative comparisons by showing how reliability-driven supervision affects sample assessment, representation structure, and facial attention behavior.

4.5.1. Reliability score analysis

A central assumption behind RDAS is that the estimated reliability score should correlate with the trustworthiness of the observed label. We examine this assumption by comparing the reliability score distributions of clean and noisy training samples on RAF-DB with 30% synthetic label noise. Because label corruption is synthetically introduced in this setting, a training sample can be identified as corrupted by comparing its observed training label with the original annotation.

To quantify this relationship, we further evaluate three teacher-derived indicators for detecting corrupted annotations: the probability assigned to the observed label, the annotated-versus-competing confidence margin, and the final RDAS reliability score. Table 7 characterizes both the sample-level score distributions and the cross-run stability of noise detection. For each run, the clean/noisy means and sample-level standard deviations are computed over the corresponding training samples, and the reported score statistics are then averaged over five runs with different random seeds. The area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC) are also evaluated over the same five runs, with corrupted annotations treated as the positive class. Because all three indicators increase with annotation reliability, their negated values are used as corrupted-label

Table 7. Quantitative evaluation of reliability indicators on the RAF-DB training set with 30% symmetric label noise over five random-seed runs. For Clean and Noisy, μ_s and σ_s denote the sample-level mean and standard deviation computed within each run; the table reports their averages over the five runs. For the AUROC and AUPRC, $\mu_r \pm \sigma_r$ denotes the mean and run-level standard deviation over the same five runs. All scores are reported on the [0, 1] scale, except for the confidence margin, which lies in [−1, 1].

	Sample statistics averaged over five runs		Five-run detection ($\mu_r \pm \sigma_r$)	
Indicator	Clean	Noisy	AUROC	AUPRC
Annotated-class confidence	0.5526 ± 0.1779	0.0980 ± 0.1256	0.9823 ± 0.0015	0.9818 ± 0.0017
Confidence margin	0.3045 ± 0.2839	−0.3278 ± 0.1754	0.9819 ± 0.0021	0.9826 ± 0.0015
RDAS reliability score	0.7252 ± 0.1615	0.0986 ± 0.1520	0.9827 ± 0.0014	0.9829 ± 0.0015

detection scores when computing the AUROC and AUPRC.

The sample-level statistics reveal a substantial separation between clean and corrupted annotations across the five runs. The average RDAS reliability score is 0.7252 for clean samples and 0.0986 for noisy samples, with corresponding average sample-level standard deviations of 0.1615 and 0.1520. Treating corrupted annotations as the positive class, the resulting detector achieves an AUROC of 0.9827 ± 0.0014 and an AUPRC of 0.9829 ± 0.0015. The run-level standard deviations do not exceed 0.0021 for any indicator, showing that the separation remains stable under different random initializations. Annotated-class confidence, confidence margin, and RDAS reliability obtain comparable detection performance, and their small numerical differences do not support a claim that one indicator clearly dominates the others. Nevertheless, they encode different evidence: annotated-class confidence measures absolute teacher support, whereas the margin measures whether the observed label is preferred over its strongest alternative. RDAS combines these two signals to determine the subsequent supervision policy, rather than treating the reliability score solely as a binary noise detector.

To examine whether mini-batch composition and the training stage affect the dynamic partition, Table 8 evaluates assignment stability and the actual noise concentration within each reliability group. For each of the five training seeds, we generate ten independent random mini-batch permutations at each evaluated batch size. Agreement, Cohen's κ, and the group-wise Jaccard indices are first averaged over all $\binom{10}{2} = 45$ permutation pairs within each seed and are then averaged across the five seeds. The fallback rate is computed as the percentage of mini-batches in which no sample naturally satisfies the upper reliability boundary, averaged over permutations and seeds.

