Region-Aware Instance Consistency Learning for Micro-Expression Recognition

Supplementary Material

6. More Experimental Details

6.1. Multiple Instances Representation

For MIR with middle sampling, we divide the ME sequence into three segments of equal length: the initial segment, the middle segment, and the end segment. The middle segment corresponds to the sampling window. Therefore, the number of sampled frames N (i.e., the length of the middle segment) can vary across different ME sequences. If the length of a ME sequence is T, then the number N can be obtained by:

$$N = \frac{T}{3}. \tag{13}$$

To simplify the construction of instance sets, we set a unified value of N for all sequences based on the frame rate prior of datasets. Specifically, the duration of a ME sequence is typically less than 0.5 seconds. Denote the frame rate as f, the number N should satisfy the Eq. (14):

$$N \leq \frac{1}{3} \cdot 0.5 \cdot f = \frac{f}{6}. \tag{14}$$

Under the CDE setting, the lowest frame rate is 100 fps from the SMIC-HS dataset. Therefore, setting N to 16 is a reasonable choice, as it ensures motion variety while satisfying the requirement of the Eq. (14). For the CAS(ME)^3 dataset, although its frame rate is 30 fps, sequence lengths vary from just a few frames to over 100 frames, which do not strictly adhere to the constraint of less than 0.5 seconds. For consistency, we also set N to 16 on the CAS(ME)^3 dataset. For all datasets, we apply full sampling (i.e., all frames are used) on sequences containing fewer than 16 frames.

Table 4. Details of three datasets under the CDE setting.

Category	CASME II	SAMM	SMIC-HS	Full
Negative	88	92	70	250
Positive	32	26	51	109
Surprise	25	15	43	83
Total	145	133	164	442

6.2. Leave-one-subject-out Cross-validation

For CASME II, SAMM and SMIC-HS, the final distribution with a total of 442 samples under the CDE setting is shown in Tab. 4. These 442 samples belong to 68 subjects. CASME II, SAMM and SMIC-HS contain 24, 28 and 16 subjects, respectively. Therefore, the LOSO cross-validation should be repeated 68 times.

Table 5. Details of 3-class, 4-class and 7-class evaluation on the CAS(ME)^3 dataset.

3-class		4-class		7-class	
Negative	438	Negative	438	Anger Fear Disgust Sadness	55 82 245 56
Positive	49	Positive	49	Happiness	49
Surprise	183	Surprise	183	Surprise	183
\	\	Others	131	Others	131
All	670	All	801	All	801

For the CAS(ME)^3 dataset, the details of 3-class, 4-class and 7-class evaluation are shown in Tab. 5. For the 3-class evaluation, the 670 samples belong to 87 subjects. For the 4-class and 7-class evaluation, the 801 samples belong to 93 subjects. Therefore, the LOSO cross-validation should be repeated 87, 93 and 93 times for 3-class, 4-class and 7-class evaluation, respectively.

7. Hyper-Parameter Analysis

7.1. Influence of Different Numbers of Instances

We conduct experiments on different numbers of motion instances. Considering that MEs are usually recorded below or equal to 200 fps, and ME sequences generally last less than 0.5 seconds, we selected 4, 8, 12, 16, 20, and 32 as different number of instances for a sequence. When the number of instances is fewer than 16, the middle sampling window is compressed inward. When it exceeds 16, the window expands outward toward both ends of the sequence. If the required number of instances exceeds the sequence length, full sampling is performed.

As shown in Fig. 8, the model achieves the best performance when the number of instances is 16. When the number of instances is less than 16, the motion information becomes insufficiently varied, resulting in less robust features. In contrast, when the number of instances is larger than 16, the model exhibits a slight decline in overall performance. Although increased instances provide greater motion variety, they simultaneously introduce additional noise that may obscure the ability of the model to perceive subtle micro-movements. Therefore, it is not appropriate to increase the number of instances excessively, as it necessitates a careful trade-off between motion variety and noise introduction.

