Noise-Robust Facial Expression Recognition via Mamba-driven Neighbor Weight Refinement

Yuzhuang Yang[1], Xiaolin Tian[1⋆], and Qigong Sun[2]

[1] Xidian University
24171214009@stu.xidian.edu.cn, xltian@mail.xidian.edu.cn
[2] SenseTime Research

Abstract. Facial Expression Recognition (FER) in the wild is inherently affected by annotation ambiguity and noisy label. Recent label distribution learning (LDL) methods alleviate this issue by exploiting neighborhood information. However, existing approaches typically rely on static or heuristic mechanisms to estimate neighbor contribution weights, limiting their capacity to model complex dependencies among neighboring samples. In this paper, we propose **M**amba-driven **N**eighbor **W**eight **R**efinement (Mamba-NWR), a robust FER framework that integrates cross-attention routing (CAR) for initial neighbor weight estimation and Mamba-based iterative refinement (MIR) for distribution aggregation, effectively handling noisy labels. Specifically, our CAR mechanism introduces a multi-route local feature projection for both the input and neighbor samples. The subsequent cross-attention-based interaction among these features serves as the foundation for computing the initial neighbor weights. Then, we employ a Mamba-based state space model to iteratively refine the contribution weights of neighboring label distributions, enabling dynamic balance between preserving historical states and absorbing new neighborhood information. The optimized neighbor distributions are then aggregated with the original logical label through a learnable fusion factor to construct the final supervision information. Extensive experiments on multiple benchmark datasets demonstrate that our approach consistently improves robustness under ambiguous and noisy annotations, achieving state-of-the-art performance.

Keywords: Facial Expression Recognition · Label Distribution Learning · Neighbor Weight Refinement

1 Introduction

Facial Expression Recognition (FER) is a fundamental task in affective computing, with wide applications in human-computer interaction and behavioral analysis [22]. Although deep neural networks and large-scale datasets have led to significant progress [10,16,18], FER still faces considerable challenges in real-world applications. The primary difficulty is mainly due to the subjective and

⋆ Corresponding author

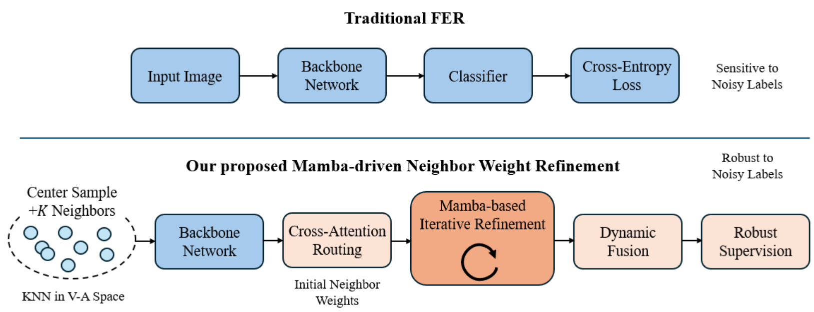

Fig. 1: Comparison between our proposed Mamba-NWR (bottom) and the traditional facial recognition framework (top).

ambiguous nature of facial expressions: the same facial appearance may be interpreted as different emotions by different observers, especially for subtle or transitional expressions [35, 41]. Furthermore, large-scale datasets are often labeled through crowd-sourcing, which inevitably introduces label noise and inconsistency [23, 29]. These issues result in unreliable supervision information during training, which limit the model's robustness and generalization [37]. Therefore, addressing expression ambiguity and label uncertainty remains a critical challenge for FER.

To alleviate the adverse effects of ambiguous expressions and unreliable annotations, a number of approaches have been proposed to improve the robustness of FER models. Early efforts focus on designing noise-tolerant loss functions or correction mechanisms to reduce the influence of corrupted labels during training [20,31]. Beyond loss-level solutions, recent studies introduce uncertainty modeling strategies that explicitly estimate the reliability of samples or predictions, enabling regulation of the learning process under noisy supervision [30]. Other works explore the use of contextual cues, relational information, or auxiliary structures to enrich supervision and capture latent correlations among facial expressions beyond individual instances [13,14]. These methods have demonstrated encouraging improvements, especially in challenging in-the-wild environments. However, most existing approaches regulate supervision using static or shallow mechanisms, which limits their ability to model complex inter-sample dependencies. As a result, such methods may struggle to dynamically adjust supervision under varying levels of ambiguity. These limitations highlight the need for more expressive and flexible frameworks that can progressively refine and regulate supervision while maintaining its reliability and stability in FER tasks.