Panel (a) shows that the dynamic partition remains stable under changes in mini-batch composition. At the default batch size of 128, the pairwise assignment agreement reaches 96.15%, with a Cohen's κ of 0.928 and group-wise Jaccard indices of 0.881, 0.958, and 0.975 for the High, Medium, and Low groups, respectively. Increasing the batch size further improves all stability measures and reduces the High-group fallback rate (HF) from 2.56% at a batch size of 64 to 0.27% at 256, indicating that more samples provide more stable estimates of the batch reliability statistics. Panel (b) further shows that the partition becomes more semantically informative during training. From epochs 10 to 60, the High-group proportion increases from 5.68% to 18.71%, while its within-group noise ratio decreases from 10.01% to 3.25%. The Medium group contracts from 64.81% to 53.15% and becomes progressively cleaner. Meanwhile, the Low-group proportion remains close to the imposed 30% corruption level, whereas its noise ratio increases from 71.39% to 95.48%. At epoch 60, the Low group contains approximately

89.56% of all corrupted annotations, while only 4.52% of its members are clean samples. Together with the increase in stage agreement from 88.53% to 94.17%, these results indicate that the relative mini-batch partition remains stable under batch reshuffling and progressively concentrates corrupted annotations in the Low-reliability group during training.

Table 8. Stability and composition of the dynamic reliability partition on RAF-DB with 30% symmetric label noise over five random-seed runs. Panel (a) evaluates sensitivity to mini-batch size at the final training stage using ten random batch permutations per seed. Agreement and κ denote pairwise exact assignment agreement and Cohen's kappa, respectively; J_{H}, J_{M}, and J_{L} denote the group-wise Jaccard indices; and HF denotes the percentage of mini-batches requiring the High-group fallback. Panel (b) reports the sample proportion (Prop.) and within-group noise ratio (Noise), averaged over the same five seeds and ten permutations, at representative training stages using a mini-batch size of 128. Stage agreement is computed between the modal sample assignments of consecutive reported stages after aligning sample identities and is then averaged over the five seeds. Agreement, HF, Prop., and Noise are reported as percentages.

(a) Sensitivity to mini-batch size at the final training stage

Batch size	Agreement	κ	J_{H}	J_{M}	J_{L}	HF
64	94.27	0.862	0.820	0.923	0.958	2.56
128	96.15	0.928	0.881	0.958	0.975	0.89
256	97.04	0.949	0.915	0.964	0.986	0.27

(b) Group composition and noise concentration across training stages

		High		Medium		Low	
Epoch	Stage agreement	Prop.	Noise	Prop.	Noise	Prop.	Noise
10	–	5.68	10.01	64.81	12.90	29.51	71.39
30	88.53	14.04	5.57	57.58	7.39	28.38	87.96
60	94.17	18.71	3.25	53.15	4.75	28.14	95.48

Consistent with the quantitative results in Table 7, Figure 3 shows a substantial but incomplete separation between clean and corrupted annotations. Corrupted annotations form a dominant peak near zero followed by a rapidly decaying tail, indicating that the EMA teacher generally assigns substantially less support to the observed corrupted label than to its strongest competing class. In contrast, the clean distribution shifts markedly to the right and contains both a broad medium-to-high-reliability mode and a second concentration near one. This pattern is consistent with the expected behavior of the reliability measure: visually unambiguous clean expressions can produce near-certain label–evidence agreement, whereas correctly annotated but difficult or ambiguous expressions may retain only moderate reliability. The dashed lines summarize the clean and corrupted distributions for the displayed seed-0 checkpoint, whereas Table 7 reports the corresponding statistics averaged over five random-seed runs. Neither summary captures the full heterogeneity within either distribution.

The limited overlap in the intermediate region is consistent with the graded nature of label reliability. Difficult but correctly annotated samples may receive limited teacher support, whereas some corrupted annotations may temporarily agree with the teacher prediction. Consequently, r_i should be interpreted as a continuous measure of compatibility between the observed label and temporally stabilized facial evidence, rather than as a perfect binary noise detector. This observation motivates the three-region RAS policy: strongly supported annotations retain hard-label supervision, intermediate cases receive teacher-guided soft correction, and strongly contradicted annotations avoid potentially harmful hard-label gradients while still contributing through consistency learning. Together with the high AUROC and AUPRC in Table 7, the empirical distributions support both the discriminative validity of the reliability

score and its use as a supervision-policy variable rather than a simple scalar weight.

