RMFER: Semi-supervised Contrastive Learning for Facial Expression Recognition with Reaction Mashup Video

Yunseong Cho[1,2] Chanwoo Kim[1] Hoseong Cho[1] Yunhoe Ku[1] Eunseo Kim[1]
Muhammadjon Boboev[1] Joonseok Lee[3,4] Seungryul Baek[1]

[1]UNIST [2]SNOW Corp. [3]Seoul National University

Contribution of the main paper. In this paper, we proposed the semi-supervised framework for the facial expression recognition(FER) task that exploits both original FER benchmarks and unlabeled datasets (i.e., proposed RMset). Semi-supervised learning is proposed to tackle the data issue in the facial expression recognition problem. It is hard to collect quality FER datasets due to the mislabeling caused by the annotator's subjectivity and the subtlety and complexity of facial expressions. To bypass the difficulty of the data collection, we tried to verify that the raw video data having rich facial expressions while not having supervised labels annotated by humans could help improve the FER accuracy. What we have done could be summarized as follows:

(1) We trained our framework, RMFER, exploiting the conventional cross-entropy loss using the original FER benchmarks.

(2) Upon this, we collected unlabeled video data called reaction mashup (RM) video, processed it, and eventually made the reaction mashup dataset (i.e., RMset), which has strong potential for improving the FER accuracy. The RM video contains multiple persons inside who are watching the same film (we call this a trigger film). In these videos, the persons in the same frame might share the same feeling as they are watching the same scene of the film. Also, the same person's snapshots in far different frames might exhibit dissimilar expressions. As the video sequence, RMset includes rich and continuous information about the natural facial expressions. Using the mentioned prior, we could effectively incorporate RMset with contrastive learning by defining positive and negative sets based on the similarity and dissimilarity assumptions.

(3) The prior inherent in the RMset is not always true: persons in the same frame could have different expressions, and the same person's snapshots in far different frames could have similar expressions. This can cause a negative impact on contrastive learning and could eventually spoil the FER accuracy. To relieve the issue, we proposed to learn the pairwise similarity between samples using inter-sample attention learning (IAL), and we improved the positive/negative sets (initially made based on RM prior) for contrastive learning, using the pairwise attention between samples in attention-based contrastive learning (ACL).

Via the proposed RMset and RMFER framework, we demonstrated that state-of-the-art accuracy can be obtained on several challenging FER benchmarks.

Content of the supplemental. In this supplemental, we offer implementation details, insights into the RMset, more qualitative results, more ablative studies, additional details of the training strategy, a discussion of the FERPlus dataset, a more in-depth qualitative analysis using the MDS plot. We hope that the content of the supplemental could relieve inquiries arising from the main paper.

1. Implementation details

We use the EfficientNet-b2 [12] as a CNN backbone and sharpness-aware minimization (SAM) [2] as our optimizer, following [11]. We additionally conduct experiments utilizing ResNet50 [3], Adam optimizer [5], and 224 × 224 image resolution to ensure a fair comparison and report the results in the Sec 3.1. Grayscale, horizontal flip, and color jitter were used for data augmentation. We conduct training in batches, considering the image set $\mathbf{x}$ mentioned in Sec. 3 of the main paper as one batch. The batch size of the benchmark dataset was set to 32. The number of the first few epochs, which use only classification and attention modules, is denoted $epoch_{\text{pre}}$. How $epoch_{\text{pre}}$ was determined can be found in Table 8 and Sec 4. In the RMset, we use 50 positives and negatives per anchor, respectively, and sample 10% of them using attention to the anchor. All facial data utilized in the framework underwent facial alignment before entry.

2. Reaction Mashup dataset (RMset)

2.1. RMset license and disclosure

We gathered the RM videos from YouTube to establish the RM dataset (RMset). YouTube explains its provisions for fair use, such as for research purposes. The collection of videos was conducted within the framework of a Creative Commons license.

Upon acceptance of the paper, we plan to publicly share the YouTube links and the code used to collect the data, along with the normalized raw data that can be used to replicate our results. Before publishing, we will normalize the pixel values of the raw data the same way we did for training our model to ensure the privacy of the individuals depicted in the videos. In addition, the RMset can be easily expanded further by utilizing the publicly available source videos and provided code. This scalability is a notable advantage over traditionally labeled datasets, which are typically more challenging to expand.

2.2. Keywords used for collection

The RM dataset (RMset) was compiled through a two-step process. Initial searches were conducted using the keyword 'reaction mashup video,' followed by specific keywords indicated in Table 1 to identify videos corresponding to various expressions. The facial expressions in the RM videos are not clearly discretized, but we can make a rough distinction based on the trigger film and the approximate human reactions. Furthermore, even if the facial expressions are not all the same in one video, the basic assumptions of our RM videos (Fig. 1 in the main paper) are satisfied. This situation does not pose a significant challenge due to our utilization of inter-sample attention and attention-based contrastive learning during the sampling process.

