Multimodal Emotion Recognition with Modality-Pairwise Unsupervised Contrastive Loss

Riccardo Franceschini*, Enrico Fini†, Cigdem Beyan†, Alessandro Conti†, Federica Arrigoni†, and Elisa Ricci†‡
*Eurecat, Centre Tecnològic de Catalunya, Cerdanyola del Valles, Spain, riccardo.franceschini@eurecat.org
†Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
{enrico.fini, cigdem.beyan, alessandro.conti-1, federica.arrigoni, e.ricci}@unitn.it
‡Fondazione Bruno Kessler (FBK), Trento, Italy

***Abstract*—Emotion recognition is involved in several real-world applications. With an increase in available modalities, automatic understanding of emotions is being performed more accurately. The success in Multimodal Emotion Recognition (MER), primarily relies on the supervised learning paradigm. However, data annotation is expensive, time-consuming, and as emotion expression and perception depends on several factors (e.g., age, gender, culture) obtaining labels with a high reliability is hard. Motivated by these, we focus on unsupervised feature learning for MER. We consider discrete emotions, and as modalities text, audio and vision are used. Our method, as being based on contrastive loss between pairwise modalities, is the first attempt in MER literature. Our end-to-end feature learning approach has several differences (and advantages) compared to existing MER methods: i) it is unsupervised, so the learning is lack of data labelling cost; ii) it does not require data spatial augmentation, modality alignment, large number of batch size or epochs; iii) it applies data fusion only at inference; and iv) it does not require backbones pre-trained on emotion recognition task. The experiments on benchmark datasets show that our method outperforms several baseline approaches and unsupervised learning methods applied in MER. Particularly, it even surpasses a few supervised MER state-of-the-art.**

I. Introduction

Emotion is a key factor driving people's actions and thoughts, and a fundamental part of the human verbal and nonverbal communication. Automated emotion recognition is an important aspect of many applications, including social assistive robots [1], smart systems to work in customer service [2], health-care [3], education [4], and automated-driving cars [5]. However, it is a highly challenging problem due to the complex nature of emotion *expression* and *perception*, which are hard to generalize as being dependent on several factors such as age [6], gender [7], cultural background [8], and personality traits [9]. Furthermore, as humans can express their emotions across various modalities (e.g., language, facial expressions, gestures, and speech), it is essential to effectively model the interactions between these modalities, containing complementary but also (possibly) redundant information [10].

The majority of works mainly concentrated on unimodal learning of emotions [11], [12], [13], i.e., processing a single modality. Although there exist breakthrough achievements by unimodal emotion recognition, due to the aforementioned multimodal nature of emotion expression, such models remain incapable in some circumstances. On the other hand, multimodal emotion recognition (MER) holds the challenges of multimodal machine learning, e.g., representing the data to be able to exploit the complementarity and redundancy of modalities, data translation among modalities, co-learning, modality alignment (e.g., capturing temporal information) and data fusion (see [10] for details). Like most intelligent systems, the advancements in deep learning have enhanced MER, particularly, by utilizing the abundance of data availability. Studies in this field (e.g., [14], [15], [16], [17]) so far, treat the learning process with the supervised way, thus require an intense labor for annotations.

This paper addresses the problem of ***perceived multimodal emotion recognition*** when the emotions are represented as ***discrete categories*** and, more importantly, we learn the features in an ***unsupervised fashion***. Motivated by the fact that contrastive learning has shown accurate and robust performance in many domains (e.g., [18], [19]), we adapt the contrastive loss function [20] to perform pairwise modality feature learning. To the best of our knowledge, this is the first time contrastive loss is adapted for MER. Our approach learns feature embeddings in an end-to-end fashion (see [21] for the definition), and differs from the prior works in terms of several aspects, which are described as follows.