In this paper, we propose a novel framework with progressive neighborhood-aware refinement for FER that aims to construct more informative and robust supervision information under ambiguous annotations, as illustrated in Fig. 1. The key idea is to model neighborhood aggregation as an iterative optimization process rather than relying on static or heuristic weighting mechanisms. Most existing neighborhood-based methods perform aggregation in a single pass, where neighbor weights are computed from initial similarity measures and directly used for distribution aggregation without any error-correction mechanism. However, under noisy and ambiguous supervision, these initial neighborhood relations are often unreliable, and static weighting may amplify noise rather than suppress it [11]. This limitation motivates the need for a dynamic refinement strategy that can progressively correct biased neighbor estimations. Given an input sample, we first identify its neighbors in the auxiliary space (Valence-Arousal Space) and obtain an initial estimate of their contribution weights using the cross-attention routing mechanism. This design provides a stable and effective initialization. Building upon this initialization, we introduce a Mamba-based state space model to dynamically refine the neighbor contribution weights in a sequential manner. At each iteration, the model takes the aggregated neighbor distribution as input and adaptively updates the weights by balancing historical state information and newly absorbed neighborhood. Through multiple iterations, the model progressively selects and emphasizes informative neighbors while suppressing unreliable ones. The optimized neighbor distribution is then combined with the original logical label via a learnable fusion factor to construct the final supervision information. The main contributions of this paper are as follows:

- We propose a novel framework with progressive neighborhood-aware refinement that formulates neighbor supervision modeling as an iterative optimization process.
- We introduce a Mamba-based state space model to dynamically refine neighbor contribution weights, providing an adaptive mechanism for supervision regulation.
- We design a cross-attention routing mechanism to initialize neighbor contribution weights, offering a stable and informative initial state.
- Extensive experiments on multiple benchmark datasets demonstrate that the proposed method consistently maintains robustness under ambiguous and noisy labels, achieving state-of-the-art performance.

2 Related Work

2.1 FER with Noisy Labels

FER in unconstrained environments is severely affected by label ambiguity and noisy annotations arising from subjective labeling and inter-class similarity. To alleviate this issue, recent studies have explored robust learning strategies under noisy supervision. ASM introduces adaptive sample mining to dynamically separate clean, ambiguous, and noisy samples based on confidence estimation and mutual learning, improving robustness during training [40]. LA-Net exploits facial landmark cues to construct neighborhood-based label distributions, thereby suppressing uncertainty in expression space without incurring extra inference cost [33]. ReSup further enhances noise suppression by modeling clean and noisy label distributions via dual-network collaboration, achieving strong performance gains [36]. In addition, consistency-based regularization methods such

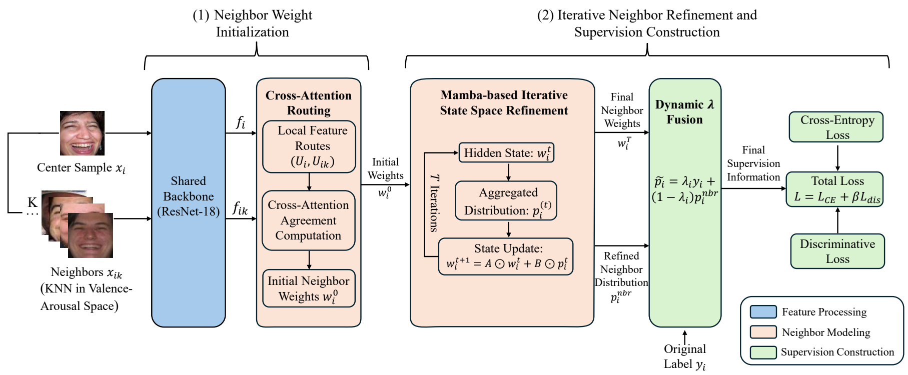

Fig. 2: Overview of the proposed framework for FER with noisy labels.

as MCR learn noise-robust representations by enforcing prediction consistency rather than relying solely on labels [26]. Other works focus on label distribution learning with uncertainty compensation to refine supervision information in FER [25]. Although these approaches highlight the effectiveness of structured supervision, the mechanisms for dynamic weighting and iterative refinement in neighbor modeling have not been sufficiently explored.

2.2 Inter-sample Relationships in FER

In FER, exploiting relationships among samples or features has proven effective for enhancing discriminative representations. Recent studies increasingly adopt neighborhood- or relation-aware mechanisms to model local and global interactions. For instance, coordinate-based neighborhood attention is proposed to capture directional dependencies among nearby features, effectively retrieving local relational cues and achieving competitive performance on benchmark datasets [21]. Other works integrate contextual information through attention-based architectures that fuse hierarchical features, enabling the modeling of inter-dependencies among facial regions and improving expression discrimination [3]. Additionally, approaches combining local feature enhancement with global correlation have been shown to alleviate challenges caused by occlusion and pose variation, implicitly reinforcing relational cues during representation learning [5]. These methods demonstrate the importance of relational modeling in real-world FER. Nevertheless, most existing approaches rely on static or heuristic aggregation strategies, lacking both dynamic adaptability and progressive refinement mechanisms to adaptively weight neighbor contributions, which motivates more flexible neighbor-aware frameworks.

3 Method

3.1 Overview

Given an input facial image $\mathbf{x}_i$, the goal of facial expression recognition under noisy labels is to learn a robust classifier $f(\cdot)$ that can tolerate annotation ambiguity while preserving discriminative representations. Instead of relying solely on the provided label $\mathbf{y}_i$, we aim to construct a more informative supervision information by exploiting neighborhood consistency among samples. Figure 2 illustrates the overall framework of the proposed method.

Specifically, for each input sample $\mathbf{x}_i$, we first retrieve its K nearest neighbors $\mathcal{N}_i = \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}\}$ in an auxiliary space using a KNN-based strategy. Each neighbor is processed by a shared backbone network and classifier to obtain its corresponding predictive label distribution. A critical challenge is to determine appropriate contribution weights $\mathbf{w}_i = [w_{i_1}, \dots, w_{i_K}]$ for different neighbors, as noisy labels and ambiguous expressions may lead to unreliable neighborhood information.