Expression	Search Keywords
Happiness, Surprise	K-pop, fan cam, action movie, hero movie, amazing football plays
Sadness	missing father, try not to cry, 911 call, touching
Fear, Surprise	horror movies trailer, try not to get scared, do not watch at night, creepiest
Disgust, Contempt	try not to look away, anime death, freakshow, racist, jocker, animal abuse
Anger	try not to get mad, racism, discrimination, 911

Table 1. Keywords used in video search

2.3. Statistics for RMset

We collected a total of 216 videos for all expressions. Additional statistics on the collected RMset are presented in Table 2. Several facial expressions may appear together in the same video; therefore, we categorized the expressions based on the keywords used to search for them. The 'video #,' 'face #,' 'frame #,' and 'image #' denote the number of videos collected related to the corresponding expressions, the number of different identities within the videos, the number of total frames for videos and total facial images obtained, respectively. Additionally, the RMset collects more negative facial expressions (disgust, anger, contempt, etc.), which are lacking in the existing FER dataset. This can be seen by comparing Table 2 and Table 10. As a result, our model significantly improves accuracy on negative facial expressions, such as 'Disgust' and 'Contempt.' This will be discussed in more detail in Sec. 5.

Expression	video #	face #	frame #	image #
Happiness, Surprise	48	702	$562,261$	$7,763,728$
Sadness	30	657	$418,141$	$6,334,245$
Fear, Surprise	64	985	561164	$7,367,635$
Disgust, Contempt	60	938	$1,392,316$	$21,492,401$
Anger	14	196	$206,817$	$2,612,570$
Total	216	$3,478$	$3,140,699$	$45,570,580$

Table 2. Statistics of the proposed RMset

3. Additional results

3.1. Quantitative comparison in the same implementation detail

In the main paper, we report the results of experiments with the SAM optimizer [2] with 260×260 images, using EfficientNet-b2 [12] as a backbone. To ensure fairness to other methods, in this section, we include the results of experiments using the Adam optimizer [5], 224×224 images, and ResNet50 [3] as the backbone. ResNet18 was not included in the comparison due to its small model size, which led to a collapse problem during contrastive learning [15]. As a result, we replaced the backbone of EAC [14] and SOFT [9] with ResNet50.

Method	AffectNet-7
EAC [14]	65.83
SOFT [9]	65.93
Ours w/o ACL, IAL	65.69
Ours w/o ACL	66.06
Ours	**66.39**

Table 3. Quantitative comparisons with state-of-the-art methods in identical implementation details (i.e., ResNet50, Adam optimizer, 224×224 image).

Our methodology achieves superior performance compared to the other two models, even when using identical implementation details such as ResNet50, Adam optimizer, and 224×224 images. To implement EAC [14] and SOFT [9], we followed the Github repositories provided in their respective papers and only made modifications to the backbone.

3.2. Comparison of using RMset with an existing dataset or data augmentation.

We conducted several additional experiments to demonstrate the effectiveness of the RMset. Table 4 shows the results of the experiments. First, we compared the baseline with the RMset trained by increasing the data with mixup [13] data augmentation. 'Ours w/o ACL, IAL' is the baseline without our additional modules, and 'Ours w/o ACL, IAL + DA (mixup)' is the baseline with mixup applied. Using mixup, we see a slight increase in performance over the baseline but less than our entire model.

Furthermore, we conducted attempts at attention-based contrastive learning (ACL) using pre-existing datasets, excluding the RMset: 'Ours (AffectNet)' uses the AffectNet same as the benchmark dataset to ACL, and 'Ours (CelebA)' applies CelebA [8] to ACL. Since the prior used in the RMset is unavailable for these two datasets, we used $\mathbf{P}_{\text{imp}}$ and $\mathbf{N}_{\text{imp}}$ for ACL by sampling only by cosine similarity from a random set of anchors, positives and negatives.

The ACL with AffectNet had the effect of reinforcing the inter-sample attention learned by IAL on the same data, but it did not help to improve performance and caused a drop in performance. Conducting ACL on CelebA did not lead to a decline in performance but it did not yield a significant improvement. This is because the CelebA is not a FER dataset. Hence, it offers a limited range of facial expressions. Additionally, attention alone does not provide optimal outcomes for contrastive learning.

On the other hand, as our RMset is a dataset created from reaction videos, we can train the model with a wide variety of subtle facial expressions using this dataset. Self-supervised learning is possible because of its assumption and ACL, even for identities not learned in the benchmark dataset. Finally, the combination of prior on our RMset and attention-based contrastive learning allows for proper sampling, which is an effective synergy.

Methods	AffectNet-7
Ours w/o ACL, IAL	66.13
Ours w/o ACL, IAL + DA (mixup)	66.3
Ours w/o ACL	66.33
Ours (AffectNet)	66.28
Ours (CelebA)	66.33
Ours	**66.85**

Table 4. Quantitative comparison of using the RMset with an existing dataset or data augmentation. 'DA' denotes data augmentation.

3.3. Semi-supervised methods with the RMset

Typically, other semi-supervised methods split the benchmark dataset into labeled and unlabeled data portions. In contrast, our model utilizes newly created unlabeled data, which may introduce unfairness due to differences in the amount of data. To address this, we applied the RMset to AdaCM [7], a semi-supervised method that can use additional unlabeled data, and reported the experimental results in Table 5. The experiment utilized $4,000$ annotated samples from RAF-DB.

Table 5 indicates that AdaCM [7] performs better without utilizing the RMset compared to when they are used, and our model outperforms AdaCM [7] regardless of the utilization of the RMset. This is due to AdaCM [7] not fully exploiting the potential of the RMset. In contrast, our semi-supervised contrastive framework is designed to effectively leverage the RMset, as evidenced by the consistent performance improvement of the ACL in Table 2 of the main paper. Therefore, simply increasing the amount of unlabeled data may not always lead to performance improvement unless the method is specifically designed to leverage the data.