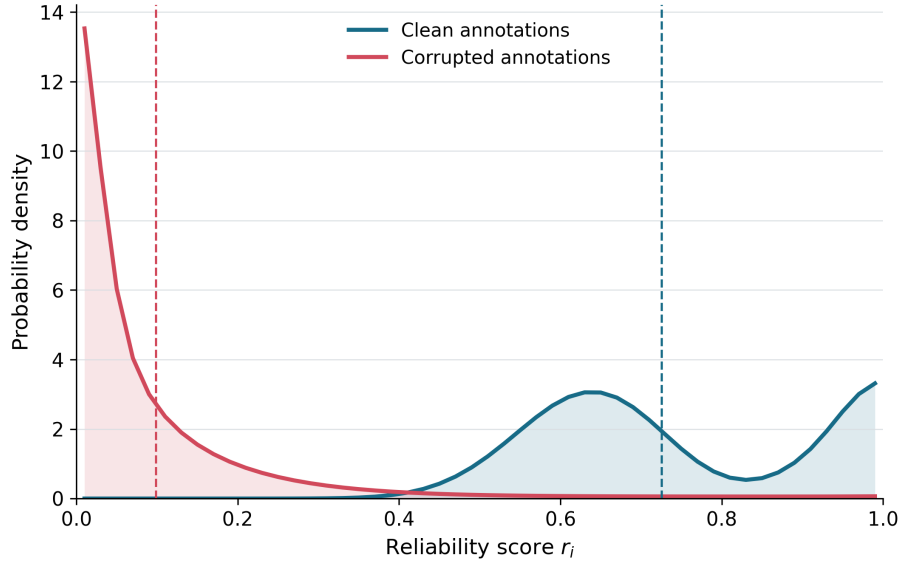


Figure 3. Empirical probability-density distributions of RDAS reliability scores for clean and corrupted annotations on RAF-DB with 30% symmetric label noise. The distributions are computed over all training samples using the final epoch-60 EMA checkpoint from training seed 0. A sample is considered corrupted when its observed training label differs from the original annotation. The dashed lines mark the clean- and corrupted-distribution means for the displayed seed-0 checkpoint.

The representative samples in Figure 4 illustrate how RDAS applies different supervision objectives according to estimated reliability. The three clean High cases have r_i between 0.933 and 0.961, and exhibit complete agreement among R, O, and T. They also satisfy the upper reliability threshold without invoking the fallback rule, so retaining the complete base supervised objective $\ell_{\text{sup},i}$ preserves strongly supported discriminative supervision together with view consistency. The Medium row contains two clean cases with R = O = T and one corrupted case. For the clean cases, the soft targets retain most of the probability mass on the shared observed–teacher class while incorporating a modest contribution from the teacher distribution. In the corrupted case, O is Happy whereas both R and T are Neutral; the target therefore assigns nonzero mass to the teacher-supported reference class without replacing the annotation by another one-hot label. This auxiliary soft correction reduces the influence of uncertain hard-label supervision while retaining useful class information.

The Low-reliability examples further illustrate that the reliability score does not constitute a perfect binary noise indicator. In the two corrupted cases, O conflicts with both R and T, and direct label-dependent fitting would reinforce the injected Surprise and Neutral labels despite teacher and visual evidence supporting Happy and Angry, respectively. RDAS therefore replaces the base supervised objective with the weak weighted consistency objective $\lambda_l \ell_{\text{jsd},i}$ for these samples. The third case is a difficult clean example: R and O are both Disgust, but T is Neutral and $r_i = 0.037$. Its presence demonstrates that a correctly annotated but ambiguous image may also receive low reliability. Rather than discarding Low samples, RDAS retains JSD consistency across views, allowing label-independent representation learning without enforcing a potentially unreliable class decision. Tables 7 and 8 report the corresponding group statistics.

