Yunkai Li
School of Electrical and Information Engineering
Zhengzhou University
Zhengzhou 450001, P.R. China
E-mail: liyunkai@zzu.edu.cn

Response to Decision Letter

Manuscript No.: TCE-2025-07-2122

"Emotion Recognition in Human–Computer Interaction Using Multimodal Adaptive Reweighting Strategy"

Dear Editor,

I am very glad to receive your notice about our manuscript and thank you very much for providing us the revision opportunity with which we can further improve the quality of our manuscript entitled "Emotion Recognition in Human–Computer Interaction Using Multimodal Adaptive Reweighting Strategy". We sincerely appreciate the professional comments and useful suggestions from the reviewers, and we believe now all the comments and suggestions have been fully addressed, comprehensively, and clearly.

We have revised the manuscript carefully by fully considering all the questions. A list of responses to the reviewers' comments is provided below for your kind consideration. For the convenience of reviewers, we have highlighted all the revisions in our revised manuscript, and the locations of the revisions in the manuscript have been marked in the response.

We hope that the revised version can be satisfactory.

Looking forward to hearing from you.

Sincerely yours,

Yunkai Li

Responses to Reviewer #1:

Comments: The paper introduces Mars (Multimodal Adaptive Reweighting Strategy), a framework for emotion recognition in human–computer interaction. It addresses the challenge of inconsistent modality quality and sample noise in multimodal emotion recognition (MER). Mars incorporates a two-stage training process combining an uncertainty-based dynamic reweighting mechanism with cross-entropy (CE) and Mars loss. This enables the model to prioritize higher-quality samples and reliable modalities over noisier data.

Q1: Mars's performance heavily depends on finely tuned parameters for modality-specific reweighting and stage transition. The need for grid search makes it computationally intensive and less user-friendly for deployment.

Response 1: Thank you very much for the comment. We acknowledge that the grid search requires intensive computation during the training stage. However, it enables systematic exploration of hyperparameter effects on the optimization of Mars, providing valuable insights for developing more efficient optimization strategies (e.g., meta-gradient methods or evolutionary algorithms). These limitations and related discussions have been supplemented in the revised manuscript. Moreover, the proposed Mars maintains a relatively low computational cost during inference, because it only requires the fusion of output probabilities across modalities in the testing stage. The comparison of computational cost between Mars and baseline approaches has also been supplemented.

In the revised manuscript, we have added a discussion on the computational cost of grid search as follows:

"Moreover, the grid search makes the training process computationally intensive, but it provides valuable insights for designing more efficient hyperparameter optimization methods. To address these challenges, we will combine Mars with meta-gradient methods or evolutionary algorithms in future work to dynamically adjust hyperparameters and reduce training complexity." **Please see Section V-H.**

The computational cost of Mars in the testing phase has been supplemented in **Table II**, and the comparison between Mars and baseline approaches has been analyzed as follows:

"Computational Cost: To evaluate the computational cost of the proposed method, we compare Mars with baseline approaches on the RAVDESS dataset, as summarized in Table II. FLOPs denote the number of floating-point operations, and inference time refers to the average time required to process a video–audio sample pair. All evaluations are conducted in model testing mode. As shown in Table II, Mars achieves the lowest computational cost among all methods, requiring only 56.23 GFLOPs and 20.19 ms of inference time. It matches the performance of decision-fusion and other uncertainty reweighting methods, while outperforming feature-fusion and model-fusion methods, including lightweight baselines such as Concatenation (57.23 GFLOPs, 24.59 ms) and more complex models such as AataNet (108.42 GFLOPs, 130.57 ms). This advantage stems from the use of decision fusion as the backbone of Mars, which only requires the fusion of output probabilities across modalities. These results demonstrate the efficiency of Mars and highlight its potential for practical deployment." **Please see Section V-C.**

TABLE II

COMPARISON OF COMPUTATIONAL COST ON RAVDESS.