Method	RAF-DB (%)
AdaCM w/o RMset	84.4
AdaCM*	82.27
Ours*	**87.13**

Table 5. Quantitative comparisons with AdaCM were performed using the RMset. The asterisk (*) indicates that the RMset was used.

3.4. Time efficiency

Our method, which samples useful data from an unlabeled dataset, differs from standard approaches in that it incorporates an unlabeled dataset as well as the benchmark dataset for training. It may raise concerns about increased training time. However, in experiments conducted under the same conditions, we found that our method takes 28 minutes per epoch on AffectNet, while the EAC [14] takes 27 minutes on the same dataset. The difference in training time is not statistically significant.

3.5. Hyper-parameter analysis

In this section, we present an analysis of our model's performance on AffectNet [10] with respect to the hyper-parameters λ_1, λ_2, and τ.

Initially, we identified the learning rate through 'Our w/o ACL, IAL' and fixed it. Then, we performed a grid search on IAL by increasing λ_1, which is the weight of L_{IAL}, from 0.1 to 1.2 in increments of 0.1 and let the scale τ have four values: 0.1, 0.25, 0.5, and 1.0.

In IAL, as τ increases, the cosine similarity between samples tends to be similar for the same facial expression and different for distinct facial expressions. Table 6 provides a difference between the average cosine similarity among samples with the same labels and those with different labels. A larger difference between these values signifies an improved separation of similarity across different labels. However, it is important to note that this difference does not directly correlate with accuracy. Table 7 demonstrates that a larger τ, which causes a larger difference of cosine similarity, does not always result in a higher accuracy. (The values of λ_1 in Table 7 are near the best accuracy for each value of τ.) Setting τ to 0.25 leads to a slight improvement in overall accuracy. The inter-class separation between labels is not too robust, reflecting the subtle and complex nature of facial expressions, making them challenging to annotate. Therefore, a value of $\tau = 0.25$, which is neither too large nor too small, was determined to be the optimal choice.

Subsequently, we fix the values of τ and λ_1 in order to find the optimal value of λ_2, the weight of L_{ACL}, in ACL. Table 7 shows that when τ is set to 0.25, the change in performance while changing the value of λ_1 is not significant in IAL. Consequently, we established the optimal values in ACL as $\tau = 0.25$ and $\lambda_1 = 1.0$. Subsequently, we systematically adjusted λ_2 across a range of tenfold increments, spanning from 0.001 to 1.0 to find the optimal value of λ_2. After several cross-validations, we found an optimal value of $\lambda_2 = 1.0$.

We conducted additional experiments to determine the interaction between λ_2 and τ. Both λ_2 and τ tend to maximize intra-class similarity and inter-class separation as they get larger. From Table 7, we can observe that small values of λ_2 do not improve performance, indicating that a certain level of contrastive learning is necessary for good classification performance. Furthermore, when we fixed $\lambda_2 = 1.0$ and varied the value of τ, the performance was still poor at the extreme values of 0.1 and 1.0 but good at the middle values of 0.25 and 0.5.

The information above demonstrates that the hyperparameters can differ based on levels of intra-class similarity and inter-class separation within the dataset. It became evident that these hyperparameters are influenced by the dataset's specific characteristics. Consequently, we followed a similar hyperparameter tuning approach for both the RAF-DB and FERPlus datasets. Hyper-parameters are summarized in Table 8.

τ	Dfference of cosine similarity
0.1	0.3083
0.25	0.4413
0.5	0.5133
1.0	0.5589

Table 6. The difference between the cosine similarity of samples with the same and different labels in AffectNet-7.

IAL parameters		Accuracy(%)	ACL parameters			Accuracy(%)
λ_1	τ		λ_1	λ_2	τ	
0	0.1	0	1.0	0.001	0.1	66.13
0		0			0.25	66.22
0		0			0.5	66.19
0		0			1.0	66.02
0.9	0.25	66.3	1.0	0.01	0.1	66.1
1.0		**66.33**			0.25	66.22
1.1		66.22			0.5	66.13
1.2		66.28			1.0	66.08
0.2	0.5	66.25	1.0	0.1	0.1	66.05
0.3		66.3			0.25	66.16
0.4		66.28			0.5	66.1
0.5		66.16			1.0	66.22
0.4	1.0	66.25	1.0	1.0	0.1	66.25
0.5		66.3			0.25	**66.85**
0.6		66.25			0.5	66.7
0.7		66.22			1.0	66.53

Table 7. Hyperparameter analysis of λ_1, λ_2, and τ from an accuracy perspective in AffectNet-7

	AffectNet		RAF-DB	FERPlus	
	7 emo	8 emo		overall	average
$epoch_{\text{pre}}$	3		10	20	
$epoch_{\text{total}}$	30		500	100	
lr	2×10^{-6}		1×10^{-4}	1×10^{-4}	1×10^{-5}
τ	0.25				
λ_1	1.0	0.4	0.7	1.0	
λ_2	1.0	0.001	0.002	0.001	0.1
γ	0.1				