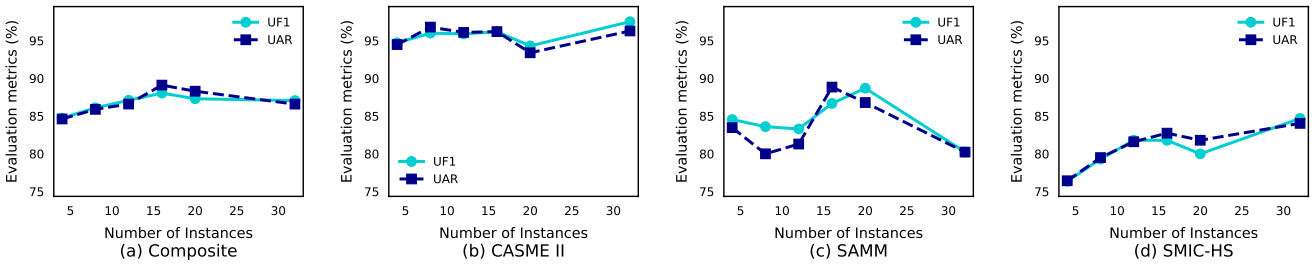

Figure 8. Hyper-parameter analysis on the number of instances under the CDE setting.

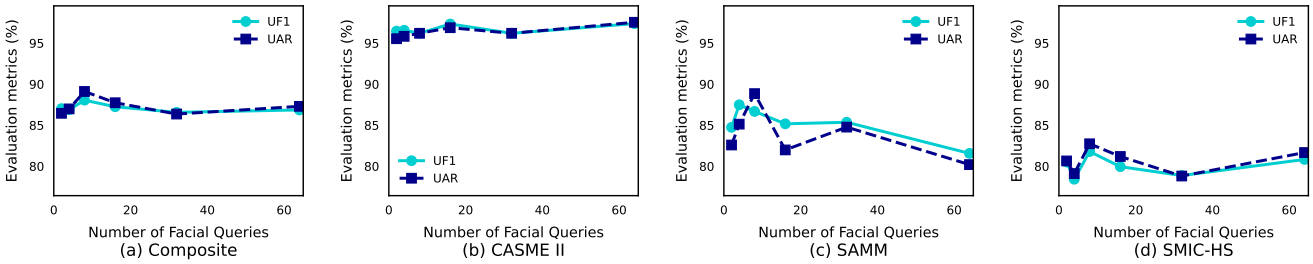

Figure 9. Hyper-parameter analysis on the number of facial queries under the CDE setting.

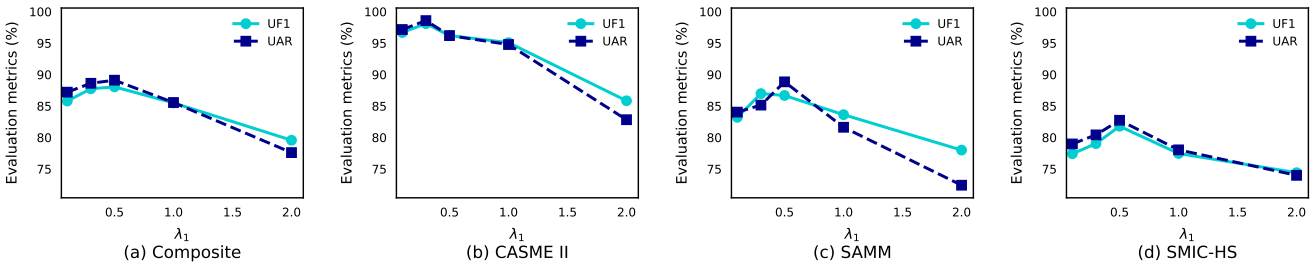

Figure 10. Hyper-parameter analysis on λ_1 under the CDE setting.