To address this challenge, we formulate neighbor contribution estimation as an iterative refinement process. We first introduce a cross-attention routing (CAR) module to generate an initial estimation $\mathbf{w}_i^{(0)}$. The CAR module performs multi-route feature projection for both the center sample and its neighbors, and computes initial neighbor weights through cross-attention-based interaction, providing a stable and effective initial state. Based on this initialization, we further propose a Mamba-based state space model to iteratively refine the neighbor weights. At iteration t, the model takes the aggregated neighborhood distribution $\mathbf{p}_i^{(t)}$ as input and updates the weight vector $\mathbf{w}_i^{(t)}$ by dynamically balancing historical state information and newly absorbed neighborhood cues. After T iterations, the optimized weights $\mathbf{w}_i^{(T)}$ are obtained and used to construct the final aggregated neighborhood distribution $\mathbf{p}_i^{(T)}$.

Finally, the optimized neighborhood distribution $\mathbf{p}_i^{(T)}$ is fused with the original logical label $\mathbf{y}_i$ through a learnable fusion factor λ_i to form the final supervision information $\tilde{\mathbf{y}}_i$ for training. The entire framework introduces no additional computational cost during inference. By jointly modeling neighborhood relationships and iterative weight refinement, the proposed method effectively mitigates the impact of noisy labels and improves robustness in facial expression recognition.

3.2 Iterative Neighbor Weight Refinement

In this section, we describe the proposed iterative neighbor weight refinement strategy, which consists of two key components: CAR-based weight initialization and Mamba-based iterative refinement. The overall objective is to learn reliable neighbor contribution weights under noisy labels by progressively refining neighborhood supervision.

Neighborhood construction in the Valence–Arousal Space Following previous studies [4, 34], we assume that facial images tend to share similar emotional semantics with their neighbors in an appropriate auxiliary space. Under this assumption, the label distribution of a given instance can be constructed by aggregating information from its neighbors, which provides a more reliable supervision information under noisy labels. A key factor in this process is the choice of the auxiliary space, which should be highly correlated with the underlying emotion space to enable effective information transfer.

Among various candidates, such as facial landmarks or action units, we adopt the valence–arousal (V–A) space as the auxiliary space. The V–A representation provides a continuous and compact description of affective states, where valence measures the positivity or negativity of an expression, and arousal reflects its activation or intensity level. This representation has been widely used to characterize the human emotional spectrum and exhibits strong correspondence with discrete emotion categories. As a result, the V–A space is particularly suitable for modeling affective similarity and constructing meaningful neighborhoods.

Formally, for each sample $\mathbf{x}_i$, we associate it with a V–A representation $\mathbf{v}_i = (v_i^{\text{val}}, v_i^{\text{aro}}) \in \mathbb{R}^2$. Since ground-truth V–A annotations are not always available across FER datasets, we adopt an existing method [27] to generate pseudo V–A values for all samples. Based on these representations, the neighborhood of $\mathbf{x}_i$ is constructed using a KNN strategy:

$$\mathcal{N}_i = \text{KNN}(\mathbf{v}_i, \{\mathbf{v}_j\}_{j=1}^{N}), \tag{1}$$

where the distance between two samples is defined as

$$d_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\|_2. \tag{2}$$

CAR-based Weight Initialization Based on the neighborhood $\mathcal{N}_i$ constructed in the V–A space, we first feed the center sample $\mathbf{x}_i$ and its K neighbors $\{\mathbf{x}_{i_k}\}_{k=1}^{K}$ into a shared feature extractor $\mathcal{F}$ to obtain deep feature representations, which are then mapped to predictive label distributions by a classifier f:

$$\mathbf{f}_i = \mathcal{F}(\mathbf{x}_i), \quad \mathbf{f}_{i_k} = \mathcal{F}(\mathbf{x}_{i_k}), \tag{3}$$

$$\hat{\mathbf{y}}_i = f(\mathbf{f}_i), \quad \hat{\mathbf{y}}_{i_k} = f(\mathbf{f}_{i_k}), \tag{4}$$

where $\hat{\mathbf{y}}_i \in \mathbb{R}^C$ denotes the predictive label distribution over C emotion classes for the center sample, and $\hat{\mathbf{y}}_{i_k} \in \mathbb{R}^C$ represents the corresponding distribution of the k-th neighbor.

To initialize neighbor weights under noisy labels, we propose a CAR mechanism. This mechanism can decompose holistic representations into multiple functional routes, thereby explicitly modeling the fine-grained semantic agreement between samples. Unlike standard self-attention which focuses on intra-sample dependencies [28], our CAR mechanism encode features as vector entities and model part–whole relationships through consistency between the center sample and its neighbors. By projecting features into different semantic routes, the model can effectively discern whether a neighbor shares consistent local affective patterns with the center sample. This routing process ensures that neighbor contributions are determined by multi-dimensional semantic alignment rather than simple proximity, providing a robust foundation for handling ambiguous annotations.

