Navigating Label Ambiguity for Facial Expression Recognition in the Wild

JunGyu Lee[1,2*], Yeji Choi[1,3*], Haksub Kim[1], Ig-Jae Kim[1,2], Gi Pyo Nam[1,2]

[1] Korea Institute of Science and Technology, Seoul, Korea
[2] AI-Robotics, KIST School, University of Science and Technology, Daejeon, Korea
[3] Yonsei University, Seoul, Korea
{jungyu0413, cyjcyj91, hskim, drjay, gpnam}@kist.re.kr

Abstract

Facial expression recognition (FER) remains a challenging task due to label ambiguity caused by the subjective nature of facial expressions and noisy samples. Additionally, class imbalance, which is common in real-world datasets, further complicates FER. Although many studies have shown impressive improvements, they typically address only one of these issues, leading to suboptimal results. To tackle both challenges simultaneously, we propose a novel framework called Navigating Label Ambiguity (NLA), which is robust under real-world conditions. The motivation behind NLA is that dynamically estimating and emphasizing ambiguous samples at each iteration helps mitigate noise and class imbalance by reducing the model's bias toward majority classes. To achieve this, NLA consists of two main components: Noise-aware Adaptive Weighting (NAW) and consistency regularization. Specifically, NAW adaptively assigns higher importance to ambiguous samples and lower importance to noisy ones, based on the correlation between the intermediate prediction scores for the ground truth and the nearest negative. Moreover, we incorporate a regularization term to ensure consistent latent distributions. Consequently, NLA enables the model to progressively focus on more challenging ambiguous samples, which primarily belong to the minority class, in the later stages of training. Extensive experiments demonstrate that NLA outperforms existing methods in both overall and mean accuracy, confirming its robustness against noise and class imbalance. To the best of our knowledge, this is the first framework to address both problems simultaneously.

Introduction

Facial expressions play a crucial role in non-verbal communication in daily life. The growing use of remote communication through video connections has heightened the importance of facial expression recognition (FER) technology. It has many valuable applications ranging from human-computer interaction in customer service to remote mental health care support. In recent years, owing to the emergence of large-scale datasets collected in the wild, such as FERPlus (Barsoum et al. 2016), RAF-DB (Li, Deng, and Du 2017), and AffectNet (Mollahosseini, Hasani, and Mahoor 2017), many deep learning-based FER methods (Wang et al. 2020b; Zhao, Liu, and Zhou 2021; Li et al. 2021; Xue, Wang, and Guo 2021; Xue et al. 2022; Zheng, Mendieta, and Chen 2023; Mao et al. 2023) have been developed rapidly and achieved impressive performance.

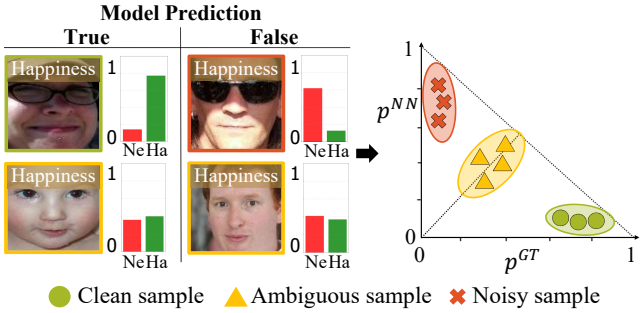

Figure 1: **Visualization of estimating sample ambiguity.** The visual analysis on the right illustrates how the correlation between the prediction scores for ground truth (GT) and nearest negative (NN) serves as a criterion for categorizing samples as clean, ambiguous, or noisy. The prediction scores for the GT and the NN are represented by green and red bars in the probability distribution on the left side, respectively (Ne: Neutral, Ha: Happiness).

One of the main challenges in FER is label ambiguity, which arises from variability in the expression and interpretation of facial emotions across individuals. This complexity makes it difficult to reliably categorize facial expressions into one of the seven basic classes (e.g., neutral, happiness, fear, etc.). Several approaches (She et al. 2021; Zhang, Wang, and Deng 2021; Zhang et al. 2022b) have attempted to mitigate this ambiguity by comparing pairwise samples or aligning predictions from multiple branches. Meanwhile, class imbalance presents another critical issue in FER, where certain expressions are overrepresented in datasets (e.g., happiness and neutral), while others (e.g., fear and disgust) are underrepresented. MEK (Zhang et al. 2024) has been proposed to address this imbalance by emphasizing minority classes through re-balanced attention consistency. However, existing methods focus on either label ambiguity or class imbalance in isolation, resulting in suboptimal performance. Moreover, studies on label ambiguity often take a narrow approach, focusing solely on noisy samples without leveraging

*These authors contributed equally.

their full potential to address class imbalance.

