sensors

MDPI

Article

Multimodal Emotion Recognition Using Modality-Wise Knowledge Distillation

Seonggyu Lee, Youngdo Ahn and Jong Won Shin *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Buk-gu, Gwangju 61005, Republic of Korea; lsqjin2022@gm.gist.ac.kr (S.L.); ayoungdo@gm.gist.ac.kr (Y.A.)
* Correspondence: jwshin@gist.ac.kr

Abstract

Multimodal emotion recognition (MER) aims to estimate emotional states utilizing multiple sensors simultaneously. Most previous MER models extract unimodal representation via modality-wise encoders and combine them into a multimodal representation to classify the emotion, and these models are trained with an objective for the final output of the MER. If an encoder for a specific modality is optimized better than others at some point of the training procedure, the parameters for the other encoders may not be sufficiently updated to provide optimal performance. In this paper, we propose a MER using modality-wise knowledge distillation, which adapts the unimodal encoders using pre-trained unimodal emotion recognition models. Experimental results on CREMA-D and IEMOCAP databases demonstrated that the proposed method outperformed previous approaches to overcome the optimization imbalance phenomenon and could also be combined with these approaches effectively.

Keywords: multimodal emotion recognition; knowledge distillation; optimization imbalance phenomenon

Academic Editors: Hongying Meng and Tong Chen

Received: 2 September 2025
Revised: 3 October 2025
Accepted: 13 October 2025
Published: 14 October 2025

Citation: Lee, S.; Ahn, Y.; Shin, J.W. Multimodal Emotion Recognition Using Modality-Wise Knowledge Distillation. *Sensors* **2025**, *25*, 6341. https://doi.org/10.3390/s25206341

1. Introduction

Emotion recognition refers to the techniques used to estimate emotional attributes or categorical emotions from various modalities such as speech, text, and video [1–5] captured by multiple sensors such as microphones and cameras, which can be applied to health care [6], automobility systems [7], and human–computer interactions [8]. Recently, multimodal emotion recognition (MER) has gained attention [9–18] as it mimics human perception of emotion utilizing multiple senses such as vision, hearing, and linguistic comprehension [19–21], and the performance of unimodal emotion recognition was not satisfactory. Although there are a few works on emotion recognition in conversation using contextual information [22–25], most of the research has focused on utterance-level emotion recognition, which estimates emotional state given an utterance [9–18]. The typical configuration of a MER model consists of unimodal encoders, which extract emotional representations from each modality; a fusion module, which combines them to produce multimodal representations; and a classifier to produce final probabilities for emotional classes. Usually, a single loss on the final output of the multimodal emotion classifier is used to train these models.

In [15], an analysis of the optimization imbalance phenomenon is provided, with an example for a simple fusion module and a linear classifier, in which the parameters of the

unimodal encoders may not be sufficiently updated in training once an encoder for another modality produces more discriminative embeddings during the training procedure.

To facilitate the learning of less-optimized encoders, on-the-fly gradient modulation with generalization enhancement (OGM-GE) was proposed in [15], which penalizes the learning rate for the parameters of the more-optimized unimodal encoder. In the same context, a prototypical modal rebalance (PMR) method is proposed in [16], which introduces separate loss terms for individual modalities using the "prototype" features for each class and entropy regularization to slow the learning of the more-optimized encoder. The authors of [17] proposed a modality-wise cosine similarity (MMCosine) loss function, which normalizes the weighted embeddings from different modalities so that one modality cannot dominate the loss function.

These approaches provide several ways to mitigate the optimization imbalance among modalities, but the loss functions for training the MER models only apply to the final MER output. In [26], an ensemble of unimodal models and a multimodal model is presented, in which a portion of the parameters is shared across different unimodal models, but the multimodal model in this approach is also trained with the loss on the final MER output. In this paper, we propose a MER with a loss function on the MER output and unimodal representations for all modalities. To help with the optimization of each unimodal encoder, we employ a knowledge distillation (KD) technique [27,28] so that the pre-trained unimodal emotion classification models guide the training of the corresponding unimodal encoders within the MER model. Experimental results on the utterance-wise emotion classification using a microphone and camera sensors show that the proposed MKD approach outperformed the previously proposed methods for MER and can be combined with previous methods to further improve the results.