Figure 4. Mechanism-focused RGB cases from the RAF-DB training set with 30% symmetric label noise, evaluated with the seed-0 EMA checkpoint at epoch 60. Rows show High-, Medium-, and Low-reliability samples. R, O, and T denote the original reference label, the observed training label after noise injection, and the EMA-teacher prediction, respectively; clean/noisy indicates whether O agrees with R. For Medium cases, mix $= a_i = \beta(1 - r_i)$, and $q(c)$ denotes the class-c entry $\tilde{y}_{i,c}$ of the teacher-guided target in Eq (24); $q(\mathrm{O} = \mathrm{T})$ is used when O and T coincide. Groups are reconstructed in deterministic batches of 128 using $\bar{r} \pm ks_r$ with $k = 1.0$; all displayed High cases satisfy the upper threshold without fallback. The displayed composition is three clean High, two clean Medium, one noisy Medium, two noisy Low, and one difficult-clean Low sample. High cases retain $\ell_{\mathrm{sup},i}$, Medium cases add $\lambda_r(t)\ell_{\mathrm{soft},i}$, and Low cases use only $\lambda_l\ell_{\mathrm{jsd},i}$.

4.5.2. Feature distribution visualization

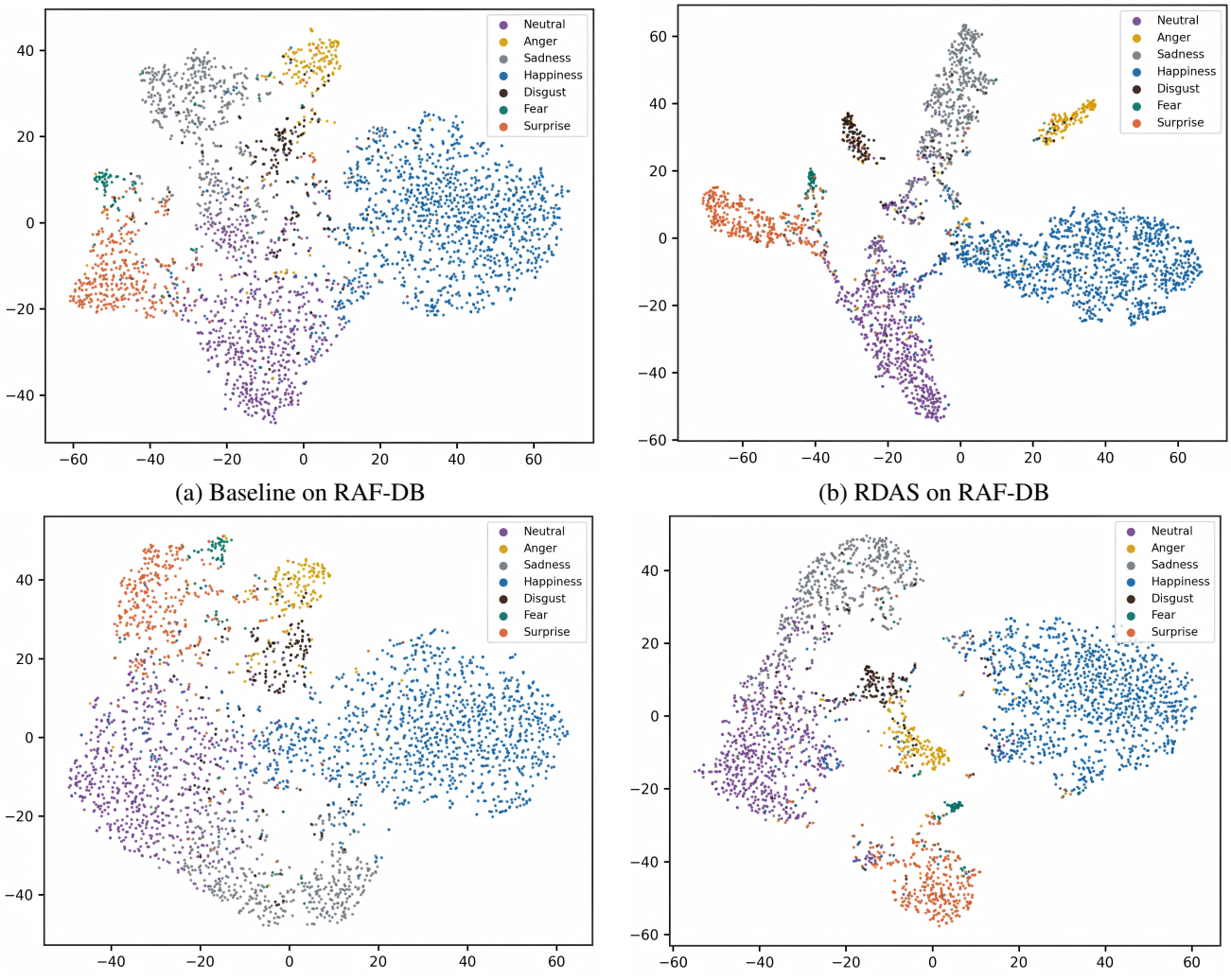
Feature distributions are visualized with t-SNE to examine whether the proposed supervision strategy leads to more discriminative representations. Four settings are compared: baseline on the original RAF-DB, RDAS on the original RAF-DB, baseline on RAF-DB with 30% label noise, and RDAS on RAF-DB with 30% label noise. The baseline uses the same ResNet-18 backbone and cross-entropy loss, whereas RDAS is trained with the proposed reliability-driven adaptive supervision. All four feature sets are extracted from the corresponding final epoch-60 checkpoints trained with seed 0. For all panels, t-SNE is applied to the penultimate-layer 512-dimensional features under identical settings: PCA pre-reduction to 50 dimensions, perplexity 40, 1000 iterations, and random seed 42.

