D^{3}FER: Dual Channel and Dual Branch Network for Robust Facial Expression Recognition under Dual Challenges

Supplementary Material

0.1. Pseudocode

Algorithm 1 outlines the pipeline of the proposed D^3FER. It begins by initializing the Query and Key branches with shared parameters and an empty memory queue. During training (lines 2–18), each epoch processes a mini-batch: it computes weakly and strongly augmented predictions from the Query branch and strong features from the Key branch, storing them in the queue (lines 5–8). In the warm-up phase (epoch $< E_{\text{warm}}$), the model is trained with standard cross-entropy loss and the Key branch is hard-copied from the Query (lines 9–13). After warm-up, the algorithm performs sample filtering and label correction using the queue (line 15), computes a multi-component loss (line 16), updates the Query branch via gradient descent (line 17), and applies momentum updates to the Key branch (line 18). The trained Key branch is then used for inference on test samples to predict emotion labels (lines 20–23).

Algorithm 1 The Algorithmic of D^3FER

Require: Query branch parameters $\boldsymbol{\theta}_{f^{\mathrm{Q}}}, \boldsymbol{\theta}_{g^{\mathrm{Q}}}$; Key branch parameters $\boldsymbol{\theta}_{f^{\mathrm{K}}}, \boldsymbol{\theta}_{g^{\mathrm{K}}}$; Training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$; Test set $\mathcal{D}_{\text{test}} = \{\mathbf{x}_i\}_{i=1}^{N_{\text{test}}}$; Memory queue $\mathcal{Q}$; Batch size B; Learning rate η; Momentum coefficient m; Number of emotion classes C; Total training epochs $E_{\max}$; Warm-up epochs E_{warm}.

1: Initialize $\boldsymbol{\theta}_{f^{\mathrm{Q}}}$, $\boldsymbol{\theta}_{g^{\mathrm{Q}}}$, $\mathcal{Q} = \emptyset$, $\boldsymbol{\theta}_{f^{\mathrm{K}}} = \boldsymbol{\theta}_{f^{\mathrm{Q}}}$, $\boldsymbol{\theta}_{g^{\mathrm{K}}} = \boldsymbol{\theta}_{g^{\mathrm{Q}}}$ randomly

// Training Phase

2: **for** epoch $= 1, 2, \ldots, E_{\max}$ **do**

3: Shuffle training data $\mathcal{D}$

4: Sample a mini-batch $\mathcal{D}_{\text{min}}$ from $\mathcal{D}$

5: Compute weak-augmented confidence on Query branch: $\{\mathbf{p}_i^{\mathrm{W,Q}}\}_{i\in\mathcal{D}_{\text{min}}}$ via Eq. (1)

6: Compute strong-augmented confidence on Query branch: $\{\mathbf{p}_i^{\mathrm{S,Q}}\}_{i\in\mathcal{D}_{\text{min}}}$ via Eq. (2)

7: Compute strong-augmented features on Key branch: $\{\mathbf{h}_i^{\mathrm{S,K}}\}_{i\in\mathcal{D}_{\text{min}}}$ via Eq. (3)

8: push$(\mathcal{Q}, \{\mathbf{p}_i^{\mathrm{S,Q}}, \mathbf{p}_i^{\mathrm{W,Q}}, \mathbf{h}_i^{\mathrm{S,K}}\}_{i\in\mathcal{D}_{\text{min}}})$

9: pop$(\mathcal{Q})$ // Remove the oldest batch

10: **if** epoch $< E_{\text{warm}}$ **then**

11: $\mathcal{L} = -\frac{1}{|\mathcal{D}_{\text{min}}|}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{\text{min}}}\left(\log \mathbf{p}_{i,y_i}^{\mathrm{W,Q}} + \log \mathbf{p}_{i,y_i}^{\mathrm{S,Q}}\right)$

12: Update $\boldsymbol{\theta}_{f^{\mathrm{Q}}} \leftarrow \boldsymbol{\theta}_{f^{\mathrm{Q}}} - \eta\nabla\mathcal{L}$, $\boldsymbol{\theta}_{g^{\mathrm{Q}}} \leftarrow \boldsymbol{\theta}_{g^{\mathrm{Q}}} - \eta\nabla\mathcal{L}$