To tackle these challenges together, we prioritize sample ambiguity as a key criterion. Drawing inspiration from dynamic weight adjustment, we assign different weights to each sample based on its estimated ambiguity. Fig. 1 illustrates the process of estimating sample ambiguity by analyzing the correlation between the prediction scores for the ground truth (GT) and the nearest negative (NN)—the class with the highest prediction score excluding the GT. When the model's prediction is true, it generates a GT score much higher than the NN score for clean samples, whereas for ambiguous samples, the scores are similar. In contrast, when the model's prediction is false, it produces the opposite result. Based on these discrepancies, we categorize the samples by facial expression class, observing that most of the ambiguous and noisy samples are concentrated in the minority class. This observation motivates us to leverage this relationship to emphasize the minority class during model training.

Building on these insights, we propose a new framework called Navigating Label Ambiguity (NLA). The core idea of NLA is to adaptively assign weights based on the estimated sample ambiguity at each training step. This strategy allows the model to progressively focus on ambiguous samples in the minority class as training progresses, while penalizing noisy samples. To achieve this, NLA comprises two main components: Noise-aware Adaptive Weighting (NAW) and consistency regularization. Specifically, NAW assigns higher weights to ambiguous samples and lower weights to noisy ones by analyzing the correlation between the prediction scores of the ground truth and the nearest negative. In particular, we apply different forms of multivariate Gaussian kernels for NAW, based on intermediate prediction results and the training epoch, ensuring that ambiguous samples within the minority class receive increased attention after the learning for the majority class has saturated. Additionally, consistency regularization further enhances the reliability of NLA by aligning the latent distributions of the original and flipped images using Jensen-Shannon Divergence.

The main contributions of our method are as follows:

- We propose a novel framework called Navigating Label Ambiguity (NLA) by leveraging adaptive weighting and consistency regularization. To the best of our knowledge, this is the first attempt to address both class imbalance and noisy labels by handling label ambiguity.
- We introduce a Noise-aware Adaptive Weighting (NAW) that dynamically assigns weights based on sample ambiguity, allowing the model to focus progressively on ambiguous samples while minimizing the impact of noise.
- NLA demonstrates superior performance in overall and mean accuracy on in-the-wild datasets. Additionally, extensive experiments under different noise and class imbalance conditions confirm the robustness of our method.

Related Work

Facial Expression Recognition

In-the-wild FER scenarios present two major challenges: 1) noisy labels, which are common in image classification but more prevalent in FER due to the inherent ambiguity of facial expressions, and 2) class imbalance resulting from varying expression frequencies. Recently, numerous methods have been proposed to alleviate these issues. For instance, SCN (Wang et al. 2020a) uses re-labeling and ranking regularization to mitigate noise, while EAC (Zhang et al. 2022a) minimizes overfitting to noisy samples by aligning Class Activation Maps (Zhou et al. 2016) of the original and flipped images. LA-Net (Wu and Cui 2023) utilizes landmark data to enhance expression features via expression-landmark interactions. Alternatively, paying more attention to ambiguity, DMUE (She et al. 2021) explores latent distributions with multiple branches based on uncertainty estimation, and RUL (Zhang, Wang, and Deng 2021) uses a multi-branch framework with feature mixup to learn from relative sample difficulty. MAN (Zhang et al. 2022b) employs a two-branch network with a co-division module to enhance discriminative ability by focusing on clean samples.

Regarding class imbalance, Face2Exp (Zeng et al. 2022) improves FER by using large-scale unlabeled face recognition data within a meta-optimization framework, while MEK (Zhang et al. 2024) emphasizes minority classes through re-balanced attention consistency and label smoothing. Despite these advances, real-world FER is still limited to improving generalization due to the combined challenges of noisy labels and class imbalance. Unlike previous works, our method addresses both issues simultaneously by dynamically exploring each sample's ambiguity.

Learning with Ambiguity

Recent approaches to handling ambiguity can be broadly categorized into small loss selection and disagreement/agreement strategies. Previous research (Arpit et al. 2017) has shown that deep neural networks initially learn simple patterns, leading to the assumption that small-loss samples are treated as clean. Inspired by this, MentorNet (Jiang et al. 2018) uses a teacher-student framework, where the student network is trained only on small-loss samples selected by the teacher network. Co-teaching (Han et al. 2018) further cross-updates two networks by exchanging small-loss samples in each mini-batch.

Alternatively, the disagreement/agreement strategy focuses on instances where predictions between two networks differ, which is similar to hard sample mining. As representative works, Decoupling (Malach and Shalev-Shwartz 2017) and Co-teaching+ (Yu et al. 2019) update networks based on these disagreements. In contrast, JoCoR (Wei et al. 2020) aligns the predictions of the two networks using Jensen-Shannon divergence to ensure agreement. Drawing ideas from this approach, we apply a consistency regularization method that aligns the probabilities between the original data and its flipped version within a single network.