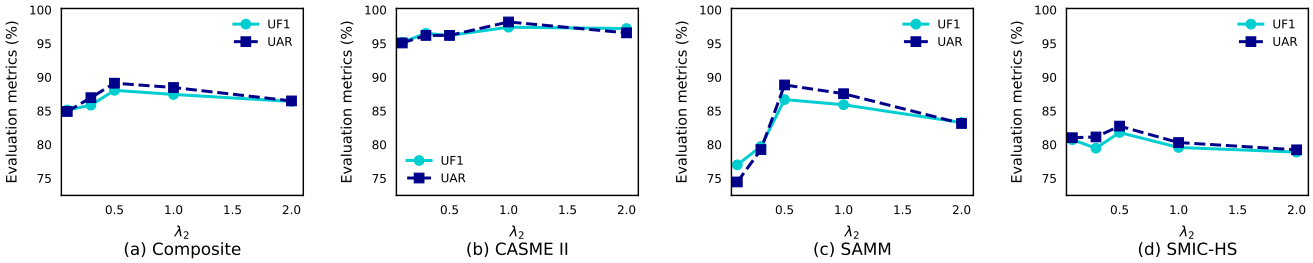

Figure 11. Hyper-parameter analysis on λ_2 under the CDE setting.

7.2. Influence of Different Numbers of Facial Queries

MRD is enforced to discover distinct facial regions by different facial queries. Fig. 9 shows the effect of different numbers of facial queries. The optimal number of facial queries is 8. If the number of facial queries is smaller, the model is at risk of failing to identify complete and meaningful facial activation regions. Conversely, if the number of facial queries is excessively large, the model may overfit to non-discriminative motion noise. This would impair its discriminative capability for critical ME features.

7.3. Influence of Different Loss Weights

The overall objective contains three loss weights $\lambda_1, \lambda_2, \lambda_3$ for the balance of classification, IRC, and MRD, respec-

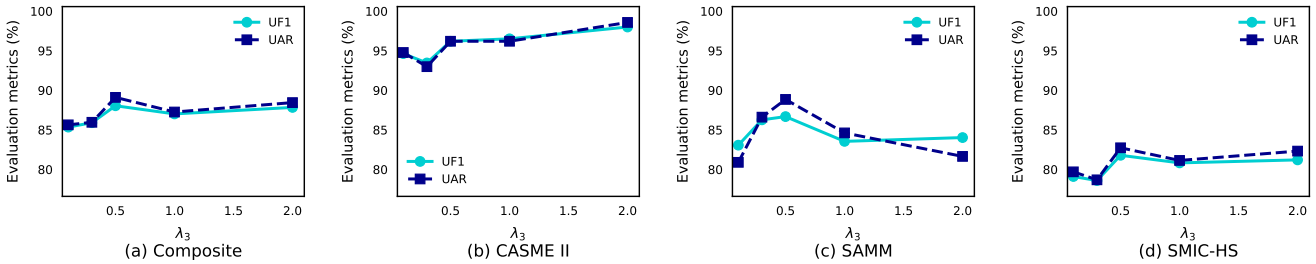

Figure 12. Hyper-parameter analysis on λ_3 under the CDE setting.