2. Method

2.1. Typical Multimodal Emotion Recognition Model and Optimization Imbalance Phenomenon

A MER model aims to classify emotion in multimodal input data $x = \{x^m\}_{m=1}^M$, where x^m represents the input feature for the m-th modality, and M is the number of modalities. Typically, a MER model f consists of unimodal encoders $\{\psi^m\}_{m=1}^M$, a fusion module $\mathcal{F}$, and a classifier $\mathcal{C}$, as shown in Figure 1.

Each encoder transforms the input features into unimodal representations, and the fusion module combines them into a multimodal representation. Then, the classifier maps the multimodal representation into a C-dimensional vector, where C is the number of emotion classes. Then, the output of the model can be described as

$$f(x) = \mathcal{C}\Big(\mathcal{F}\Big(\psi^1(x^1), \psi^2(x^2), \ldots, \psi^M(x^M)\Big)\Big). \tag{1}$$

The most common choice for the loss function is CE loss:

$$\mathcal{L}_{\mathrm{CE}}(y, \hat{y}) = -y \cdot \log \hat{y}, \tag{2}$$

where y is a one-hot vector representing the emotion label of the input sample x, and $\hat{y} = \mathrm{softmax}(f(x))$ is the logit from the MER model. In the simplest example in which $\mathcal{F}$ is a concatenation of $\psi^m(x^m) \in \mathbb{R}^{d_m}$ and $\mathcal{C}$ is a linear classifier with weight matrices $W^m \in \mathbb{R}^{C \times d_m}, m = 1, \cdots, M$ and bias $b \in \mathbb{R}^C$, (1) can be represented as

$$f(x) = \textstyle\sum_{m=1}^{M} (W^m \psi^m(x^m) + b/M) \tag{3}$$

where $W^m \psi^m(x^m) + b/M$ corresponds to the output of a classifier when only the m-th modality is present. It was reported in [15,17] that the elements of W^m for a modality with

a less-optimized encoder become small when one unimodal encoder is better optimized at some point during the training procedure if CE loss is applied to the MER output, disrupting the further optimization of the other encoders. There have been a few approaches to this optimization imbalance phenomenon, including the adjustment of the learning rate [15] and the use of modality-wise cosine similarity [17]. In this paper, we propose a different approach that can also be combined with the previously proposed methods, which is to adapt each unimodal encoder using the knowledge distillation from the pre-trained unimodal emotion recognition model.

2.2. Modality-Wise Knowledge Distillation

The overall block diagram of the proposed MER system employing modality-wise knowledge distillation to facilitate the optimization of all unimodal encoders is shown in Figure 1. We use a pre-trained unimodal emotion recognition model $\bar{f}^m$ to guide the unimodal encoder in the MER model ψ^m. Any unimodal emotion recognition model can be used as $\bar{f}^m$, but if it has the structure of unimodal encoder $\bar{\psi}$, followed by a linear classifier $\bar{C}$, it can be represented as

$$\bar{f}^m(x^m) = \bar{C}^m(\bar{\psi}^m(x^m)). \tag{4}$$

An additional classifier for each unimodal encoder in the MER model, C^m, is used for MKD to compare the emotion classification capability of the encoder with that of the pre-trained unimodal emotion recognition model $\bar{f}^m$. In the experiment, $\bar{C}^m$ and $\bar{\psi}^m$ are configured to have the same structure but different parameter values from C^m and ψ^m, respectively. The target label vector $\bar{y}^m$ obtained from the unimodal pre-trained model is given by

$$\bar{y}^m = \text{softmax}\left(\frac{\bar{f}^m(x^m)}{T_m}\right), \tag{5}$$

where T_m is a hyperparameter called temperature. Then, the objective function for MKD is given by

