SetPeER: Set-based Personalized Emotion Recognition with Weak Supervision

Minh Tran*, Yufeng Yin*, and Mohammad Soleymani, *Senior Member, IEEE*

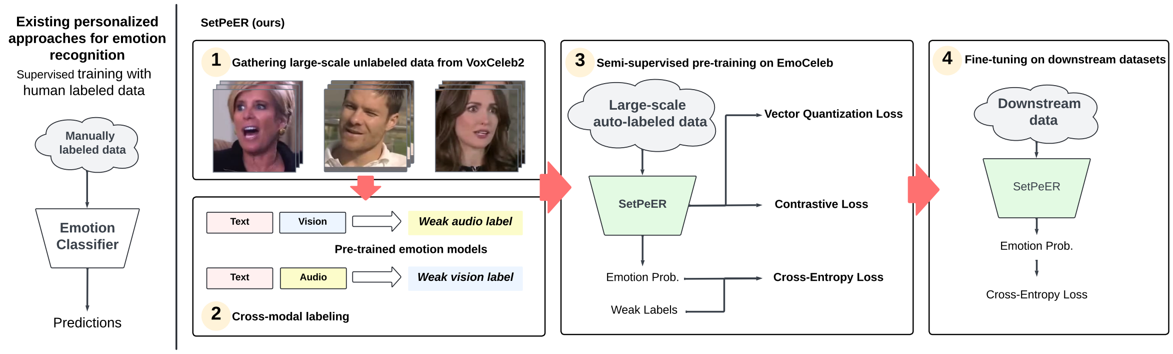


Fig. 1: Proposed framework for personalized emotion recognition. The figure illustrates the following four steps: (1) We collect unlabeled videos from VoxCeleb2 which is a diverse audiovisual dataset. (2) We use cross-modal labeling to create a large-scale weakly-labeled emotion dataset, *i.e.*, EmoCeleb. (3) We propose a novel personalization method, *i.e.*, SetPeER with set learning. The model is pre-trained on EmoCeleb and learns representative speaker embedding for personalization. (4) We fine-tune SetPeER on downstream datasets with the provided manual labels.

***Abstract*—Individual variability of expressive behaviors is a major challenge for emotion recognition systems. Personalized emotion recognition strives to adapt machine learning models to individual behaviors, thereby enhancing emotion recognition performance and overcoming the limitations of generalized emotion recognition systems. However, existing datasets for audiovisual emotion recognition either have a very low number of data points per speaker or include a limited number of speakers. The scarcity of data significantly limits the development and assessment of personalized models, hindering their ability to effectively learn and adapt to individual expressive styles. This paper introduces EmoCeleb: a large-scale, weakly labeled emotion dataset generated via cross-modal labeling. EmoCeleb comprises over 150 hours of audiovisual content from approximately 1,500 speakers, with a median of 50 utterances per speaker. This rich dataset provides a rich resource for developing and benchmarking personalized emotion recognition methods, including those requiring substantial data per individual, such as set learning approaches. We also propose SetPeER: a novel personalized emotion recognition architecture employing set learning. SetPeER effectively captures individual expressive styles by learning representative speaker features from limited data, achieving strong performance with as few as eight utterances per speaker. By leveraging set learning, SetPeER overcomes the limitations of previous approaches that struggle to learn effectively from limited data per individual. Through extensive experiments on EmoCeleb and established benchmarks, *i.e.*, MSP-Podcast and MSP-Improv, we demonstrate the effectiveness of our dataset and the superior performance of SetPeER compared to existing methods for emotion recognition. Our work paves the way for more robust and accurate personalized emotion recognition systems.**

***Index Terms*—Emotion Recognition, Personalization, Transfer Learning, Machine Learning.**

I. Introduction

EMOTION recognition is a foundational block for developing socially intelligent AI, with its applications spanning various domains from healthcare to user satisfaction assessment. Recent progress in the field has been driven by advancements in deep learning and multimodal data processing [1]–[4]. Despite these advances, there are challenges in building robust and generalizable emotion recognition. Specifically, effectively capturing individual variations in emotional behaviors while also addressing the scarcity of suitable data poses significant hurdles.

A key challenge in emotion recognition is the inherent variability of emotional expressions. Individuals exhibit diverse expressive styles shaped by factors such as culture, upbringing, personality, and situational context [5]. This variability poses a significant challenge for general-purpose emotion recognition models, resulting in inconsistent performance across speakers [6]. To address this limitation, personalized emotion recognition aims to adapt models to individual behaviors, leading to more accurate and robust performance. Several approaches have explored personalized emotion recognition for visual and speech tasks [1], [4], [7], [8]. For example, Shahabinejad *et al.* [4] jointly train a face recognition and a visual emotion recognition model, enabling the model to learn personalized emotion representations. Sridhar *et al.* [8] propose a speaker matching method to find the most similar speakers in a fixed training set to use as data augmentation to train personalized speech emotion recognition systems. However, most existing methods are limited in their applicability to other modalities, extensibility to unseen speakers, or efficiency due to the need for model re-training.

Another significant obstacle in emotion recognition, particularly for personalized approaches, stems from the scarcity of appropriate data. Prior efforts in personalized emotion recognition have predominantly focused on the speech modality due to data availability. However, these approaches were mainly trained and evaluated on a limited number of speakers [2], [9]–[11]. While recent advancements, such as the development of large pre-trained models like HuBERT [12] or Wav2Vec2 [13], have narrowed the personalization gap, challenges persist in the availability of comprehensive datasets for personalized visual emotion recognition. Although databases like MSP-Podcast [14] have been utilized for personalized speech emotion recognition [1], [8], they often lack adequate number of samples for each speakers and do not support personalized visual emotion recognition due to their unimodality. As shown in Table I, commonly used emotion recognition databases suffer from limitations such as a small number of speakers or insufficient samples per speaker. This scarcity of appropriate data not only impedes the development of robust personalized emotion recognition systems, particularly those incorporating visual cues, but also makes it challenging to rigorously evaluate such systems.

This paper addresses the aforementioned challenges in personalized emotion recognition. In this paper, emotion recognition refers to the automatic recognition of apparent emotions observed by others. From the modeling perspective, we introduce a novel approach called the Set-based Personalized Representation Learning for Emotion Recognition (**SetPeER**). This model is designed to extract personalized information from as few as eight samples per speaker. Notably, SetPeER exhibits versatility across different modalities by merely adjusting the backbone encoder architecture, *e.g.*, HuBERT [12] for audio and VideoMAE [15] for vision, and remains effective on unseen speakers without requiring any retraining of components. At the core of SetPeER is a Personalized Feature Extractor module $\mathbf{P}$ that encodes a set of utterances from the same speaker into meaningful speaker embeddings. These embeddings are then injected into pre-trained encoders to generate personalized features, thereby enhancing the model's ability to capture individual characteristics in emotional expression. Regarding data, we develop a scalable framework to weakly label in-the-wild audiovisual videos. Specifically, we use pre-trained models for text, vision, or audio-based emotion recognition to assign weak labels to a target modality from the remaining two modalities (text and audio or text and vision) to a large dataset of unlabeled data with a large number of speakers. To improve label quality, we only keep the utterances for which two modalities predict similar labels. We use the bimodal predictions for training emotion recognition models for the third modality. With the scalability of the proposed weak labeling approach, we introduce EmoCeleb-A and EmoCeleb-V, two large-scale datasets with substantially more speakers and samples per speaker than existing emotion recognition datasets.

Through extensive experiments, we validate the usefulness of EmoCeleb-A and EmoCeleb-V. First, we demonstrate the superior performance of our proposed weak labeling pipeline compared to random guessing. Moreover, our findings indicate that models trained with our dataset can surpass those trained with human-annotated data in zero-shot out-of-domain evaluations, underscoring the role of scalability and diversity in enhancing generalization capabilities. We further use EmoCeleb-A and EmoCeleb-V to both train and evaluate SetPeER, alongside established emotion recognition datasets, namely MSP-Podcast [14] for audio and MSP-Improv [16] for vision tasks. Through comprehensive evaluation, we validate our proposed model's effectiveness compared to existing personalized emotion recognition approaches.

TABLE I: Comparison of EmoCeleb with previous emotion recognition datasets. Mod indicates the available modalities, (a)udio, (v)ision, and (t)ext. TL denotes the total number of hours. # U and # S denote the number of utterances and speakers respectively. Our datasets are larger and have more speakers, with at least 50 utterances per speaker. All datasets are audio-visual except for MSP-Podcast. * We exclude samples without speaker identifications.

Dataset	Mod	TL (h)	# U	# S	Per Speaker Stats Mean	Per Speaker Stats Median
RAVDESS [17]	{a,v}	1.5	1.4K	24	60	60
AFEW [18]	{a,v}	2.5	1.6K	0.3K	5	-
HUMAINE [19]	{a,v}	4	50	4	13	-
RECOLA [20]	{a,v}	4	46	46	1	1
SEWA [21]	{a,v}	5	0.5K	0.4K	1	1
SEMAINE [22]	{a,v}	7	0.3K	21	13	6
CREMA-D [23]	{a,v}	8	7.4K	91	82	82
MSP-Improv [16]	{a,v}	9	8.4K	12	0.7K	0.7K
VAM [24]	{a,v}	12	0.5K	20	25	-
IEMOCAP [25]	{a,v}	12	10K	10	1.0K	1.0K
MSP-Face [26]	{a,v}	25	9.4K	0.4K	23	15
CMU-MOSEI [27]	{a,v,t}	66	24K	1.0K	24	4
MSP-Podcast [14]	{a,t}	71	43K	1.0K	40	12
EmoCeleb-A	{a}	159	74K	1.5K	50	50
EmoCeleb-V	{v}	162	75K	1.5K	50	50

The major contributions of this work are summarized as follows.

