SAAML: A Framework for Semi-supervised Affective Adaptation via Metric Learning

Minh Tran*
University of Southern California
Los Angeles, CA, USA
minhntra@usc.edu

Yelin Kim
Amazon Lab126
Sunnyvale, CA, USA
kimyelin@amazon.com

Che-Chun Su
Amazon Lab126
Bellevue, WA, USA
ccsu@amazon.com

Cheng-Hao Kuo
Amazon Lab126
Bellevue, WA, USA
chkuo@amazon.com

Mohammad Soleymani
University of Southern California
Los Angeles, CA, USA
soleymani@ict.usc.edu

ABSTRACT

Socially intelligent systems such as home robots should be able to perceive emotions and social behaviors. Affect recognition datasets have limited labeled data, and existing large unlabeled datasets, e.g., VoxCeleb2, suitable for pre-training, mostly contain neutral expressions, limiting their application to affective downstream tasks. We introduce a novel Semi-supervised Affective Adaptation framework via Metric Learning (SAAML) to adapt pre-trained audiovisual models (e.g., AV-HuBERT) to expressive behaviors associated with emotions and social communication. The proposed framework automatically retrieves a large number of emotional excerpts (> 100 hours) from the VoxCeleb2 dataset via metric learning from two emotion recognition datasets (MSP-IMPROV and CREMA-D), and learns domain-invariant emotion-aware representations. Experimental results show that fine-tuning the proposed affect-aware AV-HuBERT (AW-HuBERT) improves the emotion recognition accuracy by 3-6% compared to fine-tuning the original pre-trained models. We further validate the effectiveness of the AW-HuBERT on human-centered visual understanding tasks, namely, facial expression recognition, video highlight detection, and continuous emotion recognition. The proposed approach consistently outperforms AV-HuBERT and delivers competitive performance compared to the existing methods. With this work, we demonstrate the effectiveness of adaptive pre-training for existing models on domain-specific data to enhance their performance for human-centered tasks.

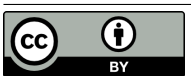
CCS CONCEPTS

• **Computing methodologies** → **Computer vision representations**.

KEYWORDS

affective computing, domain adaptation, pre-training, audiovisual learning

*This work was partly done during an internship at Amazon Lab126.



MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612286

ACM Reference Format:
Minh Tran, Yelin Kim, Che-Chun Su, Cheng-Hao Kuo, and Mohammad Soleymani. 2023. SAAML: A Framework for Semi-supervised Affective Adaptation via Metric Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612286

1 INTRODUCTION

With the popularity of socially intelligent systems such as voice assistants and home robots, affect-aware technologies have become a key component to facilitate human-machine interactions. Machines with a better understanding of humans' emotions are able to generate intelligent responses to enhance users' engagement [61] and trust [25]. In response, significant progress has been made to infer emotions and sentiments from spoken language (textual) [35, 66], visual behaviors (facial expressions and gestures) [42, 79], voice [54], or multimodal signals [59]. A major difficulty in training robust affective behavior analysis systems lies in the relatively small size of existing datasets [11, 40, 53] due to the large cost associated with manual annotations. To address this issue, representations from large pre-trained models have been widely used in state-of-the-art affective behavior analysis systems [29, 52, 86].

There is often a large discrepancy between the pre-training data and the target domain, which potentially limits application of a pre-trained encoders on domain-specific tasks. In the field of Natural Language Processing (NLP), Gururangan *et al.* [26] present Domain Adaptive Pre-training (DAPT), which resumes the pre-training process of the encoder on large amount of unlabeled domain-specific data, to boost performance on domain-relevant downstream tasks. The technique has been successfully applied to improve pre-trained language encoders' performance in various NLP tasks such as machine translation [64], question-answering [85], and sentiment analysis [83]. However, DAPT has been relatively under-explored for visual and video understanding applications. Recently, Kim *et al.* [37] show the potential gain of DAPT for pre-trained vision models on the DomainNet benchmark with 6 domains (Clipart, Infograph, Painting, Quickdraw, Real, and Sketch).

In this work, we propose **Semi-supervised Affective Adaptation via Metric Learning (SAAML)**, a DAPT-based framework to adapt a pre-trained audiovisual encoder to the affective domain (Fig. 1). Given the limited size of affective datasets, we propose

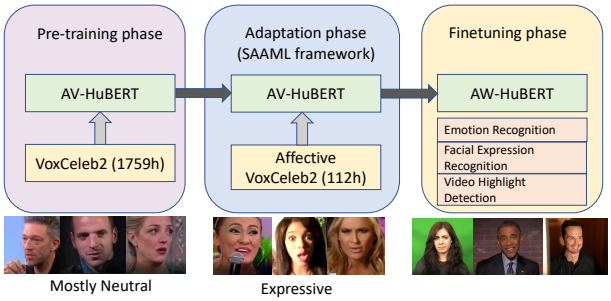


Figure 1: An overview of the process to obtain and evaluate AW-HuBERT. We pre-train AV-HuBERT encoder [68] on the VoxCeleb2 dataset [15], then perform adaptive pretraining with the SAAML framework, and evaluate the adapted model on different affective-related tasks. Details of the proposed SAAML framework can be found in Figure 2 and Figure 3.

using metric learning to generate affect-aware embeddings using labeled emotion recognition (ER) datasets and retrieve large-scale affective data from unlabeled videos. To further improve DAPT performance, we propose adding an emotion classification loss along with a domain adversarial loss. We apply SAAML on a pre-trained AV-HuBERT encoder [68] to produce an affect-aware version of the model, which we call **Affect-Aware HuBERT (AW-HuBERT)**.

We validate the capabilities of AW-HUBERT as a novel audiovisual encoder as follows. First, we demonstrate consistent improvements in emotion recognition accuracy on two audiovisual ER datasets, namely MSP-IMPROV [12] and CREMA-D [13]. Since the ultimate goal of ER is to detect emotions reliably in uncontrolled environments, we extend AW-HuBERT to a more challenging setting with the facial expression recognition in-the-wild using the AffWild2 dataset [40] and demonstrate superior performance over prior work [88]. Second, we show the promising results for using AW-HuBERT for the (affective) video highlight detection, with the Video2GIF [27] and PHD-GIF [21] datasets. Finally, we demonstrate the robustness of AW-HuBERT compared to recent methods for arousal and valence regression with the RECOLA [63] and SEMAINE [51] datasets.

The main contributions of this work are as follows.

- We develop SAAML to retrieve a large amount of emotional audiovisual video excerpts via metric learning for domain adaptation on pre-trained audiovisual encoders.
- We evaluate and show the improvement achieved by SAAML in widely used multimodal emotion recognition datasets, namely MSP-IMPROV [12] and CREMA-D [13], using the AW-HuBERT [68] encoder.
- We demonstrate that AW-HuBERT can be generalized to other tasks within the affective domain, including Facial Expression Recognition in-the-wild (AffWild2 dataset [40]), (Affective) Highlight Detection (Video2GIF [27] and PHD-GIF [21] datasets), and Continuous Emotion Recognition (RECOLA [63] and SEMAINE [51] datasets).

2 RELATED WORK

Self-supervised Audiovisual Representation Learning effectively extracts semantically rich, cross-modal associations by learning the underlying correspondence between the audio and visual signals [3, 4, 41, 55, 57]. For instance, Owen *et al.* [57] propose using audio and video temporal synchronization to train a multimodal network. Morgado *et al.* [55] use contrastive learning to optimize cross-modal discrimination between audio and video signals. Lee *et al.* [43] introduce a parameter-efficient pre-trained audiovisual Transformer, trained with a contrastive objective and a mid-level fusion strategy to match audiovisual information. For self-supervised audiovisual speech representation learning, Shi *et al.* introduce AV-HuBERT [68], an extension of the audio HuBERT [32], to model interactions between lip movements and speech from videos. We provide a detailed description of AV-HuBERT in Section 3. In this work, we re-train AV-HuBERT with video recordings of full faces, and show that the re-trained model is more robust for emotion recognition tasks. Most similar to our work, Tran *et al.* [75] introduce an audiovisual encoder for emotion recognition that learns on masked audiovisual features reconstruction. The model is tailored for emotion recognition downstream tasks due to the feature selection process, in which the selected features (OpenFace [7] for visual and TRILL [69] for audio) are demonstrated to be highly effective for emotion recognition. On the other hand, the main focus of our work is on the data selection process, *i.e.,* creating a diverse affective-related dataset, while our chosen encoder is end-to-end that requires no prior feature extraction.