Methods		Computational Cost FLOPs (G)	Computational Cost Inference Time (ms)
Feature Fusion	Concatenation	57.23	24.59
	CBP [32]	61.32	46.18
	TFN [19]	58.10	26.88
Decision Fusion	CE/ Multiplication [15]	56.23	20.19
Model Fusion	MMTM [22]	86.71	110.43
	MSAF [23]	57.25	84.15
	ERANNs [33]	64.86	92.73
	CFN-SR [34]	57.37	86.24
	AataNet[35]	108.42	130.57
	MAE-DFER [36]	89.70	116.35
Uncertainty Reweighting	FL [26]/ M3ER [16]/ **Mars**	56.23	20.19

Q2: While the method down-weights low-confidence samples, it lacks a robust mechanism to differentiate between genuinely noisy data and inherently difficult cases, which could suppress valuable learning signals.

Response 2:

Thank you very much for the helpful comment. The proposed uncertainty weighting mechanism computes a dynamically scaled loss for each sample pair, enabling the model to automatically prioritize information from more reliable modalities while de-emphasizing less reliable ones from a sample perspective. This method treats both noisy data and inherently difficult cases as low-confidence samples. Although it improves the training robustness to some extent, it may also suppress valuable learning signals. Therefore, we have discussed this issue in the conclusion and proposed corresponding directions for future research.

In the revised manuscript, the discussion and the future extension of the current work have been supplemented as follows:

"Despite the promising performance of the proposed Mars framework, several limitations remain for further investigation. Notably, the uncertainty reweighting strategy, while effective in suppressing the influence of low-confidence data, still struggles to differentiate between genuinely noisy data and inherently difficult cases. This limitation may lead to suboptimal learning behavior, especially in complex or ambiguous emotional contexts. Future work will focus on enhancing the robustness of the uncertainty modeling mechanism by developing more fine-grained criteria or auxiliary models to better differentiate between difficult and noisy samples." **Please see Section VI.**

Q3: Although no extra parameters are added to the model, modality-specific adjustments to loss scaling introduce an extra layer of complexity during implementation.

Response 3:

Thank you very much for the helpful comment. Although the modality-specific loss scaling adjustments introduce additional complexity during the training stage, the proposed Mars framework does not require the calculation of loss weights during the testing stage, relying solely on the fusion of output probabilities across modalities. As a result, it maintains a relatively low computational cost compared to other fusion approaches during implementation.

In the revised manuscript, the computational cost of Mars in the testing phase has been supplemented in **Table II**, and the comparison between Mars and baseline approaches has been analyzed as follows:

"Computational Cost: To evaluate the computational cost of the proposed method, we compare Mars with baseline approaches on the RAVDESS dataset, as summarized in Table II. FLOPs denote the number of floating-point operations, and inference time refers to the average time required to process a video–audio sample pair. All evaluations are conducted in model testing mode. As shown in Table II, Mars achieves the lowest computational cost among all methods, requiring only 56.23 GFLOPs and 20.19 ms of inference time. It matches the performance of decision-fusion and other uncertainty reweighting methods, while outperforming feature-fusion and model-fusion methods, including lightweight baselines such as Concatenation (57.23 GFLOPs, 24.59 ms) and more complex models such as AataNet (108.42 GFLOPs, 130.57 ms). This advantage stems from the use of decision fusion as the backbone of Mars, which only requires the fusion of output probabilities across modalities. These results demonstrate the efficiency of Mars and highlight its potential for practical deployment." **Please see Section V-C.**

TABLE II

COMPARISON OF COMPUTATIONAL COST ON RAVDESS.

Methods		Computational Cost	
		FLOPs (G)	Inference Time (ms)
Feature Fusion	Concatenation	57.23	24.59
	CBP [32]	61.32	46.18
	TFN [19]	58.10	26.88
Decision Fusion	CE/ Multiplication [15]	56.23	20.19
Model Fusion	MMTM [22]	86.71	110.43
	MSAF [23]	57.25	84.15
	ERANNs [33]	64.86	92.73
	CFN-SR [34]	57.37	86.24
	AataNet[35]	108.42	130.57
	MAE-DFER [36]	89.70	116.35
Uncertainty Reweighting	FL [26]/ M3ER [16]/ **Mars**	56.23	20.19

Q4: While speech and visual cues often have temporal dependencies, the Mars framework doesn't explicitly leverage temporal models like RNNs, LSTMs, or transformers to model emotional evolution over time.