- **A large-scale weakly-labeled dataset.** We introduce EmoCeleb, a new dataset for personalized speech emotion recognition created using cross-modal labeling. This dataset comprises over 150 hours of speech from approximately 1,500 speakers, with each speaker having at least 50 utterances. This resource provides valuable data for both pretraining and evaluating personalized emotion recognition models. **EmoCeleb will be publicly released upon acceptance of this paper.**
- **A novel personalization method.** We propose a novel approach, SetPeER, for personalization that leverages set learning. Our method effectively learns a representative speaker embedding using only eight unlabeled utterances from a given speaker, enabling rapid adaptation to unseen individuals.
- **Extensive evaluation.** We conduct thorough experiments demonstrating the validity and utility of EmoCeleb. Furthermore, we demonstrate the effectiveness of SetPeER in personalized emotion recognition by visualizing the learned speaker embedding distributions.

II. Related Work

A. Emotion Recognition Databases

Access to expansive, natural databases that capture different facets of emotional expression is essential for improving emotion recognition. Table I presents some of the widely used emotion recognition databases. Generally, emotion recognition datasets can be categorized into three main types. Acted databases constitute the first type, where speakers are directed to express specific emotions while reciting predetermined sentences. This method is employed in various databases such as RAVDESS [17] and CREMA-D [23]. The second and most prevalent type consists of datasets captured within controlled laboratory environments. Participants are typically instructed to engage in interactions surrounding a given topic or to respond to emotion-inducing videos. Notable examples of this type include HUMAINE [19], SEWA [21], IEMOCAP [25], and MPS-Improv [16]. Lastly, the third type comprises fully natural utterances sourced from real-world settings, such as YouTube, and subsequently annotated through crowdsourcing. Datasets falling into this category include CMU-MOSEI [27], MSP-FACE [26], and MSP-Podcast [14]. Arguably, datasets of the third type are optimal for developing generalized emotion recognition systems applicable across diverse environments. Their potential is particularly promising as in-the-wild utterances are readily accessible online. However, the expense associated with human annotations often impedes large-scale development efforts, especially in personalized emotion recognition, where both the dataset size and the number of samples per speaker are crucial. As illustrated in Table I, existing large-scale emotion recognition datasets typically suffer from a scarcity of utterances per speaker. This paper aims to bridge the gap by leveraging the wealth of in-the-wild data to construct a large-scale weakly-labeled dataset customized for training and evaluating personalized emotion recognition systems and explore the trade-off between annotation accuracy and automated labeling.

B. Personalized Emotion Recognition

Various modalities have been investigated for personalized emotion recognition, *e.g.*, physiological signals [28]–[30], speech [1], [8], [9], [31], and facial expressions [4], [7]. Bang *et al*. [31] introduce a framework for robust personalized speech emotion recognition, which incrementally provides a customized training model for a target user via virtual data augmentation. Their method is evaluated on IEMOCAP [25] with ten speakers. Zhao *et al*. [28], [29] explore the impact of personality on emotional behavior through physiological signals using graph learning. Their studies are conducted on the ASCERTAIN dataset [32], which comprises data from 58 subjects. Zen *et al*. [33] propose an SVM-based vision regression model to learn the relationship between a user's sample distribution and the parameters of that individual's classifier and use the learned model to transfer to new users with unseen distributions. Chen *et al*. [9] develop a two-layer fuzzy random forest using features extracted from openSMILE [34] and train on different categories of people generated via a fuzzy C-means clustering algorithm. They demonstrate a potential performance gain in four subjects. Shahabinejad *et al*. [4] introduce an innovative attention mechanism tailored for facial expression recognition (FER). This mechanism generates an attention map using a face recognition (FR) network, thereby personalizing the FER process with FR features. However, their method relies on a single image for personalization, which raises concerns about the reliability of the personalization. Barros *et al*. [35] propose a Grow-When-Required network that learns person-specific features on seen speakers via a conditional adversarial autoencoder. In another work, Barros *et al*. [36] introduce a set of layers designed to learn both clusters of general facial expressions and individual behaviors through online learning and affective memories. However, the method is not applicable to unseen speakers. Barros *et al*. [37] presents Contrastive Inhibitory Adaption (CIAO) to adapt the last layer of facial encoders to model nuances in facial expressions across different datasets.

Most prior studies are either limited by the number of subjects available in the existing emotion datasets or rely on a single data point for personalization, compromising the learned personalized features' reliability and hindering their applicability to unseen speakers. Two notable exceptions are the studies by Sridhar *et al*. [8] and Tran *et al*. [1] that utilize MSP-Podcast [14], benefiting from its extensive range of subjects. However, the dataset is limited to the audio modality. Sridhar *et al*. [8] propose to find speakers in the training set whose acoustic patterns closely match those of the testing speakers to create an adaptation set. The approach needs additional training (model adaptation) at inference time, limiting its applicability to unseen speakers. Tran *et al*. [1] present PAPT, a personalized adaptive pre-training method, where the model is pre-trained with learnable speaker embeddings in a self-supervised manner and personalized label distribution calibration, which adjusts the predicted label distribution using label statistics from similar training speakers. PAPT has demonstrated superior effectiveness in personalization compared to Sridhar *et al*.'s method [1] while eliminating the necessity for retraining on new speakers.

C. Set Learning

Set representation learning extracts meaningful embeddings invariant to permutations for set inputs. DeepSets [38] operates by independently processing elements within a set and subsequently aggregating them using operations such as minimum, maximum, mean, or sum. Set Transformers [39] explore self-attention to model interactions between elements of a set. In addition to designing permutation-invariant modules for set encoding, alternative set-learning methodologies have emerged. These include methods that learn set representations by minimizing the disparity between an input set and a trainable reference set through bipartite matching [40] or optimal transport [41]. In this paper, we expand the scope of set representation learning to the realm of personalization, in which we aim to extract meaningful information about an individual based on a set of data samples. By leveraging personalized information, we enhance the encoding of the individual's data.

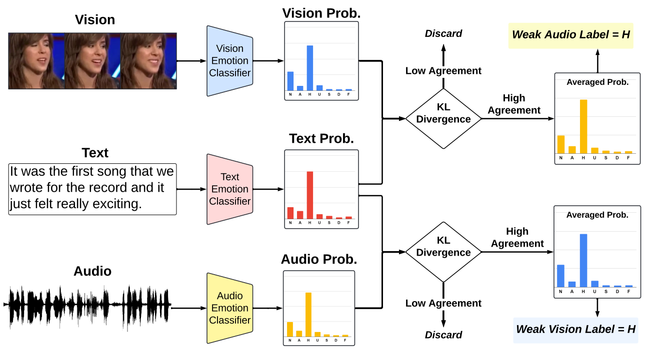


Fig. 2: Overview of cross-modal labeling: (i) Emotion recognition with two modalities (vision+text or audio+text) to provide weak supervision. (ii) Weak labels are retained when two modalities are in sufficient agreement (measured by KL divergence). (iii) Inference results are averaged to generate a weak label for the target modality (audio or vision).

III. EmoCeleb Dataset

Existing datasets for audiovisual emotion recognition have few speakers or lack enough data points per individual. This motivates us to develop a novel dataset via cross-modal labeling, *i.e*., utilizing information from one or more modalities to annotate another. Our approach enables the development of a large-scale emotion dataset with weak labels suitable for training and evaluating personalized emotion recognition systems.

Figure 2 provides an overview of the EmoCeleb dataset generation process. To enhance the utility of the dataset, we diverge from previous approaches such as the one by Albanie *et al*. [42], which utilizes a single modality input for cross-modal distillation (from vision to audio). Instead, we perform emotion recognition using two modalities to provide weak supervision. Weak labels are retained only when the emotion recognition results from both modalities agree. In particular, with the three modalities (vision, audio, and text), we perform cross-modal labeling in two directions: combining vision and text to label audio (EmoCeleb-A) and leveraging audio and text to label vision (EmoCeleb-V).

A. Labeling Dataset

We perform weak labeling on **VoxCeleb2** [43], which is an audiovisual dataset for speaker recognition. VoxCeleb2 includes interview videos of celebrities from YouTube, which contains **over 1M utterances** with **more than 6K speakers**. The dataset is roughly gender balanced (61% male), and the speakers span a wide range of ethnicities, accents, professions and ages [43]. The dataset provides the identities and apparent gender information for the speakers, but it does not have any emotion labels. We only use the English portion[1] of VoxCeleb2, which contains 1,326 hours of content.

[1]https://github.com/facebookresearch/av_hubert/blob/main/avhubert/preparation/data/vox-en.id.gz

TABLE II: Number of utterances in each class for EmoCeleb.