i) Modality exploitation. Our method leverages different modalities in a contrastive learning framework. Given a data sample represented in terms of multiple modalities, our aim is to push the embeddings of two modalities of the same sequence to be close to each other while pulling the embeddings of the same two modalities of different sequences to be apart. Note that the sequences that are being pulled apart can be from the same class. But, herein we *do not use the class labels*, thus we only aim to make the representations of the same sequence across modalities similar (as close as possible) to each other.

ii) Data translation & **co-learning.** We contrast the feature embedding of one modality with another modality when both are belonging to the *same data sample*. This can be seen as an analogy to performing *data translation* and ultimately *co-learning*. Unlike existing contrastive learning approaches (e.g., [22], [19]), we do not require data spatial augmentation (e.g., random crops, blurs or color distortions). Also, different from approaches [18], [23], [24] relying on heavy data augmentations as well as large number of batch sizes and epochs, our method is much more affordable.

iii) Modality alignment. The outputs of different sensors

might have different (but fixed) sample rates. However, this is not valid for text, which makes obtaining word-aligned sequences not so obvious [16]. Still, multimodal data alignment is an imperative step to perform an effective MER for several methods (e.g., [25], [26]), resulting in the real-world application of such methods challenging. In contrast, our method does not require *perfectly aligned modalities*. We considered both aligned samples and a mixture of aligned/misaligned samples in our experiments (Sec. IV-A).

iv) Data fusion. It is applied here only at inference via the concatenation of learned feature representations. This is different from the MER state-of-the-art (SOTA) applying data fusion *both* in training and testing [27], [26], [28], [29].

v) Data labelling. Our method is free from data labeling cost by being an *unsupervised feature learning* approach. Note that there exist a few number of unsupervised approaches in the same and/or related topics, e.g., speech emotion recognition [30], [31], facial emotion recognition [32], facial expression intensity estimation [33], and multimodal sentiment and emotion analysis [25]. However, our method involves the deep architectures either pre-trained on tasks different from emotion recognition (e.g., action recognition) or *not* pre-trained. This aspect introduces a potential to apply the proposed method to the related downstream tasks, e.g., multimodal sentiment analysis and social interaction analysis, without the need of customization. Some approaches (e.g., [34], [35], [36]), instead, could supply the desired performance (e.g., outperforming the best of all methods of comparison time) if and only if they are pre-trained on large emotion datasets having the *same emotion labels* as in the test set.

To validate the effectiveness of our method, experiments were realized on two multimodal emotion datasets. Results show that the proposed method outperforms prior unsupervised MER approaches and several baselines. Moreover, despite performing unsupervised feature learning, our method even surpasses some of the fully-supervised MER methods. To summarize, the main contributions of this study are: (1) presenting a novel unsupervised multimodal feature learning approach, (2) being the first study adapting the contrastive loss for MER, and (3) improving the emotion recognition results compared to unsupervised feature learning MER SOTA. The code of the proposed method is available at https://github.com/ricfrr/mpuc-mer.

II. Related Work

Several methods for multimodal emotion recognition (MER) were proposed, as detailed in the recent survey papers: [37], [38]. In this section, our summary is regarding *discrete* MER research modeling text, visual and acoustic modalities, as we tested our method on that context. Early works adapt classifiers like SVMs, Linear and Logistic Regression [39], [40] while, by the time bigger datasets were developed, deep learning architectures were also explored. For example, [27] is based on CNNs, and [26], [28] use RNNs. Some recent studies [41], [14], [16] adopt Transformers.

Ghaleb et al. [42] apply deep metric learning in which a LSTM component models the variations of the emotions as a function of time. That is different from late fusion of modalities [27], [28] or building temporal features to extract global information by assuming that emotions are expressed simultaneously [26]. Late fusion is favorably applied by concatenating the learned features of all modalities in [27], [28] or with a pairwise scheme in [26]. Instead, the authors of M3ER [29] propose a data-driven multiplicative fusion method to combine the modalities, which learns to emphasize the more reliable cues and suppresses the others by integrating Canonical Correlation Analysis as a pre-processing step. Differently, Zadeh et al. [43] present Graph-MFN, which synchronizes the multimodal sequences by storing intra-modality and cross-modality interactions through time with a graph structure. Attention mechanism has been exploited by several works as well [44], [45], [41], [46], [47], [15], [17], [48], [21], [49]. For example, Dai et al. [21] present MESM that is composed of sparse cross-modal attention mechanism attached to the joint learning of multimodal features.