Concretely, the CAR introduces a multi-route local feature projection for both the input and neighbor samples. The feature vector $\mathbf{f}_i$ is first transformed into a set of local feature routes:

$$\mathbf{U}_i \in \mathbb{R}^{N_l \times d_l}, \tag{5}$$

where N_l denotes the number of local routes and d_l is the dimensionality of each route. The same transformation is applied to each neighbor feature $\mathbf{f}_{i_k}$, yielding

$$\mathbf{U}_{i_k} \in \mathbb{R}^{N_l \times d_l}. \tag{6}$$

Each route is then normalized to ensure stable feature:

$$\mathrm{norm}(\mathbf{U}) = \frac{\|\mathbf{U}\|^2}{1 + \|\mathbf{U}\|^2} \frac{\mathbf{U}}{\|\mathbf{U}\|}. \tag{7}$$

These local features are subsequently projected into high-level semantic spaces through a shared transformation matrix $\mathbf{W_{n,m}} \in \mathbb{R}^{N_l \times N_h \times d_l \times d_h}$:

$$\hat{\mathbf{U}}_i^{(n,m)} = \mathbf{W}_{n,m} \mathbf{U}_i^{(n)}, \quad \hat{\mathbf{U}}_{i_k}^{(n,m)} = \mathbf{W}_{n,m} \mathbf{U}_{i_k}^{(n)}, \tag{8}$$

where $n \in \{1, \ldots, N_l\}$ indexes the local routes and $m \in \{1, \ldots, N_h\}$ indexes the high-level semantic categories. Here, N_h and d_h denote the number and dimensionality of high-level semantic categories, respectively. Each local route casts a vote for every high-level semantic categories through the transformation matrix $\mathbf{W}_{n,m}$.

The subsequent cross-attention-based interaction among these features serves as the foundation for computing the initial neighbor weights. For each local route n, the agreement between the center prediction and the neighbor prediction is computed as:

$$e_{i_k}^{(n)} = \frac{\langle \hat{\mathbf{U}}_i^{(n)}, \hat{\mathbf{U}}_{i_k}^{(n)} \rangle}{\sqrt{d_h}}, \tag{9}$$

and normalized across neighbors:

$$\alpha_{i_k}^{(n)} = \frac{\exp(e_{i_k}^{(n)})}{\sum_{j=1}^{K} \exp(e_{i_j}^{(n)})}. \tag{10}$$

The initial neighbor weight for $\mathbf{x}_{i_k}$ is then computed by averaging routing coefficients across all local routes:

$$w_{i_k}^{(0)} = \frac{1}{N_l} \sum_{n=1}^{N_l} \alpha_{i_k}^{(n)}, \tag{11}$$

which is subsequently normalized to satisfy $\sum_{k=1}^{K} w_{i_k}^{(0)} = 1$.

Through the CAR mechanism, the model is able to identify neighbors that not only lie close in the V–A space but also exhibit high semantic agreement with the center sample at a fine-grained multi-route level.

Mamba-based iterative refinement Most neighborhood-based methods operate in a single pass. They compute weights based on initial, often noisy similarity measures and then perform direct aggregation without any error-correction mechanism. This reliance on potentially unreliable initial neighborhood relations can amplify errors. To further optimize neighbor weights in a dynamic and context-aware manner, we employ a Mamba-based iterative refinement mechanism [7]. Using the neighborhood information as input, it iteratively updates these weights over multiple steps. This process refines the reliability of each neighbor through repeated contextual reasoning. Even if the initial weights are significantly biased due to noise, Mamba can gradually correct them in subsequent iterations by integrating global information from all neighbors, thereby mitigating the influence of anomalous samples. This results in more accurate pseudo-labels that are consistent with the broader context.

Let $\{w_{i_k}^{(0)}\}_{k=1}^{K}$ denote the initial neighbor weights obtained from the CAR. Using these weights, we first construct an initial aggregated neighbor label distribution:

$$\mathbf{p}_i^{(0)} = \sum_{k=1}^{K} w_{i_k}^{(0)} \hat{\mathbf{y}}_{i_k}, \tag{12}$$

This distribution summarizes the semantic consensus among neighbors weighted by their initial reliability. We project both $\mathbf{w}_i^{(t)}$ and $\mathbf{p}_i^{(t)}$ into the same dimension d_z.

Correspondingly, we initialize the Mamba hidden state using the neighbor weights:

$$\mathbf{w}_i^{(0)} = [w_{i_1}^{(0)}, \ldots, w_{i_K}^{(0)}]^\top. \tag{13}$$

At iteration t, the Mamba module updates the hidden state according to:

$$\mathbf{w}_i^{(t+1)} = \mathbf{A} \odot \mathbf{w}_i^{(t)} + \mathbf{B} \odot \mathbf{p}_i^{(t)}, \tag{14}$$

where $\mathbf{A}$ and $\mathbf{B}$ are learnable parameters controlling state retention and input absorption, respectively, and $\odot$ denotes element-wise multiplication. Intuitively, $\mathbf{A}$ determines how much of the previous neighbor weighting is preserved, while $\mathbf{B}$ controls how the current semantic distribution influences the weight update.