The t-SNE maps in Figure 5 show that the baseline forms several recognizable clusters on the

original RAF-DB, although overlap remains between visually similar categories. After 30% label noise is introduced, the baseline feature space exhibits substantially greater inter-class overlap, reflecting the adverse effect of corrupted supervision on representation learning. In contrast, RDAS produces more compact and better-separated feature distributions under both the original and noisy-label settings. In particular, under 30% label noise, RDAS maintains clearer class separation than the noisy-label baseline, suggesting that reliability-driven adaptive supervision can preserve discriminative structure even when the training labels are partially corrupted.

4.5.3. Attention visualization

Grad-CAM is used to examine whether RDAS changes the facial regions emphasized by the model. We compare the baseline and RDAS on the original RAF-DB and on RAF-DB with 30% label noise using the corresponding final epoch-60 checkpoints trained with seed 0. One fixed test image is selected for each of the seven expression categories before comparison and is reused unchanged across all four



(a) Baseline on RAF-DB

(b) RDAS on RAF-DB

(c) Baseline on RAF-DB with 30% noise

(d) RDAS on RAF-DB with 30% noise

Figure 5. t-SNE visualization on RAF-DB. RDAS yields more compact clusters and clearer class separation under 30% label noise.

settings. All attention maps are generated from these fixed images using the layer3 feature block, the ground-truth expression category as the target class, and identical overlay settings. Row (1) gives the input images; rows (2) and (3) correspond to the baseline and RDAS trained on the original RAF-DB; rows (4) and (5) correspond to the baseline and RDAS trained with 30% noisy labels.