$$\mathcal{L}_{\text{MKD}}(\{\bar{y}^m, \hat{y}^m\}_{m=1}^M) = \sum_{m=1}^M [\lambda_m \mathcal{L}_{\text{CE}}(\bar{y}^m, \hat{y}^m)], \tag{6}$$

where $\hat{y}^m = \text{softmax}(C^m(\psi^m(x^m)))$ is the logit from classifier C^m for the m-th unimodal representation, and λ_ms are the hyperparameters for the weights. For each iteration in training, we first update ψ^m and C^m using the MKD loss in (6), and then we update the whole model's parameters with the same minibatch using the MER loss, such as the CE loss in (2). In this way, each unimodal encoder has a chance to be updated according to the guidance of the corresponding pre-trained model and thus will not be less-optimized. In the inference phase, only MER model f is used, while C^m, $\bar{C}^m$, and $\bar{\psi}^m$ do not need to be kept.

2.3. Combination with Other Regularization Methods

We can combine MKD with previously proposed regularization methods [15,17]. In OGM-GE [15], which can be applied only for bimodal applications with $M = 2$, the learning rates for two modality encoders are adjusted based on the discrepancy between the contributions of the modalities to the final output. colorredThe discrepancy ρ_t^m, where $m \in \{1, 2\}$ for the t-th minibatch $\mathcal{B}_t$, is computed as

$$\rho_t^1 = \sum_{i \in \mathcal{B}_t} \frac{s_i^1}{s_i^2}, \quad \rho_t^2 = 1/\rho_t^1, \tag{7}$$

where s_i^m for the i-th sample x_i^m with one-hot label vector y_i is

$$s_i^m = y_i \cdot \text{softmax}(W^m \psi^m(x_i^m) + b/2). \tag{8}$$

Then, a scaling factor k_t^m is applied to the learning rate for the parameters regarding modality m:

$$k_t^m = \begin{cases} 1 - \tanh(\alpha \rho_t^m), & \text{if } \rho_t^m > 1, \\ 1, & \text{otherwise}, \end{cases} \tag{9}$$

where $\alpha > 0$ is a hyperparameter. For generalization enhancement, the parameter updates are perturbed by Gaussian noises with appropriate covariance matrices. In [17], MMCosine loss was adopted instead of CE loss so that one modality could not dominate the loss function. The MMCosine loss is given by

$$\mathcal{L}_{\text{MMCosine}}(y, x) = -\log \frac{\exp\left(s \sum_{m=1}^{M} \cos \theta_k^m\right)}{\sum_{j=1}^{C} \exp\left(s \sum_{m=1}^{M} \cos \theta_j^m\right)}, \tag{10}$$

where $\cos \theta_j^m = \frac{W_j^m \psi^m(x^m)}{\|W_j^m\| \|\psi^m(x^m)\|}$ in which W_j^m is the j-th row of W^m, $s > 0$ is a hyperparameter, and k is the index of the emotion label for the current input. MMCosine loss can be used for an arbitrary number of modalities, unlike OGM-GE [15]. The training procedure combining MKD with other regularization methods is described in Algorithm 1.