Supervised Metric Learning is a commonly used method in information retrieval, face recognition, and speaker verification. The goal of metric learning is to train embedding models, by introducing loss functions that capture similarities between data samples. Metric learning techniques have been used to help models generate more discriminative features and enhance emotion recognition performance. Gao *et al.* [20] use a triplet loss to transform a feature encoder into a pseudo-Siamese network to improve knowledge transfer for emotion recognition. Dai *et al.* [16] combine the softmax cross-entropy loss with a metric learning loss (*i.e.,* center loss) to boost the discriminative power of the trained models and increase emotion classification accuracy. Huang *et al.* [34] propose a triplet framework with hard-positive mining to learn from variable-length inputs with an LSTM network. Lu *et al.* [50] introduce an aggregate ranking loss to compare adjacent frames for video inputs and learn meaningful representations for Action Unit detection. Han *et al.* [28] propose a novel cross-modal emotion embedding framework that transfers information from an auxiliary modality to a target modality with triplet losses. Abdou *et al.* [1] integrate gaze information into cross-modal training and leverage triplet losses to generate gaze-aware facial feature representations for emotion recognition. To the best of our knowledge, there has been no prior work using metric learning to retrieve emotional expressions.

Unsupervised Domain Adaptation (UDA) aims to bridge the performance gap when a model is deployed on an unlabeled target domain different from the source domain (training data). UDA methods can generally be divided into three categories. The first category focuses on feature space alignment, reducing the domain shift between the source and target domains with respect to metrics such

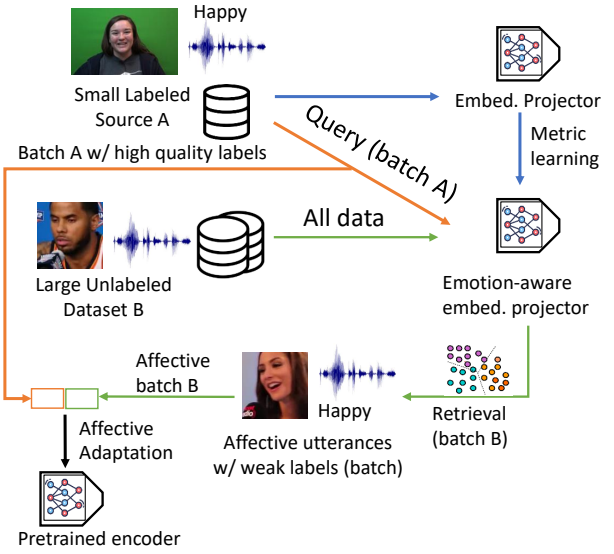


Figure 2: The SAAML framework's overview. We first train an emotion-aware embedding projector using metric learning with annotated samples from a source dataset A (blue arrows). Embeddings extracted from the embedding projector are used to perform retrieval from a large unlabeled dataset B (green arrows), given queries from A (orange arrows), in a batch-wise manner. Both samples from A and B are then used for affective adaption on a pre-trained encoder. Figure 4 provides details on the affective adaptation process.

as statistical moments [58], Maximum Mean Discrepancies (MMD) [49, 76], and correlation [72]. The second category uses adversarial-based techniques for UDA, which typically involves a domain discriminator to enhance domain confusion. Ganin *et al.* [19]introduce a gradient reversal layer to adversarially train a feature extractor jointly with a domain discriminator and produce domain-invariant features. Hoffman *et al.* [30] propose a GAN-based method to generate indistinguishable source and target images. Another line of UDA studies are reconstruction-based [10, 22], which assumes domain-invariant features can be achieved via data reconstruction. Performing Domain Adaptation on pre-trained weights, *i.e.,* domain-adaptive pre-training (DAPT) is relatively under-explored. Gururangan *et al.* [26] show that the continuation of the pre-training process on domain-specific data helps boost the performance of pre-trained language encoders on downstream tasks of the same domain. To the best of our knowledge, we are the first work to explore adaptive pre-training of multimodal audiovisual encoders with applications in vision. We choose affect as our domain of interest as it is a challenging domain in which large datasets are unavailable due to annotation costs. Despite starting with a narrowly defined domain (a small labeled dataset), our framework can produce encoders useful to a broader range of related tasks with more challenging settings.

3 METHOD

Objective We assume access to an annotated emotion recognition source dataset A and an unlabeled dataset B containing a large number of utterances. We aim to perform domain adaptation for a pre-trained encoder M using A and B to obtain M' such that fine-tuning M' for unseen affective-related datasets U yields better performance than M. Figure 2 provides an overview of the proposed SAAML framework, which includes three steps, namely, training an emotion-aware embedding projector, generating pairwise (a, b) similarity distances, and performing step-wise affective adaptation.

Training Embedding Projector. The main goal of this step is to train an Embedding Projector that can produce discriminative emotion-aware embeddings for audiovisual inputs, which are later used for large-scale retrieval of expressive utterances. The architecture for this step involves two components, namely a pretrained encoder to extract high-level representations from the raw audio-visual inputs, followed by a MLP-based Embedding Projector that maps the extracted representations into a common space that is discriminative of emotion classes via metric learning.

- *Pretrained Encoder:* The pretrained encoder takes raw image frames and audio signals as inputs and produces high-level representations.

In this paper, we use the Audiovisual Hidden Unit BERT (AV-HuBERT) architecture [68] as our pre-trained encoder. As a high-level overview, the AV-HuBERT model contains separate light encoders to extract low-level features from raw audio and visual inputs, which are then concatenated with modality dropout [68] in the audiovisual fusion layer. The fused audiovisual features are then forwarded to the Transformer to generate high-level representations. During the pre-training process, the model uses a separate *Feature Extractor* followed by K-means clustering to generate the Hidden Units, *i.e.*, pseudo-labels, for self-supervised learning. Some parts of the audiovisual fusion representation and the corresponding hidden units are masked. The pre-training task is to predict the masked cluster assignments. The HuBERT model is trained for several iterations to improve the quality of the Feature Extractor and generated pseudo-labels. We direct readers to [32, 68] for more detailed discussions of the HuBERT architectures.