Learning with Imbalanced Data

Training samples in wild datasets often exhibit imbalanced class distributions, leading to models biased toward majority classes. One basic approach to address this issue is rebalancing class distributions using different sampling frequencies for each class. Decoupling (Kang et al. 2019) eval-

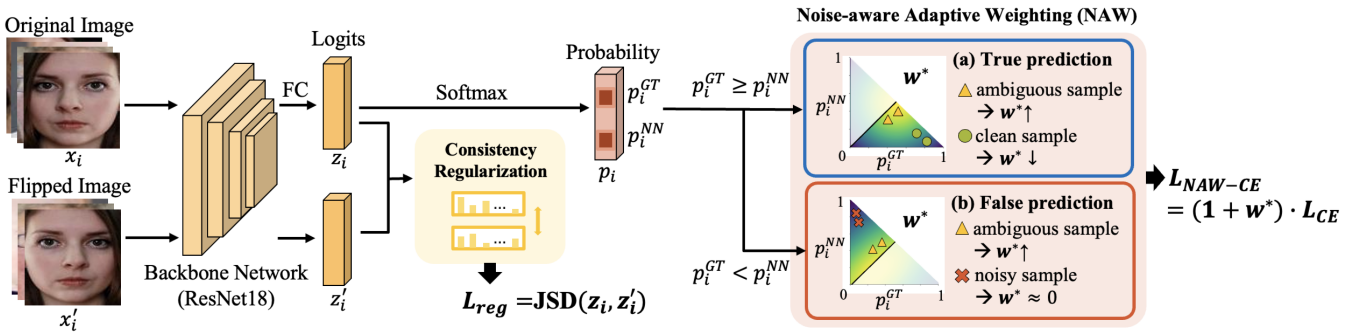

Figure 2: **The framework of Navigating Label Ambiguity (NLA).** NLA consists of two main components: 1) a Noise-aware Adaptive Weighting (NAW), which dynamically assigns weights to each sample based on the intermediate prediction scores for GT and NN, and 2) consistency regularization using pairs of original and horizontally flipped images.

uates various sampling strategies and finds that progressively balanced sampling is particularly effective. Dynamic Curriculum Learning (DCL) (Wang et al. 2019) introduces an adaptive sampling strategy that starts with random sampling and later shifts focus to minority classes.

Another approach involves re-weighting the loss for each class to ensure balanced contributions. CB loss (Cui et al. 2019a) uses the effective number, which is inversely proportional to class size, ensuring equal loss contribution. Balanced Softmax (Ren et al. 2020) adjusts predicted logits based on label frequencies by considering class priors before calculating the final loss. Our approach combines these strategies by proposing loss re-weighting based on sample ambiguity, rather than class distribution, allowing the model to prioritize ambiguous samples in the minority class during later training stages.

Method

In this section, we provide a brief review of the preliminaries related to our work. We then introduce our proposed method, Navigating Label Ambiguity (NLA), which comprises two main components: a noise-aware adaptive weighting and consistency regularization. Finally, we present the overall training objectives for our networks. The framework of our proposed method is illustrated in Fig. 2.

Preliminaries

Given a K-class, N-sample FER dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the i-th face image and $y_i \in \{1, \cdots, K\}$ represents its corresponding ground-truth label, respectively. Generally, a feature extractor based on a deep neural network utilizes a fully connected layer to predict the logit z_i for a given x_i and calculates the probability of the k-th class using a softmax function defined as:

$$p_i^k = \frac{e^{z_{ik}}}{\sum_{j=1}^{K} e^{z_{ij}}}, \tag{1}$$

where z_{ik} and z_{ij} represent the output logits for classes k and j, respectively. Then, most multi-class classification models are optimized by minimizing the cross-entropy loss, defined as follows:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \log p_i^{y_i}. \tag{2}$$

However, cross-entropy loss often produces suboptimal results in noisy and imbalanced datasets because it treats all samples with equal importance (Ghosh, Kumar, and Sastry 2017). It also tends to be biased towards easy samples, which usually belong to majority classes, weakening gradients for minority classes (Lin et al. 2017).

Noise-aware Adaptive Weighting (NAW)

Our goal is to adaptively assign weights to the loss function for each sample based on their ambiguity within the training pipeline. Following the motivation described in Fig. 1, the sample ambiguity can be estimated by considering the correlation between the prediction scores for the ground truth (GT) and the nearest negative (NN). These prediction scores for the GT and the NN are defined as:

$$p_i^{GT} = p_i^{y_i}, \qquad p_i^{NN} = \max_{k \neq y_i} p_i^k. \tag{3}$$

Building on this, we introduce the noise-aware adaptive weighting (NAW) through a multivariate Gaussian kernel that takes these two prediction scores as inputs, with a predetermined mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, formulated as follows:

$$w^*(\mathbf{p}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = C \cdot \exp\left(-\frac{1}{2}(\mathbf{p}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{p}_i - \boldsymbol{\mu})\right), \tag{4}$$

where

$$\mathbf{p}_i = \begin{bmatrix} p_i^{GT} \\ p_i^{NN} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix},$$

and $C = (2\pi\sqrt{|\boldsymbol{\Sigma}|})^{-1}$ is the normalizing constant. For convenience of presentation, we let $w^*(\mathbf{p}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = w^*(\mathbf{p}_i)$.