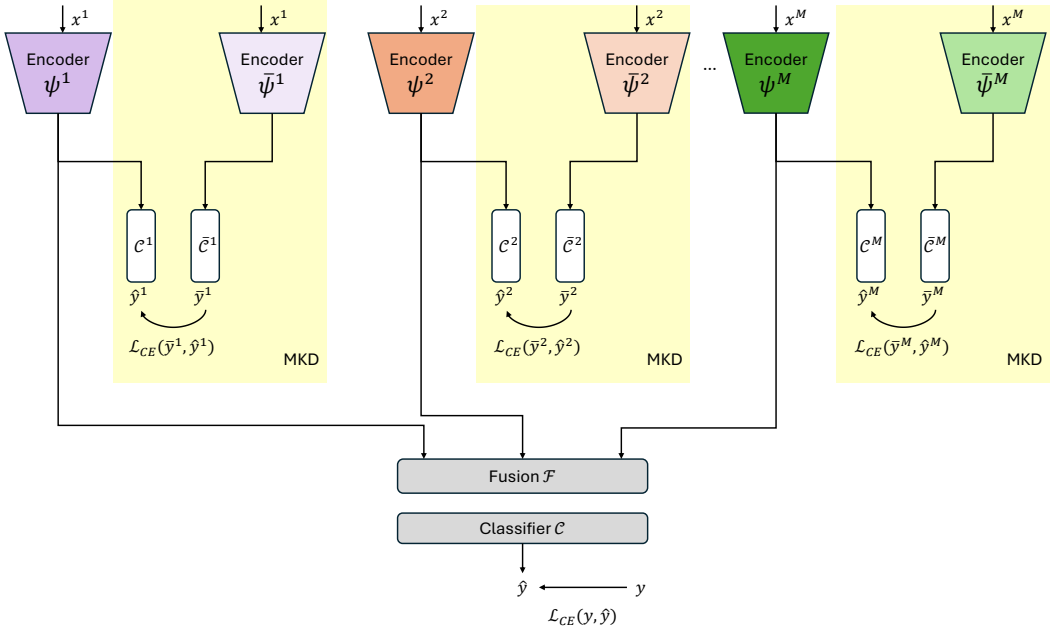

Figure 1. The overall block diagram for the proposed multimodal emotion recognition using modality-wise knowledge distillation (MKD). Yellow shaded blocks are used in the training phase only.

Algorithm 1: MKD with optional imbalance mitigation methods (OGM-GE or MMCosine)

Input : Initialized multimodal model f, modality-wise classifiers $\{\mathcal{C}^m\}_{m=1}^M$, pre-trained unimodal models $\{\bar{f}^m\}_{m=1}^M$;choice $\mathcal{R} \in \{\text{None, OGM-GE, MMCosine}\}$

Output: Trained multimodal model f

while *stopping criterion is not satisfied* **do**
  for $t = 1$ *to the number of minibatches* **do**
    (1) Compute $\mathcal{L}_{\text{MKD}}$ on minibatch $\mathcal{B}_t$.
    (2) Update $\{\psi^m, \mathcal{C}^m\}_{m=1}^M$ using $\nabla_{\{\psi^m, \mathcal{C}^m\}} \mathcal{L}_{\text{MKD}}$.
    (3) Set the loss: if $\mathcal{R} = \text{MMCosine}$, use $\mathcal{L} = \mathcal{L}_{\text{MMCosine}}$; otherwise, use $\mathcal{L} = \mathcal{L}_{\text{CE}}$.
    (4) If $\mathcal{R} = \text{OGM-GE}$ (bimodal case, $M = 2$):
      Compute discrepancy ρ_t^m (for $m \in \{1, 2\}$) on $\mathcal{B}_t$ and scaling factor k_t^m);
      apply gradient modulation (and perturbation) to the updates of ψ^m.
    (5) Update f using $\nabla_f \mathcal{L}$ (with OGM-GE modulation if applicable).

3. Experiments

3.1. Experimental Configurations

To demonstrate the performance of MKD in utterance-wise emotion recognition, two multimodal emotional databases were employed, which were the audio–visual dataset CREMA-D [29] and the audio–visual–text dataset IEMOCAP [30]. CREMA-D consists of 7422 video clips, which were recorded by 91 professional actors expressing pre-determined emotions for 12 pre-defined sentences. A total of 6698 clips were used as the training set, while 744 samples comprsied the test set. We considered six emotion classes for this dataset, which were angry, happy, sad, neutral, disgust, and fear. IEMOCAP is a dyadic conversation dataset performed by 5 male and 5 female actors including both scripted scenarios and improvisations. Three annotators assigned the emotion class labels to each speech utterance in IEMOCAP. We used 7 classes including angry, excited, happy, sad, frustrated, surprised, and neutral, and the number of utterances was 7487 [14]. We performed 5-fold cross-validation, dividing the data into a training set, validation set, and test set in an 8:0.5:1.5 ratio, as in [9,14,31]. We trained the model 10 times with different random initializations and report the performance averaged over 10 random seeds and 5-fold cross-validation for the IEMOCAP dataset, as in [9,14,31].