Table 8. Hyper parameters in the RMFER

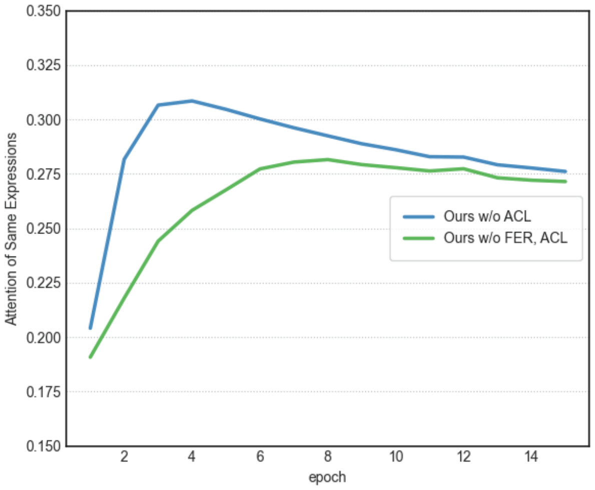


Figure 1. Average inter-sample attention value between samples with the same label over epochs.

4. Training strategy in IAL and ACL

Why the classification module is trained together in IAL pre-training. In the IAL pre-training phase, ensuring that inter-sample attention is being adequately trained is of utmost importance. We introduce a novel measure for evaluating and quantifying inter-sample attention. This measure represents the mean inter-sample attention value among samples sharing the same label within the test set. Fig. 1 shows it graphed over epochs. 'Ours w/o ACL' is IAL pre-training with classification and attention modules, and 'Ours w/o ACL, FER' uses only an attention module in IAL. As shown in Fig. 1, using both modules results in higher inter-sample attention than using only the attention module. Also, as shown in Table 9, better performance is guaranteed when the classification module is trained together.

When to add and train the contrastive module. The contrastive module of ACL reinforces the inter-sample attention learned by IAL, which maximizes intra-class similarity and inter-class separation, as shown in Fig. 4. Therefore, we add the contrastive module and RMset to perform ACL together starting from epoch 3, when inter-sample attention is moderately learned (See Fig. 1.) We do not start at four epochs when inter-sample attention is maximized because excessive IAL can break the classification module. Therefore, we start ACL at three epochs when inter-sample attention converges. As we saw in hyper-parameter tuning, different datasets have different degrees of intra-class similarity and inter-class separation. Therefore, the epoch at which attention converges is different for each dataset.

How long to train the model. Since the RMset is a huge dataset, the performance improvement takes time after starting attention-based contrastive learning. In AffectNet-7, 'Ours w/o ACL' converges after 15 epochs, however in the case of 'Ours', the performance improvement on the test set continues until 30 epochs. The total training epoch for each dataset is reported as $epoch_{\text{total}}$ in Table 8. Since RAF-DB is a smaller dataset compared to AffectNet-7, more epochs are required to fully train the RMset. FERPlus tends to converge faster than the other datasets, and it required fewer training epochs than RAF-DB.

FER using SimCLR with the RMset. We reproduced 'SimCLR' [1], by first performing the contrastive learning on the RMset and performing the finetuning on the benchmark dataset. The results (AffectNet-7) are shown in Table 9. The performance is 66.39%, and this is better than 'Our w/o ACL, IAL' that is not using the RMset, however it is worse than 'Ours'. From the results, we can extract two lessons: (1) RMset is effective for improving the performance: even in the 'SimCLR' framework, the RMset can improve the performance. (2) 'SimCLR' framework is effective; however 'Ours' is better alternative to 'SimCLR' when involving the RMset.

Why the attention module is still being learned in the ACL. In Table 9, 'Ours w/o IAL' denotes the ACL training that exclusively trains the classification and contrastive modules, excluding the attention module. The performance of 'Ours w/o IAL' is better than 'Ours w/o ACL', but it gets better when trained together with the attention module in ACL. This is because the inter-sample attention of the benchmark dataset is gradually forgotten, and the possibility of a collapse in contrastive learning, which is the worst case, increases.

Methods	AffectNet-7
SimCLR [1]	66.39
Ours w/o ACL, IAL	66.13
Ours w/o ACL, FER	63.39
Ours w/o IAL	66.42
Ours	**66.85**

Table 9. Performance on AffectNet-7 with different module combinations

5. Discussion of FERPlus

Our framework exhibits a decline in overall accuracy across the IAL and ACL of FERPlus evaluation due to the excessive class imbalance in the test set. In this section, we discuss the difference between overall and average accuracy, why overall accuracy decreased and average accuracy increased in FERPlus, and the effectiveness of using RMset to mitigate class imbalance in the benchmark dataset.

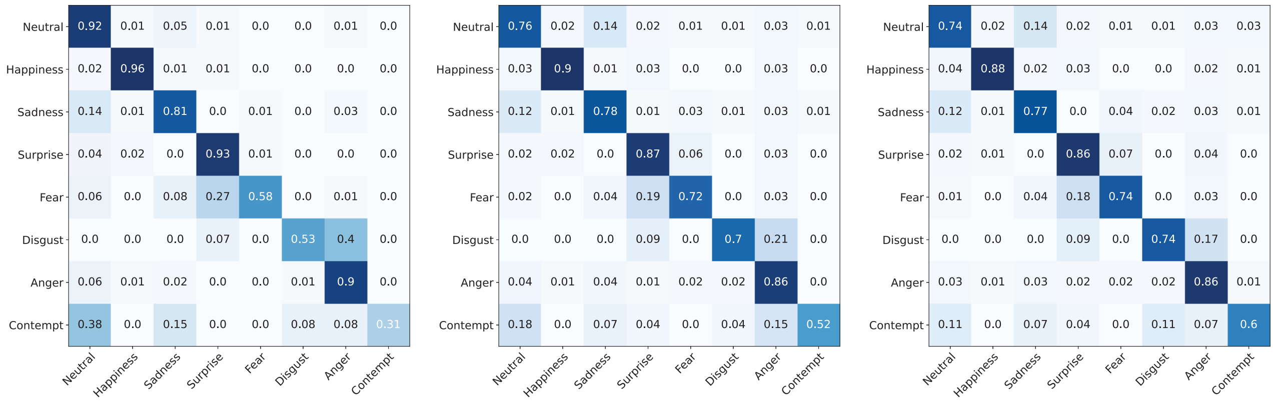
5.1. Overall accuracy and average accuracy