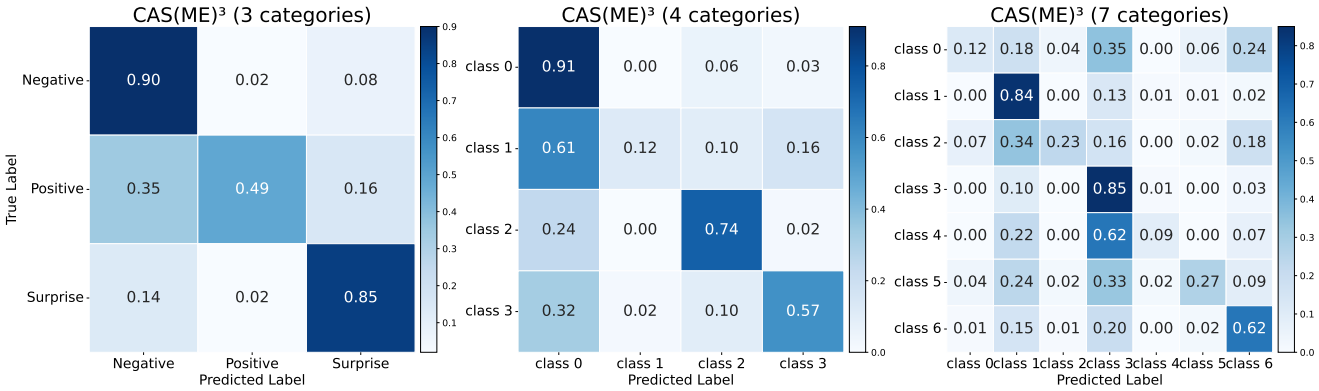

Figure 13. The confusion matrices on the CAS(ME)^3 dataset (From left to right: 3-class evaluation, 4-class evaluation, 7-class evaluation). For the 4-class evaluation, labels {0, 1, 2, 3} correspond to {Negative, Positive, Surprise, Others}. For the 7-class evaluation, labels {0, 1, 2, 3, 4, 5, 6} correspond to {Happiness, Surprise, Sadness, Disgust, Fear, Anger, Others}.

tively. To show the influence of these three loss weights on performance, we evaluated them from 0.1 to 2 under the CDE setting.

Fig. 10 shows that the optimal λ_1 is around 0.5. When λ_1 approaches zero, the guidance from ground-truth labels reduces, resulting in a decrease in performance. As λ_1 increases, classification takes precedence, compromising the effectiveness of both IRC and MRD modules and leading to degraded performance.

Fig. 11 demonstrates that the optimal value of λ_2 is approximately 0.5. The performance first increases along λ_2 and then decreases. If λ_2 is too small, the model will overfit to certain activation regions. If λ_2 is set larger than 0.5, the performance decreases slightly since the attention consistency loss outweighs the classification loss.

Fig. 12 illustrates that the model achieves the best performance when λ_3 is equal to 0.5. When $\lambda_3 < 0.5$, the MRD module does not sufficiently constrain the model to identify more meaningful regions. When $\lambda_3 > 0.5$, it disrupts the learning process of both classification and IRC.

8. Visualization

8.1. Visualization of Confusion Matrices

Fig. 13 shows the confusion matrices of Ra-ICL on the CAS(ME)^3 dataset. For the confusion matrix with 3 categories, we find that the model performs better on the negative and surprise classes but worse on the positive class. In the confusion matrix with four categories, we also observed that the positive class performed poorly. This may stem from the fewest samples of the positive class in the dataset, leading to biased model predictions. Meanwhile, the result shown in the confusion matrix with seven categories demonstrates that the model performs better in the surprise, disgust, and others classes. The three classes correspond to the three classes with the highest number of samples in the CAS(ME)^3 dataset.

8.2. Visualization of Attention Heatmaps

We conducted visualizations of attention heatmaps on a ME sequence using Grad-CAM [31]. Fig. 14 shows a ME sequence with four facial activation regions: A, B, C, and D. The third row illustrates the attention regions of the model when equipped with the IRC module but without the MRD module. It is evident that the model consistently focuses