- *Embedding Projector.* The *Embedding Projector* maps the high-level general-purpose features generated by the pre-trained encoders into affect-aware embeddings. To this end, we train the *Embedding Projector* in a supervised manner using annotated samples from A. We try different metric learning losses offered by [56] and choose the Multi-Similarity loss [81] that provides embeddings with the best quality of emotion clusters. We provide a comparative study between the Multi-Similarity loss and other metric learning losses with respect to classification accuracy and cluster quality in the *Appendix* (Section C1). The Multi-Similarity loss is:

$$L_{MS} = \frac{1}{m}\sum_{i=1}^{m}\{\frac{1}{\alpha}log[1+\sum_{k\in P_i} e^{-\alpha(S_{ik}-\lambda)}]\} + \{\frac{1}{\beta}log[1+\sum_{k\in N_i} e^{\beta(S_{ik}-\lambda)}]\} \quad (1)$$

where α, β, λ are fixed hyper-parameters, S_{ik} denotes the cosine similarity score of the pair (x_i, x_k), and P_i and N_i denote the index set of positive and negative pairs for an anchor x_i. As in [81], we only use hard positives and negatives to train our *Embedding Projector*. In particular, a pair (x_i, x_j) with labels (y_i, y_j) is chosen

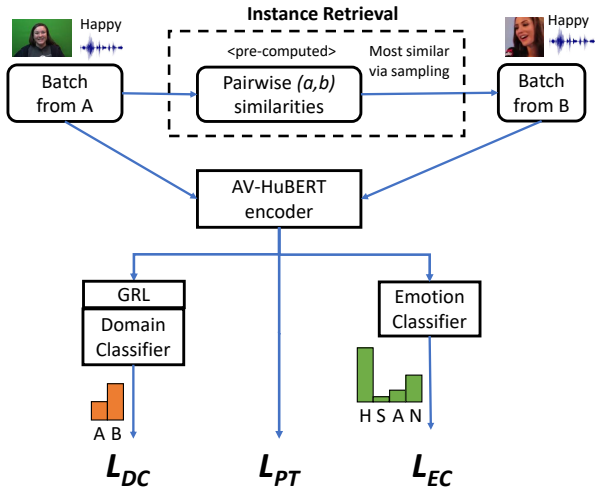


Figure 3: Details of the affective adaptation phase. Given a batch of samples from A, we retrieve relevant samples from B via sampling from a pre-computed pairwise similarity matrix. The AV-HuBERT encoder is then optimized with respect to a domain classification loss with a Gradient Reversal Layer (GRL) for domain-invariant representations, an emotion classification loss, and a pre-train loss.

with respect to a margin ϵ if

$$S_{ij}^{-} > min_{y_k=y_i} S_{ik} - \epsilon \tag{2}$$

OR

$$S_{ij}^{+} < max_{y_k \neq y_i} S_{ik} + \epsilon \tag{3}$$

Following Dai *et al.* [16], our *Embedding Projector* is optimized with a combined loss between the metric learning loss (L_{MS}) and the standard cross-entropy loss to enhance the discriminative power of the model while maintaining a strong predictive ability:

$$L_{EP} = L_{MS} + \lambda_{CE} L_{CE} \tag{4}$$

where λ_{CE} controls the trade-off between the metric learning loss and the prediction task.

Generating pairwise *(a,b)* similarity. The main goal of this step is to compute pairwise distances between utterances in A and B using the trained Embedding Projector. Having the utterance similarities pre-computed would allow the framework to efficiently retrieve and weakly label relevant utterances from B to perform affect adaptation. In particular, the trained *Embedding Projector* is used to extract embeddings for each utterance in A and B to form $\{e_i^A\}_{i=1}^{|A|}$ and $\{e_j^B\}_{j=1}^{|B|}$. Finally, a pairwise similarity matrix S between each utterance from A and B is computed, *i.e.*, $S_{ij} = Sim(e_i^A, e_j^B)\ \forall (i,j) \in [|A|] \times [|B|]$. We use the cosine similarity metric as our distance metric in this study.

Affective adaptation of the Encoder. The main goal of this step is to produce an affect-aware version of the Pretrained Encoder. As shown in Fig. 3, this step includes three components, including a pretrained encoder for affect adaptation, trained jointly with a domain classifier to generate domain-invariant features, and an emotion classifier to enhance the encoder's awareness of emotional signals. At each learning step, the framework takes in a batch of annotated samples from A and uses the pre-computed similarity scores from the previous step to retrieve relevant samples from B via an *Instance Retrieval* module, which are then used to perform adaptation according to three loss functions, namely a pre-training loss, a domain adversarial loss, and an emotion classifier loss.

- *Instance Retrieval.* The *Instance Retrieval* module selects a relevant utterance from B given the projected embedding e_i^A of an utterance from A based on S. We want samples from B with closer projected embeddings to e_i^A to have a higher chance of being selected. Hence, given utterance an utterance in A with its embedding e_i^A and its emotion label l_i^A, the module selects an utterance in B with embedding e_j^B based on a Softmax distribution

$$\mathbb{P}(e_j^B \mid e_i^A) = \frac{e^{S_{ij}/T}}{\sum_k e^{S_{ik}/T}} \tag{5}$$

where T denotes the temperature of the softmax function. Given the large number of utterances in B, we use a low temperature parameter so that the sampling function focuses the most similar utterances from B. It is important to note that we prefer retrieval via sampling over the k-NN approach to avoid retrieving low-quality samples while keeping diversity in the retrieved samples. The *Instance Retrieval* module outputs the selected utterance u_j^B in B and its **pseudo-label** $l_j^B = l_i^A$.

As shown in Figure 3, the *Instance Retrieval* step is performed jointly with the adaptation phase. Specifically, at each training step with a batch k of annotated samples drawn from A, our framework uses the *Instance Retrieval* module to retrieve a batch of k related samples with pseudo-labels from B (one-to-one retrieval). The combined batch of $2k$ samples are then used for affective adaptation.

- *Emotion Classifier.* The *Emotion Classifier* is trained jointly with the pre-trained encoder to optimize emotion classification accuracy. It processes the features generated by the pre-trained encoder to produce emotion predictions, which are then used to compute L_{EC}. We use the categorical cross-entropy loss for L_{EC}, in which the pseudo-labels are used for the retrieved utterances from B and annotated labels are used for utterances from A.

- *Domain Classifier.* The *Domain Classifier* predicts whether an utterance is from A or B with a binary cross-entropy loss L_{DC}. We want the pre-trained encoder to produce domain-invariant features. To achieve this goal, we use a Gradient Reversal Layer (GRL) [19], a commonly used method for training domain-invariant neural networks. The GRL acts as an identity transform during the forward pass, but it scales the received gradients a factor of $-\lambda_{GRL}$ during the backward pass. To help the pretrained encoder learn domain-invariant features, we train it jointly with the *Domain Classifier*, and introduce the GRL between the encoder and the *Domain Classifier*.

- *Pretrain Loss L_{PT}:* Gururangan *et al.* show the benefits of a second-phase pre-training on domain-specific data [26]. Following this idea, we add the pre-training loss L_{PT} to the loss function to continue the pre-training process of the encoder, which is computed as the cross-entropy loss over the masked timesteps

$$L_{PT} = -\sum_{t \in M} \log p_f(z_t) \tag{6}$$

where M denotes the set of masked positions, f is the masked prediction model that maps the representation outputs of the Transformer to a prediction distribution over the hidden unit entries, and z_t denotes the K-means hidden unit at time step t.

This concludes all components of the loss function we propose to adapt the pre-trained encoder. We combine the three losses to adapt our encoder for affective applications.

$$L = L_{EC} + L_{DC} + L_{PT} \tag{7}$$

Desgin of SAAML in light of semi-supervised learning literature. Semi-supervised learning is a well-studied area of research. However, the sheer size of the pre-training dataset (>1000 hours of audiovisual content) is a huge challenge that prevents the adoption of most semi-supervised and advanced pseudo-labeling techniques (e.g., iterative pseudo-labeling) due to both space and time constraints. Moreover, many semi-supervised learning methods focus on finding high-quality (easy) samples within the target domain(s) to improve the correctness of pseudo-labeling [14, 44], which is already handled via our high-confidence retrieval module (Equation 5). Finally, existing semi-supervised learning methods generally assume a clearly-defined target domain (*e.g.*, the unlabeled target dataset). This assumption does not hold in our case, as we use a metric learning approach to retrieve relevant samples to the affective domain. Hence, the retrieval component is the core of SAAML, as a strong retrieval method also results in more accurate pseudo-labeling and reduces the need for more advanced pseudo-labeling techniques. Existing semi-supervised domain adaptation methods further require a subset of the target domain to contain ground-truth labels [44, 65], which is not applicable in our case. Nevertheless, we compare our Naïve pseudo-labeling approach with several recent semi-supervised methods using the retrieval samples from our metric learning models in the *Appendix* (Section B). We demonstrate that our method achieves competitive performance compared to more advanced techniques while delivering simplicity.