As explained in the main paper, we employed two metrics to evaluate our model's performance: 'overall accuracy' and 'average accuracy.' Overall accuracy quantifies the proportion of correctly predicted samples out of the total, regardless of the accuracy for each individual class. This measure is particularly useless when dealing with imbalanced test sets, as it gives more weight to classes with larger sample sizes. In contrast, average accuracy computes the mean accuracy across all classes, ensuring that each class is considered equally, irrespective of any test set imbalances. It can be expressed as follows:

$$\text{Acc}_{\text{overall}} = \frac{\sum_{i=1}^{C} \text{Acc}_i \times N_i}{\sum_{i=1}^{C} N_i} \tag{1}$$

$$\text{Acc}_{\text{average}} = \frac{\sum_{i=1}^{C} \text{Acc}_i}{C} \tag{2}$$

Where C is the number of classes, N_i is the number of samples in the i-th class, and Acc_i is the accuracy in the i-th class. When the test set is balanced, there is no difference between overall accuracy and average accuracy. However, if the test set is imbalanced, these two metrics will be different in their values. Moreover, the greater the imbalance within the test set, the greater the difference.



(a) KTN

(b) Ours w/o ACL, IAL

(c) Ours

Figure 2. Confusion matrices of KTN [6], Ours w/o ACL, IAL, and Ours

5.2. Discrepancy between overall accuracy and average accuracy

Our approach demonstrates that the overall accuracy of FERPlus tends to decline with the inclusion of IAL and ACL, while the average accuracy sees an improvement.

This is due to two reasons: Firstly, FERPlus exhibits a notably imbalanced test set when compared to other datasets. Not only is it more imbalanced than AffectNet, which maintains an equal number of samples for all classes in the test set, but it also surpasses the imbalance seen in RAF-DB, which is highly imbalanced. As indicated in Table 10, the 'Disgust' and 'Contempt' classes, in particular, have relatively fewer samples in the FERPlus test set compared to other datasets.

The second reason is that our approach aims to enhance performance across all classes by learning inter-sample attention, rather than focusing solely on one majority class. It is crucial to balance the classes within the training batch to effectively learn inter-sample attention in the IAL, ensuring the model learns inter-sample attention between all facial expressions. This approach results in less variation in accuracy across classes. Additionally, as the RMset is designed to balance facial expression classes as much as possible after utilizing the benchmark dataset, the ACL further enhances performance across these classes. The details of how the RMset is designed to balance facial expression classes during training are discussed in Sec. 5.3.

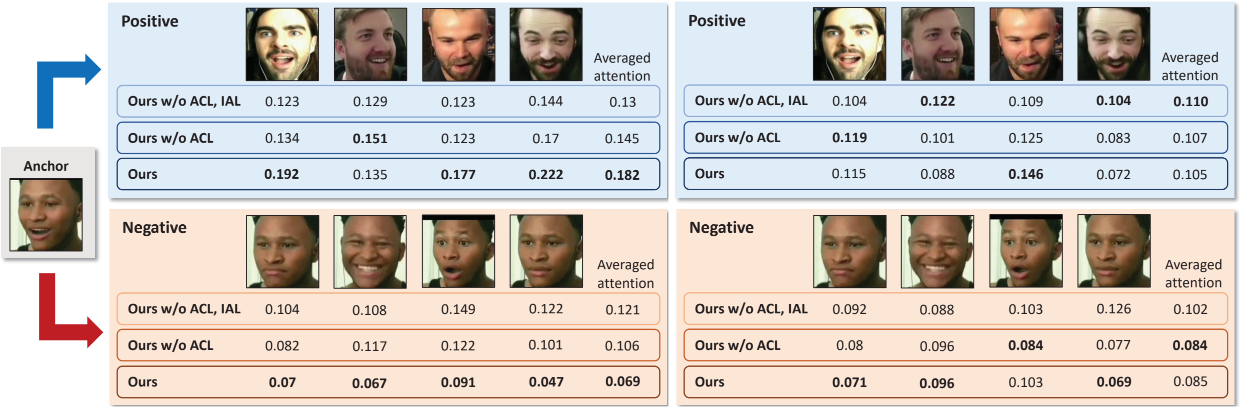
Therefore, as the learning progresses, the average accuracy increases while overall accuracy decreases because of the imbalance of the test set. In contrast, for other models with high overall accuracy (such as KTN [6] and FER-VT [4]), the average accuracy is inferior to that of our framework, and there is a significant discrepancy in accuracy between classes. Therefore, our model is robust in terms of average accuracy on FERPlus. The corresponding confusion matrix is presented in Fig. 2.