Then, using the above weights, we can define a NAW-based cross-entropy loss, $\mathcal{L}_{\text{NAW-CE}}$, as follows:

$$\mathcal{L}_{\text{NAW-CE}} = (1 + w^*(\mathbf{p}_i)) \cdot \mathcal{L}_{CE}. \tag{5}$$

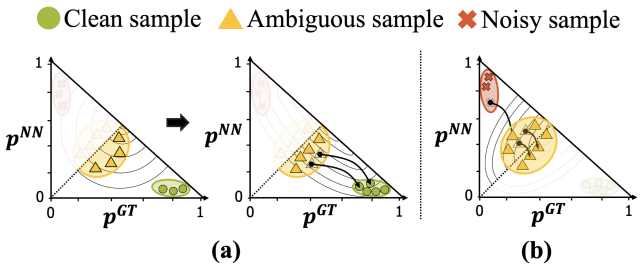

Figure 3: **Visualization of the effect of NAW by prediction results.** This figure illustrates how NAW enhances the model's ability to distinguish between clean, ambiguous, and noisy samples throughout the training process. (a) shows the results when the prediction is true, and (b) shows the results when the prediction is false.

Geometrically, the mean vector determines the center of the Gaussian kernel, where the weight is maximized, and the covariance matrix controls the shape of its contour. For example, if the covariance matrix is an identity matrix, the contours are isotropic (circular), resulting in weights that decrease at the same rate in all directions. If not, the contours take an elliptical shape, with weights decreasing gradually along the major axis and more sharply along the minor axis. Leveraging these properties, as shown in Fig. 3, we design two different forms of NAW with distinct mean vectors and covariance matrices, depending on whether the intermediate prediction is true or false, as follows.

When the prediction is true. In this case, the samples can be considered as either a clean sample (where p_i^{GT} is much higher than p_i^{NN}), or an ambiguous sample (where p_i^{GT} and p_i^{NN} are both close to 0.5). To ensure that the ambiguous samples receive the highest weight, while the relatively easy ones are assigned lower weights, we set the mean vectors in the true case, $\boldsymbol{\mu}_t = [0.5, 0.5]^T$. As training progresses, the model becomes more discriminative on the ambiguous samples, resulting in more samples being classified as clean. At this point, if an isotropic Gaussian kernel is applied throughout the entire learning process, information about clean samples may be lost due to their lower emphasis. To prevent this situation, we use a covariance scheduler (CS) that is exponentially adjusted at each epoch, modifying the contour from isotropic to an elongated ellipse along the $y = -x$ line. Fig. 3 (a) visually depicts this process. The covariance matrix in the true case is defined as follows:

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma_{11} & \text{CS} \cdot \sigma_{12} \\ \text{CS} \cdot \sigma_{21} & \sigma_{22} \end{bmatrix},$$

with

$$\text{CS}(e, E) = 1 - \exp\left(\frac{\text{-10} \cdot e}{E}\right), \tag{6}$$

where e represents the current training epoch and E denotes the total number of epochs.

When the prediction is false. In this case, the samples can be considered as either a noisy sample (where p_i^{GT} is much lower than p_i^{NN}), or an ambiguous sample (where p_i^{GT} and p_i^{NN} are similar). Since the ambiguous samples in the false case have been observed to have a lower p_i^{NN} compared to those in the true case, we set the mean vector in the false case, $\boldsymbol{\mu}_f = [0.3, 0.15]^T$. Additionally, to minimize the impact of noise, we experimentally set the covariance matrix, $\boldsymbol{\Sigma}_f$, so that its contour forms an elliptical shape elongated along the $y = x$ axis. As shown in Fig. 3 (b), NAW enhances the model's ability to distinguish ambiguous samples that were previously misclassified, while gradually increasing the weight of samples that were initially categorized as noise but are beneficial for learning. As training progresses to the later stages, most of the remaining ambiguous samples in the false cases belong to minority classes. NAW pays increased attention to these samples, improving the model's discriminative ability for minority classes.

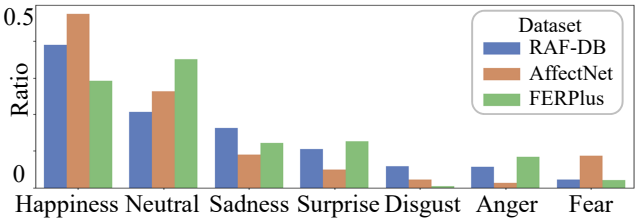

Figure 4: Imbalanced distribution of training samples in the wild FER dataset.

Consistency Regularization

Consistency regularization is one of the effective way to enhance a model's reliability in the presence of label ambiguity. Inspired by this, we apply a regularization technique to ensure consistent predictions across different views of the same sample. This is particularly essential for our method, as NAW assigns the weights based on the prediction scores for each sample at each epoch. To achieve this, we adopt Jensen-Shannon Divergence (JSD), which measures the alignment of two distributions, using output logits of the original image, x_i, and its horizontally flipped image, x_i'. Let z_i and z_i' be the output logits from x_i and x_i', respectively. Then, the consistency regularization loss can be defined as follows:

$$\mathcal{L}_{\text{reg}} = D_{KL}(z_i \parallel \frac{z_i + z_i'}{2}) + D_{KL}(z_i' \parallel \frac{z_i + z_i'}{2}), \tag{7}$$

where $D_{KL}(\cdot||\cdot)$ denotes the Kullback–Leibler (KL) divergence.