4 EXPERIMENTS

Datasets In this study, we interchangeably use the MSP-IMPROV [12] and CREMA-D datasets as A and U, and use the VoxCeleb2 dataset [15] as B in Section 3. *MSP-IMPROV* [12] is an acted audiovisual emotion recognition database containing recordings of dyadic interactions between people. The conversations are designed to trigger realistic emotions. The dataset contains 8, 450 utterances (around 9.5 hours) recorded in 6 sessions from 12 actors. Each utterance is annotated with at least 5 evaluators on 5 emotion categories, *i.e.*, happiness, sadness, anger, neutral, and other, along with a five-point Likert scale on Valence (1-negative to 5-positive), Arousal (1-calm to 5-excited), and dominance (1-weak to 5-strong).*CREMA-D* [13] is an acted audiovisual emotion recognition dataset with 6 emotional states, *i.e.*, happiness, sadness, anger, disgust, fear, and neutral. The dataset contains 7, 442 video recordings (around 1.5 hours) from 91 actors speaking a fixed set of 12 sentences with the 6 types of emotions. The annotations are collected via crowdsourcing from 2443 annotators on emotion categories and intensity level, which consists of Low, Medium, High or Unspecified. The emotion classes are balanced for CREMA-D. This work focuses on the recognition of four emotions using these databases, *i.e.*, happiness, sadness, anger, and neutral.

VoxCeleb2 is the largest publicly available audiovisual speech dataset, containing 2700 hours of video recordings from more than 6000 celebrity speakers. The dataset is constructed from around 150K videos from YouTube, segmented into utterances with a filtering pipeline involving face tracking and active speaker verification. The large scale of the dataset makes it one of the most promising datasets to pre-train audiovisual representation learning models for human social behavior [68]. We only use the English portion of the VoxCeleb2 dataset, which totals to 1326 hours of speech.

- *Facial Expression Recognition.* We use the AffWild2 dataset [40] that was used in the Affective Behavior Analysis in-the-wild (ABAW) Challenge at CVPR 2022. The dataset contains 548 videos collected from Youtube with per-frame annotations for 6 basic emotions (happiness, sadness, anger, surprise, disgust, fear), plus a neutral state, and another category "other" that represents states other than the 6 basic ones. In total, the dataset contains more than 2.6M annotated frames from 431 speakers.
- *Highlight Detection in Human-Centric Videos.* We use two in-the-wild datasets to validate the performance of AV-HuBERT for video highlight detection, namely, Video2GIF [27] and PHD-GIF [21]. We focus on the task of GIF generation with potential applications in photojournalism or advertising [27]. Both the Video2GIF and PHD-GIF datasets contain videos with GIF excerpt annotations. Because not all videos in Video2GIF and PHD-GIF are human-centric (*e.g.*, some videos contain animals), we use a HOG-based face detection model [38] to extract a subset of the data containing human faces (the face detection model can reliably detect human faces in more than 90% of the frames). We end up with 4945 GIFs from the Video2GIF dataset and 8132 GIFs from the PHD-GIF dataset for the task. Each video in Video2GIF and PHD-GIF are split into five second segments. A segment is considered a highlighted segment if the overlap with a GIF is at least 60% (3 seconds).
- *Continuous Emotion Recognition.* We use the RECOLA [63] and SEMAINE [51] datasets to validate the performance of our models on the task of continuous emotion recognition (arousal/valence regression). RECOLA contains audiovisual recordings of dyadic interactions between 23 French-speaking subjects. The videos are recorded and annotated on arousal and valence intensity at 25 Hz and last around 5 minutes each. SEMAINE [51] is a large audiovisual database built upon agents (i.e., operators)-users interactions in simulated settings. The dataset contains 959 conversations from 150 participants. Each conversation is in English, recorded and annotated by at least 2 annotators at 50Hz and typically lasts around 5 minutes. We provide pre-processing and more implementation details in the *Appendix* (Section A3).

AV-HuBERT As mentioned, the currently published AV-HuBERT model was pre-trained to capture interactions between lip movements and human voices for audiovisual speech recognition downstream tasks. This is not optimal for affective-related tasks, in which the whole face is relevant to capture facial expressions. Therefore, we retrain the AV-HuBERT model from scratch with the visual modality captured the speakers' faces. We call this model Face AV-HuBERT (FAV-HuBERT). We follow the same pre-training settings as in [68], with the exception that we only train the model for three iterations instead of five iterations. We see no further improvement in ER accuracy after the third iteration. The model was trained on 32 Tesla-V100 GPUs for approximately 10 days.

Table 1: ER performance for pre-trained AV-HuBERT models (MSP.: MSP-IMPROV, CRE.: CREMA-D).

Model	Dataset	Duration	ER Acc. (%)	
			MSP.	CRE.
A-HuBERT [32]	LibriSpeech	960h	64.14	77.78
AV-HuBERT [68]	LRS3+Vox2	1759h	65.27	85.47
FAV-HuBERT	Vox2	1326h	**68.35**	**87.61**

Table 1 compares the fine-tuning performance of our FAV-HuBERT model with existing pre-trained models on the four-class emotion classification tasks with the CREMA-D and MSP-IMPROV datasets. First, we can see that the AV-HuBERT models outperform the Audio-HuBERT model [32] by a significant margin on both CREMA-D (7.69%) and MSP-IMPROV (1.13%) datasets. This illustrates the benefits of adding visual information to the performance of the ER systems. Moreover, we can also observe that FAV-HuBERT achieves better performance compared to the published pre-trained AV-HuBERT checkpoint (lip movements + acoustic) even though it was trained on less data (no pre-training on the LRS3 dataset [2]). This confirms the importance of capturing visual information from the whole face for affective downstream tasks.

Baselines We apply the affective adaptation framework for both the audio-only and audiovisual settings with the pre-trained audio HuBERT [32] and Audiovisual HuBERT [68] encoders, respectively. We report the results with the Audio-HuBERT encoder in the *Appendix* (Section D). Because there are no similar methods on affective adaptation for pre-training audio or audiovisual encoders, we compare our affect-aware AV-HuBERT (AW-HuBERT) performance against the original pre-trained encoders (without any domain adaptation). To demonstrate the need to retrieve a large amount of data to improve generalization across tasks within the affective domain, we explore adaptive pre-training with small yet carefully annotated datasets, *i.e.*, task adaptive pre-training (TAPT) [26]. In particular, TAPT resumes the pre-training process with limited data on a specific dataset. We test two settings of TAPT: within-task adaptive pre-training (TAPT), where the encoder is adapted and tested on the same dataset, and cross-task adaptive pre-training (CTAPT), where the encoder is adapted and tested on different datasets. We further compare our performance with Tran *et al.*'s pre-trained emotion-centric encoder [75], which was designed and validated on the same ER datasets. Finally, we compare our method with the state-of-the-art performance on the MSP-IMPROV and CREMA-D dataset using the AuxFormer architecture [23]. To make a fair comparison, we fine-tune the entire AuxFormer model, along with its audio encoder (Audio-HuBERT *base* [32]) and visual encoder (TimeSformer [8] first pre-trained on the AffWild2 dataset [40]), which results in around $200M$ trainable parameters. We also perform an ablation study to explore the contributions of individual loss terms. Specifically, we train models denoted -L_{PT}, -L_{DC}, -L_{EC} to highlight models trained without the pre-training loss, domain classifier loss, and emotion classification loss, respectively.