Dataset	Neutral	Anger	Happiness	Surprise	Total
EmoCeleb-A	45,288	3,682	21,466	3,664	74,100
EmoCeleb-V	39,774	6,909	19,168	9,259	75,110

B. Unimodal Emotion Recognition

1) Vision: For vision-based analysis, we utilize Masked Auto-Encoder (MAE) [44] as the backbone. We begin by initializing the MAE with a pre-trained checkpoint[2], which is trained on the EmotionNet dataset [45]. Subsequently, we perform supervised training on the Aff-Wild2 dataset [46] for frame-level emotion recognition. We perform frame-level inference for every utterance in the VoxCeleb2 dataset and employ average pooling to aggregate the results, thereby obtaining utterance-level logits for categorical emotions.

2) Audio: In the audio domain, we adopt an open-source model[3] based on WavLM [47] trained on the MSP-Podcast dataset [14] for speech emotion recognition.

3) Text: The VoxCeleb2 dataset [43] does not provide transcripts. Thus, we first use Whisper [48] for speech recognition. Then, we employ an open-source model[4] for text emotion recognition. This model is built upon RoBERTa [49] and has been trained on diverse text emotion datasets sourced from Twitter, Reddit, student self-reports, and television dialogue utterances, *e.g*., GoEmotions [50], AIT-2018 [51], MELD [52], and CARER [53].

The aforementioned methods produce logits corresponding to Ekman's six basic emotions [54], namely, anger, disgust, fear, happiness, sadness, and surprise, in addition to neutral.

4) Comparison to State-of-the-Art Methods: The models employed for unimodal labeling demonstrate performance that is near the state-of-the-art methods within their respective domains and datasets (MSP-Podcast [14] for audio and Aff-Wild2 [46] for vision). For the MSP-Podcast dataset, Naini et al. [55] conducted a comprehensive evaluation of several self-supervised learning frameworks, including HuBERT [12], Wav2Vec2 [13], Data2Vec [56], and WavLM [47], and concluded that WavLM achieved the best performance with significantly superior results. In our study, we utilized the same fine-tuned WavLM checkpoint reported in their work. Regarding the Aff-Wild2 dataset, a central benchmark for the ABAW challenges at CVPR [57]–[59], leading approaches consistently utilize Masked Autoencoder (MAE) pre-training on large-scale facial datasets [60], [61]. For example, Zheng et al. [60], winners of the 6th ABAW Challenge at CVPR 2024 for facial expression recognition, pre-trained their model on a private dataset containing millions of face images and fine-tuned it on Aff-Wild2, achieving state-of-the-art performance with an F1 score of 49.5 for the visual modality. As the official weights were not publicly available, we replicated their MAE pre-training process and achieved an F1 score of 42.6. While our performance falls slightly short of the reported best performance, we tried our best to replicate the state-of-the-

[2]https://github.com/AIM3-RUC/ABAW4
[3]https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Categorical
[4]https://huggingface.co/j-hartmann/emotion-english-roberta-large

TABLE III: Unsupervised domain adaptation experiments on generic domain.

	CMU-MOSEI (A)		CMU-MOSEI (V)	
	ACC	F1	ACC	F1
FT on [14], [46]	52.2	34.7	37.4	24.4
FT + UDA [62]	46.2	31.9	35.8	23.1

art recipe. For the text modality, we employ a model trained on an extensive and diverse corpus of text-based emotion recognition tasks. While no standardized benchmark exists for evaluating emotion recognition in spoken language, we believe the comprehensive scope of the combined corpus provides a robust and generalized checkpoint.

Given the ultimate goal of deploying these unimodal emotion recognition models in generalized settings (e.g., YouTube videos), we investigate whether adapting to a more generic domain, such as YouTube videos, can improve generalization performance. Specifically, we utilize an unsupervised domain adaptation framework [62], where the source data comprises the datasets used to train the models (Aff-Wild2 for visual and MSP-Podcast for audio). We select random videos from the VoxCeleb2 dataset for the target domain, which provides a representative sample of generic YouTube content. In addition to fine-tuning the models for emotion recognition on the source datasets, we incorporate a domain classifier coupled with a gradient reversal layer to encourage the generation of domain-invariant feature representations. To assess the effectiveness of domain adaptation, we evaluate the models on the CMU-MOSEI dataset, which is also curated from YouTube videos.

The results are presented in Table III. Notably, we observe that applying unsupervised domain adaptation [62] to a generic domain reduces generalization performance. We hypothesize that this occurs because the generic domain (*e.g.,* YouTube videos) lacks distinctive features that the model can leverage for learning while introducing the domain classifier branch may add noises that adversely affect the learning process of the emotion classifier.

5) Cross-modal Labeling: We illustrate the labeling process using the vision + text $\rightarrow$ audio direction as a representative example. The approach used in the alternate direction (audio + text $\rightarrow$ vision) is identical.

For a given utterance x, we independently generate the logits for categorical emotions with vision and text, denoted as h_v and h_t, respectively. Weak labels are retained only when the two modalities are in agreement. We compute the Kullback-Leibler (KL) divergence between the inference results from both modalities. If the KL divergence exceeds 1, we discard the data point. If the KL divergence is less than or equal to 1, we average the inference results to formulate a weak label for the audio

$$\hat{h}_a = \frac{1}{2}(h_v + h_t). \tag{1}$$

Because of the high agreement between the two distributions, significant information loss is avoided, making simple averaging an effective and straightforward method to merge predictions. The threshold of 1 is selected based on agreement with ground truth labels in CMU-MOSEI dataset and the balance of label

TABLE IV: Fine-tuning performance on MSP-Podcast-4 of models pre-trained on EmoCeleb-A using different probability distribution distance metrics.

	ACC	F1
EmoCeleb-A w/ KL-divergence	49.4	49.9
EmoCeleb-A w/ JS-divergence	49.2	50.0

distribution. The predicted category $\hat{y}_a$ is then obtained by selecting the argument with the maximum value in $\hat{h}_a$.

We compare two probability distribution distance metrics—Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence. By adjusting the JS divergence threshold, we ensure the number of training samples (in EmoCeleb-A) is approximately equal to that obtained using KL divergence. Our analysis reveals that both metrics retrieve largely overlapping samples, with about 80% of those selected by KL divergence also appearing in the JS divergence set. To further verify the quality of the samples filtered by the two metrics, we first pre-train a HuBERT-base model [12] on the EmoCeleb-A datasets collected using KL-divergence and JS-divergence, then fine-tune the resulting models on the MSP-Improv dataset [16]. The experimental results are presented in Table IV, showing the average outcomes over five runs. Statistical significance tests ($p < 0.05$) indicate no significant differences between the two metrics. Consequently, we choose KL divergence as our preferred metric.

C. Post-processing

EmoCeleb exhibits a highly imbalanced distribution of labels, particularly with a sparse representation of disgust, fear, and sadness. This scarcity is likely attributed to the nature of the VoxCeleb2 dataset, which predominantly comprises interview videos featuring celebrities. Within such contexts, expressions of these three emotions are uncommon. Thus, we remove these three emotion classes. Furthermore, to ensure that each speaker has sufficient utterances for effective training and evaluation of personalized emotion recognition models, we also discard speakers with fewer than 50 utterances.

After this procedure, we have over 150 hours of content for both directions of cross-modal labeling. Specifically, EmoCeleb-A contains 1,480 speakers with 74,100 utterances, and EmoCeleb-V includes 1,494 speakers with 75,110 utterances. Importantly, each speaker contributes a minimum of 50 utterances. Detailed statistics of the two datasets are provided in Tables I and II. Examples of emotions in EmoCeleb are shown in Figure 3.

Figure 4 illustrates the distribution of Gini coefficient values [63] for speakers in the EmoCeleb-A and MSP-Podcast datasets, offering insights into the diversity of emotional class distributions across speakers. To ensure a fair comparison, we exclude speakers from the MSP-Podcast dataset with fewer than 50 samples, matching the lower sample limit for speakers in the EmoCeleb dataset. This filtering results in 73 speakers from MSP-Podcast compared to 1.5K speakers in EmoCeleb-A.

The Gini coefficient quantifies the inequality in class distributions, with higher values indicating greater imbalance. For the

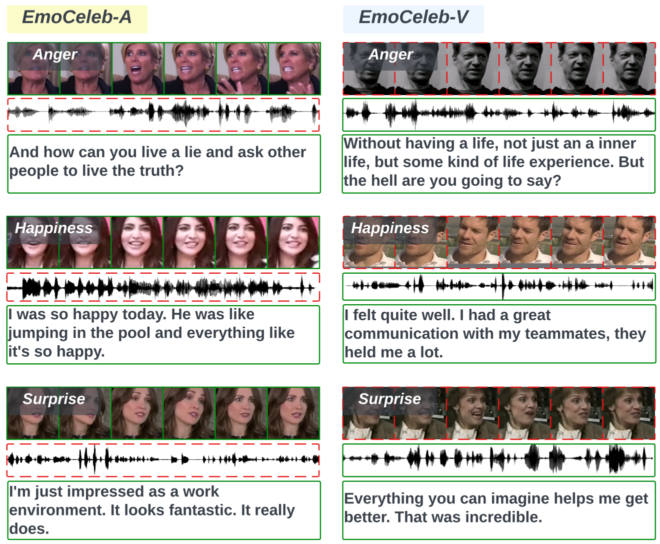


Fig. 3: Examples of emotional expressions in EmoCeleb-A and EmoCeleb-V. Green solid lines denote the modalities used for cross-modal labeling, while red dashed lines refer to the target modalities. The weak labels reflect the examples' emotions

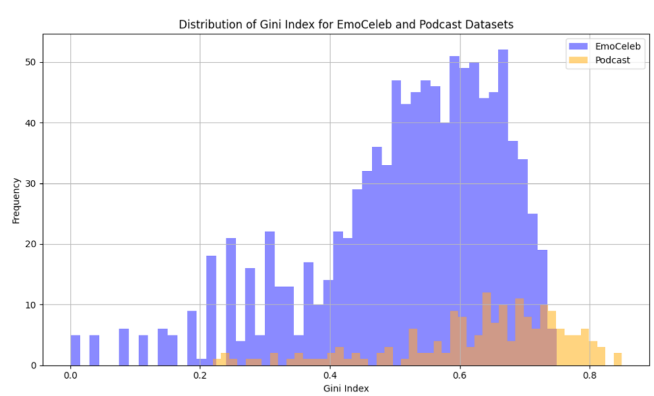


Fig. 4: Distributions of the Gini Coefficients between EmoCeleb-A and MSP-Podcast dataset (Avg Gini value for MSP-Podcast=0.63, Avg Gini value for EmoCeleb-A is 0.51).