Using the updated weights, a new aggregated neighbor label distribution is computed:

$$\mathbf{p}_i^{(t+1)} = \sum_{k=1}^{K} w_{i_k}^{(t+1)} \hat{\mathbf{y}}_{i_k}, \tag{15}$$

Through this iterative process, neighbor weights are progressively refined under the guidance of the evolving aggregated semantic distribution. Unlike conventional approaches that update distributions based on fixed weights, our formulation treats neighbor weights as latent dynamic variables and allows them to adapt according to global semantic consistency. After T iterations, the final neighbor weights $\mathbf{w}_i^{(T)}$ are used to construct the refined neighbor distribution, which serves as a robust supervision information under noisy labels.

3.3 Dynamic Fusion and Loss Function

After obtaining the refined neighbor weights $\mathbf{w}_i^{(T)}$ through the Mamba-based iterative optimization, we construct the final aggregated neighbor label distribution as:

$$\mathbf{p}_i^{\mathrm{nbr}} = \sum_{k=1}^{K} w_{i_k}^{(T)} \hat{\mathbf{y}}_{i_k}, \tag{16}$$

Instead of relying solely on either the original logical label of center sample or the neighbor-induced distribution, we dynamically fuse the two using a learnable parameter $\lambda_i \in [0, 1]$:

$$\tilde{\mathbf{p}}_i = \lambda_i \, \mathbf{y}_i + (1 - \lambda_i) \, \mathbf{p}_i^{\mathrm{nbr}}. \tag{17}$$

The parameter λ_i is produced by a prediction head and is optimized jointly with the rest of the network via back-propagation.

Intuitively, λ_i controls the relative trust between the center sample and its neighbors. A larger λ_i indicates that the model places more confidence in the original logical label of center sample, which is desirable when the neighbors are less reliable or semantically inconsistent. Conversely, a smaller λ_i encourages the model to rely more on the aggregated neighbor distribution, which is beneficial when the original logical label of center sample is incorrect but its neighbors provide consistent semantic evidence. This adaptive fusion allows the model to flexibly balance self-information and neighborhood consensus under varying noise conditions.

The fused distribution $\tilde{\mathbf{p}}_i$ is then used as the supervision information for training. Specifically, we adopt the cross-entropy loss between the fused distribution and the network prediction:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{N} \sum_{c=1}^{C} \tilde{p}_{i,c} \log \hat{y}_{i,c}, \tag{18}$$

where $\hat{y}_{i,c}$ denotes the predicted probability of class c for sample $\mathbf{x}_i$.

To further enhance class separability and mitigate the adverse effects of noisy labels, we additionally incorporate a discriminative loss term following previous work [2, 6]. Given the feature representation $\mathbf{f}_i$ and the class prototype $\boldsymbol{\mu}_c$, the discriminative loss is defined as:

$$\mathcal{L}_{\mathrm{dis}} = \sum_{i=1}^{N} \left(\|\mathbf{f}_i - \boldsymbol{\mu}_{y_i}\|_2^2 - \frac{1}{C-1} \sum_{c \neq y_i} \|\mathbf{f}_i - \boldsymbol{\mu}_c\|_2^2 \right), \tag{19}$$

where y_i is the class index of the i-th sample while $\boldsymbol{\mu}_{y_i}$ and $\boldsymbol{\mu}_c$ denotes the feature centroids of the y_i-th and c-th classes, respectively. This loss encourages features to be close to their corresponding class centers while being far away from other class centers, thereby improving discriminability under noisy supervision.

The overall loss function is given by:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \beta \mathcal{L}_{\mathrm{dis}}, \tag{20}$$

where β is a balancing coefficient controlling the contribution of the discriminative loss.

4 Experiments

4.1 Datasets

We evaluate our method on three widely used FER benchmarks: RAF-DB [15], AffectNet [18], and FERPlus [1]. RAF-DB contains approximately 30,000 in-the-wild facial images with large variations in pose, illumination, and occlusion, annotated by multiple trained annotators into seven basic emotion categories. Following the standard protocol, it is split into 12,271 training images and 3,068 testing images. AffectNet is a large-scale dataset with over 400,000 Internet images annotated with both discrete emotion labels and valence–arousal values. We use its manually annotated subset, comprising about 287,000 training samples and 3,500 test samples, and focus on the seven basic emotions for consistency. FERPlus extends FER2013 by providing more reliable annotations through crowd-sourcing, where each image is labeled by ten annotators. We convert the voting results of the seven basic emotions into soft label distributions and use the standard split of 28,709 training images and 3,589 testing images.

4.2 Implementation details

We resize all input facial images to 224×224 pixels and apply standard data augmentation techniques, including random horizontal flipping and random cropping, to enhance the diversity of the training data. We adopt ResNet-18 [10] as the backbone network, which is initialized with ImageNet pre-trained weights. For neighborhood construction in the auxiliary V–A space, the number of nearest neighbors is fixed to $K = 8$ for all experiments. In the proposed CAR mechanism, the N_l is set to 16, and the N_h is 7. The dimensionalities of local routes and high-level concepts are fixed to $d_l = 8$ and $d_h = 16$, respectively. For the Mamba-based iterative neighbor weight refinement module, the number of iterations is set to $T = 3$. The dynamic fusion parameter λ is initialized to 0 and optimized jointly with other network parameters via back-propagation. The balancing coefficient of discriminative loss is fixed to $\beta = 0.1$. The model is optimized using the Adam optimizer [12] with a batch size of 32. All hyper-parameters are kept fixed across different datasets to ensure fair comparison.