Loss Functions

In summary, the final loss function of our network is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{NAW-CE}} + (1 - \lambda) \cdot \mathcal{L}_{\text{reg}}, \tag{8}$$

where λ is a weighting factor.

Experiments

In this section, we evaluate the effectiveness of NLA on three in-the-wild FER benchmarks, considering variations in label noise and class imbalance. Additionally, we verify the contribution of each component of NLA through comprehensive ablation studies and provide visualization analysis.

Method	Conference	Overall	Mean	Happiness	Neutral	Sadness	Surprise	Disgust	Anger	Fear
Baseline	-	87.42	78.53	95.44	88.53	85.56	83.59	58.75	78.40	59.46
CB	CVPR 19	88.04	79.26	95.11	90.74	84.73	86.93	64.38	73.46	59.46
BBN	CVPR 20	87.39	78.19	94.59	**91.62**	84.94	84.80	61.88	77.78	52.70
RUL	NeurIPS 21	88.66	81.66	95.78	87.06	86.19	**89.36**	65.00	83.33	64.86
EAC	ECCV 22	89.05	81.09	95.27	88.97	90.17	87.84	61.25	83.33	60.81
MEK	NeurIPS 23	89.77	82.44	**96.37**	89.56	89.33	87.84	66.89	80.86	66.22
NLA(Ours)	-	**89.93**	**83.87**	95.70	88.97	**90.38**	87.23	**70.00**	**84.57**	**70.27**

Table 1: **Comparison with other methods on RAF-DB using pre-trained ResNet-18 as backbone.** We achieve significant improvement in minority classes (e.g., Disgust, Anger, and Fear), leading to the best performance in both overall and mean accuracy (**Bold**: best, underline: second best).

Method	Conference	Overall(=Mean)	Happiness	Neutral	Sadness	Surprise	Disgust	Anger	Fear
BBN	CVPR 20	60.76	87.00	57.10	**66.80**	54.90	30.10	58.30	**71.10**
RUL	NeurIPS 21	61.56	90.50	62.40	64.70	60.80	34.20	**69.30**	49.00
EAC	ICCV 22	65.17	**91.40**	64.50	65.70	61.60	45.80	66.30	60.90
MEK	NeurIPS 23	65.73	86.20	59.00	64.20	57.80	**61.90**	66.50	64.50
NLA(Ours)	-	**67.06**	88.60	**65.60**	63.60	**64.20**	61.20	60.40	65.80

Table 2: **Comparison with other methods on AffectNet using pre-trained ResNet-18 as backbone.** Our method achieves over 60% prediction accuracy across all classes, yielding the best overall accuracy (**Bold**: best, underline: second best).

Experimental Settings

Datasets. **RAF-DB** (Li, Deng, and Du 2017) contains 30,000 face images labeled with 7 basic and compound expressions by 40 trained annotators. We use a subset with 7 basic expressions, comprising 12,271 training images and 3,068 test images. **FERPlus** (Barsoum et al. 2016), an extension of FER2013 (Goodfellow et al. 2013), provides 8 expression labels with the addition of Contempt, labeled by 10 annotators. For a fair comparison, we use the same 7 basic classes, totaling 28,709 training images and 3,589 test images. **AffectNet** (Mollahosseini, Hasani, and Mahoor 2017), the largest FER dataset, contains 450,000 face images with 7 basic expressions and a contempt label. In our experiments, we use only the 7 basic classes, which include 283,901 training images and 3,500 test images. As shown in Fig. 4, the training samples in all datasets are notably imbalanced.

Implementation Details. We utilize ResNet-18 (He et al. 2016) pre-trained on MS-Celeb-1M (Guo et al. 2016), following previous works (Zhang et al. 2022a; Wu and Cui 2023), for a fair comparison. All face images are aligned and cropped based on three landmarks (Wang, Bo, and Fuxin 2019), and then resized to 224×224. For network training, we use the Adam optimizer (Kingma and Ba 2014) with a weight decay of 0.0001 and employ the ExponentialLR scheduler (Li and Arora 2019) with a gamma value of 0.9 and an initial learning rate of 0.0001. We set λ to 0.5, the batch size to 32, and the maximum training epoch to 60, with the best performance observed at epoch 40. We determine each Σ based on the diagonal element $\sigma_{11} = 0.8$ in both cases, with the ratio of the major to minor axis being 2:1 and 6:1 for the true and false cases, respectively. All experiments are conducted on a single NVIDIA A100.