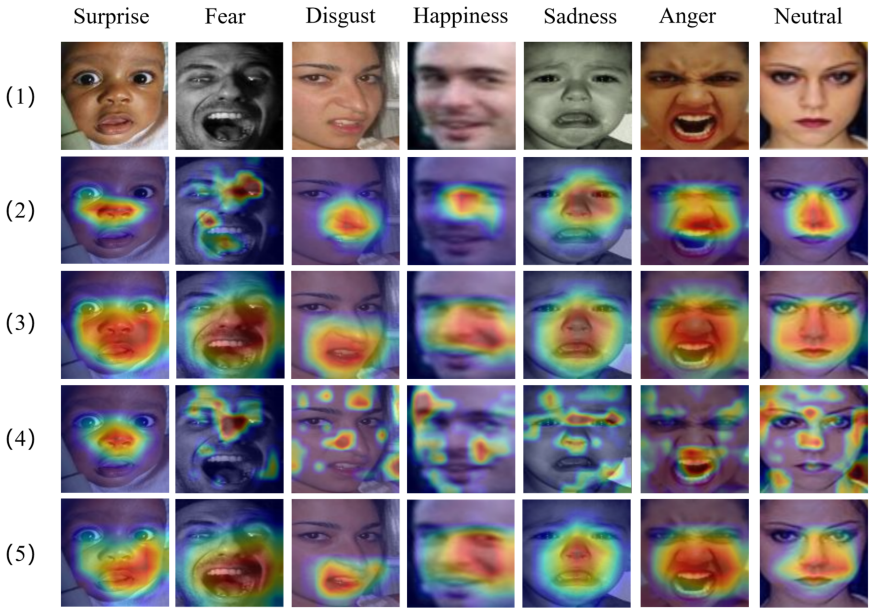


Figure 6. Grad-CAM visualization on RAF-DB. Row (1) shows input images. Rows (2) and (3) correspond to the baseline and RDAS trained on the original RAF-DB, respectively. Rows (4) and (5) correspond to the baseline and RDAS trained on RAF-DB with 30% label noise, respectively.

The Grad-CAM comparison in Figure 6 indicates that the baseline trained with noisy labels produces more spatially diffuse attention maps. Under the same noisy-label setting, RDAS yields more coherent activation patterns concentrated around expression-relevant facial regions, including the mouth, nose, eyes, and eyebrows. Although Grad-CAM provides qualitative rather than causal evidence, these observations suggest that RDAS promotes more spatially coherent localization of expression-related cues under noisy supervision.

5. Conclusions

This paper has proposed RDAS, a reliability-driven adaptive supervision framework for noisy-label facial expression recognition. The key idea is to treat sample reliability as a supervision-policy variable rather than a simple loss weight. Based on this formulation, RDAS integrates complex-enhanced expression representation, EMA-stabilized label reliability modeling, and adaptive supervision for high-, medium-, and low-reliability samples. Reliable samples preserve discriminative hard-label learning, ambiguous samples are guided by soft correction, and unreliable samples are prevented from dominating optimization with misleading hard-label gradients.

Experiments on RAF-DB, AffectNet, and FERPlus demonstrate that RDAS improves recognition performance under original supervision and uniformly corrupted labels. Additional experiments with

class-dependent asymmetric noise on RAF-DB show that the dynamic partition consistently outperforms both the baseline and the fixed-partition variant, with its advantage increasing as the corruption rate rises. The ablation studies support the complementary roles of representation enhancement, reliability modeling, and adaptive supervision, while the sensitivity analysis shows that a moderate dynamic threshold offers a better balance among reliable learning, soft correction, and noisy-label suppression. Reliability distributions, t-SNE embeddings, and Grad-CAM visualizations provide complementary qualitative evidence consistent with more stable reliability estimation, better-structured feature embeddings, and more coherent expression-related attention under noisy supervision.