13: $\boldsymbol{\theta}_{f^{\mathrm{K}}} \leftarrow \boldsymbol{\theta}_{f^{\mathrm{Q}}}$, $\boldsymbol{\theta}_{g^{\mathrm{K}}} \leftarrow \boldsymbol{\theta}_{g^{\mathrm{Q}}}$

14: **end if**

15: Perform sample filtering and label correction on $\mathcal{D}_{\text{min}}$ and queue via Eqs. (11)–(15)

16: Compute total loss $\mathcal{L}$ via Eqs. (18)–(21)

17: Update $\boldsymbol{\theta}_{f^{\mathrm{Q}}} \leftarrow \boldsymbol{\theta}_{f^{\mathrm{Q}}} - \eta\nabla\mathcal{L}$, $\boldsymbol{\theta}_{g^{\mathrm{Q}}} \leftarrow \boldsymbol{\theta}_{g^{\mathrm{Q}}} - \eta\nabla\mathcal{L}$

18: Momentum update Key branch: $\boldsymbol{\theta}_{f^{\mathrm{K}}} \leftarrow m\boldsymbol{\theta}_{f^{\mathrm{K}}} + (1-m)\boldsymbol{\theta}_{f^{\mathrm{Q}}}$, $\boldsymbol{\theta}_{g^{\mathrm{K}}} \leftarrow m\boldsymbol{\theta}_{g^{\mathrm{K}}} + (1-m)\boldsymbol{\theta}_{g^{\mathrm{Q}}}$ via Eqs. (5)–(6)

19: **end for**

Ensure: Trained Key branch parameters $\boldsymbol{\theta}_{f^{\mathrm{K}}}, \boldsymbol{\theta}_{g^{\mathrm{K}}}$

// Inference Phase

20: **for** $\mathbf{x}_i \in \mathcal{D}_{\text{test}}$ **do**

21: Compute classification confidence via Eq. (22)

22: Predict emotion label via Eq. (23)

23: **end for**

0.2. Implementation details

Following [3, 4], we generate weak and strong augmentations for each image in both original and synthetically noisy subsets of the three datasets. Weak augmentation uses random 4-pixel cropping and horizontal flipping (probability 0.5); strong augmentation adds RandAugment [2], randomly applying two transformations (e.g., contrast, rotation, translation, color inversion) on top of the weak augmentation.

Our method is implemented in PyTorch on Ubuntu 18.04, using a single NVIDIA GeForce RTX 3090 GPU with 64 GB RAM. For training, we set the total number of epochs to 100 for RAF-DB and FERPlus, with warm-up schedules of 10 and 5 epochs, respectively. On AffectNet, we apply oversampling to mitigate class imbalance and prevent overfitting, training for a maximum of 10 epochs with 1 warm-up epoch. The Adam optimizer is used with an initial learning rate of 0.0005, which is decayed exponentially by a factor of 0.98 per epoch.

0.3. dataset descriptions

FER datasets RAF-DB [6] is one of the most widely used in-the-wild FER datasets, comprising approximately 30,000 facial images collected from the Internet and meticulously annotated by 40 trained human annotators. The dataset consists of two subsets: basic and compound expressions. Following standard practice in in-the-wild FER research, we use only the basic subset, which contains 12,271 training and 3,068 test images labeled with seven basic expressions: surprise, fear, disgust, happiness, sadness, anger, and neutral.

FERPlus [1], also known as FER+, is an extended version of the FER2013 dataset [5] introduced in the ICML

2013 Representation Learning Challenge. All images in FERPlus were collected via Google Search and include 28,709 training, 3,589 validation, and 3,589 test grayscale images of size 48×48. Each image was annotated by ten annotators and assigned to one of eight expression categories. Compared to RAF-DB, FERPlus adds a "contempt" class.