Comparison with Existing Methods

We conduct evaluations on the RAF-DB and AffectNet benchmarks, comparing the performance of our proposed method with existing FER methods that use pre-trained ResNet-18 as the backbone, including CB (Cui et al. 2019a), BBN (Zhou et al. 2020), KTN (Li et al. 2021), RUL (Zhang, Wang, and Deng 2021), EAC (Zhang et al. 2022a), MEK (Zhang et al. 2024), and LA-Net (Wu and Cui 2023). As shown in Table 1, NLA outperforms other methods on RAF-DB, achieving an overall accuracy of 89.93% and a mean accuracy of 83.87%. Notably, while several methods, including RUL and EAC, have steadily improved overall accuracy by addressing noisy labels, their improvements in mean accuracy are relatively limited. A closer examination reveals that these improvements are primarily driven by enhanced performance in majority classes such as 'Happiness' and 'Neutral', while performance in minority classes such as 'Fear' and 'Disgust' remains suboptimal. This bias highlights the need for approaches that address both noisy labels and class imbalance in in-the-wild datasets. In contrast, NLA surpasses the current state-of-the-art method, MEK, by approximately 4% in minority classes and is the first to achieve over 70% accuracy in these classes, demonstrating its robustness in addressing class imbalance.

Table 2 presents the performance comparison on AffectNet, where the test set is class-balanced, making overall accuracy equal to mean accuracy. Similar to the previous results, RUL and EAC achieve high accuracy in the majority class, 'Happiness', but still suffer from performance degradation in the minority classes. Although MEK improves accuracy for minority classes through re-balancing techniques, it sacrifices accuracy in majority classes, 'Happiness' and 'Neutral'. Conversely, NLA maintains high accuracy in the

Method	Noise(%)	RAF-DB	AffectNet	FERPlus
Baseline	10	81.01	57.24	83.29
SCN	10	82.18	58.58	84.28
RUL	10	86.17	60.54	86.93
EAC	10	88.02	61.11	87.03
LA-Net	10	88.75	62.85	88.02
NLA(Ours)	10	**88.83±0.11**	**63.52±0.08**	**88.20±0.07**
Baseline	20	77.98	55.89	82.34
SCN	20	80.10	57.25	83.17
RUL	20	84.32	59.01	85.05
EAC	20	86.05	60.29	86.07
LA-Net	20	87.12	61.72	86.85
NLA(Ours)	20	**87.60±0.13**	**63.25±0.04**	**87.64±0.2**
Baseline	30	75.50	52.16	79.77
SCN	30	77.46	55.05	82.47
RUL	30	82.06	56.93	83.90
EAC	30	84.42	58.91	85.44
LA-Net	30	85.33	60.82	86.01
NLA(Ours)	30	**86.71±0.16**	**62.48±0.14**	**86.97±0.04**

Table 3: Comparison of overall accuracy with other methods under different noise ratios.

Method	Imbalance	Overall	Mean
Baseline	50	83.28	64.69
BBN	50	85.01	71.57
EAC	50	87.09	73.02
MEK	50	87.65	77.11
NLA(Ours)	50	**87.97±0.26**	**78.05±0.14**
Baseline	100	80.96	55.12
BBN	100	83.44	67.92
EAC	100	85.79	69.80
MEK	100	86.47	73.06
NLA(Ours)	100	**87.98±0.30**	**73.34±0.51**
Baseline	150	80.11	56.53
BBN	150	82.92	65.49
EAC	150	84.13	68.66
MEK	150	85.20	70.33
NLA(Ours)	150	**86.34±0.05**	**70.49±0.57**

Table 4: Comparison with other methods under different imbalances on RAF-DB.

majority class while achieving over 60% accuracy across all classes, leading to the best overall accuracy.

Different Noise Levels. We evaluate the robustness of NLA across three noise levels on the FERPlus, RAF-DB, and AffectNet datasets. Following previous studies (Zhang, Wang, and Deng 2021; Zhang et al. 2022a), we corrupt 10%, 20%, and 30% of the training labels by randomly flipping them to other categories. Experiments are conducted using five different random seeds, and we report the mean and standard deviation of the overall accuracy. As shown in Table 3, NLA consistently outperforms other methods across all noise levels, achieving significant improvements over the baseline with gains ranging from 5.38% to 10.01%. This highlights the effectiveness of our method in handling noisy labels, even under extreme noise levels, through noise-aware adaptive weighting. Furthermore, compared to the state-of-the-art LA-Net, our method achieves average improvements of 0.93%, 1.57%, and 0.83% on RAF-DB, AffectNet, and FERPlus, respectively. Considering that LA-Net employs an additional landmark-based backbone, the gains from our single-backbone model are particularly notable.

Different Imbalance Factors. To examine robustness against severe class imbalance, we follow established methods in handling imbalance in image classification (Cao et al. 2019; Cui et al. 2019b) by creating varying degrees of imbalance in the RAF-DB dataset using factors of 50, 100, and 150. The imbalance factor is the ratio of the number of training samples in the largest class to the number in the smallest class. We also conduct experiments using five random seeds, reporting the mean and standard deviation of overall and mean accuracy. Table 4 shows NLA's superior performance across all imbalance factors. Compared to MEK, the current state-of-the-art in handling imbalance, NLA outperforms it in both overall accuracy and mean accuracy. This indicates that NLA effectively discriminates between minority and majority classes without bias.