5.3. Balanced learning with the RMset

Table 10 illustrates that FERPlus exhibits a severe imbalance, with Fear, Disgust, and Contempt represented in small proportions. As shown in Fig. 2 (a), the performance of KTN [6] on expression labels with limited samples, such as Fear, Disgust, and Contempt, is notably weak. However, as the test set labels are also imbalanced, the overall accuracy does not fully capture this imbalance.

On the other hand, our model learns to avoid significant differences in accuracy between classes in both IAL and ACL. In the IAL, we ensured class balance within each training batch, as previously explained. In the ACL, we extended this approach by incorporating the RMset. The RMset is specifically designed to include a significant number of negative facial expressions, including categories like 'Disgust' and 'Contempt,' which were notably underrepresented in existing labeled datasets. As a result, Fig. 2 (b) demonstrates a significantly smaller performance deviation between expressions with limited and abundant samples compared to KTN [6]. (We enforced in-batch balance even when training only the classification module without IAL and ACL.) Furthermore, Figure 2 (c) demonstrates that our complete model, which integrates IAL and ACL, consistently maintains performance that is on par (within a 1-2% deviation) with the baseline for facial expressions other than Disgust and Contempt. Notably, the model significantly improves performance for expressions with very few samples, such as Disgust (with a 4% increase) and Contempt (with an 8% increase), improving average accuracy.

This also means that when RMset is expanded, RM-FER can learn additional facial expressions based on the benchmark dataset and compensate for the lack of facial ex-



(a) Ours

(b) Ours w/o self-masking

Figure 3. Visualization of attention for the samples of the RMset. As we involve IAL and ACL, the attention becomes better: Attention needs to be higher for similar expressions; while becoming less for different expressions with the same identity.

pressions in the existing training benchmark dataset without annotation. See Table 10, you can check the number of samples in training sets for all cases. In many FER datasets, positive and easily obtainable data such as 'Happiness' are abundant, while negative emotions like 'Disgust' and 'Anger' are less readily collected and thus less common. However, the RMset is a dataset that notably encompasses a higher quantity of these negative reactions.

Expression	FERPlus		RAF-DB		AffectNet-7	
	Training	Testing	Training	Testing	Training	Testing
Neutral	10,309	1,262	2,524	680	74.833	500
Happiness	7,528	928	4,772	1,185	134,304	
Sadness	3,515	444	1,982	478	25,441	
Surprise	3,562	444	1,290	329	14,078	
Fear	652	93	281	74	6,374	
Disgust	**191**	**23**	717	160	3,801	
Anger	2,467	325	705	162	24,873	
Contempt	**165**	**27**	-	-	3,746	

Table 10. The number of samples for each facial expression in the training and testing sets of FERPlus, RAF-DB, and AffectNet. The number of fear, disgust, and contempt facial expressions in the FERPlus is significantly lower in the training and testing sets than in other facial expressions and datasets.

6. Additional ablation study

6.1. Effects on γ.

An ablation study was conducted on the hyperparameter γ used for sampling the improved $\mathbf{P}_{\text{imp}}$ and $\mathbf{N}_{\text{imp}}$. Table 11 shows results on testing sets when γ is manipulated from 0 to 1.0. As the gamma ratio decreases from 1.0 to 0.1, the accuracy becomes robust by effectively filtering out unnecessary samples; if γ is less than 0.1, the accuracy becomes less as insufficient samples are used for the training.

We also added the 'ratio of filtered samples that have the same label with the anchor sample' result from 'Ours (AffectNet)' in Table 4, by varying the γ in Table 12. Naturally, a smaller γ leads to a higher proportion of correctly filtered samples. However, our experimentation found that 0.1 was the optimal value. When γ falls below 0.1, it results in an insufficient number of samples for effective contrastive learning. Furthermore, considering that this ratio is calculated from the entire AffectNet test set, the combination of ACL with the prior information from the RMset yields a significantly enhanced positive and negative set, thereby contributing to overall performance improvement. Additionally, samples are better filtered according to semantic classes by involving ACL modules.

γ	RAF-DB		AffectNet-7
	overall	average	
0.05	91.17	85.26	66.39
0.1	**91.33**	**85.59**	**66.85**
0.5	91.04	85.04	66.33
1.0	91	84.32	65.99

Table 11. Ablation study of γ

γ	Ours w/o ACL, SM	Ours w/o ACL	Ours w/o SM	Ours
0.05	41.7	45.0	44.2	50.3
0.1	41.2	42.8	41.8	46.0
0.5	23.2	23.7	22.3	23.6

Table 12. The ratio of filtered samples having the same label as the anchor sample about ACL, SM, and γ in AffectNet-7. 'SM' denotes self-masking.

6.2. Effects of self-masking softmax on attention

Self-masking softmax has a more positive impact on inter-sample attention learning by preventing selfreferencing. In the main paper, we looked at this from a performance perspective. However, here, we evaluate this aspect using two quantitative measures, the cosine similarity and filtering ratio, in addition to presenting qualitative results in Fig. 3.

Firstly, considering the difference between the average cosine similarity among samples that share the same labels and those that have different labels similar to Table 6, Table 13 illustrates that the implementation of self-masking, both in the IAL and ACL, leads to an increase in the cosine similarity between samples sharing the same label, while simultaneously reducing the cosine similarity between samples with different labels.