EmoCeleb dataset (blue), Gini values are primarily concentrated between 0.2 and 0.6, with a notable peak around 0.4–0.5. This distribution reflects a relatively balanced emotional class representation among speakers, consistent with the curated nature of the dataset. In contrast, the MSP-Podcast dataset (orange) exhibits a broader range of Gini values, with a significant portion exceeding 0.6. The mean Gini coefficient for MSP-Podcast is 0.63, compared to 0.51 for EmoCeleb, further highlighting the per-speaker emotion diversity of EmoCeleb.

D. Label Quality Evaluation

We evaluate our weak labels through (i) a comparison with human annotations and (ii) a comparison of the utility of labels for model training with existing emotion recognition datasets. To maintain consistency with the label space of EmoCeleb, our analysis focuses on four emotions: anger, happiness, surprise, and neutral.

TABLE V: Weak label generation with one or more modalities. V, A, T stand for vision, audio, and text respectively.

	CMU-MOSEI		MSP-Face	
	ACC	F1	ACC	F1
Random	30.5	24.9	27.4	24.6
V	37.4	24.4	29.6	26.3
A	52.2	34.7	34.3	28.9
T	43.6	33.7	36.6	36.2
V+T	50.8	36.4	43.4	41.2
A+T	57.2	42.2	41.8	38.4
Human	70.8	49.6	69.4	69.2

TABLE VI: Comparison with existing emotion datasets. Accuracy (ACC %, ↑) and F1-score (F1 %, ↑) are the evaluation metrics. Model trained with EmoCeleb outperforms RAVDESS and CMU-MOSEI which are manually labeled.

Train dataset \ Test dataset	IEMOCAP		MSP-Face		MSP-Face	
	Audio		Audio		Vision	
	ACC	F1	ACC	F1	ACC	F1
Random	35.5	24.5	27.4	24.6	27.4	24.6
RAVDESS [17]	31.3	28.0	21.0	16.5	12.9	6.7
CMU-MOSEI [27]	39.1	29.9	27.7	20.8	32.3	18.5
MSP-Podcast [14]	53.8	38.0	39.1	34.7	-	-
EmoCeleb	48.1	31.9	35.7	30.1	33.5	26.9

1) Comparison with human annotations: The objective of this experiment is to evaluate the performance of the proposed cross-modal labeling pipeline compared to random guessing and human performance. To ensure the availability of ground-truth labels, we use existing emotion recognition datasets with annotations obtained through crowdsourcing. Specifically, we apply the labeling process to two well-established emotion datasets, CMU-MOSEI [27], and MSP-Face [26], and compare the generated weak labels against the ground-truth annotations provided by these datasets.

Since both datasets include annotations from multiple annotators, we also assess the performance of a single annotator's judgments relative to the consensus ground truth. The results of these evaluations, summarized in Table V, display expected behaviors: the weak-labeling pipeline produces label quality that is significantly better than random guessing yet still lags behind the quality of manually annotated labels.

We also provide the contribution of each modality in the labeling process in table V. The contribution of each modality to prediction accuracy varies depending on the evaluation datasets. However, combining two modalities to generate labels consistently outperforms using any single modality, as expected. This finding motivates our approach to leveraging multiple modalities to generate higher-quality weak labels.

2) Comparison with existing emotion datasets: The objective of this experiment is to evaluate the position of EmoCeleb-A/V within the dataset hierarchy presented in Table I, specifically regarding its usefulness as a source dataset for emotion recognition models. EmoCeleb-A/V offers superior diversity

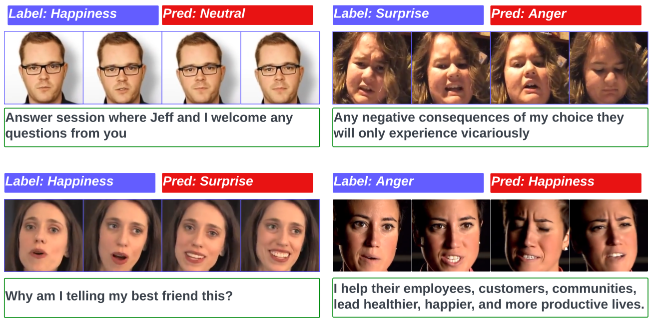


Fig. 5: Examples of CMU-MOSEI dataset [27] samples that the proposed weak labeling pipeline produced wrong labels given video and text inputs.

and scale compared to existing datasets but falls behind in label quality compared to human annotations (as demonstrated in III-D1).

We conduct a zero-shot evaluation experiment to benchmark our dataset against existing emotion recognition datasets. Specifically, we train emotion recognition models (HuBERT [12] for audio and VideoMAE [15] for vision) on one of the source datasets and evaluate their performance on different target datasets. The source datasets used in our experiments include RAVDESS [17], CMU-MOSEI [27], MSP-Podcast [14], and the proposed EmoCeleb-A/V. The target datasets selected for evaluation are IEMOCAP [25] and MSP-Face [26]. For all datasets, we limit the emotion categories to the four present in EmoCeleb-A/V: *happiness*, *anger*, *neutral*, and *surprise*.

Detailed results of the zero-shot transfer experiments are presented in Table VI. Notably, we exclude IEMOCAP [25] from the vision model evaluations due to its non-frontal face views, which introduce a significant domain gap, affecting model generalization. Similarly, MSP-Improv [16] is excluded as a target dataset for audio because it lacks the "surprise" emotion class, complicating comparative analysis.

The results reveal that EmoCeleb-A/V not only significantly surpasses random guessing (Random) but also outperforms two established emotion datasets, RAVDESS and CMU-MOSEI. These findings underscore the efficacy and utility of our weakly-labeled dataset as a valuable resource for pre-training emotion recognition models.

3) Failure cases analysis: Fig. 5 presents four instances where our labeling pipeline, applied to the CMU-MOSEI [27] dataset, yields incorrect emotion labels given video and text inputs. These errors come from the conflicting emotional cues expressed through different modalities. For example, in the top left case, the speaker's neutral expression clashes with the happiness implied by the uttered language, "Jeff and I welcome any questions." Conversely, in the bottom left case, the woman's happy expression contradicts the surprised emotion conveyed in the text. These failure cases indicate that our pipeline struggles to reconcile conflicting multimodal cues. While adjusting the KL divergence threshold could potentially reduce such inconsistencies, this approach presents a trade-off. A lower threshold might improve multimodal alignment but could also lead to a smaller dataset with a less balanced label distribution, potentially hindering overall performance.

IV. Method

The goal of the SetPeER is to learn personalized representations for emotion recognition using a set of K utterances from a single speaker. Drawing inspiration from recent advancements in set-based representation learning, our approach focuses on deriving personalized speaker representations from the input set of utterances. These personalized representations are then conditioned on the features generated by deep encoders, as illustrated in Figure 6. SetPeER comprises two main components: a multi-layer backbone encoder $\mathbf{E}$ designed to produce high-level representations from audio/visual input signals and a lightweight personalized feature extractor $\mathbf{P}$ aimed at generating personalized representations from input sets.

A. Backbone Encoder

The backbone encoder $\mathbf{E}$ produces high-level feature representations from raw audio or video inputs. Although SetPeER is applicable to many backbone encoders with transformer architecture, we adopt the widely-used HuBERT [12] and VideoMAE [15] as the backbone encoders for our audio and vision modalities, respectively. As a high-level overview, both architectures consist of two main components: a feature extractor E_0 to extract low-level representations from raw audio or video inputs and a deep encoder $\mathrm{E}^{'}$ to extract high-level representations from the extracted low-level features. For HuBERT [12], E_0 consists of several layers of 1D Convolutional Neural Networks (1D-CNN) to generate features at 20ms audio frames from raw waveforms sampled at 16kHz. For VideoMAE [15], E_0 is a space-time cube embedding that maps 3D raw video tokens to patches with a pre-specified channel dimension. The deep encoder $\mathrm{E}^{'}$ for both architectures are a stack of N transformer encoder layers [64], *i.e.*, $\mathrm{E}^{'} = \{\mathrm{E}_1, \ldots, \mathrm{E}_N\}$, where the output of the i-th layer E_i is $x_i = \mathrm{E}_i(x_{i-1})$ for $i \in [1, 2, \ldots, N]$ and x_0 is the features produced by E_0. The output of the last layer x_N is temporally mean-pooled and fed to linear layers to produce the emotion classification predictions.