Implementation details We use a multi-layer perceptron as *Embedding Projector*, *Emotion Classifier*, and *Domain Classifier* architectures. Specifically, the *Embedding Projector* contains a Mean Pooling layer followed by two dense layers of sizes $\{256, 32\}$ to produce 32-d embeddings. To train the *Embedding Projector*, we use $\alpha = 2, \lambda = 1, \beta = 50, \epsilon = 0.5$ as suggested in the original paper [81] and $\lambda_{CE} = 0.5$ via hyper-parameter tuning (details available in Section C2). Our *Embedding Projector* achieves better classification accuracy of $1-3\%$ on MSP-IMPROV and CREMA-D datasets compared to a classification-only model, consistent with prior findings [16]. We use the *Cosine similarity* as our similarity metric for the generated embeddings. The *Emotion Classifier* and *Domain Classifier* contain a Mean Pooling layer followed by two dense layers of sizes $\{100, 4\}$ and $\{100, 2\}$. For the *Instance Retrieval* module, we use a temperature $T = 0.1$. For the pre-training loss L_{PT}, we use the same masking strategy and objective as HuBERT [32, 68].

Training and Evaluation Details During the adaptation phase, we train the network for 50 epochs with a linear warmup scheduler using the Adam optimizer. The learning rate for the HuBERT encoder is set to $1e^{-5}$ and the learning rate for other components is set to $1e^{-3}$. We dynamically set the λ_{GRL} for the GRL similar to Ganin *et al.* [19] to suppress noisy signals from the *Domain Classifier* in the early training stage. During the evaluation stage, we simply add a Mean Pooling layer followed by a dense layer on top of the HuBERT encoders to make emotion predictions. We fine-tune the models for 5 epochs with a learning rate of $1e^{-5}$ using the Adam optimizer. For the MSP-IMPROV dataset, we perform a session-based speaker-independent cross validation (six-folds) and report the averaged results. For the CREMA-D dataset, we perform speaker-independent split of the train-validation-test set according to a ratio of 70%-10%-20%. We report the Weighted and Unweighted classification accuracy (A and UA) as our evaluation metrics. We provide more details in the *Appendix* (Section A).

Table 2: Results for AV-HuBERT adaptations on CREMA-D and MSP-IMPROV (A: Accuracy; UA: Unweighted Accuracy).

Method	CREMA-D		MSP-IMPROV	
	A	UA	A	UA
Tran *et al.* [75]	83.46	83.29	65.29	59.41
AuxFormer [23]	91.62	91.10	70.28	62.97
FAV-HuBERT	87.61	87.34	68.35	61.05
TAPT-HuBERT	92.84	92.78	70.46	63.95
CTAPT-HuBERT	90.39	90.52	68.02	60.83
AW-HuBERT	**93.65**	**93.65**	**71.80**	**65.72**
-L_{DC}	92.53	92.42	70.95	64.88
-L_{PT}	88.06	88.02	70.32	63.13
-L_{EC}	92.83	92.86	70.93	64.38

5 RESULTS AND DISCUSSION

5.1 Quantitative results

Results from different adaptations of FAV-HuBERT encoder are available in Table 2. We report the weighted and unweighted classification accuracy as our evaluation metrics. Except for the baseline, which involves no domain adaptation, the results for CREMA-D are from adapted models with MSP-IMPROV being the source dataset, and the results for MSP-IMPROV are from adapted models with CREMA-D being the source dataset.

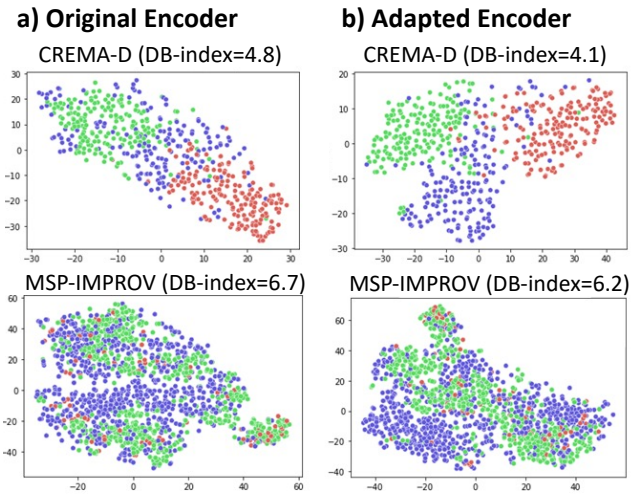


Figure 4: t-SNE visualization of the generated embeddings with emotion categories (Anger: red, Happiness: green, Sadness: blue) from CREMA-D and MSP-IMPROV.

Results in Table 2 show that the proposed adapted models consistently outperform the baselines by a considerable margin. In particular, SAAML helps to boost the unweighted accuracy (6.04% on CREMA-D and 4.67% on MSP-IMPROV) compared to the original AV-HuBERT. This demonstrates the potential of affective adaptation of pre-trained encoders to improve ER performance. Furthermore, because we use a subset of the Voxceleb2 data to adapt the audiovisual encoder, we can negate the possibility that the improvement is due to exposure to more unseen data. TAPT baselines show consistent improvement when the pre-trained encoder is adapted and tested on the same dataset (TAPT), but the generalization is limited when tested on unseen data (CTAPT), partially due to the risk of overfitting when tuning a large model on a small amount of data.

From Table 2, we can also see that dropping a loss term reduces the performance of the adapted models, which indicates the relevance of all components in the proposed loss. The pre-training loss L_{PT} is consistently the most important loss term to the Unweighted Accuracies of the models (except for the CREMA-D results on the audio-only setting). This demonstrates the importance of continuing the pre-training process (DAPT) instead of adapting to emotion-labels only, to prevent catastrophic forgetting.

5.2 Qualitative results

We visualize the mean-pooled embeddings by the pre-trained and adapted AV-HuBERT models using t-SNE [77]. Figure 4 shows the projected embeddings for samples annotated with *happiness*, *sadness*, or *anger* with *medium* or *high* intensity from the CREMA-D dataset and samples annotated with *happiness*, *sadness*, or *anger* from the MSP-IMPROV dataset. Figure 4(a) shows the embeddings extracted from the pre-trained AV-HuBERT model, and Figure 4(b) illustrates the embeddings for the adapted model with (MSP-IMPROV source dataset). A cluster quality metric, the Davies-Bouldin index [17], is reported to quantify the quality of embedding clusters with respect to the emotion labels. A lower DB index suggests a better separation between clusters. The embedding space in Figure 4(b) is more linearly separable compared to 4(a) for both datasets, demonstrating the adapted model is capable of learning, generalizing, and generating affect-aware features.

Table 3: The performance of different models on the official validation set of AffWild2.

Method	Modality	Mean F1
Zhang et al. [88]	A+V	32.6
FAV-HuBERT	A+V	36.08
Zhang et al. [88]	A+V+T	39.4
AW-HuBERT	A+V	38.52
AW-HuBERT	A+V+T	**41.31**

5.3 Generalizability on Other Tasks

To further illustrate the usefulness of AW-HuBERT, we apply the encoder to three tasks: Facial Expression Recognition (FER) In-the-wild, (Affective) Highlight Detection in videos, and Continuous Emotion Recognition. With the FER and continuous ER tasks, we want to test the robustness of the learned representation in more challenging settings with a wider range of expressions compared to the MSP-IMPROV and CREMA-D datasets. With the Highlight Detection task, we demonstrate the applicability of the method on social and human-centered tasks without emotion labels.

5.3.1 *Facial Expression Recognition In-the-Wild.*

Similar to Section 4, we add a Mean pooling layer followed by a fully-connected layer on top of the AV-HuBERT model to make facial expression predictions. We compare the fine-tuned AV-HuBERT models with the winner of the ABAW challenge [88]. Following the challenge's evaluation criteria, we report the average F1-score over the 8 expression categories. Because the wining model uses audio, visual, and textual information for prediction, we also report the performance of AV-HuBERT in combination with a frozen RoBERTa model [48] (late fusion), which extracts textual features from transcripts obtained via speech recognition.