Table 1: Comparison with state-of-the-art methods under symmetric noise.

Method	RAF-DB			AffectNet			FERPlus		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
Baseline	81.01	77.98	75.50	57.24	55.89	52.16	83.29	82.34	79.77
SCN	82.18	80.10	77.46	58.58	57.25	55.05	84.28	83.17	82.47
DMUE	83.19	81.02	79.41	61.21	59.06	56.88	-	-	-
RUL	86.17	84.32	82.06	60.54	59.01	56.93	86.93	85.05	83.90
EAC	88.02	86.05	84.42	61.11	60.29	58.91	87.03	86.07	85.44
SOFT	89.05	87.86	86.08	61.16	61.35	59.50	-	-	-
FENN	85.45	83.41	81.84	59.49	58.21	57.09	-	-	-
LA-Net	88.75	87.12	85.33	62.85	61.72	60.82	88.02	86.85	86.01
LDLVA	87.98	86.81	85.85	64.37	63.89	62.57	-	-	-
MCR	89.28	87.61	86.08	62.33	61.35	59.50	-	-	-
Ours	**89.97**	**88.38**	**87.62**	**66.18**	**65.66**	**63.91**	**88.11**	**87.62**	**87.15**

4.3 Comparison with State of the Arts

Evaluation on Symmetric Noise We conduct extensive experiments to evaluate the noise robustness of our approach by artificially corrupting the training labels with symmetric noise at varying rates: 10%, 20%, and 30%. The corrupted labels are generated by randomly allocating the original labels to other emotion classes with equal probability. We compare our method against competitive noise-robust approaches: MCR [26], LDLVA [13], LA-Net [33], FENN [8], SOFT [17], EAC [39], RUL [38], DMUE [24], and SCN [32] on both RAF-DB, AffectNet, and FERPlus datasets, as shown in Tab. 1. The results show that our method consistently surpasses MCR, which represents the current best-performing method in this comparison. Specifically, on the RAF-DB dataset, our approach outperforms MCR by 0.69%, 0.77%, and 1.54%, under 10%, 20%, and 30% noise ratios, respectively. On the AffectNet dataset, we achieve improvements of 1.81%, 1.77%, and 1.34% over LDLVA at the corresponding noise levels. The experimental results validate that our method effectively handles label noise and maintains stable performance under all ratios of noise.

Table 2: Comparison with state-of-the-art methods under asymmetric noise on RAF-DB.

Method	**10%**	**20%**	**30%**
DMUE	74.18	73.67	66.09
SOFT	88.01	86.35	81.92
Ours	89.30	87.44	83.56

Table 3: Ablation study of different neighbor weight initialization strategies under varying noise ratios on three FER datasets.

Datasets	Methods	10%	20%	30%
RAF-DB	V-A Distance	88.04	87.51	85.96
	Feature Similarity	88.43	87.45	86.16
	CAR(Ours)	**89.97**	**88.38**	**87.62**
AffectNet	V-A Distance	60.15	58.93	57.20
	Feature Similarity	65.01	63.89	60.43
	CAR(Ours)	**66.18**	**65.66**	**63.91**
FERPlus	V-A Distance	84.73	83.49	82.21
	Feature Similarity	86.80	85.77	84.64
	CAR(Ours)	**88.11**	**87.62**	**87.15**

Evaluation on Asymmetric Noise Real-world noisy labels often exhibit asymmetric patterns, where certain emotion classes are more likely to be confused with other specific categories rather than being evenly distributed across all classes [9, 19]. To simulate this realistic scenario, we design asymmetric noise experiments where label injection follows class-dependent transition probabilities. We construct an asymmetric noise transition matrix by analyzing common confusion patterns in FER. Specifically, we train a standard ResNet-18 model with clean data and construct a matrix based on its top-1 error. Based on this analysis, we identify the most common misclassified pairs and design a transition matrix where each emotion class has a higher probability of being misclassified to the most easily confused corresponding classes. For example, the "fear" expression is mainly mislabeled as "surprise", while "disgust" is mainly confused with "anger".

We benchmark our method against SOFT and DMUE under asymmetric noise rates of 10%, 20%, 30%. Asymmetric noise presents greater challenges than symmetric noise as models must simultaneously discriminate confusable sentiment pairs and maintain robustness against systematic labeling biases. Our experimental results show that our method maintains excellent performance under asymmetric noise conditions, as shown in Tab. 2. On RAF-DB, we achieve 1.29%, 1.09% and 1.64% improvements over SOFT at 10%, 20% and 30%noise rates, respectively. Our proposed approach achieves consistent improvements under different asymmetric noise rates, indicating that our method can effectively handle label noise commonly found in real applications.

4.4 Ablation Studies

Effect of CAR-based Weight Initialization This experiment evaluates the effectiveness of different neighbor weight initialization strategies under noisy labels. We compare the proposed CAR initialization with two commonly used alternatives: (1) V–A distance-based initialization, which assigns weights ac-

Table 4: Ablation study on the effect of Mamba-based weight refinement under different noise ratios.