B. Personalized Feature Extractor

The objective of $\mathbf{P}$ is to produce personalized embeddings given a set of utterances. As mentioned in Section II-C, a key property of set-based learning is *permutation invariance*, *i.e.*, the output for a set remains the same regardless of the ordering of the input. We follow the previous work [38], [39] and use permutation-invariant modules to build the personalized feature extractor $\mathbf{P}$. Specifically, $\mathbf{P}$ consists of several linear layers to reduce the dimensionality of the inputs, a transformer encoder layer (without positional encoding), and a Vector Quantization module [65] to discretize the learned representations into meaningful concepts. Formally, we want to extract personalized features for each encoder layer in $\mathrm{E}^{'}$, given a set of utterances of the same speaker $\mathcal{S}_x = \{x^1, x^2, \ldots, x^K\}$. For the first encoder layer, $\mathbf{P}$ takes as inputs p_0 the temporally mean-pooled features extracted from E_0 while for the remaining layers, $\mathbf{P}$ takes as inputs p_l the temporally mean-pooled features extracted from

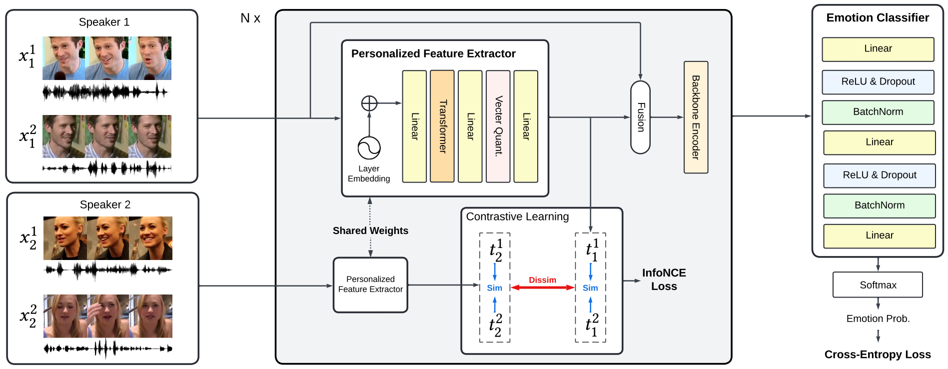


Fig. 6: SetPeER overview. The personalized feature extractor **P** generates layer-specific personalized embeddings from the input and feeds the embeddings to the backbone encoder layer. These personalized embeddings serve as contextual cues for the current layer, aiding in generating more targeted features. The weights of **P** are shared across layers. Additionally, we apply contrastive learning for embeddings generated from **P** to enhance the consistency in producing personalized speaker embeddings.

the l-th layer E_l in $\mathrm{E}^{'}$. In other words, $p_1 = \mathrm{Pool}(\mathrm{E}_0(\mathcal{S}_x))$ and $p_l = \mathrm{Pool}(x_{l-1})$. The dimension of p_l is $\mathcal{R}^{K\times D}$, where K is the size of the input set and D is the feature dimension. Given p_l, **P** extracts the speaker embeddings for the set as follows.

1) Dimensionality reduction: We want to keep the parameter count of SetPeER analogous to the original encoders to demonstrate the effectiveness of the proposed method. Hence, we first use a linear layer L_1 to reduce the dimensionality of the inputs from D to C and share **P** across all layers of **E** using a learnable layer embedding $\Phi(\cdot)$

$$q_l = \mathrm{L}_1(p_l + \Phi(l)). \tag{2}$$

2) Contextualized feature learning: Then, we leverage a transformer encoder layer T [64] to generate contextualized representations for the set of processed vectors. We do not add any positional encoding to the q_l to ensure permutation invariance.

$$r_l = \mathrm{T}(q_l). \tag{3}$$

3) Personalized embedding generation: Next, we average the produced contextualized representations to generate a single vector representing the set and use another linear layer L_2 to resize the generated embedding to a target output dimension of size $Q \times C$, where Q denotes the number of embeddings per speaker we want to extract.

$$s_l = \mathrm{L}_2(\mathrm{Pool}(r_l, dim = 0)). \tag{4}$$

4) Quantized Speaker Representation Codebooks: VQ-VAE is a popular technique for learning a quantized codebook of image elements, facilitating the autoregressive synthesis of images. We extend Vector Quantization (VQ) [65] to create personalized speaker embeddings with two main motivations. First, certain individual attributes, such as race, gender, and age, are inherently discrete. Moreover, VQ facilitates the creation of compact and generalized feature representations by filtering out irrelevant information from the continuous space.

For VQ, we use a discrete codebook $\mathcal{Z} = \{z_i\}_{i=1}^{M}$ where $z_i \in \mathcal{R}^C$ to generate Q embeddings from s_l, where M denotes the number of entries in the codebook. In particular, we first reshape s_l into $\mathcal{R}^{Q\times C}$. Then, for each personalized vector of size C in s_l, we look up the nearest entry j in $\mathcal{Z}$ and output the corresponding representation z_j for the entry. During back-propagation, we use a straight-through gradient as in [65]. Finally, we use a linear layer L_3 to map the produced speaker embeddings from C to D.

$$t_l = \mathrm{L}_3(\mathrm{VQ}(s_l)) \in \mathcal{R}^{Q\times D}. \tag{5}$$

C. Training Scheme

In section IV-B, we present the personalized embedding extraction process of **P** for a single speaker. In each training step, SetPeER receives B sets of labeled utterances, each representing a speaker and consisting of K utterances. We utilize **P** to derive personalized speaker embeddings for every layer of E. These embeddings are concatenated with contextualized features extracted from the previous layer (or features from E_0 for the first layer), thereby integrating personal information into the features generated at each layer. This technique is commonly called *Prefix Tuning* [66].

$$x_l = \mathrm{E}_l([t_l\ ;\ x_{l-1}]), \tag{6}$$

where E_l is the l-th layer of E'. We later show the difference in fusion strategies between the extracted personalized embeddings and the deep, contextualized features in an ablation study. Finally, the encoder's output is temporally mean-pooled and fed into a linear head to predict emotions relative to the ground-truth labels using the cross-entropy loss $\mathcal{L}_{CE}(\tilde{y}, y)$.

Consistency-aware embedding generation. Ideally, **P** should produce identical outputs given two sets of utterances from the same speaker. Hence, to enhance the consistency of **P** in producing personalized speaker embeddings, we propose to use contrastive representation learning. Specifically, given the input $p \in \mathcal{R}^{K\times D}$ for the personalized feature extractor, we split it into two equal subsets p^1 and $p^2 \in \mathcal{R}^{\frac{K}{2}\times D}$. We use **P** to extract the speaker embeddings t^1 and t^2 for these two sets.

TABLE VII: Speech emotion recognition on EmoCeleb-A and downstream datasets. Accuracy (ACC %, ↑) and F1-score (F1 %, ↑) are the evaluation metrics. Average accuracy (A-ACC %, ↑) and average F1-score (A-F1 %, ↑) across speakers are also reported. * denotes statistical significance ($p < 0.05$) based on 5 runs.

	EmoCeleb-A				MSP-Podcast-4				MSP-Podcast-8				MSP-Improv			
Method	ACC	A-ACC	F1	A-F1	ACC	A-ACC	F1	A-F1	ACC	A-ACC	F1	A-F1	ACC	A-ACC	F1	A-F1
HuBERT [12]	47.7	51.2	46.5	41.8	47.3	46.2	49.0	43.3	24.5	25.5	22.4	21.4	54.1	54.0	51.8	51.5
HuBERT-PT	-	-	-	-	49.4	46.1	49.9	42.0	24.4	26.8	23.9	24.1	56.0	55.7	53.8	53.3
Sridhar *et al.* [8]	48.3	52.0	46.9	41.8	49.0	46.7	49.5	42.2	25.0	27.5	24.8	24.2	55.7	54.9	53.1	52.5
PAPT [1]	48.6	53.4	47.1	42.1	50.0	48.3	50.9	43.5	25.2	27.4	24.8	24.4	56.2	56.0	53.6	53.4
SetPeER (ours)	**50.1***	**54.4**	**49.0***	**45.4***	**51.7***	**51.1***	**52.6***	**46.9***	**26.2**	**28.4***	**26.1***	**25.0**	**57.2**	**57.6***	**54.0**	**53.9**

TABLE VIII: Visual emotion recognition on EmoCeleb-V and MSP-Improv. SetPeER surpasses the baseline methods across all evaluated metrics. * denotes statistial significance ($p < 0.05$) based on 5 runs.