Response 4:

Thank you for the helpful comment. In this study, the input samples from the RAVDESS and FETE datasets were segmented into 1-second clips for emotion recognition. Under this short-time setting, we employed established single-modality feature extractors, namely a 3D CNN (ResNeXt50) for visual features and a 1D CNN for audio features, which have been widely adopted in prior work and are

effective in capturing short-term spatiotemporal patterns. We agree with the reviewer that models such as RNNs, LSTMs, and Transformers are more powerful for modeling long-range temporal dependencies and emotional dynamics. Accordingly, future work will extend our approach to longer sequences and explicitly incorporate temporal modeling mechanisms.

In the revised manuscript, the discussion and the future extension of the current work have been supplemented as follows:

"Since speech and visual cues often exhibit strong temporal dependencies, we will also incorporate temporal modeling and cross-modal temporal dependency analysis to better capture the dynamic evolution of emotions." **Please see Section VI.**

Q5: Some recent multimodal fusion frameworks (especially using transformers or self-supervised learning) are missing. Compare to newer and more recent state of the art works.

Response 5:

Thank you very much for the helpful suggestion. We have compared the proposed Mars framework with several recent multimodal fusion methods in **Table I**, including transformer-based approaches (e.g., AataNet, 2025) and self-supervised learning methods (e.g., MAE-DFER, 2023).

The detailed observations have been supplemented as follows:

"Model fusion strategies exhibit superior performance compared to both feature fusion methods and decision fusion methods. The versatility of modality interaction at any level of representation for unimodal models enables model fusion to leverage complementary relationships across modalities and adaptively emphasize more crucial features. Furthermore, Transformer-based approaches (e.g., CFN-SR, AataNet) demonstrate superior performance among model fusion methods, outperforming both intermediate fusion methods (e.g., MMTM, MSAF) and self-supervised learning approaches (e.g., MAE-DFER). This superiority can be attributed to the self-attention and cross-attention mechanisms' capacity to effectively capture complex dependencies across multiple modalities."

"Among the various uncertainty weighting strategies, focal loss demonstrates the least favorable results, performing even worse than the baseline method (Averaging). In contrast, the outcomes of M3ER and Mars exhibit significant improvements compared to the baseline method. Notably, our proposed Mars achieves the highest accuracy among all multimodal fusion techniques. On the *RAVDESS* dataset, Mars outperforms the current best-performing method, CFN-SR, by 0.9% in accuracy and 1.04% in F1 score. On the *FETE* dataset, Mars outperforms M3ER by 2.34% in accuracy and 3.22% in F1 score. The performance enhancements with Mars can be attributed to several advantages. Firstly, Mars effectively mitigates the influence of low-quality samples and data noise by adaptively down-weighting misclassified instances, allowing the model to focus on more reliable instances. Secondly, it avoids the vanishing gradient issue that can occur with M3ER, thereby ensuring stable training dynamics and improved convergence. Additionally, due to the limited scale of the two datasets used in this study and the absence of large-scale multimodal pre-training, transformer-based and self-supervised learning methods cannot fully realize their potential, whereas Mars exhibits more robust performance under these data constraints." **Please see Table I and Section V-C.**

TABLE I

CLASSIFICATION PERFORMANCE COMPARISON ON RAVDNESS AND FETE DATASETS

Standards	Models	RAVDNESS			FETE		
		Accuracy (%)	F1-score (%)	#Params	Accuracy (%)	F1-score (%)	#Params
Single Visual	ResNeXt50 / C3D	62.99	62.75	25.88 M	64.05	62.75	178.68 M
Single Audio / Tactile	1D CNN / d3D	56.53	56.13	0.03 M	66.01	64.28	1.10 M
Feature Fusion	Concatenation	72.67	72.33	26.87 M	75.23	73.10	180.72 M
	CBP [32]	72.96	72.61	51.03 M	74.79	71.14	204.82 M
	TFN [19]	73.44	73.14	27.97 M	75.68	73.80	181.80 M
Decision Fusion	Averaging (CE) [15]	72.06	71.83	25.92 M	73.07	70.47	179.79 M
	Multiplication [15]	72.55	72.30	25.92 M	74.35	70.80	179.79 M
Model Fusion	MMTM [22]	73.12	73.01	31.97M	76.51	74.12	185.86 M
	MSAF [23]	74.86	74.68	25.94 M	75.89	73.67	179.82 M
	ERANNs [33]	74.80	74.60	28.46 M	75.13	73.44	183.95 M
	CFN-SR [34]	75.76	75.54	25.95 M	75.67	74.11	179.86 M
	AataNet [35]	75.24	75.17	364.1 M	75.88	73.12	518.62 M
	MAE-DFER [36]	74.91	74.56	85.16 M	75.11	73.20	239.23 M
Uncertainty Reweighting	Focal loss [26]	71.29	71.03	25.92 M	72.18	70.76	179.79 M
	M3ER β=0.2 [16]	74.03	73.97	25.92 M	76.80	74.23	179.79 M
	Mars (ours)	**76.66**	**76.58**	25.92 M	**79.14**	**77.45**	179.79 M