Table 3 shows the performance of the proposed method on the Affwild2 dataset. We can see that AW-HuBERT boosts the performance of FAV-HuBERT by 2.44% (from 36.08% to 38.52%). With only audio and visual information, both FAV-HuBERT and AW-HuBERT outperform the previous winner of the ABAW Challenge at CVPR2022 by a fairly large margin (3.47% and 5.92%, respectively). When combined with textual features extracted from the RoBERTa language encoder [48], our model's performance is further improved by 2.78% to 41.31% F1-score, outperform Zhang *et al.*'s approach by 1.91%. More importantly, AW-HuBERT is less dependent on textual information when text inputs are missing (2.78% drop in F1-score compared to 6.8%).

5.3.2 *Highlight Detection in Human-Centric Videos.*

Following prior work on highlight detection [27, 31, 84], we freeze our AV-HuBERT encoders and only use them for feature extraction. We use the Set-based Learning Module (SL-Module) proposed by Xu *et al.* [84]. Using the extracted features from the encoder, the SL-module uses a Transformer encoder followed by a scoring layer to estimate the highlight score distribution for the inputs. We compare our model with two supervised video highlight detection methods, *i.e.*, Video2GIF [27] and the SL-module [84] itself with the widely

Table 4: Highlight detection results (mAP) on the Video2GIF and PHD-GIF datasets.

	V2GIF	PHD-GIF
Video2GIF [27]	12.46	15.53
SSL [84] w/ C3D [74]	14.32	17.41
SSL [84] w / FAV-HuBERT	14.87	18.32
SSL [84] w/ AW-HuBERT	**16.21**	**20.14**

Table 5: Continuous ER results (CCC) on the official validation set of the RECOLA dataset.

	Arousal	Valence	Avg.
TS-SATCN [33]	0.659	**0.690**	0.675
SS-VAERR [36]	0.675	0.626	0.651
FAV-HuBERT	0.676	0.637	0.657
AW-HuBERT	**0.701**	0.653	**0.677**

Table 6: Continuous ER results on the SEMAINE dataset.

	Arousal	Valence	Avg.
AV-GaS [6]	0.587	0.642	0.615
FAV-HuBERT	0.584	0.661	0.623
AW-HuBERT	**0.593**	**0.664**	**0.629**

used C3D feature extractor. As the evaluation metric, we report the mean Average Precision (mAP) of the detected highlights.

Notably, our training data are a subset of PHD-GIF and Video2GIF (10%), and our results cannot be directly compared to the existing works. Hence, we use a recent architecture baseline for highlight detection, *i.e.*, SSL [84] and re-run the baselines on the selected face-centric subset of the datasets. Moreover, our main goal is affective representation learning from pre-trained encoders. Therefore, we compare our extracted features with the pre-trained C3D encoder [74], a widely used feature extractor for video highlight detection.

Table 4 shows the performance of the proposed method on the highlight detection task on the Video2GIF and PHD-GIF datasets. We can see that both FAV-HuBERT and AW-HuBERT outperform Video2GIF, and SSL with pre-trained C3D features [74], which demonstrates the effectiveness of the trained AV-HuBERT models in extracting facial features. We can also observe that AW-HuBERT consistently boosts AV-HuBERT performance across three datasets (1.34% on Video2GIF, and 1.82% on PHD-GIF), which confirms our expectation that the affect-aware model is more robust for expressive facial behaviors.

5.3.3 *Continuous Emotion Recognition.*

We use the same architecture for regression analysis as in section 5.3.1 (Facial Expression Recognition) to make per-frame predictions. Since we use a sliding window, there can be multiple predictions for a single frame during inference. We average all the overlapping predictions for each frame to get the final predictions. We use the widely used Concordance Correlation Coefficient (CCC) as our evaluation metric for continuous emotion recognition. We use the CCC loss to train our models. For RECOLA, we use the official train/validation set to train/test our models as in [36], with 10% of the train set being used for hyper-parameter tuning. For SEMAINE, we perform a 5-fold cross-validation as in [6] to evaluate our models as there are no official splits. We compare our method with *AV-GaS* [6] on the SEMAINE dataset, and *TS-SATCN* [33] and SS-VAERR [36] on the RECOLA dataset.

Table 5 and Table 6 show the experimental results on the RECOLA and SEMAINE datasets, respectively. For both datasets, we can see that both FAV-HuBERT and AW-HuBERT achieve competitive performance compared to state-of-the-art architectures. Furthermore, AW-HuBERT also consistently outperforms FAV-HuBERT on both arousal and valence regression, with a boost of 2.5% and 1.6% (RECOLA), and 0.9% and 0.3% (SEMAINE). This validates the usefulness of the SAAML framework. More importantly, the robustness of both FAV-HuBERT and AW-HuBERT on the RECOLA dataset further demonstrates the usefulness of the models in multi-lingual settings (despite being pre-trained on English-only data).

5.4 Limitations

The superior performance of AW-HuBERT is limited to tasks close to affective understanding with visible faces. We can see performance drop as we move from emotion recognition benchmarks, recorded in controlled lab environments, to in-the-wild data, and finally to the highlight detection tasks in which the GIFs are not guaranteed to contain strong facial expressions (*e.g.*, the GIF can come from an interesting pose). Moreover, we only perform experiments on datasets in which faces carry important predictive signals. Hence, whether the framework can be extended to other human behaviors (e.g., poses or gestures) is an open question. Lastly, we only evaluate the effectiveness of SAAML on the AV-HuBERT architecture, mainly because it is the first architecture proposed and pre-trained on video recordings of speech. Although there are other audiovisual pre-trained encoders for different types of tasks [43, 87], we lack the computational resources to retrain them on large-scale datasets of human social behavior.

6 CONCLUSION

We propose a novel SAAML framework to adapt existing pre-trained encoders for affective downstream tasks. The framework utilizes metric learning to retrieve a large amount of emotional audiovisual speech from unlabeled video recordings to adapt pre-trained encoders to the affective domain. Experimental results show that SAAML consistently boosts the performance of the original encoder on a wide range of affective tasks, including audiovisual emotion recognition, facial expression recognition in-the-wild, (affective) highlight detection and continuous emotion recognition. We demonstrate the usefulness of exposing pre-trained encoders to affective data to enhance their awareness of social signals.

ACKNOWLEDGMENTS

This work was partly sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. 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] Ahmed Abdou, Ekta Sood, Philipp Müller, and Andreas Bulling. 2022. Gaze-enhanced Crossmodal Embeddings for Emotion Recognition. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–18.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496* (2018).

[3] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*. 609–617.

[4] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*. 435–451.

[5] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[6] Decky Aspandi, Federico Sukno, Bjorn W Schuller, and Xavier Binefa. 2022. Audio-Visual Gated-Sequenced Neural Networks for Affect Recognition. *IEEE Transactions on Affective Computing* (2022).

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[9] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems* 32 (2019).

[10] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems* 29 (2016).

[11] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[12] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[13] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[14] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6912–6920.

[15] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*. 1086–1090. https://doi.org/10.21437/Interspeech.2018-1929

[16] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. 2019. Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7405–7409.

[17] David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), 224–227.

[18] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition* 130 (2022), 108777.

[19] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[20] Yuan Gao, Jiaxing Liu, Longbiao Wang, and Jianwu Dang. 2021. Metric Learning Based Feature Representation with Gated Fusion Model for Speech Emotion Recognition.. In *Interspeech*. 4503–4507.

[21] Ana Garcia del Molino and Michael Gygli. 2018. Phd-gifs: personalized highlight detection for automatic gif creation. In *Proceedings of the 26th ACM international conference on Multimedia*. 600–608.

[22] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*. Springer, 597–613.

[23] Lucas Goncalves and Carlos Busso. 2022. AuxFormer: Robust approach to audiovisual emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7357–7361.

[24] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*. IEEE, 381–385.

[25] Felix Grundmann, Kai Epstude, and Susanne Scheibe. 2021. Face masks reduce emotion-recognition accuracy and perceived closeness. *Plos one* 16, 4 (2021), e0249792.

[26] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8342–8360.