As for the features and model structures, we followed the configurations in [15] and [14] for CREMA-D and IEMOCAP, respectively. For CREMA-D, we used one frame sampled from a video clip as a visual feature and a spectrogram of 257×299 as an audio feature. For IEMOCAP, 300-dimensional GloVe embeddings of transcriptions were employed as text features. Also, 40-dimensional MFCC features with their first and second derivatives were used as the audio features, while 2048-dimensional Resnet-101 embeddings of a cropped face image at 3 Hz frame rates were utilized as the visual features. For the model structure, we used ResNet18 as the encoder for CREMA-D, as in [15], and the fusion was made by a simple concatenation, which outperformed other fusion methods such as sum, FiLM [32], and Gated [33]. For IEMOCAP, we utilized the tri-modal self-attention model in [14], in which the encoders contained convolution layers, bi-directional GRU, and multi-head attention layers. The outputs of three encoders were fused as in [14], where the encoder outputs for frames were temporally averaged first, and then the mean and the

standard deviations of the temporal averages for three modalities were concatenated to form the fused vector.

The performance of the proposed MKD was compared with that of the original MER systems in [14,15], using previous approaches for mitigating the optimization imbalance phenomenon including OGM-GE [15], PMR [16], and MMCosine [17], the ensemble of unimodal emotion recognition models (denoted Uni-sum), and the ensemble of unimodal and multimodal emotion recognition models (denoted All-sum) [26]. To verify the performance improvement was from the modality-wise KD rather than KD itself, we implemented Self-KD [34] in which the softmax output of a pre-trained MER model with the same structure provided the target for the MER output. We also tested the use of modality-wise CE loss (MWCE) instead of MKD loss using the ground-truth one-hot label vector to show that MKD was indeed effective compared with guiding the unimodal encoders with additional losses. Additionally, we applied MKD along with OGM-GE [15], PMR [16], and MMCosine [17], and we tested the ensemble of MKD and unimodal models (denoted All-sum (MKD)). It is noted that OGM-GE and PMR are essentially bimodal methods and thus were not tested for IEMOCAP in our experiment.

For the models applied to CREMA-D, the learning rate schedule and the stopping criterion followed the code and methods in [15] except for MKD+ [16], for which we used the same configuration as in [16]. For IEMOCAP, we followed the optimizer, learning rates, and stopping criterion in [14]. For MKD, the parameters used in the experiments were $\lambda_m \in \{0.001, 0.1, 1, 1.5, 4.5, 5, 8\}$ and $T_m \in \{0.001, 0.01, 0.05, 1, 3.7\}$, depending on the database and combined approaches. We first crudely determined the parameter values using a grid search for a wide range of parameters, and then we adjusted them one-by-one to optimize the performance for the validation set for IEMOCAP and the test set for CREMA-D. The accuracy (ACC) and unweighted accuracy (UA) were used as evaluation metrics for CREMA-D and IEMOCAP, respectively, as in many previous works on the corresponding databases. ACC is the ratio of the number of correct predictions and test samples. UA is an average of the accuracies for individual emotional classes.

3.2. Results

Table 1 summarizes the performance of the MER systems and the number of parameters needed in the inference phase for CREMA-D and IEMOCAP. It is noted that this work focused on utterance-wise emotion recognition for six emotions for CREMA-D and seven emotions for IEMOCAP, while some other works focused on four-emotion classification or emotion recognition in conversation. In Table 1, the highest accuracy (ACC) and unweighted average (UA) values for each dataset are shown in bold. We can see that the adoption of the proposed MKD improved the performance of not only the original MER models but also the MER models with the previously proposed OGM-GE [15], PMR [16], MMCosine [17], and All-sum [26]. The best ACC for CREMA-D was 69.3%, which was achieved when the proposed MKD was combined with MMCosine loss, while the best UA of 68.7% was obtained for IEMOCAP when the ensemble of unimodal models and the MER model with MKD was employed. Although the performance improvement achieved with All-sum (MKD) over the best previously proposed method, All-sum [26], for IEMOCAP was not very high, it was statistically significant with a p-value of 0.037. All three previously proposed approaches that deal with optimization imbalance phenomenon improved the performance of MER on CREMA-D, and MMCosine [17] performed the best. However, MMCosine did not improve the performance on IEMOCAP, although we tried several fusion methods and achieved the best performance with concatenation fusion. This result may have been due to the modality imbalance for IEMOCAP being less severe, as can be seen in Table 2, and thus ignoring the magnitude of the weighted embedding for each