There are a lot of attempts applying end-to-end learning [27], [26], [50], [51], but only [21] compared a fully end-to-end method (defined as jointly optimizing feature extraction and feature learning stages [21]) with the two-phase pipelines (i.e., feature extraction is independent from multimodal learning). Indeed, it is very common in the MER litreature to apply the feature extraction step separately. This is performed on each modality by using either hand-crafted formulations [52], [53], [29], [26], [27], [43]) and/or deep learning architectures [42], [26], [27]. As example of acoustic features; Log-Mel spectrogram [27], pitch, voiced/unvoiced segmenting features [26], [43], [29], MFCCs [28], [26], [43], [29], features extracted from SoundNet [42]) can be given. On the other hand, various backbones such as VGG16 [28], I3D [42], FaceNet [42] as well as facial features; facial landmarks and facial action units extracted by OpenFace [43], [29] are among the most popular visual features. For text, Glove embeddings [54] have been frequently utilized [26], [43], [29], [41], [55], [16], [14], while Transformers are used as the backbone [41], [55], [16], [14], [49] or LSTMs are trained with the extracted word embeddings [43], [29].

Among the aforementioned approaches, [45], [55] use text and audio, [53], [48], [42], [27], [28], [50], [52] use video and audio, and all others use text, audio and video together. It is worth noting that these techniques are all ***supervised***. Recently, Khare et al. [49] investigated the usage of large unlabeled multimodal datasets for pre-training a cross-modal transformer, which is then fine-tuned for the emotion recognition task. In detail, the VoxCeleb dataset [56], composed of 1.1 million videos that are associated to emotions [57], is used to pre-train the multimodal transformer. Then, the decoder layer is removed, and an average pooling and additional fully connected layers are added to fine-tune the model for emotion recognition task. Unlike [49], we do not rely on auxiliary large-scale datasets to pre-train our model, and both the feature learning and inference are performed on the same datasets,

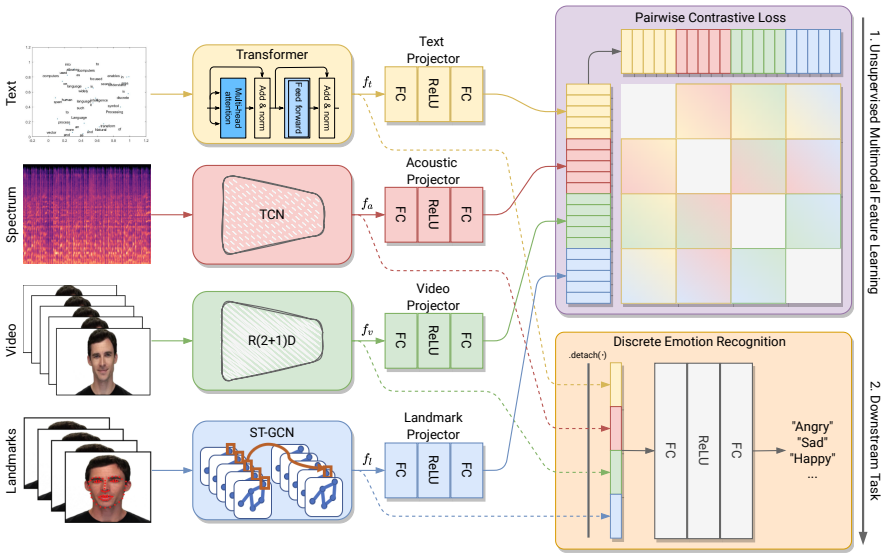

Fig. 1. **Summary of our approach.** We first learn the multimodal features in an unsupervised fashion, then the downstream task (discrete emotion recognition) is performed. We jointly train, each possible pair of modalities' backbone using contrastive loss in order to predict the correct pairings of a batch of training examples. The final loss is the average of all losses calculated. During inference, f_t, f_a, f_v, f_l are extracted before the projection layers (i.e., fc+RELU+fc) and concatenated, then feed to a linear classifier for emotion recognition.