Responses to Reviewer #2:

Q1: The necessity of an Ablation Study to analyze the contribution of each core component. The current manuscript achieves impressive performance by combining two novel ideas: the new Mars loss function and the staged training strategy. However, because these two components are applied simultaneously, it is difficult to ascertain the individual contribution of each or to determine if both are necessary for the observed improvements. To clearly demonstrate the value of each proposed component, we strongly recommend including an ablation study with the following experiments:

- Experiment 1: Baseline (using only the standard loss function, as currently presented).

- Experiment 2: Mars loss function only (applied from the beginning of training, without the staged strategy).

- Experiment 3: The final proposed model (as currently presented).

Response 1:

Thank you very much for the helpful suggestion. In the revised manuscript, we have added an ablation study to analyze the contribution of each core component (new subsection in **Section V-D**).

The experimental results and analysis are as follows:

"Table III summarizes the ablation study results on the two datasets. The Baseline model, which adopts only the standard loss function, achieves accuracies of 72.06% on RAVDNESS and 73.07% on FETE. When incorporating the Mars loss function, i.e., applying the uncertainty weighting mechanism from the beginning of training without the staged strategy, the performance improves to 73.39% and 75.71% on RAVDNESS and FETE, respectively. This improvement highlights the effectiveness of the uncertainty weighting mechanism in enhancing model robustness. However, compared with the full Mars framework, the improvement is relatively modest, since applying the Mars loss at the early stages of training makes the model overly sensitive to local noise, thereby hindering its generalization capability. By contrast, the full proposed Mars, which integrates both the uncertainty weighting mechanism and the staged training framework, achieves the highest accuracies of 76.66% on RAVDNESS and 79.14% on FETE. These findings demonstrate that each component contributes positively to performance, while the staged strategy plays a crucial role in mitigating noise sensitivity and fully realizing the potential of Mars." **Please see Table III and Section V-D.**

TABLE III

ABLATION EXPERIMENT RESULTS.

Method	Accuracy (%)	
	RAVDNESS	FETE
Baseline	72.06	73.07
Mars loss function only	73.39	75.71
Proposed Mars	76.66	79.14

Q2: A discussion on the generalizability and sensitivity of the hyperparameter L. The study demonstrates that L=10 is the optimal switching point for the staged training. However, this value was determined under a specific set of conditions (the RAVDESS dataset, a particular model architecture,

and a total of 30 training epochs). This raises an important question regarding the method's generalizability: How would the optimal value of L change with different datasets, model architectures, or total training durations? If this hyperparameter requires a new search for each use case, the universality of the methodology could be somewhat limited. Therefore, we advise the authors to add a discussion in the "Experiments" or "Conclusion" section acknowledging that L is a sensitive hyperparameter that may require tuning under different conditions and that this represents a practical limitation of the current methodology.

Response 2:

Thank you very much for the helpful suggestion. As noted, our current experiments demonstrate that L = 10 yields the optimal switching point under the specific conditions of the RAVDESS dataset, our chosen model architecture, and 30 training epochs. We fully acknowledge that this optimal value may vary when different datasets, model architectures, or training durations are employed. Consequently, L should be regarded as a sensitive hyperparameter that may require re-tuning for different scenarios, which indeed represents a practical limitation of the current methodology.