Method	**RAF-DB**			**AffectNet**			**FERPlus**		
	10%	**20%**	**30%**	**10%**	**20%**	**30%**	**10%**	**20%**	**30%**
Without Mamba	88.15	86.81	86.43	65.29	64.90	62.72	87.23	86.74	85.96
With Mamba (Ours)	89.97	88.38	87.62	66.81	65.66	63.91	88.11	87.62	87.15

cording to the Euclidean distance in the auxiliary V–A space, and (2) feature similarity-based initialization, where weights are computed using cosine similarity between deep features. For fair comparison, all variants share the same neighbor set constructed in the V–A space, while the subsequent Mamba-based refinement and dynamic λ fusion are kept identical.

We conduct experiments on three widely used FER datasets under three different noise ratios (10%, 20%, and 30%). The results are summarized in Table 3.

From the results, we observe that distance-based and feature similarity-based initialization strategies exhibit noticeable performance degradation as the noise ratio increases across all datasets. This is mainly because these methods rely on coarse similarity measures and are sensitive to noisy supervision. In contrast, the proposed CAR initialization consistently achieves superior performance under all noise levels. By enabling fine-grained route-level comparison between the center sample and its neighbors, our method can effectively identify reliable neighbors that provide consistent semantic evidence, resulting in a more robust initialization under noisy labels.

Effect of Mamba-based Weight Refinement This experiment evaluates the effectiveness of the proposed Mamba-based iterative neighbor weight refinement module. We first compare the full model with Mamba refinement against a variant without Mamba, where neighbor weights are initialized by the CAR mechanism and kept fixed during training. All other components, including dynamic λ fusion and loss functions, remain unchanged to ensure fair comparison.

Experiments are also conducted on three FER datasets under three noise ratios (10%, 20%, and 30%). From Table 4, we observe that incorporating Mamba-based refinement consistently improves performance across all datasets and noise levels. Furthermore, it maintains high accuracy even at high noise levels, indicating that our method is still applicable in high-noise environments.

To further analyze the influence of the number of refinement iterations, we evaluate the model with different numbers of Mamba iterations T. The results on RAF-DB are reported in Table 5, where all other settings are fixed.

We find that performance improves steadily as the number of refinement iterations increases, and reaches the best performance at $T = 3$. Further increasing the number of iterations leads to marginal gains or slight degradation, which suggests that excessive refinement may introduce redundant updates. Based on this observation, we set the number of Mamba iterations to $T = 3$ in all experiments.

Table 5: Effect of the number of Mamba refinement iterations T on RAF-DB.

Iterations T	**10% Noise**	**20% Noise**	**30% Noise**
$T = 1$	87.68	87.22	86.36
$T = 2$	88.19	87.73	87.15
$T = 3$	89.97	88.38	87.62
$T = 4$	89.21	88.10	87.34

Table 6: Ablation study on the effect of dynamic λ fusion on RAF-DB.

Fusion Strategy	**10% Noise**	**20% Noise**	**30% Noise**
Fixed $\lambda = 1$ (Center Only)	84.28	82.09	80.12
Fixed $\lambda = 0$ (Neighbors Only)	88.35	87.51	86.74
Learnable λ (Ours)	89.97	88.38	87.62

Effect of Dynamic $\boldsymbol{\lambda}$ Fusion This experiment investigates the impact of the proposed dynamic λ fusion strategy, which adaptively balances the contribution of the center prediction and the aggregated neighbor distribution. To evaluate its effectiveness, we conduct experiments on the RAF-DB dataset under three different noise ratios (10%, 20%, and 30%). We compare the proposed learnable λ with two fixed fusion strategies: (1) $\lambda = 1$, where only the center prediction is used for supervision, and (2) $\lambda = 0$, where supervision relies solely on the aggregated neighbor distribution. All other components of the model are kept unchanged for fair comparison.

As shown in Table 6, both fixed fusion strategies suffer from noticeable performance degradation under noisy labels. Relying solely on the center prediction is vulnerable to corrupted annotations, while using only the neighbor distribution may introduce bias when neighbors are unreliable. In contrast, the proposed learnable λ consistently achieves superior performance across all noise levels. These results demonstrate that dynamically balancing self-information and neighborhood consensus is crucial for robust facial expression recognition under noisy supervision.

5 Conclusion

This paper proposes a robust FER framework for learning with noisy labels. Using the valence–arousal space, we construct meaningful neighborhoods for reliable auxiliary supervision. CAR selectively identifies informative neighbors at the route level, while Mamba-based iterative refinement progressively optimizes neighbor weights and aggregated label distributions. A learnable fusion parameter further balances the original label and neighborhood consensus. Extensive experiments on multiple FER benchmarks under symmetric and asymmetric noise demonstrate consistent improvements over state-of-the-art methods, validating the effectiveness and robustness of our approach.