Although RDAS improves robustness under controlled symmetric and class-dependent asymmetric label noise, these predefined corruption processes do not fully reflect real-world FER annotation uncertainty. In practice, FER datasets may involve instance-dependent errors, annotator-specific disagreement, and open-set or mixed-expression ambiguity, none of which is directly evaluated in the present experiments. Future work should therefore assess RDAS on benchmarks that preserve annotator-level label distributions and naturally occurring annotation uncertainty, as well as across stronger backbone architectures. Quantitative evaluation of reliability calibration and attention localization may also provide further insight into model behavior under ambiguous supervision.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the maintainers of the public FER datasets used in this study. This work was supported by the Jiangsu Province Marine Science and Technology Innovation Project under Grant No. JSZRHYKJ202201, the Jiangsu Provincial Water Conservancy Science and Technology Project under Grant No. 2020058, the Lianyungang "521 Project" Research Funding Project under Grant No. LYG06521202131, and the Lianyungang Key Research and Development Program (Industrial Foresight and Key Technologies) under Grant No. CG2527.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. R. W. Picard, *Affective Computing*, MIT Press, 1997. https://doi.org/10.7551/mitpress/1140.001.0001
2. M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, *IEEE Trans. Pattern Anal. Mach. Intell.*, **22** (2000), 1424–1445. https://doi.org/10.1109/34.895976
3. Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Trans. Pattern Anal. Mach. Intell.*, **31** (2009), 39–58. https://doi.org/10.1109/TPAMI.2008.52

4. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., Challenges in representation learning: A report on three machine learning contests, in *Neural Information Processing*, (2013), 117–124. https://doi.org/10.1007/978-3-642-42051-1_16
5. E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in *Proceedings of the ACM International Conference on Multimodal Interaction*, (2016), 279–283. https://doi.org/10.1145/2993148.2993165
6. S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2584–2593. https://doi.org/10.1109/CVPR.2017.277
7. A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild, *IEEE Trans. Affect. Comput.*, **10** (2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923
8. D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, et al., A closer look at memorization in deep networks, in *Proceedings of the International Conference on Machine Learning*, (2017), 233–242.
9. K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 6897–6906. https://doi.org/10.1109/CVPR42600.2020.00693
10. Y. Zhang, C. Wang, X. Ling, W. Deng, Relative uncertainty learning for facial expression recognition, in *Advances in Neural Information Processing Systems*, **34** (2021), 17616–17627.
11. Y. Zhang, C. Wang, W. Deng, Learn from all: Erasing attention consistency for noisy label facial expression recognition, in *Proceedings of the European Conference on Computer Vision*, (2022), 418–434. https://doi.org/10.1007/978-3-031-19809-0_24
12. Y. Tan, H. Xia, S. Song, Robust consistency learning for facial expression recognition under label noise, *Vis. Comput.*, **41** (2025), 2655–2667. https://doi.org/10.1007/s00371-024-03558-1
13. H. Xia, C. Su, S. Song, Y. Tan, Dual-consistency constraints network for noisy facial expression recognition, *Image Vis. Comput.*, **148** (2024), 105141. https://doi.org/10.1016/j.imavis.2024.105141
14. X. Zhang, Y. Lu, H. Yan, J. Huang, Y. Gu, Y. Ji, et al., ReSup: Reliable label noise suppression for facial expression recognition, *IEEE Trans. Affect. Comput.*, **16** (2025), 2006–2019. https://doi.org/10.1109/TAFFC.2025.3549017
15. N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le, 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), 6088–6097. https://doi.org/10.1109/WACV56688.2023.00603
16. J. Zheng, B. Li, S. Zhang, S. Wu, L. Cao, S. Ding, Attack can benefit: An adversarial approach to recognizing facial expressions under noisy annotations, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **37** (2023), 3660–3668. https://doi.org/10.1609/aaai.v37i3.25477
17. Y. Li, H. Liu, D. Jiang, J. Liang, Facial expression recognition with label-noisy under dual-branch noise extraction and suppression, *IEEE Trans. Affect. Comput.*, **16** (2025), 1514–1525. https://doi.org/10.1109/TAFFC.2024.3519359
18. J. Ye, D. Liu, C. Wang, H. Huang, L. Ying, L. Zhang, et al., CNLA: Collaborative noisy label adaptive learning for facial expression recognition, *Inf. Sci.*, **719** (2025), 122436. https://doi.org/10.1016/j.ins.2025.122436

19. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
20. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision*, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
21. K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, *IEEE Trans. Image Process.*, **29** (2020), 4057–4069. https://doi.org/10.1109/TIP.2019.2956143
22. Y. Fan, J. C. K. Lam, V. O. K. Li, Facial expression recognition with deeply-supervised attention network, *IEEE Trans. Affective Comput.*, **13** (2022), 1057–1071. https://doi.org/10.1109/TAFFC.2020.2988264
23. F. Xue, Q. Wang, G. Guo, TransFER: Learning relation-aware facial expression representations with transformers, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 3601–3610. https://doi.org/10.1109/ICCV48922.2021.00358
24. Z. Wen, W. Lin, T. Wang, G. Xu, Distract your attention: Multi-head cross attention network for facial expression recognition, *Biomimetics*, **8** (2023), 199. https://doi.org/10.3390/biomimetics8020199
25. C. Zheng, M. Mendieta, C. Chen, POSTER: A pyramid cross-fusion transformer network for facial expression recognition, in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, (2023), 3146–3155. https://doi.org/10.1109/ICCVW60793.2023.00339
26. J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, POSTER++: A simpler and stronger facial expression recognition network, *Pattern Recognit.*, **157** (2025), 110951. https://doi.org/10.1016/j.patcog.2024.110951
27. F. Ma, B. Sun, S. Li, Transformer-augmented network with online label correction for facial expression recognition, *IEEE Trans. Affective Comput.*, **15** (2024), 593–605. https://doi.org/10.1109/TAFFC.2023.3285231
28. Z. Wu, 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), 20698–20707. https://doi.org/10.1109/ICCV51070.2023.01892
29. C. Y. Che, H. M. Sun, Y. X. Chen, S. Yang, R. S. Jia, NLFER: Multi-branch attention cross-fusion for robust facial expression recognition amidst noisy labels, *Vis. Comput.*, **42** (2026), 1. https://doi.org/10.1007/s00371-025-04210-2
30. J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, T. Mei, Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6248–6257. https://doi.org/10.1109/CVPR46437.2021.00618
31. J. Lee, Y. Choi, H. Kim, I. J. Kim, G. P. Nam, Navigating label ambiguity for facial expression recognition in the wild, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **39** (2025), 4517–4525. https://doi.org/10.1609/aaai.v39i4.32476
32. D. Li, W. Xiong, T. Luo, L. Zhang, 3WAUS: A novel three-way adaptive uncertainty-suppressing model for facial expression recognition, *Inf. Sci.*, **677** (2024), 120962. https://doi.org/10.1016/j.ins.2024.120962
33. Y. Apedo, H. Tao, A weakly supervised pavement crack segmentation based on adversarial learning and transformers, *Multimedia Syst.*, **31** (2025), 266. https://doi.org/10.1007/s00530-025-01850-1

34. Y. Apedo, H. Tao, W. Gao, C. Xie, S. Zhao, Unsupervised domain adaptation for crack segmentation via cross-domain stylization and dual adversarial feature learning, *J. Comput. Civ. Eng.*, **40** (2026), 04025146. https://doi.org/10.1061/JCCEE5.CPENG-7223
35. H. Zhou, H. Tao, Q. Zhang, C. Xie, B. Li, MRKD-PBCL: Multi-level region-wise knowledge distillation and prototype balanced contrastive learning for class incremental semantic segmentation, *Neurocomputing*, **679** (2026), 133285. https://doi.org/10.1016/j.neucom.2026.133285
36. A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in *Advances in Neural Information Processing Systems*, **30** (2017), 1195–1204.
37. Y. Zhang, Y. Li, L. Qin, X. Liu, W. Deng, Leave no stone unturned: Mine extra knowledge for imbalanced facial expression recognition, in *Advances in Neural Information Processing Systems*, **36** (2023), 14414–14426. https://doi.org/10.52202/075280-0634
38. Y. Cui, M. Jia, T. Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 9268–9277. https://doi.org/10.1109/CVPR.2019.00949
39. B. Zhou, Q. Cui, X. S. Wei, Z. M. Chen, BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 9719–9728. https://doi.org/10.1109/CVPR42600.2020.00974

AIMS Press