modality would have resulted in information loss rather than modality balancing. In contrast, the ensemble approach denoted All-sum [26] produced less of a performance improvement than these three approaches on CREMA-D, and the performance of the ensemble of MKD with unimodal models was inferior to that of MKD. As can be seen from the performance of the unimodal models on CREMA-D in Table 2, the visual cues in CREMA-D were much weaker than the audio cues, which led to the poor performance of the simple ensemble of audio and visual modalities. MWCE and Self-KD [34] improved the MER performance on both of the databases but not as much as MKD. We confirmed that the performance improvement of the proposed MKD was not just due to the additional loss function for unimodal encoders or to the KD with soft labels.

Additionally, we checked if MKD actually mitigated the optimization imbalance phenomenon by examining the performance of each unimodal encoder in the MER model by attaching a linear classifier, as in [35]. Specifically, we froze the unimodal encoders in the MER models and trained a linear classifier for each of the unimodal encoders to identify emotional classes. Table 2 shows the ACCs (%) and UAs (%) of the unimodal models trained with a single modality and unimodal encoders in the MER model with and without MKD. It can be seen that each of the unimodal encoders in the MER model without MKD was not as good as the unimodal models for all cases, while the unimodal encoders in the MER model with MKD performed better than the unimodal models by utilizing both the KD from the unimodal models and the gradient affected by other modalities.

Table 3 presents the ablation study results for MKD in which the modality-wise knowledge distribution was applied to a subset of the unimodal encoders. The experimental results show that the performance was improved by applying MKD for each of the unimodal encoders one-by-one on both of the databases and for every modality. The performance improvement achieved by adding MKD for one more modality was almost the same for any modality in the experiment with IEMOCAP, although the performance of the unimodal pre-trained models was not the same.

Table 1. Performance of multimodal emotion recognition and the number of parameters for the proposed and comparison systems on CREMA-D and IEMOCAP databases.

Method	**CREMA-D**		**IEMOCAP**	
	#Param	**ACC (%)**	**#Param**	**UA (%)**
Multimodal	22.4M	55.9	1.6M	64.20
OGM-GE [15]	22.4M	62.2 *	-	-
PMR [16]	22.9M	61.8 *	-	-
MMCosine [17]	22.9M	66.4 *	1.6M	61.80
Uni-sum [26]	22.4M	55.9	1.6M	63.60
All-sum [26]	44.7M	60.3	3.3M	68.00
MWCE	22.4M	60.8	1.6M	65.70
Self-KD [34]	22.4M	60.3	1.6M	64.40
MKD	22.4M	67.7	1.6M	67.50
MKD+ [15]	22.4M	68.4	-	-
MKD+ [16]	22.9M	67.9	-	-
MKD+ [17]	22.4M	**69.3**	1.6M	66.90
All-sum (MKD)	44.7M	67.1	3.3M	**68.70**

* indicates the result from the original work.

Table 2. Modality-wise emotion recognition accuracies for unimodal models and unimodal encoders in multimodal models with and without MKD attached to linear classifiers.

Method	CREMA-D		IEMOCAP		
	Audio	Visual	Audio	Visual	Text
Unimodal	57.5	27.3	45.1	53.2	51.3
Multimodal	57.0	18.6	43.2	50.8	50.6
MKD	62.5	29.2	45.7	54.7	52.0

Table 3. Performance of multimodal emotion recognition with and without modality-wise knowledge distillation for individual unimodal encoders.