Settings	Components					Overall	Mean
	CE	NAW-CE	$\mathcal{L}_1$	CAM	JSD		
(a)	✓					87.42	78.53
(b)		✓				88.17	81.40
(c)	✓				✓	87.78	81.01
(d)		✓	✓			89.18	82.87
(e)		✓		✓		89.15	83.02
(f)		✓			✓	**89.93**	**83.87**

Table 5: Ablation study results on RAF-DB.

Ablation Studies

In Table 5, we present the results of an ablation study on the RAF-DB dataset, analyzing the contributions of each component within the NLA framework using a ResNet18 backbone. We compare the model using standard cross-entropy loss (CE) with the proposed NAW-based cross-entropy loss (NAW-CE). In setting (b), performance improves by 0.75% in overall accuracy and 2.87% in mean accuracy over the baseline (a), highlighting NAW's effectiveness in adjusting sample weights. Additionally, when regularization techniques are incorporated, as shown in settings (c) through (f), the model's performance improves even further. In these settings, $\mathcal{L}1$ represents the L1 regularization loss, and CAM refers to Class Activation Maps (Zhou et al. 2016). Notably, setting (f), which corresponds to our method, achieves the highest performance with an overall accuracy of 89.93% and a mean accuracy of 83.87%. This significant enhancement underscores how Jensen-Shannon Divergence (JSD) amplifies the benefits of NAW by enforcing consistent distributional regularization. We also evaluate the framework under conditions with 30% noise and a class imbalance factor of 150. Replacing CE with NAW-CE leads to significant improvements—19.55% in mean accuracy under noise and 7.97% under class imbalance. These findings demonstrate that the proposed components of NLA significantly improve the model's robustness and reliability.

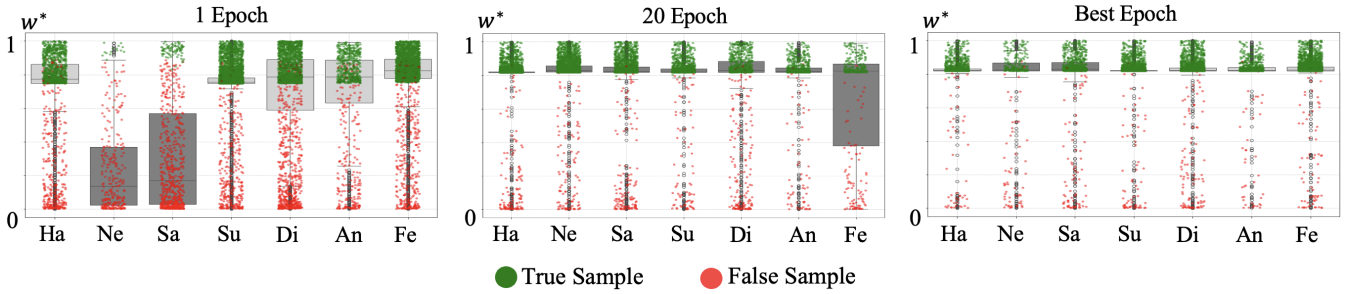

Figure 5: **Visualization of the training process of our method.** This figure demonstrates how our method enhances discriminative ability by adaptively assigning weights to each sample through NAW.

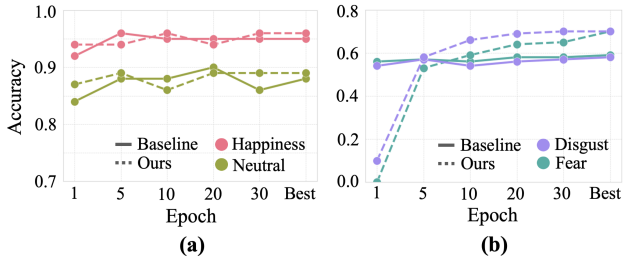

Figure 6: Comparison of test accuracy across epochs for majority classes (a) and minority classes (b) on RAF-DB.

Visualization Analysis

We visualize the training process of our method to analyze how the model develops class-specific discriminative ability. In Fig. 5, the y-axis represents the noise-aware adaptive weight, w^*, assigned to each sample, and the x-axis shows the different classes (Ne: Neutral, Ha: Happiness, Sa: Sadness, Su: Surprise, An: Anger, Di: Disgust, Fe: Fear). The gray boxes indicate the Inter-Quartile Range, illustrating the sample distribution for each class. In the first epoch, samples in majority classes receive higher weights, while those in minority classes receive less attention. However, as training progresses, our model gradually assigns higher weights to minority classes while maintaining the higher weights to the majority classes. Fig. 6 further supports this behavior, as (b) shows that our model initially exhibits low accuracy for minority classes but improves over time, whereas the baseline performance saturates. These results demonstrate that the proposed model effectively handles class imbalance by adaptively adjusting weights based on sample ambiguity.