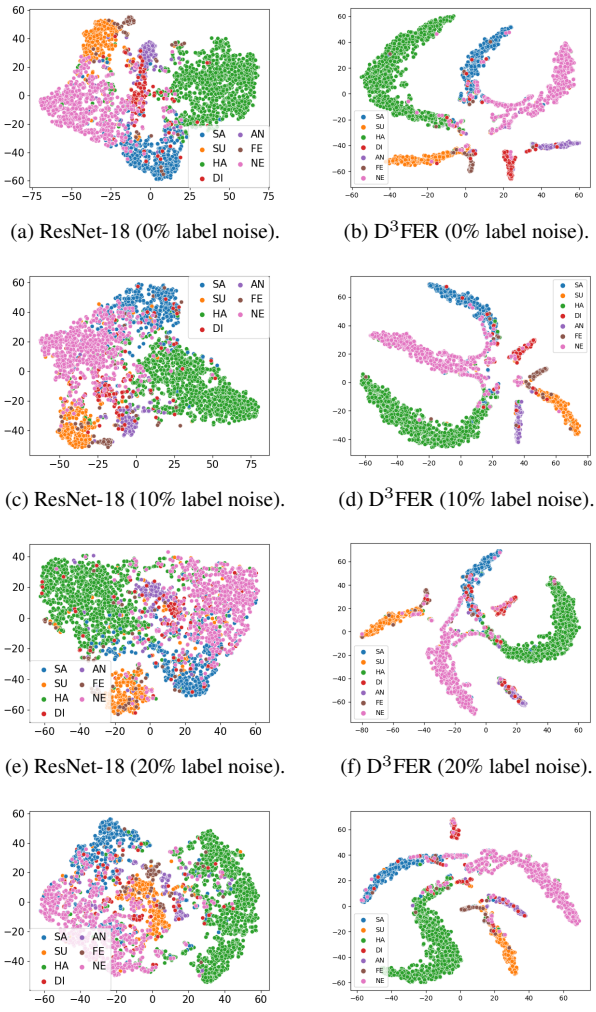
AffectNet [7] is currently the largest in-the-wild facial expression dataset, containing over 1 million facial images collected from the web, of which approximately 450,000 are manually annotated across 11 expression categories. It provides two benchmark subsets: AffectNet-7 (seven classes) and AffectNet-8 (eight classes, including "contempt"). Specifically, AffectNet-7 comprises 283,901 training and 3,500 test images, while AffectNet-8 includes 287,568 training and 4,000 test images.

Occlusion and Pose Variation Subsets To evaluate robustness against occlusion and pose variations, we test our method on multiple test subsets with occlusion or pose annotations (e.g., Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet, and Pose-AffectNet), manually curated following the protocol in [11]. The occlusion subsets cover six types: mask, glasses, bottom-half, top-half, left-side, and right-side occlusions. Pose subsets categorize images based on whether pitch or yaw angles exceed $30°$ or $45°$.

Synthetic Label Noise Datasets Following [10], we inject synthetic label noise into the training sets of RAF-DB, FERPlus, and AffectNet-7 by randomly selecting a fixed percentage of samples and assigning them random incorrect labels (e.g., at 30% noise level, 30% of training labels are corrupted). The test sets remain unchanged and noise-free, enabling evaluation of model robustness under realistic label noise conditions.

0.4. Feature Distribution Visualization

To further validate the effectiveness of D^{3}FER, we visualize feature embeddings of the RAF-DB test set using t-SNE [9], as shown in Figure 1. The visualization includes 3,068 images across seven expression classes, comparing the baseline ResNet-18 with our D^{3}FER. Each color denotes a category of facial expression. Under clean labels (Figure 1(a), ResNet-18 produces scattered features with significant inter-class overlap, indicating limited discriminability. In contrast, D^{3}FER (Figure 1(b) yields tightly clustered and well-separated features, owing to its dynamic queue-based contrastive learning module. As label noise increases, ResNet-18's embeddings become increasingly entangled, revealing high sensitivity to noisy supervision. Remarkably, D^{3}FER maintains clear inter-class boundaries and strong intra-class compactness even under substantial noise, demonstrating superior robustness.



(a) ResNet-18 (0% label noise). (b) D^{3}FER (0% label noise).

(c) ResNet-18 (10% label noise). (d) D^{3}FER (10% label noise).

(e) ResNet-18 (20% label noise). (f) D^{3}FER (20% label noise).

(g) ResNet-18 (30% label noise). (h) D^{3}FER (30% label noise).

Figure 1. Feature distributions of ResNet-18 and D^{3}FER on RAF-DB under varying levels of label noise.