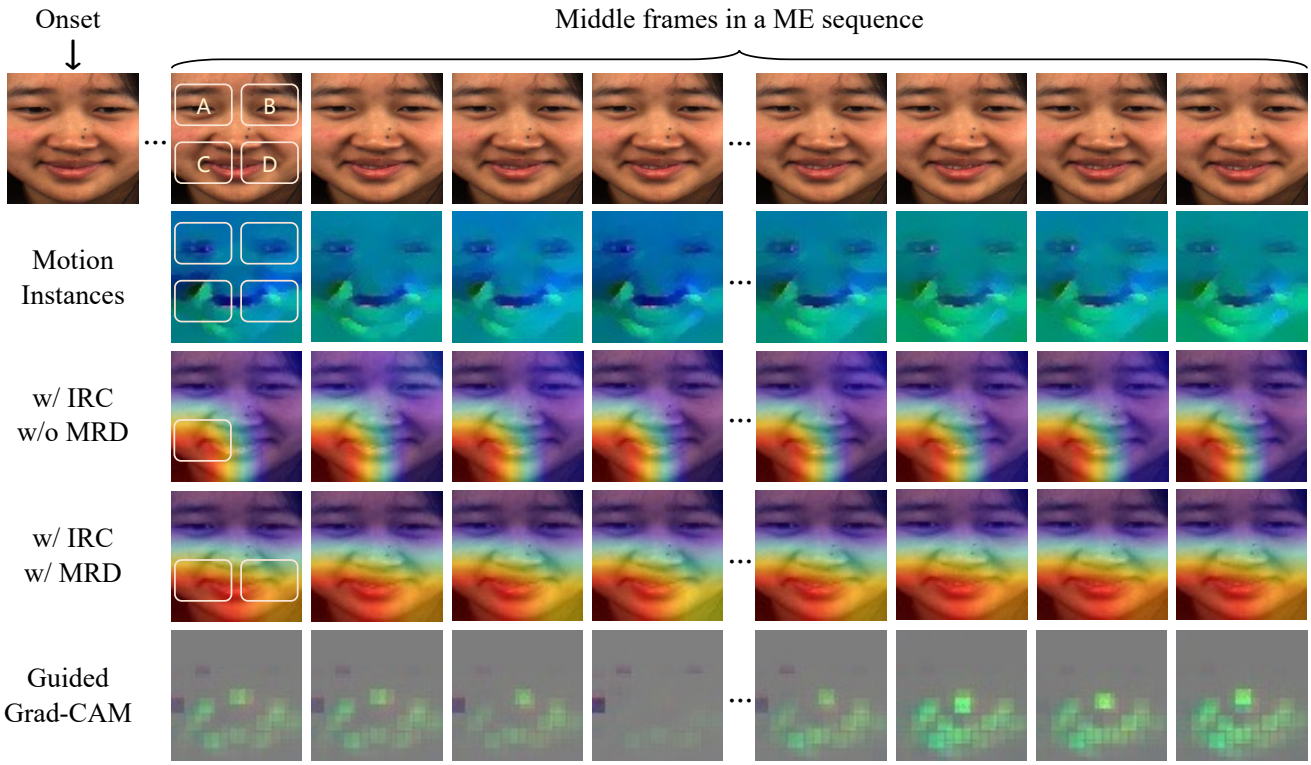

Figure 14. The visualization of attention heatmaps. The first row: RGB frames of a ME sequence. The second row: Motion instances. The third row: Attention heatmaps of the model with (w/) the IRC module but without (w/o) the MRD module. The fourth row: Attention heatmaps of the model with both IRC and MRD modules. The fifth row: Guided Grad-CAM of each motion instance.

on region C across all motion instances, indicating that IRC effectively enforces visual attention consistency for all instances within an instance set. However, without MRD, the attention region fails to cover region D. The fourth row represents the attention regions when both IRC and MRD modules are implemented. We observe that the model not only maintains consistent attention regions across all samples in the set but also expands its focus to the region D. Note that A and B indicate eye movements, which should be considered noise in this sample. In both the third and fourth rows, the attention of the model avoids regions A and B. This demonstrates the capability of the model to effectively distinguish between noise and subtle motions that contribute to classification. The fifth row presents the Guided Grad-CAM [31, 34] for each motion instance. It is illustrated that pixels in regions C and D are critical for classification.