Moreover, Fig. 7 compares the baseline model and our NLA in handling ambiguous samples across different facial expression classes. The top row shows ambiguous images from the RAF-DB dataset misclassified by the baseline model, with bar graphs below each image displaying predicted probabilities: green for GT and red for NN. The baseline model often assigns similar probabilities to GT and NN, causing misclassification. In contrast, the bottom charts show that our model significantly increases the GT probability, creating a larger margin over the NN and resulting in correct classifications. This figure clearly demonstrates

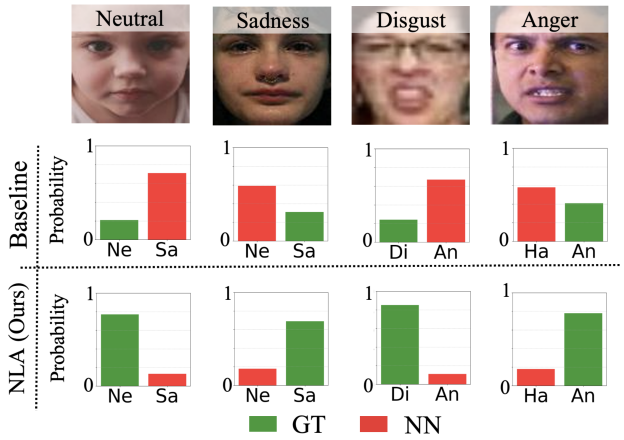

Figure 7: **Comparison of results for ambiguous samples between the baseline and our model.** The baseline model confuses the prediction probabilities between the GT and NN, whereas our model correctly predicts by improving the prediction probability for the GT with a large margin.

the effectiveness of NLA in accurately predicting ambiguous samples by enhancing the model's discriminative ability and reducing confusion with the NN.

Conclusion

In this paper, we propose a novel framework, Navigating Label Ambiguity (NLA), which addresses label ambiguity to mitigate both noise and class imbalance in facial expression recognition (FER). To the best of our knowledge, this is the first attempt to tackle both problems within a single framework. Our approach employs Noise-aware Adaptive Weighting (NAW) and consistency regularization to dynamically adjust weights based on sample ambiguity, enabling the model to focus on ambiguous samples while reducing the impact of noise. Extensive experiments demonstrate that NLA achieves superior overall and mean accuracy across multiple FER datasets, proving its robustness. Furthermore, while NLA is designed for FER, its ability to address label ambiguity in noisy or imbalanced data indicates promising applicability to a broader range of tasks, which we leave as an avenue for future work.

Acknowledgments

This research was supported by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea(NRF) funded by the Ministry of Science ICT and Korea Customs Service(2022M3I1A1095154), and supported by KIST Institutional Program (Project No. 2E33001).

References

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, 233–242. PMLR.

Barsoum, E.; Zhang, C.; Ferrer, C. C.; and Zhang, Z. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, 279–283.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019a. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019b. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, 117–124. Springer.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 87–102. Springer.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, 2304–2313. PMLR.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, H.; Wang, N.; Ding, X.; Yang, X.; and Gao, X. 2021. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30: 2016–2028.

Li, S.; Deng, W.; and Du, J. 2017. 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*, 2852–2861.

Li, Z.; and Arora, S. 2019. An exponential learning rate schedule for deep learning. *arXiv preprint arXiv:1910.07454*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Malach, E.; and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". *Advances in neural information processing systems*, 30.

Mao, J.; Xu, R.; Yin, X.; Chang, Y.; Nie, B.; and Huang, A. 2023. POSTER++: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186.

She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; and Mei, T. 2021. 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*, 6248–6257.

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020a. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6906.

Wang, K.; Peng, X.; Yang, J.; Meng, D.; and Qiao, Y. 2020b. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29: 4057–4069.

Wang, X.; Bo, L.; and Fuxin, L. 2019. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6971–6981.

Wang, Y.; Gan, W.; Yang, J.; Wu, W.; and Yan, J. 2019. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5017–5026.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13726–13735.

Wu, Z.; and Cui, J. 2023. LA-Net: Landmark-Aware Learning for Reliable Facial Expression Recognition under Label Noise. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20698–20707.

Xue, F.; Wang, Q.; and Guo, G. 2021. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3601–3610.

Xue, F.; Wang, Q.; Tan, Z.; Ma, Z.; and Guo, G. 2022. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International conference on machine learning*, 7164–7173. PMLR.

Zeng, D.; Lin, Z.; Yan, X.; Liu, Y.; Wang, F.; and Tang, B. 2022. Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20291–20300.

Zhang, Y.; Li, Y.; Liu, X.; Deng, W.; et al. 2024. Leave No Stone Unturned: Mine Extra Knowledge for Imbalanced Facial Expression Recognition. *Advances in Neural Information Processing Systems*, 36.

Zhang, Y.; Wang, C.; and Deng, W. 2021. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34: 17616–17627.

Zhang, Y.; Wang, C.; Ling, X.; and Deng, W. 2022a. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision*, 418–434. Springer.

Zhang, Z.; Sun, X.; Li, J.; and Wang, M. 2022b. MAN: Mining ambiguity and noise for facial expression recognition in the wild. *Pattern Recognition Letters*, 164: 23–29.

Zhao, Z.; Liu, Q.; and Zhou, F. 2021. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3510–3519.

Zheng, C.; Mendieta, M.; and Chen, C. 2023. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3146–3155.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. 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*, 9719–9728.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.