0.5. Attention Visualization

To investigate the differences in facial region attention between the Query and Key branches, we visualize their attention-weighted activation maps using Grad-CAM [8] on RAF-DB, as shown in Figure 2. As observed, the Key branch demonstrates superior attention characteristics: its activation is spatially balanced and covers a broader range of facial regions, capturing both discriminative local cues (e.g., eyes, mouth) and holistic structures (e.g., cheeks, forehead), reflecting stronger global contextual awareness. In contrast, the Query branch focuses narrowly on a few salient regions, often overlooking subtle yet informative areas, which limits its ability to handle ambiguous or complex expressions; its attention maps also exhibit higher variability across samples, indicating reduced stability under

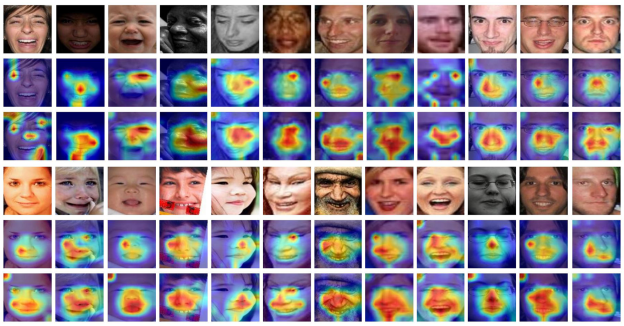


Figure 2. Attention maps of the Query and Key branches. The first row displays the original input images, while the second and third rows show the attention maps generated by the Query and Key branches, respectively. Darker regions indicate higher attention weights.

occlusion or noise. This enhanced robustness of the Key branch stems from its momentum-based parameter update, which smooths optimization trajectories and mitigates attention bias from noisy or outlier samples. Consequently, it learns to attend consistently to semantically meaningful facial regions, yielding more discriminative and robust features for facial expression recognition—particularly beneficial for reliable inference in real-world conditions.

1. Analysis of FER Confusion matrices

We visualize the confusion matrices of D^{3}FER across different datasets in Figure 3, where the vertical axis represents the true classes, and the horizontal axis indicates the predicted classes by the model. These matrices not only reveal the recognition accuracy for each emotion category but also highlight the most commonly misclassified emotions, thereby indicating directions for model enhancement. Specifically, on the RAF-DB dataset, while the model excels at recognizing "happiness" and "surprise," the confusion between "anger" and "disgust" suggests significant feature overlap between these two emotions. Similarly, in the FERPlus dataset, the misclassification between "sadness" and "fear" points to their inherent similarities, suggesting potential improvements through the incorporation of contextual information or more detailed annotation strategies. Furthermore, results from AffectNet-7 and AffectNet-8 indicate that the model faces greater challenges with nuanced emotional distinctions, especially with categories such as "neutral," "happy," and "surprise." The higher confusion rates among these categories may be attributed to the intrinsic difficulty of annotating subtle expressions and the distribution of samples within these datasets, highlighting areas for further investigation.

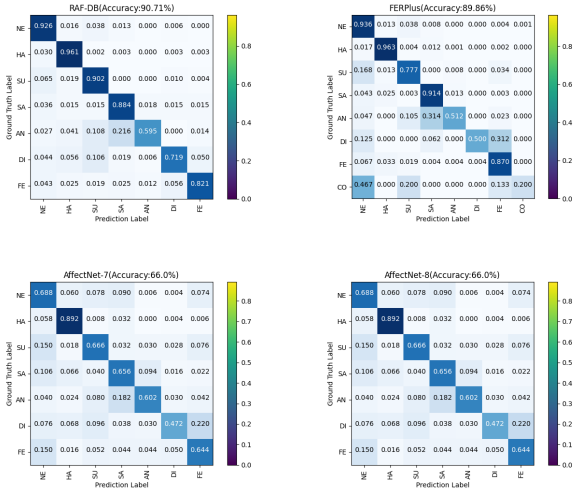


Figure 3. D^{3}FER confusion matrices on RAF-DB, FERPlus, and AffectNet.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *In Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 1

[2] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 1

[3] Darshan Gera and S Balasubramanian. Noisy annotations robust consensual collaborative affect expression recognition. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3585–3592, 2021. 1

[4] Darshan Gera, Bobbili Veerendra Raj Kumar, Naveen Siva Kumar Badveeti, and S Balasubramanian. Dynamic adaptive threshold based learning for noisy annotations robust facial expression recognition. *Multimedia Tools and Applications*, 83(16):49537–49566, 2024. 1

[5] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *In Proceedings of the 20th International Conference on Neural Information Processing*, pages 117–124. Springer, 2013. 1

[6] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017. 1

[7] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence,

and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 2

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *In Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2

[9] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11):2579–2605, 2008. 2

[10] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 2

[11] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2