which are much smaller than the VoxCeleb dataset [56]. Our learned features are frozen such that we do not apply any fine-tuning as in [49]. This is an important difference because some studies [58], [59] have shown that, compared to using frozen features that are learned in an unsupervised fashion, fine-tuning can bring up to 17.5% improvement for the downstream task. However, following the fine-tuning approach would not keep the feature learning methodology *"entirely unsupervised"*, as it requires the labels of the downstream task. Moreover, our model is applicable with different modality combinations, whereas text is an anchor modality in [49].

The MER litreature is very limited in terms of ***fully unsupervised feature learning*** approaches. Very recently, a Convolutional Autoencoder architecture is presented in [25]. Despite being very different from our method in terms of the architecture, [25] is still our "direct competitor" by having the following common aspects with the proposed method: *i)* performing unsupervised feature learning without fine-tuning, *ii)* being independent to the number of modalities and modality combinations, and *iii)* not being task-specific.

III. Our Approach

An overview of our approach is given in Fig. 1. First, the multimodal features are learned with an unsupervised way (Sec. III-B). Then, the downstream task (discrete emotion recognition) is performed (Sec. III-C). Sec. III-A describes the modalities and Sec. III-D presents the implementation details.

A. Modalities

The modalities and backbones we utilize are described as follows.

► **Text.** The word vectors are extracted from transcripts with the Glove word embeddings [54], following the procedure in [43]. As the backbone, we use the Transformer in [60], which is one of the SOTA architectures of language processing.

► **Visual.** We rely on two sources of visual data. One of them is the **facial images** extracted by MTCNN face detector [61] (unless faces are supplied by the dataset) from RGB video frames. As the backbone associated to the facial images, the R(2+1)D architecture [62] pre-trained on Kinetics-400 dataset [63] is used. The other visual data is the **facial landmarks** detected by the method in [64] (unless it is provided by the dataset used), and the associated backbone is Spatio-Temporal Graph Neural Network (ST-GCN) [65].

► **Acoustic.** Mel-spectograms are extracted with the same procedure and settings in [66], [41], [55] with Librosa Python Library [67] using 80 filter banks and by selecting one frame for every 16 frames. The dimension of the mel-spectograms is fixed to 128. We adapt Time Convoluted Network (TCN) [68] such that it takes mel-spectrograms as the input.

As seen, each modality has its own backbone, which have been chosen as being the SOTA architectures for diverse applications of language, visual and acoustic data processing.

B. Unsupervised Multimodal Feature Learning

The proposed method includes separate multi-layer projection heads onto each backbone defined in Sec. III-A. All projection heads have the same structure such that they are composed of fully-connected layers (fc), where the first layer is followed by a ReLU activation function ($fc_1+ReLU+fc_2$). This structure is motivated by SimCLR [18], which shows that a nonlinear projection head contributes to the performance more than a linear projection head, and its contribution is even more compared to not including any projection layer.

We adapt the CLIP fashion [69] training, *without using any labels* of the downstream task (i.e., emotion recognition).

Given a data sample represented by a sequence of observations in multiple modalities, our aim is to make the embeddings of two modalities of the same sequence (*positives*) close to each other, and make the embeddings of the same two modalities of different sequences (*negatives*) apart from each other. This is repeated for all possible pairs of modalities. Notice that negative samples might belong to the same class (*i.e.* exhibit the same emotion). However, herein, we assume that the class labels are not available, and we resort to instance discrimination with contrastive learning which encourages the model to produce invariant representations and align the latent spaces of all the modalities.