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. p. 279–283. ICMI '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2993148.2993165, https://doi.org/10.1145/2993148.2993165
2. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 302–309 (2018). https://doi.org/10.1109/FG.2018.00051
3. Chen, J., Chen, N.: A framework for facial expression recognition combining contextual information and attention mechanism. Journal of Information Processing Systems (2024), https://jips-k.org/pub-reader/1187, published 31 Aug 2024
4. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13981–13990 (2020). https://doi.org/10.1109/CVPR42600.2020.01400
5. Chen, Z., Yan, L., Wang, H., Adamyk, B.: Improved facial expression recognition algorithm based on local feature enhancement and global information association. Electronics **13**(14) (2024). https://doi.org/10.3390/electronics13142813, https://www.mdpi.com/2079-9292/13/14/2813
6. Farzaneh, A.H., Qi, X.: Discriminant distribution-agnostic loss for facial expression recognition in the wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1631–1639 (2020). https://doi.org/10.1109/CVPRW50498.2020.00211
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. ArXiv **abs/2312.00752** (2023), https://api.semanticscholar.org/CorpusID:265551773
8. Gu, Y., Yan, H., Zhang, X., Wang, Y., Ji, Y., Ren, F.: Toward facial expression recognition in the wild via noise-tolerant network. IEEE Transactions on Circuits and Systems for Video Technology **33**(5), 2033–2047 (2023). https://doi.org/10.1109/TCSVT.2022.3220669
9. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in neural information processing systems. pp. 8527–8537 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
11. Iscen, A., Valmadre, J., Arnab, A., Schmid, C.: Learning with Neighbor Consistency for Noisy Labels . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4662–4671. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.00463, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00463
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
13. Le, N., Nguyen, K.T.N., Tran, Q.D., Tjiputra, E., Le, B., Nguyen, A.: Uncertainty-aware label distribution learning for facial expression recognition. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 6077–6086 (2022), https://api.semanticscholar.org/CorpusID:252407727

14. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)
15. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing **28**(1), 356–370 (2019)
16. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593 (2017). https://doi.org/10.1109/CVPR.2017.277
17. Lukov, T., Zhao, N., Lee, G.H., Lim, S.N.: Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. p. 648–665. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19775-8_38, https://doi.org/10.1007/978-3-031-19775-8_38
18. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. In: IEEE Transactions on Affective Computing. vol. 10, pp. 18–31. IEEE (2017)
19. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1944–1952 (2017)
20. Patrini, G., Rozza, A., Menon, A.K., Nock, R., Qu, L.: Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach . In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2233–2241. IEEE Computer Society, Los Alamitos, CA, USA (Jul 2017). https://doi.org/10.1109/CVPR.2017.240, https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.240
21. Peng, C., Li, B., Zou, K., Zhang, B., Dai, G., Tsoi, A.C.: An innovative neighbor attention mechanism based on coordinates for the recognition of facial expressions. Sensors **24**(22), 7404 (2024). https://doi.org/10.3390/s24227404
22. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(6), 1113–1133 (2015). https://doi.org/10.1109/TPAMI.2014.2366127
23. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition . In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6244–6253. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.00618, https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00618
24. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: 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 (CVPR). pp. 6248–6257 (June 2021)
25. Shin, H., Lee, B., Ku, B., Ko, H.: Noisy label facial expression recognition via face-specific label distribution learning. Image and Vision Computing **143**, 104901 (2024). https://doi.org/https://doi.org/10.1016/j.imavis.2024.104901, https://www.sciencedirect.com/science/article/pii/S0262885624000040

26. Tan, Y., Xia, H., Song, S.: Mining label-free consistency regularization for noisy facial expression recognition. Complex & Intelligent Systems **11** (12 2024). https://doi.org/10.1007/s40747-024-01722-7
27. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nature Machine Intelligence **3** (01 2021). https://doi.org/10.1038/s42256-020-00280-0
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
29. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6896–6905 (2020). https://doi.org/10.1109/CVPR42600.2020.00693
30. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6896–6905 (2020), https://api.semanticscholar.org/CorpusID:211258635
31. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 322–330 (2019), https://api.semanticscholar.org/CorpusID:201070534
32. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 322–330 (2019)
33. Wu, Z., Cui, J.: LA-Net: Landmark-Aware Learning for Reliable Facial Expression Recognition under Label Noise . In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20641–20650. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.01892, https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01892
34. Xu, N., Liu, Y.P., Geng, X.: Label enhancement for label distribution learning. IEEE Transactions on Knowledge and Data Engineering **33**(4), 1632–1643 (2021). https://doi.org/10.1109/TKDE.2019.2947040
35. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
36. Zhang, X., Lu, Y., Yan, H., Huang, J., Gu, Y., Ji, Y., Liu, Z., Liu, B.: ReSup: Reliable Label Noise Suppression for Facial Expression Recognition . IEEE Transactions on Affective Computing **16**(03), 2006–2019 (Jul 2025). https://doi.org/10.1109/TAFFC.2025.3549017, https://doi.ieeecomputersociety.org/10.1109/TAFFC.2025.3549017
37. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
38. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan,

J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 17616–17627. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/9332c513ef44b682e9347822c2e457ac-Paper.pdf

39. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI. p. 418–434. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19809-0_24, https://doi.org/10.1007/978-3-031-19809-0_24
40. Zhang, Z., Sun, X., An, L., Wang, M.: Asm: Adaptive sample mining for in-the-wild facial expression recognition. In: Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV 2023, Xiamen, China, October 13–15, 2023, Proceedings, Part V. p. 291–302. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-981-99-8469-5_23, https://doi.org/10.1007/978-981-99-8469-5_23
41. Zhou, Y., Xue, H., Geng, X.: Emotion distribution recognition from facial expressions. In: Proceedings of the 23rd ACM International Conference on Multimedia. p. 1247–1250. MM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2733373.2806328, https://doi.org/10.1145/2733373.2806328