In Table 12, adding self-masking (SM) to the framework improves filtering performance. When γ is 0.05, the performance improvement from 'Ours w/o ACL, SM' to 'Ours w/o ACL' is about 3.3%, and the improvement from 'Ours w/o SM' to 'Ours' is 6.1%. This suggests that self-masking leads to better filtering, which could be one of the outcomes of learning proper inter-sample attention. Moreover, the combination of self-masking and ACL results in an enhanced filtering performance, making this combination a favorable choice.

Fig. 3 (a) visualizes the positive and negative sets of attention to anchor presented in the main paper. In this case, (a) uses a model with self-masking. In contrast, (b) uses a model without self-masking. As a result, the difference in attention value between models, which was relatively strong in (a), is weaker in (b). This suggests that self-masking strengthens IAL and ACL.

Methods	Dfference of cosine similarity
Ours w/o ACL, SM	0.3725
Ours w/o ACL	0.4426
Ours w/o SM	0.3904
Ours	0.4716

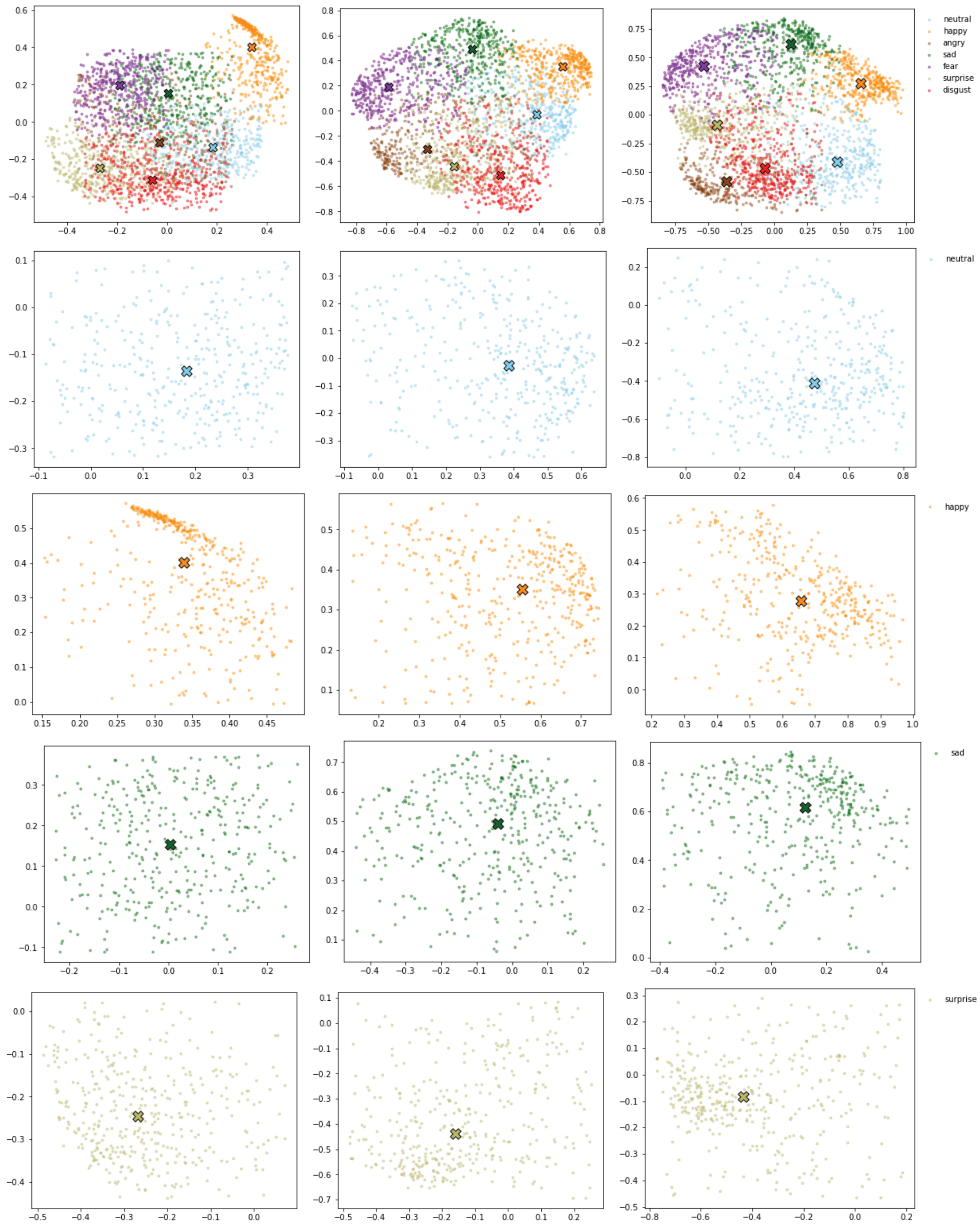
Table 13. The difference between the cosine similarity based on self-masking AffectNet-7