[27] Michael Gygli, Yale Song, and Liangliang Cao. 2016. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1001–1009.

[28] Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller. 2019. EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *IEEE Transactions on Affective Computing* 12, 3 (2019), 553–564.

[29] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.

[30] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*. Pmlr, 1989–1998.

[31] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. 2020. Mininet: Multiple instance ranking network for video highlight detection. In *European Conference on Computer Vision*. Springer, 345–360.

[32] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[33] Min Hu, Qian Chu, Xiaohua Wang, Lei He, and Fuji Ren. 2021. A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video. *IEEE Signal Processing Letters* 28 (2021), 698–702.

[34] Jian Huang, Ya Li, Jianhua Tao, Zhen Lian, et al. 2018. Speech emotion recognition from variable-length inputs with triplet loss function.. In *Interspeech*. 3673–3677.

[35] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7360–7370.

[36] Marija Jegorova, Stavros Petridis, and Maja Pantic. 2023. SS-VAERR: Self-Supervised Apparent Emotional Reaction Recognition from Video. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[37] Donghyun et al. Kim. 2022. A broad study of pre-training for domain generalization and adaptation. In *ECCV*.

[38] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

[39] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

[40] Dimitrios Kollias. 2022. 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*. 2328–2336.

[41] Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems* 31 (2018).

[42] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10143–10152.

[43] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. 2020. Parameter Efficient Multimodal Transformers for Video Representation Learning. In *International Conference on Learning Representations*.

[44] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. 2021. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2505–2514.

[45] Lu Liu and Robby T Tan. 2021. Certainty driven consistency loss on multi-teacher networks for semi-supervised learning. *Pattern Recognition* 120 (2021), 108140.

[46] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 212–220.

[47] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on Machine Learning-Volume 48*. 507–516.

[48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[49] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2208–2217.

[50] Liupei Lu, Leili Tavabi, and Mohammad Soleymani. 2020. Self-supervised learning for facial action unit recognition through temporal consistency. In *BMVC*.

[51] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions*

on affective computing 3, 1 (2011), 5–17.

[52] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 1359–1367.

[53] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[54] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6922–6926.

[55] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12475–12486.

[56] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. PyTorch Metric Learning. arXiv:2008.09164 [cs.CV]

[57] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 631–648.

[58] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.

[59] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. *Proc. Interspeech 2021* (2021), 3400–3404.

[60] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*. 135–152.

[61] Jaime A Rincon, Angelo Costa, Paulo Novais, Vicente Julian, and Carlos Carrascosa. 2019. A new emotional robot assistant that facilitates human interaction and persuasion. *Knowledge and Information Systems* 60, 1 (2019), 363–383.

[62] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*. 3–13.

[63] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. 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)*. IEEE, 1–8.

[64] Raphael Rubino and Eiichiro Sumita. 2020. Intermediate self-supervised learning for machine translation quality estimation. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4355–4360.

[65] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8050–8058.

[66] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 3687–3697.

[67] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[68] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2021. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations*.

[69] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764* (2020).

[70] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* 33 (2020), 596–608.

[71] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[72] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[73] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6398–6407.

[74] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[75] Minh et al. Tran. 2022. A Pre-Trained Audio-Visual Transformer for Emotion Recognition. In *ICASSP*. IEEE.

[76] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).

[77] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[79] Chu Wang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2019. Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 56–60.

[80] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. 2017. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*. 2593–2601.

[81] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5022–5030.

[82] Si Wu, Jichang Li, Cheng Liu, Zhiwen Yu, and Hau-San Wong. 2019. Mutual learning of complementary networks via residual correction for improving semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6500–6509.

[83] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2324–2335.

[84] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. 2021. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7970–7979.

[85] Qinyuan Ye, Belinda Z Li, Sinong Wang, Benjamin Bolte, Hao Ma, Wen-tau Yih, Xiang Ren, and Madian Khabsa. 2020. Studying strategically: Learning to mask for closed-book QA. *arXiv preprint arXiv:2012.15856* (2020).

[86] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10790–10797.

[87] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. 2021. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems* 34 (2021), 7025–7040.

[88] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. 2022. Transformer-based Multimodal Information Fusion for Facial Expression Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2428–2437.

A MORE IMPLEMENTATION DETAILS

For the MSP-IMPROV and CREMA-D results, the AW-HuBERT models are adapted on a single source dataset (*i.e,* MSP-IMPROV as the source dataset when AW-HuBERT is later tested on CREMA-D and vice versa). For the remaining tasks, we use both the CREMA-D and MSP-IMPROV datasets as the source dataset to perform adaptation with the SAAML framework and create a single AW-HuBERT model.

A.1 Facial Expression Recognition

For the visual modality, we use the cropped and aligned frames provided with the AffWild2 dataset [40]. In particular, the frames are cropped according to the detected face bounding boxes and aligned via a similar transformation on detected facial landmarks. For the textual modality, we first transcribe the provided audio files via IBM Watson Speech-to-text to get transcriptions with timestamps and use a frozen RoBERTa language encoder [48] to extract linguistic information from video segments.

We perform classification at a frame level. Specifically, cropped frames, audio, and spoken words from two seconds around the current frames are chosen as the context for the classification of the current frame. Without the textual information, our classifier is simply the AV-HuBERT [68] architecture, followed by a mean-pooling layer and a fully-connected layer to make facial expression predictions. To make it easier to compare to prior work [88], *e.g.*, removing the textual modality, we test our models' performance on the Validation set of AffWild2 [40], and use 20% of the Train set for validation. We train our models using an Adam optimizer [39] with a learning rate of $1e^{-5}$ for the AV-HuBERT encoder and $1e^{-3}$ for the fully-connected prediction layer for 5 epochs. We use a batch size of 64 and train each model on 4 V100 GPUs to speed up the training process, which took us around 12 hours per model. We use a weighted dataset sampler to reduce the impact of class imbalance. Following prior work on Video Highlight Detection that uses frozen encoders to extract features from video segments [21, 27, 84], we also use our FAV-HuBERT encoder for feature extraction only. However, unlike the MSP-IMPROV, CREMA-D, and AffWild2 datasets that only contain frames of one active speaker, there can be multiple speakers appearing in each video in the Video2GIF and PHD-GIF datasets. To address this issue, we split the extracted frames from the Video2GIF and PHD-GIF datasets into patches of size 88×88 (the input image size for the AV-HuBERT architecture), extract the features for each sequence of patches, and later concatenate the extracted features to generate a representation for the whole video segment.

As mentioned, we use the SL-module [84] to make predictions on whether a segment is highlighted or not, given the extracted representations from pre-trained encoders. Other than the feature extractor, the SL-module consists of a Transformer encoder [78] followed by a scoring model C of fully-connected layers to predict the highlight score for each video segment. Following Xu *et al.* [84], we use a Transformer encoder with 5 layers and 8 self-attention heads. The scoring model C consists of FC layers of size $\{d_{enc}, 1024, 256\}$ with a ReLU activation in-between each FC layer. $d_{enc} = 4096$ for C3D [74] pre-trained on UCF101 dataset [71] feature extractors and $d_{enc} = 768 \times p$ for FAV-HuBERT feature extractors with p being the number of image patches per frame. The SL-module is optimized based on a KL-divergence loss between the softmax-normalized predicted video highlight score distribution and the ground-truth distribution. As in [84], we use an SGD optimizer with a learning rate of $1e^{-3}$, momentum of 0.9, and weight decay of $5e^{-4}$ to train our highlight detection models for 20 epochs.

A.2 Continuous Emotion Recognition

Preprocessing. For both datasets, we follow the pre-processing pipeline of AV-HuBERT [68] to obtain cropped and aligned faces. We downsample the video recordings to 25FPS and audio recordings to 16kHZ to match the sampling rate of AV-HuBERT pre-training. For both datasets, we use a window of 200 continuous frames (8s) and a step size of 100 frames. We provide an empirical analysis of the effect of window sizes on performance in section B3. For the SEMAINE dataset, we focus our analysis on the users' Solid SAL recordings [51] with at least 6 raters to ensure high-quality annotations. We follow the *evaluator weighted estimator*-based [24] method, as done in [62], to aggregate ground-truth emotional labels.