	EmoCeleb-V				MSP-Improv			
Method	ACC	A-ACC	F1	A-F1	ACC	A-ACC	F1	A-F1
VideoMAE [15]	38.6	33.6	36.7	27.0	52.8	52.0	49.9	48.2
VideoMAE-PT	-	-	-	-	54.1	54.3	52.7	51.7
Sridhar *et al.* [8]	39.0	33.5	37.6	27.4	54.2	54.2	52.5	52.1
PAPT [1]	39.2	33.4	38.0	27.2	54.5	54.7	53.0	52.8
SetPeER (ours)	**39.2**	**34.2***	**38.7***	**28.0**	**57.5***	**55.6***	**56.6***	**55.2***

We enhance **P**'s ability to extract consistent features with an InfoNCE contrastive loss [67].

$$\mathcal{L}_{NCE} = -\frac{1}{B}\sum_{i=1}^{B}\log[\frac{\exp(t_i^1 \cdot t_i^2/\tau)}{\sum_{i\neq k}\exp(t_i^1 \cdot t_k^2/\tau) + \exp(t_i^1 \cdot t_i^2/\tau)}], \tag{7}$$

where B represents the number of speakers we use for training in one batch and τ stands for the temperature parameter.

Overall, SetPeER is optimized with the following loss function with hyper-parameters λ_1, λ_2, and λ_3

$$\mathcal{L} = \lambda_1\mathcal{L}_{CE} + \lambda_2\mathcal{L}_{NCE} + \lambda_3\mathcal{L}_{VQ}, \tag{8}$$

where $\mathcal{L}_{VQ}$ is the commitment loss associated with Vector Quantization. More details on the commitment loss are in [65].

V. Experiments

A. Implementation and Training Details

All methods are implemented in PyTorch [68]. We provide code in the supplementary materials. The code and datasets will be published upon acceptance.

1) Model architecture: We adopt HuBERT-base [12] and VideoMAE-tiny [15] as our audio and vision encoders, respectively. These models are widely used foundational backbones across various audio and vision tasks. It is important to note we aim to develop and validate a general model suitable for personalization across various backbone architectures. Consequently, we select two widely used backbones across various audio and vision tasks but with a relatively small number of parameters for efficient training. Both architectures consist of 12 transformer encoder layers with the feature dimension $D = 768$ for HuBERT and $D = 384$ for VideoMAE. We use the same personalization network **P** for both audio and vision experiments, in which $C = 64$, $Q = 4$, and $M = 512$. This results in ~ 400K additional trainable parameters, about 0.5% of the number of parameters of HuBERT-base [12] and 1.2% of the number of parameters of VideoMAE-tiny [15].

2) Model training: We optimize the network weights using the AdamW optimizer [69] on a single NVIDIA Quadro RTX8000 GPU. The weight decay is $1e^{-4}$. The gradient clipping is 1.0. We train all the models for 100 epochs with a learning rate of $3e^{-5}$. We set $\lambda_1 = 1.0, \lambda_2 = 0.1, \lambda_3 = 0.1$ for training loss weights. It is important to note that the parameters λ_1, λ_2, and λ_3 are selected through hyper-parameter tuning. Setting λ_2 too high prioritizes learning speaker-specific attributes at the expense of overall emotion recognition ability, leading to degraded performance. Conversely, setting λ_2 too low encourages the model to focus on general emotion recognition without considering personalized features, which also results in reduced performance, as demonstrated empirically in Section V-C.

To facilitate set learning, our data loaders are designed at the speaker level. Specifically, during training, a batch comprises B speakers, each composed of a set of K utterances randomly drawn from all utterances of the corresponding speaker. Consequently, within an epoch, SetPeER encounters every speaker in the dataset, though not necessarily all utterances. During testing, we conduct inference on one speaker at a time, *i.e.*, $B = 1$, accommodating varying numbers of utterances (K) per speaker. However, we ensure the model never encounters more than K utterances within a single batch. In all our experiments, we set $B = 8$ and $K = 8$ during training.

3) Experiment overview & Notations: We utilize EmoCeleb-A/V in two experimental scenarios: (1) as a new personalized emotion recognition (ER) evaluation dataset, leveraging its suitability as a benchmark due to its large number of speakers, substantial samples per speaker, and diverse emotional labels per speaker, and (2) as a pre-training dataset for personalization. In the latter scenario, we first train SetPeER and baseline models on the EmoCeleb-A/V dataset, followed by fine-tuning these pre-trained personalized models on existing personalized ER benchmarks (*e.g.,* MSP-Podcast [14] for audio and MSP-Improv [16] for vision). The primary experimental results are provided in Tables VII and VIII. The first four columns of these tables present results when using EmoCeleb-A/V as the evaluation dataset, while the remaining columns show results when EmoCeleb-A/V is used for pre-training personalized models, which are subsequently evaluated on other datasets. We provide details of the used baselines in Section V-C1.

We evaluate SetPeER across multiple datasets in this study (as detailed in Section V-B), each featuring a distinct number of emotional categories. For the MSP-Podcast dataset, we additionally report results on the four emotional categories that overlap with EmoCeleb-A/V (happiness, anger, neutral, and surprise). This allows us to benchmark the effectiveness of EmoCeleb-A/V as a pre-training dataset under both same-emotion-class (MSP-Podcast-4) and different-emotion-class (MSP-Podcast-8) settings.

B. Datasets

We divide EmoCeleb into train, validation and test sets with a distribution ratio of 70%, 10%, and 20%, respectively, on a speaker-independent basis. This means speakers included in the training set are excluded from the validation and test sets to ensure no overlap. Additionally, we perform experiments on two benchmark emotion datasets, *i.e.*, MSP-Improv [16] and MSP-Podcast [14] to evaluate the utility of our weakly-labeled dataset and the effectiveness of our proposed method.

While MSP-Podcast has been used in prior research on personalized speech emotion recognition [1], [8], no suitable dataset has emerged with both high-quality visual data and a diverse pool of speakers for audiovisual emotion recognition experiments. Existing datasets like CMU-MOSEI and MSP-Face offer visual information with a large speaker pool; however, CMU-MOSEI lacks speaker identity labels, while MSP-Face yields performance akin to random guessing [26]. Consequently, for visual evaluation, we opted for MSP-Improv alongside EmoCeleb-V. Although MSP-Improv features a small number of speakers (12), it remains a popular choice in current visual and audio-visual emotion recognition literature.

- **MSP-Improv** is an acted audiovisual emotional database that explores emotional behaviors during acted and improvised dyadic interaction [16]. The dataset consists of 8,438 turns (over 9 hours) of emotional sentences, categorized into four primary emotions: neutral, happiness, sadness, and anger. The corpus has six sessions, and each session has one male and one female speaker (12 in total). We use sessions $1-4$ as the training set, session 5 as the validation set, and session 6 as the testing set.
- **MSP-Podcast** is the largest corpus for speech emotion recognition in English. The dataset contains speech segments from podcast recordings. Each utterance in the dataset is annotated using crowd-sourcing with continuous labels of arousal, valence and dominance in addition to the categorical emotions. In this paper, we exclude any samples that lack speaker identification. This refinement process yields a total of 42,541 utterances, encompassing over 71 hours of emotional speech content. The corpus provides the official data split and has eight emotion classes: neutral, happiness, sadness, anger, surprise, fear, disgust, and contempt. We conduct the downstream evaluation in two ways: (i) we use the subset with the four emotion categories in EmoCeleb (MSP-Podcast-4); (ii) we use all the eight emotion categories (MSP-Podcast-8).

C. Experimental Results

1) Quantitative Analysis: We pre-train SetPeER on EmoCeleb and then fine-tune it on the downstream datasets with supervised emotion recognition. Thus, we report the model performance on both EmoCeleb and downstream datasets. Accuracy (ACC %, ↑) and F1-score (F1 %, ↑) are the evaluation metrics. Additionally, we report the average accuracy (A-ACC ↑) and the average F1-score (A-F1 ↑) across speakers.

Four baseline methods are implemented and compared. We do not benchmark our method against existing approaches tuned for maximum within-domain performance with more complex backbone architectures, as the backbone in SetPeER can be interchangeable.

- **Vanilla backbones.** We train HuBERT / VideoMAE on EmoCeleb and downstream datasets with the official checkpoints.
- **Pre-trained backbones (PT)** This baseline represents the performance of the backbone models with an intermediate pre-training stage on the collected EmoCeleb-A/V datasets. This serves as an ablation to isolate the contribution of our dataset creation method alone, allowing us to distinguish the impact of dataset pre-training from that of our proposed architectural improvements.
- **PAPT** [1] trains speaker embeddings on an extensive set of training speakers in a self-supervised fashion. These embeddings are then incorporated into the generated features via prefix tuning for personalized emotion recognition. In the testing phase on unseen speakers, the method identifies the most closely aligned speakers from the training set and uses the corresponding trained embeddings to generate personalized features. SetPeER differs from PAPT in two key aspects: Firstly, while PAPT requires two iterations for personalization, our model can be trained directly with labels, bypassing the need for initial self-supervised training. Secondly, PAPT relies on a diverse and large training speaker set for matching unseen speakers, whereas our dataset performs well with fewer speakers (see Table VIII). Efficiency-wise, our model eliminates the need to match each test speaker with every training speaker, substantially reducing inference time. Nevertheless, as far as we know, PAPT remains the only personalization method for unseen speakers without retraining any components. For a fair comparison, we initialize the backbone encoder of PAPT with our pre-trained backbone (PT) on EmoCeleb.
- **Sirdhar *et al.* [8]** propose performing speaker matching, in which for each unseen speaker, the method finds the closest speakers in the train set and retrains the original model with more weights on the selected speakers for more personalized predictions. For fair comparisons, we use features extracted from HuBERT [12] pre-trained on EmoCeleb. It is important to note that the method

TABLE IX: Demographics (EmoCeleb-A). # S: #
(%) per demographic category.