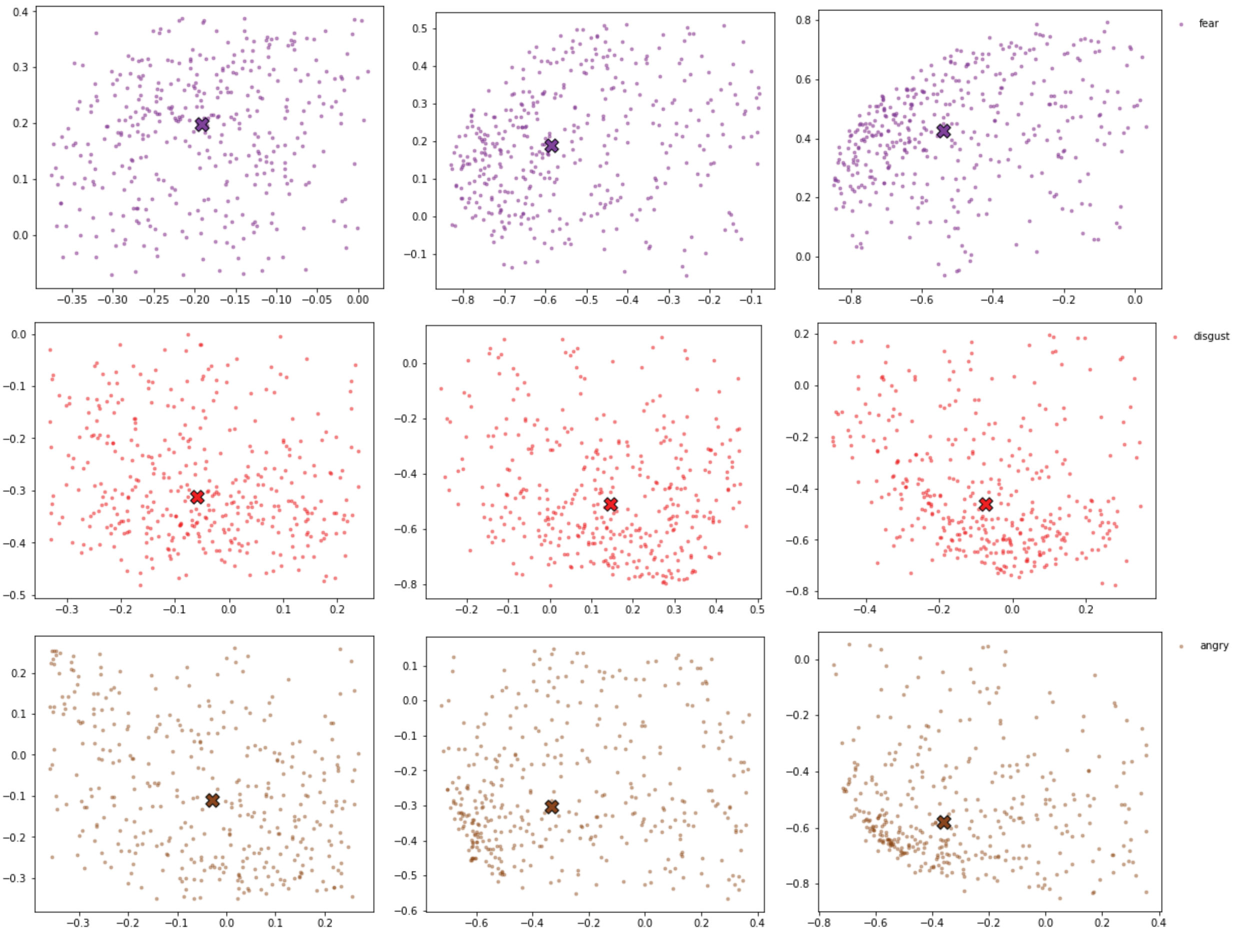
7. MDS plots

The main paper presented the MDS plot on AffectNet-7 simultaneously for all classes. In this supplemental, we visualized the MDS plots for three models (i.e., 'Ours w/o IAL, ACL,' 'Ours w/o ACL,' and 'Ours') in the "class-wise manner" in Fig. 4 to make them better visible. In the first row of Fig. 4, we visualized the MDS plot for all classes; while in the remaining rows, we visualized it class by class. In each class, we could find the points gathered as IAL and ACL were used (from the left to the right column). Also, the average distance between the center and the samples in each expression class is presented in Table 14 to provide the numerical measure for the distribution. We could concretely conclude that the distribution becomes better and better as more learning strategies of our framework (i.e., IAL, ACL) are used.

expression category	Ours w/o ACL, IAL	Ours w/o IAL	Ours
Neutral	0.369	0.339	**0.337**
Happiness	0.223	0.199	**0.181**
Sadness	0.397	0.336	**0.274**
Surprise	**0.240**	0.338	0.274
Fear	0.256	0.261	**0.246**
Disgust	0.327	**0.309**	0.347
Anger	0.491	0.493	**0.375**
Average	0.329	0.325	**0.291**

Table 14. Mean of distance from the center in MDS plot


neutral
happy
angry
sad
fear
surprise
disgust
neutral
happy
sad
surprise



(a) Ours w/o ACL, IAL

(b) Ours w/o ACL

(c) Ours

Figure 4. MDS plot of (a) Ours w/o ACL, IAL, (b) Ours w/o ACL, (c) Ours for each class in AffectNet-7. The mean distance from the center is presented in Table 14.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[2] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 2021.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[6] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *TIP*, 2021.

[7] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *CVPR*, 2022.

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[9] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *ECCV*, 2022.

[10] 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*, 2017.

[11] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022.

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[14] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: erasing attention consistency for noisy label facial expression recognition. In *ECCV*, 2022.

[15] Kai Zheng, Yuanjiang Wang, and Ye Yuan. Boosting contrastive learning with relation knowledge distillation. In *AAAI*, 2022.