We use the same architecture as in the Facial Expression Recognition task, in which our model consists of the pre-trained AV-HuBERT model(s) followed by a fully-connected layer to make frame-level predictions. However, we predict labels for all frames in a given segment (window size of 50 for the SEWA and 200 for the SEMAINE dataset). We use the standard CCC loss ($L_{CCC} = 1 - CCC$) and train two models independently (one for arousal and one for valence estimation) with the Adam optimizer. We perform grid-search hyper-parameter tuning with the learning rate ranging from $1e-3$ to $1e-6$, the weight decay ranging from $1e-3$ to $1e-5$, and the batch size ranging from 1 to 128. We set the maximum training epoch to be 100 (patience=20). We use a linear scheduler with the number of warmup steps being 5% of the total training steps. We empirically determine that 100 is the optimal window size.

B COMPARISON WITH ADVANCED SEMI-SUPERVISED METHODS

Method	CREMA-D		MSP-IMPROV	
	A	UA	A	UA
FixMatch [70]	**94.28**	**94.31**	71.57	65.19
CDAC [44]	92.10	92.07	71.18	65.31
SAAML	93.65	93.65	**71.80**	**65.72**

Table 7: Comparison with other semi-supervised methods.

In semi-supervised learning, pseudo-labeling is generally improved via one of the three methods, namely combining the predictions of multiple models [18, 45, 60, 82], combining multiple predictions of the same model during the training process such as iterative pseudo-labeling [5, 14], and input augmentations [9, 44, 70]. The first two approaches usually suffer from space and time constraints on large target datasets. Hence, we compare SAAML with augmentation-based semi-supervised learning methods due to the relatively lighter computational resources required. In particular, we compare our approach with two methods, namely *FixMatch* [70] and *CDAC* [44].

Sohn *et al.* [70] proposed *FixMatch*, which trains a model to match the pseudo-labels of weakly-augmented and strongly-augmented

unlabeled inputs. We adapt the idea of *FixMatch* into our framework, in which our emotion classifier is trained to predict the pseudo-labels of weakly-augmented inputs given strongly-augmented inputs. We only apply augmentations on visual inputs, using the same types of image augmentation introduced by Sohn *et al.* to our video frames. We use the same hyperparameters as in the original paper without any tuning.

Li *et al.* [44] introduce Cross-Domain Adaptive Clustering (CDAC) for semi-supervised domain adaptation. At the core of the method is the adversarial adaptive clustering loss that first computes pairwise similarities between features of unlabeled samples in a batch and forces the pseudo-labels for the samples with pairwise feature similarity to be consistent. In addition, the method combines three other types of loss: a standard cross-entropy loss for the labeled samples, a pseudo-label loss that predicts the pseudo-labels of original unlabeled samples given augmented inputs, and an L2-based consistency loss that encourage feature similarities between augmented versions of the same input. We follow the same experimental settings as in the original paper and only apply augmentation to the visual inputs. Since we do not have labels on the target dataset (*i.e.,* VoxCeleb2), we drop parts of CDAC that involve labeled samples from the target domain.

Table 7 shows the performance comparison between SAAML and the semi-supervised learning methods. We can see that more advanced pseudo-labeling methods do not necessarily lead to better performance. In particular, performing adaptive pre-training with CDAC leads to slightly reduced performance while *FixMatch* results in better performance on CREMA-D but worse performance on MSP-IMPROV (with small margins). We attribute this observation to the fact that our retrieved unlabeled samples are of high confidence (Equation 5), so it is relatively easy to re-match the pseudo-labels of these samples to the labels of the source samples despite the involvement of augmentations. Furthermore, as our (retrieved) samples already formed clusters in the embedding space via metric learning, most of the losses in CDAC (adversarial adaptive clustering loss, consistency loss, and pseudo-label loss) become trivial to learn, which could negatively impact the adaptation process. Therefore, our naive pseudo-labeling method delivers both simplicity and effectiveness.

C HYPER-PARAMETER TUNING

C.1 Choice of metric-learning loss

	MSP-IMPROV		CREMA-D	
	Acc. ↑	DB-idx ↓	Acc. ↑	DB-idx ↓
TripletLoss [67]	58.9	6.5	70.2	5.4
AngularLoss [80]	60.4	5.3	76.1	4.6
CircleLoss [73]	59.7	7.6	73.9	5.2
LM-SoftmaxLoss [47]	62.5	5.2	74.3	3.9
SphereFaceLoss [46]	62.8	4.9	75.7	4.1
MultiSimLoss [81]	**64.3**	**4.8**	**76.5**	**3.7**

Table 8: Metric learning loss comparison.

Table 8 shows the performance comparison between the chosen Multi-Similarity loss [81] and other types of metric learning loss provided by the *pytorch-metric-learning* library [56]. For efficiency, hyper-parameter tuning is performed when the pre-trained encoder is frozen. We can observe that on both the MSP-IMPROV and CREMA-D datasets, the Multi-Similarity loss outperforms other types of loss by a considerable margin on both the classification accuracy (1.5% on MSP-IMPROV and 0.4% on CREMA-D dataset) while maintaining more robust cluster quality with the lowest DB-index on both datasets. These findings motivate our usage of the Multi-Similarity loss to perform expressive speech retrieval. It is worth noting that the DB-indexes in Table 8 are better than in Figure 4 due to within-corpus training.

C.2 Choice of λ_{CE}

λ_{CE}	MSP-IMPROV		CREMA-D	
	Acc. ↑	DB-idx ↓	Acc. ↑	DB-idx ↓
0	62.3	4.4	73.1	3.5
0.25	63.5	4.5	74.6	3.5
0.5	64.3	4.8	76.5	3.7
0.75	64.4	4.9	77.2	3.9
1	64.7	5.1	77.9	4.1

Table 9: Performance comparison with different values of λ_{CE}.

Table 9 shows the performance comparison of the *Embedding Projector* with different values of λ_{CE}. For efficiency, hyper-parameter tuning is performed when the pre-trained encoder is frozen. We can observe that as the value of λ_{CE} increases, the models' classification accuracy increases at the cost of worse cluster quality, as the cross-entropy classification loss L_{CE} gains more weight with respect to the overall loss. Generally, we can observe the best trade-off at around $\lambda_{CE} = 0.5$ for both the MSP-IMPROV and the CREMA-D datasets, as increasing λ_{CE} gains marginal accuracy while decreasing λ_{CE} receives results in steep cluster quality drops.

D AFFECTIVE ADAPTATION FOR AUDIO-HUBERT

Method	CREMA-D		MSP-IMPROV	
	A	UA	A	UA
Audio-HuBERT	77.78	77.5	64.14	58.56
Audio AW-HuBERT	**83.95**	**84.01**	**66.55**	**62.60**
-L_{DC}	82.87	82.49	64.87	60.40
-L_{PT}	82.41	82.86	65.52	59.55
-L_{EC}	81.64	81.49	65.93	61.36

Table 10: Results for Audio-HuBERT adaptations.

We provide the adaptation results for the Audio-HuBERT encoder [32] in Table 10. We can see that the Audio AW-HuBERT models show consistent improvements compared to the original Audio HuBERT encoder on both datasets, which further demonstrates the usefulness of the SAAML framework in a single-modality setting. However, unlike the Audio-visual ablation results, the pre-training loss L_{PT} seems to contribute less to the performance boost in the audio-only setting. In particular, dropping the pre-training loss only results in a drop of 1.54% and 1.03% on Weighted Accuracy for CREMA-D and MSP-IMPROV, respectively, *i.e.*, it is no longer the most important loss term out of the three.