Group	# S	HuBERT	PAPT
Caucasian (American)	452	47.2	48.1
Caucasian (European)	535	47.7	48.2
Caucasian (Australian)	63	39.6	40.3
East Asian	36	35.9	35.9
South Asian	122	45.4	**47.9**
African	52	44.9	45.7
African American	108	39.3	**42.1**

required model re-training and is not directly
with PAPT [1] and SetPeER.

The audio and vision performances are provide
VII and VIII, respectively. Results in both table
our proposed method outperforms all other competi
across various metrics. We can observe that H
consistently outperforms HuBERT across both
visual experiments. This underscores the suitabi
datasets, EmoCeleb-A and EmoCeleb-V, not only
evaluation datasets for personalization but also as promising resources for large-scale pre-training in emotion recognition tasks. SetPeER further boosts the performance of HuBERT-PT by a large margin, especially in the per-speaker metrics (A-ACC and A-F1), demonstrating the effectiveness of the proposed personalized feature extraction pipeline. Compared to PAPT [1], we not only demonstrate superior performance overall but also remain effective on the MSP-Improv dataset with a limited number of training speakers (ten speakers). On the other hand, PAPT only achieves marginal improvements over HuBERT-PT on the MSP-Improv dataset for both audio and visual modalities, indicating its limitations when confronted with a small pool of training speakers.

A demographic breakdown of the EmoCeleb-A dataset and per-demographic performance is provided in Table IX. We prompted Chat-GPT to collect celebrities' demographics. The demographic distribution is unbalanced (due to VoxCeleb's language). However, the dataset is still relatively diverse compared to existing ones. The model's performance with demographic info compared to the baselines is provided; our method outperforms the baselines for most groups.

2) Qualitative Analysis: To understand the information learned in speaker embedding, we inspect the information learned by the personalized encoder **P**. In particular, we investigate the relation between the extracted speaker embeddings and gender, which is the only demographic information available for the MSP-Podcast dataset [17]. Figure 7 displays the 2D T-SNE visualizations [70] of speaker embeddings from MSP-Podcast, with each point representing an utterance[5]. Colors denote gender, with blue representing male and orange representing female. It is evident that SetPeER can generate linearly separable features with respect to gender, even without explicit gender labels. This not only showcases SetPeER's capability in producing useful personalized features but also underscores the significance of gender in emotion recognition, aligning with the literature [71].

[5]We cannot produce the T-SNE plot for our visual model due to the limited speaker pool of MSP-Improv.

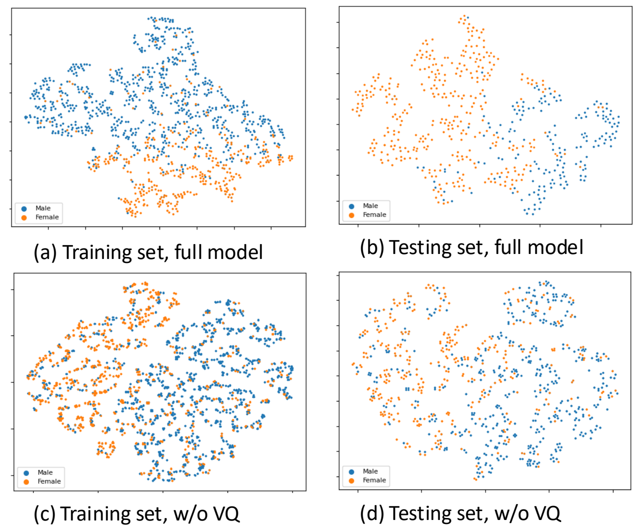


Fig. 7: t-SNE visualizations of speaker embeddings from MSP-Podcast. Blue points represent male speakers and orange points indicate female speakers. Representations by SetPeER (first row) show clear separation w.r.t. gender.

TABLE X: Ablations for SetPeER. Fusion refers to fusing speaker embedding with audiovisual features for personalized emotion recognition. A and V stand for audio and vision modalities respectively.

Modules	MSP-Podcast-4 (A) ACC	F1	MSP-Improv (V) ACC	F1
SetPeER	51.7	52.6	57.3	56.7
$-\mathcal{L}_{NCE}$	51.3	51.6	55.4	54.9
$-VQ$	51.0	51.8	56.4	55.1
Fusion Strategy				
Prefix [66]	51.7	52.6	57.3	56.7
Addition	50.9	51.4	53.8	48.7
Cross-attn [72]	48.2	49.5	55.9	52.8

3) Ablations: We perform extensive ablation studies to demonstrate the effectiveness of each component, as shown in Table X.

- **Contrastive loss $\mathcal{L}_{NCE}$.** Removing the contrastive loss leads to notable performance degradation, with approximately a 2% decrease in both accuracy and F1 score on MSP-Improv (V). This underscores the importance of maintaining uniform representations across various inputs from the same speaker.
- **Vector Quantization.** Quantizing personalized speaker embeddings proves to be effective, increasing the F1 metric by 1% on the MSP-Podcast-4 (A) and 1.8% on the MSP-Improv (V) dataset. Furthermore, in Figures 7(a) and 7(b), we can see a clear degradation in cluster quality when a model is trained without the VQ module.
- **Fusion strategy.** Information in speaker embedding is fused with the input to provide personalized emotion

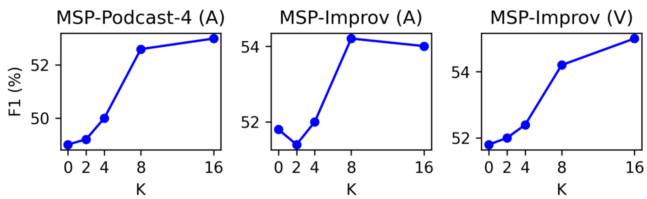


Fig. 8: Impact of set size K on performance. Larger set sizes lead to higher performance.

TABLE XI: SetPeER performance on the MSP-Podcast-4 dataset with different weak labeling strategies.

	Acc	A-Acc	F1	A-F1
$A \to A$	47.6	46.8	48.8	42.4
$V \to A$	49.2	48.6	50.1	44.5
$T \to A$	48.3	47.8	49.5	42.9
$T + V \to A$	**51.7**	**51.1**	52.6	**46.9**
$A + V \to A$	51.0	50.2	50.9	45.2
$T + A \to A$	51.5	50.9	**52.8**	46.1
$A + V + T \to A$	50.1	47.8	49.8	43.5

recognition. This work used Prefix Tuning for this purpose. In addition to Prefix Tuning [66], which temporally prepend t_l with x_{l-1}, we explore two other fusion strategies, namely addition and cross-attention [72]. For addition, we set $Q = 1$ and directly add t_l to x_{l-1}. For cross-attention, we adapt the cross-attention formulation proposed by Tsai *et al.* [72], where keys and values are x_{l-1} and queries are t_l. Overall, Prefix Tuning exhibits notably superior performance compared to the other two fusion strategies. The discrepancy likely arises because Addition is constrained by a fixed number of embeddings ($Q = 1$), whereas Cross-attention suffers from information loss.

- **Set size K.** In Figure 8, we investigate the impact of set size (K) on SetPeER's learning process. Ideally, a larger set size enables more precise construction of personalized information, leading to more accurate predictions. However, the practicality of having a large set size is often limited by the availability of samples per speaker. Therefore, finding the optimal value for K that balances performance and practicality is crucial. As expected, SetPeER becomes increasingly effective as K increases, yet the returns appear to diminish at $K = 8$.

4) Effects of Cross-modality labeling: We investigate the impact of different modality labeling strategies on the quality of pre-trained personalized emotion recognition models. Specifically, we use audio as the target label modality and construct pre-training datasets by combining weak labels from audio, video, and text. Using the proposed SetPeER architecture, we pre-train on the curated weak labels and subsequently fine-tune the models on the MSP-Podcast-4 dataset to identify the modality labeling strategies that achieve the best performance.

The experimental results are presented in Table XI. Notably, the proposed cross-modal labeling strategy, which uses two auxiliary modalities to label the target modality, produced the most effective pre-training checkpoints. This outcome can be attributed to several factors. For single-modality labeling (*e.g.*, $V \to A$), the resulting labels tend to be of lower quality due to their reliance on the performance of unimodal emotion recognition models, which often degrade significantly in out-of-domain settings. Conversely, labeling strategies that include the target modality (*e.g.*, $A + T \to A$) predominantly capture easy samples—cases where the speech emotion recognition model agrees with another modality—which limits their effectiveness. Prior work has shown that training mostly on easy-to-learn samples does not result in robust model performance [73]. Cross-modal labeling strikes a balance between these limitations. In particular, $V + T \to A$ avoids the pitfalls of low-quality labels by retaining only high-agreement samples between visual and textual emotion recognition models. At the same time, it reduces the likelihood of retrieving overly simplistic samples, as an emotion that is evident in audio and text may not necessarily be straightforward in speech emotion recognition.

VI. Limitations

Our work has several limitations. First, our cross-modal labeling pipeline does not explicitly consider the relationships between different emotions, treating all misclassifications equally. This could lead to suboptimal label quality as some misclassifications might be more acceptable than others, *e.g.*, misclassifying sadness as neutral might be less problematic than misclassifying anger as happiness). In future work, we plan to explore alternative metrics that account for inter-emotion relationships, potentially leading to more nuanced consistency assessments between modalities.