Audio	Visual	Text	CREMA-D	IEMOCAP
✗	✗	✗	55.9	64.2
✓	✗	✗	63.5	66.1
✗	✓	✗	62.9	66.1
✗	✗	✓	-	65.9
✓	✓	✗	67.7	66.4
✓	✗	✓	-	66.4
✗	✓	✓	-	66.2
✓	✓	✓	-	67.5

4. Conclusions

In this paper, we propose a MER method using modality-wise knowledge distillation that utilizes pre-trained unimodal emotion recognition models to overcome the optimization imbalance phenomenon. In the proposed method, each unimodal encoder in the MER model is trained to produce similar logits with an extra linear classifier for the logits than a pre-trained unimodal model produces, while the whole MER model including unimodal encoders is also updated with the same minibatch using the loss for the MER. By guiding the unimodal encoders within the MER model using pre-trained unimodal models, each unimodal encoder can be trained well even though another unimodal encoder provides much more useful information for MER at a certain point in the training process. Experimental results showed that MKD outperformed previous approaches for addressing the optimization imbalance phenomenon, and the combination of MKD and those techniques could further improve the performance. The best performance was obtained when MKD was combined with MMCosine loss on the CREMA-D datase, and when using the ensemble of MER with MKD and unimodal models on the IEMOCAP dataset. Further investigation is needed to find out why different combinations are more effective on different databases. Future works include the application of MKD with more MER models and multimodal classifiers in other domains.

Author Contributions: Coceptualization, S.L., Y.A. and J.W.S.; methodology, S.L. and Y.A.; validation, S.L.; formal analysis, S.L. and J.W.S.; investigation, S.L.; resources, S.L.; data curation, S.L.; writing, S.L., Y.A. and J.W.S.; original draft preparation, S.L. and Y.A.; review, J.W.S.; supervision, J.W.S.; project administration, J.W.S.; funding acquisition, J.W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP), funded by the Korean government (MSIT), (RS-2022-II220989(2022-0-00989), Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling (RS-2025-25443882), and S.A.M.A.N.T.H.A: Sentiment Audio Machine for Alive Natural Talking & Human Affection).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets mentioned in this paper can be downloaded using the following links: CREMA-D https://github.com/CheyneyComputerScience/CREMA-D (accessed on 25 August 2025), IEMOCAP https://sail.usc.edu/iemocap/ (accessed on 25 August 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kim, E.; Shin, J.W. DNN-based Emotion Recognition Based on Bottleneck Acoustic Features and Lexical Features. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6720–6724.
2. Kossaifi, J.; Toisoul, A.; Bulat, A.; Panagakis, Y.; Hospedales, T.M.; Pantic, M. Factorized Higher-Order CNNs with an Application to Spatio-Temporal Emotion Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6060–6069.
3. Ahn, Y.; Lee, S.J.; Shin, J.W. Cross-Corpus Speech Emotion Recognition Based on Few-Shot Learning and Domain Adaptation. *IEEE Signal Process. Lett.* **2021**, *28*, 1190–1194. [CrossRef]
4. Ahn, Y.; Lee, S.J.; Shin, J.W. Multi-Corpus Speech Emotion Recognition for Unseen Corpus Using Corpus-Wise Weights in Classification Loss. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 131–135.
5. Ahn, Y.; Han, S.; Lee, S.; Shin, J.W. Speech Emotion Recognition Incorporating Relative Difficulty and Labeling Reliability. *Sensors* **2024**, *24*, 4111. [CrossRef] [PubMed]
6. Zhang, Y.; Chen, M.; Huang, D.; Wu, D.; Li, Y. iDoctor: Personalized and Professionalized Medical Recommendations Based on Hybrid Matrix Factorization. *Future Gener. Comput. Syst.* **2017**, *66*, 30–35. [CrossRef]
7. Katsis, C.D.; Rigas, G.; Goletsis, Y.; Fotiadis, D.I. Emotion Recognition in Car Industry. In *Emotion Recognition*; Wang, W., Ed.; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2015; Chapter 20, pp. 515–544.
8. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
9. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118.
10. Chen, B.; Cao, Q.; Hou, M.; Zhang, Z.; Lu, G.; Zhang, D. Multimodal Emotion Recognition with Temporal and Semantic Consistency. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3592-3-603. [CrossRef]
11. Sun, L.; Liu, B.; Tao, J.; Lian, Z. Multimodal Cross- and Self-Attention Network for Speech Emotion Recognition. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 4275–4279.
12. Yang, D.; Huang, S.; Liu, Y.; Zhang, L. Contextual and Cross-Modal Interaction for Multi-Modal Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2022**, *29*, 2093–2097. [CrossRef]
13. Middya, A.I.; Nag, B.; Roy, S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowl.-Based Syst.* **2022**, *244*, 108580. [CrossRef]
14. Rajan, V.; Brutti, A.; Cavallaro, A. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4693–4697.
15. Peng, X.; Wei, Y.; Deng, A.; Wang, D.; Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 8238–8247.
16. Fan, Y.; Xu, W.; Wang, H.; Wang, J.; Guo, S. PMR: Prototypical Modal Rebalance for Multimodal Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20029–20038.
17. Xu, R.; Feng, R.; Zhang, S.-X.; Hu, D. MMCosine: Multi-Modal Cosine Loss Towards Balanced Audio-Visual Fine-Grained Learning. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
18. Xie, J.; Wang, J.; Wang, Q.; Yang, D.; Gu, J.; Tang, Y.; Varatnitski, Y.I. A multimodal fusion emotion recognition method based on multitask learning and attention mechanism. *Neurocomputing* **2023**, *556*, 126649. [CrossRef]
19. Sebe, N.; Cohen, I.; Huang, T.S. Multimodal emotion recognition. In *Handbook of Pattern Recognition and Computer Vision*; World Scientific: Singapore, 2005; pp. 387–409.
20. Haq, S.; Jackson, P.J.B. Multimodal Emotion Recognition. In *Machine Audition: Principles, Algorithms and Systems*; Wang, W., Ed.; IGI Global: Hershey, PA, USA, 2011; pp. 398–423.