In the revised manuscript, we have added a discussion on the generalizability and sensitivity of the hyperparameter L as follows:

"In addition, we observed that the hyperparameter L plays a crucial role in the staged training strategy. Specifically, our experiments show that L=10 yields the best performance under the current setting (i.e., the RAVDESS dataset, the adopted model architecture, and 30 training epochs). However, it should be noted that the optimal value of L is sensitive to the experimental configuration and may shift when different datasets, model architectures, or training durations are considered. This sensitivity implies that re-tuning L could be necessary in new scenarios, which represents a practical limitation of the current methodology. Moreover, the grid search makes the training process computationally intensive, but it provides valuable insights for designing more efficient hyperparameter optimization methods. To address these challenges, we will combine Mars with meta-gradient methods or evolutionary algorithms in future work to dynamically adjust hyperparameters and reduce training complexity." **Please see Section V-H.**

Responses to Reviewer #3:

Comments: This paper proposes a novel multimodal adaptive reweighting strategy (Mars) to address noise issues and modality conflicts in emotion recognition. The experimental design is rigorous, and the results show significant improvements on both the RAVDESS and FETE datasets. However, some concerns need to be addressed before it is considered for publication, here are my comments.

Q1: While the literature review covers relevant works, it could benefit from including more studies on EEG-based emotion recognition. Relevant references are provided below for consideration:

[1] Fusing Frequency-Domain Features and Brain Connectivity Features for Cross-Subject Emotion Recognition. IEEE Transactions on Instrumentation and Measurement, 2022.

[2] Emotion Recognition and Dynamic Functional Connectivity Analysis Based on EEG. IEEE Access, 2019.

Response 1:

Thank you very much for the comment and suggestion. In the related studies of emotion recognition, the works "Fusing Frequency-Domain Features and Brain Connectivity Features for Cross-Subject Emotion Recognition" and "Emotion Recognition and Dynamic Functional Connectivity Analysis Based on EEG" have been introduced in the revised manuscript.

Specifically, Ref. [2] has been cited in the following context:

"In general, emotions can be inferred from various modalities, such as facial expressions, speech, body movements, touch gestures, and electrophysiological signals [2], [3]." **Please see Section I**.

In addition, Ref. [1] has been introduced as:

"Chen et al. [20] proposed an approximate empirical kernel map fusion method for electroencephalography-based emotion recognition, which enhances both the effectiveness and efficiency of fusing features with different dimensions." **Please see Section II-A.**

Q2: **"**Low computational complexity" lacks specific data support. Could quantitative indicators be provided to enhance credibility?

Response 2:

Thank you very much for the helpful suggestion. The computational cost of Mars in the testing stage has been supplemented in **Table II**, and the comparison between Mars and baseline approaches has been analyzed as follows:

"Computational Cost: To evaluate the computational cost of the proposed method, we compare Mars with baseline approaches on the RAVDESS dataset, as summarized in Table II. FLOPs denote the number of floating-point operations, and inference time refers to the average time required to process a video–audio sample pair. All evaluations are conducted in model testing mode. As shown in Table II, Mars achieves the lowest computational cost among all methods, requiring only 56.23 GFLOPs and

20.19 ms of inference time. It matches the performance of decision-fusion and other uncertainty reweighting methods, while outperforming feature-fusion and model-fusion methods, including lightweight baselines such as Concatenation (57.23 GFLOPs, 24.59 ms) and more complex models such as AataNet (108.42 GFLOPs, 130.57 ms). This advantage stems from the use of decision fusion as the backbone of Mars, which only requires the fusion of output probabilities across modalities. These results demonstrate the efficiency of Mars and highlight its potential for practical deployment." **Please see Section V-C**.

TABLE II

COMPARISON OF COMPUTATIONAL COST ON RAVDESS.

Methods		Computational Cost	
		FLOPs (G)	Inference Time (ms)
Feature Fusion	Concatenation	57.23	24.59
	CBP [32]	61.32	46.18
	TFN [19]	58.10	26.88
Decision Fusion	CE/ Multiplication [15]	56.23	20.19
Model Fusion	MMTM [22]	86.71	110.43
	MSAF [23]	57.25	84.15
	ERANNs [33]	64.86	92.73
	CFN-SR [34]	57.37	86.24
	AataNet[35]	108.42	130.57
	MAE-DFER [36]	89.70	116.35
Uncertainty Reweighting	FL [26]/ M3ER [16]/ **Mars**	56.23	20.19

Q3: "Facilitating natural human-computer interaction (HCI)" contains an obvious grammatical error. It is recommended to check the full text sentence by sentence to avoid similar issues.

Response 3:

Thank you very much for the comment and suggestion. We have replaced the hyphen in "human-computer interaction" with an en dash "–" throughout the entire manuscript. Besides, we have revised the rest of the manuscript carefully to avoid similar mistakes.