Second, our current cross-modal labeling approach is limited to generating single-modality labels using two other modalities, *e.g.*, vision + text → audio. This is because transferring from a single modality to multimodal labels resulted in low-quality annotations, *e.g.*, text → audio + vision. Developing more robust cross-modal transfer techniques that can reliably generate multi-modal labels could further improve the quality and utility of our dataset.

Finally, SetPeER requires a sufficient number of utterances per speaker (at least eight in our experiments) for effective set learning. This limits its applicability to scenarios where fewer utterances are available per speaker. Addressing this limitation by developing techniques that can effectively learn from limited data is an important direction for future research.

VII. Conclusions

In this study, we introduce SetPeER, a modality-agnostic framework designed for personalized emotion recognition. Our approach leverages cross-modal labeling to curate a large dataset for both training and evaluating personalized emotion recognition models. We present an innovative personalized architecture, enhanced with set learning, which is adept at efficiently learning distinctive speaker features. Through comprehensive experiments, we showcase the utility of the EmoCeleb dataset and the superior efficacy of the proposed method for personalized emotion recognition, outperforming baseline models on the MSP-Podcast and MSP-Improv benchmarks.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. 2211550. The work was also sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. Results presented in this paper were obtained using CloudBank [74], which is supported by the National Science Foundation under award #1925001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] M. Tran, Y. Yin, and M. Soleymani, "Personalized adaptation with pre-trained speech encoders for continuous emotion recognition," *Proc. INTERSPEECH 2023*, 2023.

[2] N. Jia and C. Zheng, "Two-level discriminative speech emotion recognition model with wave field dynamics: A personalized speech emotion recognition method," *Computer Communications*, vol. 180, pp. 161–170, 2021.

[3] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6442–6446, IEEE, 2022.

[4] M. Shahabinejad, Y. Wang, Y. Yu, J. Tang, and J. Li, "Toward personalized emotion recognition: A face recognition based attention method for facial emotion recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–5, IEEE, 2021.

[5] D. Matsumoto and H. C. Hwang, "The cultural bases of nonverbal communication," in *APA handbook of nonverbal communication*, APA handbooks in psychology®, pp. 77–101, Washington, DC, US: American Psychological Association, 2016.

[6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] M. Rescigno, M. Spezialetti, and S. Rossi, "Personalized models for facial emotion recognition through transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35811–35828, 2020.

[8] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, 2022.

[9] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[10] J.-B. Kim and J.-S. Park, "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition," *Engineering applications of artificial intelligence*, vol. 52, pp. 126–134, 2016.

[11] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, "Speech emotion recognition adapted to multimodal semantic repositories," in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 31–35, IEEE, 2018.

[12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[14] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[15] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10078–10093, 2022.

[16] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[17] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[18] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, p. 34, 2012.

[19] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, *et al.*, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*, pp. 488–500, Springer, 2007.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8, IEEE, 2013.

[21] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.

[22] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *2010 IEEE international conference on multimedia and expo*, pp. 1079–1084, IEEE, 2010.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[24] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*, pp. 865–868, IEEE, 2008.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[26] A. Vidal, A. Salman, W.-C. Lin, and C. Busso, "Msp-face corpus: a natural audiovisual emotional database," in *Proceedings of the 2020 international conference on multimodal interaction*, pp. 397–405, 2020.

[27] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

[28] S. Zhao, G. Ding, J. Han, and Y. Gao, "Personality-aware personalized emotion recognition from physiological signals.," in *IJCAI*, pp. 1660–1667, 2018.

[29] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–18, 2019.

[30] S. Rayatdoost, Y. Yin, D. Rudrauf, and M. Soleymani, "Subject-invariant eeg representation learning for emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3955–3959, 2021.

[31] J. Bang, T. Hur, D. Kim, T. Huynh-The, J. Lee, Y. Han, O. Banos, J.-I. Kim, and S. Lee, "Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments," *Sensors*, vol. 18, no. 11, p. 3744, 2018.

[32] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[33] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, "Unsupervised domain adaptation for personalized facial emotion recognition," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 128–135, 2014.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[35] P. Barros, G. Parisi, and S. Wermter, "A personalized affective memory model for improving emotion recognition," in *International Conference on Machine Learning*, pp. 485–494, PMLR, 2019.

[36] P. Barros, E. Barakova, and S. Wermter, "Adapting the interplay between personalized and generalized affect recognition based on an unsupervised neural framework," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1349–1365, 2020.

[37] P. Barros and A. Sciutti, "Ciao! a contrastive adaptation mechanism for non-universal facial expression recognition," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, IEEE, 2022.

[38] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in neural information processing systems*, vol. 30, 2017.

[39] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International conference on machine learning*, pp. 3744–3753, PMLR, 2019.

[40] K. Skianis, G. Nikolentzos, S. Limnios, and M. Vazirgiannis, "Rep the set: Neural networks for learning set representations," in *International conference on artificial intelligence and statistics*, pp. 1410–1420, PMLR, 2020.

[41] D. dan Guo, L. Tian, M. Zhang, M. Zhou, and H. Zha, "Learning prototype-oriented set representations for meta-learning," in *International Conference on Learning Representations*, 2021.

[42] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 292–301, 2018.

[43] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[44] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[45] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570, 2016.

[46] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.

[47] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[48] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, pp. 28492–28518, PMLR, 2023.

[49] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[50] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[51] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17, 2018.

[52] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019.

[53] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "Carer: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3687–3697, 2018.

[54] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[55] A. R. Naini, M. A. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12031–12035, IEEE, 2024.

[56] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, pp. 1298–1312, PMLR, 2022.

[57] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2328–2336, 2022.

[58] D. Kollias, P. Tzirakis, A. Baird, A. Cowen, and S. Zafeiriou, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5889–5898, 2023.

[59] D. Kollias, P. Tzirakis, A. Cowen, S. Zafeiriou, I. Kotsia, A. Baird, C. Gagne, C. Shao, and G. Hu, "The 6th affective behavior analysis in-the-wild (abaw) competition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4587–4598, 2024.

[60] W. Zhang, F. Qiu, C. Liu, L. Li, H. Du, T. Guo, and X. Yu, "Affective behaviour analysis via integrating multi-modal knowledge," *arXiv preprint arXiv:2403.10825*, 2024.

[61] W. Zhou, J. Lu, C. Ling, W. Wang, and S. Liu, "Boosting continuous emotion recognition with self-pretraining using masked autoencoders, temporal convolutional networks, and transformers," *arXiv preprint arXiv:2403.11440*, 2024.

[62] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, pp. 1180–1189, PMLR, 2015.

[63] C. Gini, "Measurement of inequality of incomes," *The economic journal*, vol. 31, no. 121, pp. 124–125, 1921.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[65] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[66] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

[67] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[68] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NeurIPS Autodiff Workshop*, 2017.

[69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, (New Orleans, LA, USA), OpenReview, 2019.

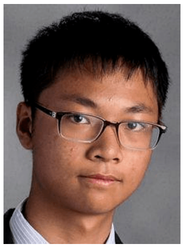
[70] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[71] M. Sidorov, A. Schmitt, E. Semenkin, and W. Minker, "Could speaker, gender or age awareness be beneficial in speech-based emotion recognition?," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 61–68, 2016.

[72] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, p. 6558, NIH Public Access, 2019.

[73] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

[74] M. Norman, V. Kellen, S. Smallen, B. DeMeulle, S. Strande, E. Lazowska, N. Alterman, R. Fatland, S. Stone, A. Tan, K. Yelick, E. Van Dusen, and J. Mitchell, "Cloudbank: Managed services to simplify cloud access for computer science research and education," in *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions*, PEARC '21, (New York, NY, USA), Association for Computing Machinery, 2021.



Minh Tran is currently a fifth year Ph.D. candidate in computer science from the University of Southern California under the supervision of Prof. Mohammad Soleymani. He received his bachelor's degree in Computer Science from University of Rochester in 2020. His research interests are Affective Computing, Self-supervised Learning, and Machine Learning.

Yufeng Yin is currently a machine learning engineer at Google Search Notification team. He received his Ph.D. in computer science from the University of Southern California in 2024 under the supervision of Prof. Mohammad Soleymani. He received his bachelor's degree in Computer Science from Tsinghua University in 2019. His research interests are Affective Computing, Multimodal Learning, and Machine Learning.

Mohammad Soleymani (S'04, M'12, SM'19) is a research associate professor of computer science with the USC Institute for Creative Technologies. He received his PhD in computer science from the University of Geneva in 2011. From 2012 to 2014, he was a Marie Curie fellow at Imperial College London. Prior to joining ICT, he was a research scientist at the Swiss Center for Affective Sciences, University of Geneva. His main line of research involves machine learning for emotion recognition and behavior understanding. He is a recipient of the Swiss National Science Foundation Ambizione grant and the EU Marie Curie fellowship. He has served on multiple conference organization committees and editorial roles, most notably as associate editor for the IEEE Transactions on Affective Computing (2015-2021), general chair for ICMI 2024 and ACII 2021 and technical program chair for ACM ICMI 2018 and ACII 2017. From 2019-2022, he served as the president of the Association for the Advancement of Affective Computing (AAAC). He is a member of the ACM, the AAAC and a Senior Member of the IEEE.