21. Geetha, A.V.; Mala, T.; Priyanka, D.; Uma, E. Multimodal Emotion Recognition with deep learning: Advancements, challenges, and future directions. *Inf. Fusion* **2024**, *105*, 102218.
22. Fu, Y.; Yuan, S.; Zhang, C.; Cao, J. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics* **2023**, *12*, 4714. [CrossRef]
23. Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; Xu, B. A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. *IEEE Trans. Multimed.* **2023** *26*, 776–788. [CrossRef]
24. Chen, F.; Shao, J.; Zhu, S.; Shen, H.T. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10761–10770.
25. Zhang, X.; Cui, W.; Hu, B.; Li, Y. A Multi-Level Alignment and Cross-Modal Unified Semantic Graph Refinement Network for Conversational Emotion Recognition. *IEEE Trans. Affect. Comput.* **2024**, *15*, 1553–1566. [CrossRef]
26. Sari, L.; Singh, K.; Zhou, J.; Torresani, L.; Singhal, N.; Saraf, Y. A multi-view approach to audio-visual speaker verification. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6194–6198.
27. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531. [CrossRef]
28. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
29. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef] [PubMed]
30. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
31. Ahn, C.S.; Kasun, L.L.C.; Sivadas, S.; Rajapakse, J.C. Recurrent multi-head attention fusion network for combining audio and text for speech emotion recognition. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 744–748.
32. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2018; Volume 32, pp. 1–9.
33. Kiela, D.; Grave, E.; Joulin, A.; Mikolov, T. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2018; Volume 32, pp. 1–9.
34. Yun, S.; Park, J.; Lee, K.; Shin, J. Regularizing class-wise predictions via self-knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13876–13885.
35. Alain, G.; Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv* **2016**, arXiv:1610.01644.