The errors have been revised in our revised manuscript:

"The perception of human emotions is critical for developing natural human–computer interaction (HCI) systems and holds significant potential across a wide range of applications, including consumer electronics, social robotics, mental health monitoring, the entertainment industry, and virtual reality [1]." **Please see Section I.**

Q4: In the caption of Fig. 1, the descriptions of w_v and w_a are insufficient, and specific explanations of their dynamic adjustment mechanisms need to be supplemented.

Response 4:

Thank you very much for the helpful suggestion. The descriptions of w_v and w_a and the specific explanations of their dynamic adjustment mechanisms have been supplemented in our revised manuscript as follows:

"Fig. 1. Overview of the multimodal emotion recognition network using Mars. (a) Network architecture: Multimodal inputs are processed through feature extractors and classifiers to compute modality-specific losses l_v and l_a. Their corresponding weights w_v and w_a are dynamically adjusted across training stages. (b) Staged training strategy: In the first stage (where the current epoch $j < L$, and L denotes the switching point), both weights are fixed to 1 ($w_v = w_a = 1$), ensuring balanced contributions from each modality. In the second stage ($j \geq L$), the weights are adaptively updated according to $w_v = -e^{-\gamma v(1-pv)}$ and $w_a = -e^{-\gamma a(1-pa)}$, where p_v and p_a denote the confidence scores of the visual and audio modalities, respectively. This mechanism emphasizes high-contribution samples while suppressing low-contribution ones, thus improving the robustness of multimodal fusion." **Please see Fig.1.**

Q5: The peak positions of the loss distributions for the audio modality and visual modality in Fig. 7 are not labeled, making the key features unintuitive; moreover, the main text does not analyze why Mars reduces the impact of the audio modality by dynamically adjusting its weights.

Response 5:

Thank you very much for the helpful suggestion. The peak positions of the loss distributions for the audio modality and visual modality in Fig. 7 have been labeled. **Please see Fig.7.**

Experimental results demonstrate that Mars reduces the influence of the audio modality by dynamically adjusting its weights. This is because audio signals are more susceptible to background noise, speaker variability, and prosodic ambiguity, resulting in unstable representations and lower prediction confidence. Thus, the adaptive weighting mechanism reduces the influence of the audio modality and prioritizes the more reliable visual modality.

The analysis has been supplemented in the revised manuscript as follows:

"The loss distribution statistics of the last epoch for both CE and Mars loss functions are illustrated in Fig. 7. As shown in Fig. 7(a), the head of the distribution for the audio modality is noticeably higher than that of the visual modality when using CE loss, whereas the opposite trend is observed in Fig. 7(b) under the Mars loss, indicating that Mars reduces the impact of the audio modality by dynamically adjusting its weights. This behavior is attributed to the higher proportion of low-confidence samples in the audio modality of the RAVDESS dataset. Compared with visual samples, audio signals are more susceptible to background noise, speaker variability, and prosodic ambiguity, resulting in unstable representations and lower prediction confidence. Consequently, the adaptive weighting mechanism decreases the contribution of audio modality while assigning greater weights to the more reliable visual modality. This selective emphasis effectively suppresses modality-specific noise and improves the robustness of multimodal fusion." **Please see Section V-F.**

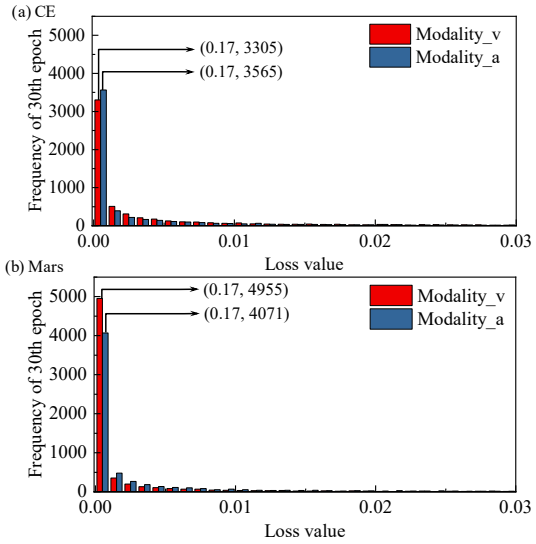

Fig. 7. Frequency distribution statistics of sample losses for the last epoch on RAVDNESS using (a) CE and (b) Mars loss functions.