More formally, the contrastive loss function for a pair of modalities (m,n) has the following form:

$$L_i^{m,n} = -log \frac{\exp(\text{sim}(z_i^m, z_i^n)/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{[i \neq j]} \exp(\text{sim}(z_i^m, z_j^n)/\tau)}, \quad (1)$$

where z denotes the embedding after the projection, i, j are indices of samples in the current batch of size N, τ is the temperature parameter (scalar), $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$, and $sim(u, v) = \frac{u^\top v}{||u|| \ ||v||}$ denotes the dot product between ℓ_2-normalized vectors u and v (i.e., cosine similarity). Eq. (1) is computed across all samples i in the batch, resulting in $L^{m,n} = \sum_{i=1}^{N} L_i^{m,n}$. In addition, we minimize this loss for each possible pairs of modalities. Notice that, since the negatives are drawn from only one modality (see denominator in Eq. (1)), the loss is asymmetric, *i.e.*, $L^{m,n}$ is not equal to $L^{n,m}$. Therefore, our final loss function (Eq. (2)) includes the loss obtained from all the permutations of two elements drawn with replacement from the set of modalities $\mathcal{M}$:

$$L_{final} = \frac{\sum_{(m,n) \in \mathcal{M} \times \mathcal{M}} \mathbb{1}_{[m \neq n]} L^{m,n}}{|\mathcal{M}|(|\mathcal{M}| - 1)}. \quad (2)$$

Note that, we found empirically that only contrasting different modalities (*i.e.* when $m \neq n$) produces better representations. In addition, we perform temporal augmentations (see Sec. III-D for details) to the sequences in order to avoid overfitting and improve performance.

C. Discrete Emotion Recognition

Following the common practice [18], [70], in order to perform the downstream task (i.e., discrete emotion recognition), we discard the projection layers (described in Sec. III-B) and use the 512-dimensional feature representation extracted from each backbone. The extracted features are concatenated (e.g., for 3 modalities, the combined vector holds 3×512 number of features) and given to a prediction layer, that shares the same design with the projection heads (i.e., fc+RELU+fc) where its output is the emotion classes. The aforementioned prediction layers are trained with the emotion labels using the cross entropy loss and a variant of it (see Sec. IV-B for details).

D. Implementation Details

The training is performed with the SGD optimizer with the momentum of 0.9 and the weight decay of 0.001. All models are trained with the batch size of 32 (or 64) while the batch size of our downstream task is 64 (or 128). The learning rate is initialized as 0.001. We create a linear scheduler to vary the learning rate over the training process such that at every 5 epochs for CMU-MOSEI [43] and every 100 for RAVDESS [71], we multiply the learning rate with 0.9 (notice that RAVDESS dataset is much smaller than CMU-MOSEI). We do not apply any "spatial" data augmentation (e.g., random crops, blurs or color distortions), but data sampling can have overlapping sequences. For example, a video segment from t to t+10, and another video segment from t+5 to t+15 can be used in the same training. This is referred as augmentation in the temporal dimension. We set the number of epochs to 2000, but we also define a *patience parameter* such that: if after 100 consecutive epochs the validation performance does not change, then we stop the training. In practice, the maximum number of epochs was never been reached because the patience parameter stopped the training before. The temperature scalar τ is taken as 0.07.

TABLE I
RESULTS OF THE PROPOSED AND BASELINE METHODS ON RAVDESS DATASET [71] IN TERMS OF ACCURACY (ACC).

Methods	Actor Split	Facial Images	Acoustics	Facial Landmarks	ACC (%)
Unimodal	✓	✓			60.80
Unimodal	✓		✓		58.50
Unimodal	✓			✓	62.05
Late Fusion	✓	✓	✓	✓	64.10
Attention Mec.	✓	✓	✓	✓	65.40
Ours	✓	✓	✓		63.78
Ours	✓		✓	✓	**77.10**
Ours	✓	✓	✓	✓	**78.54**
Unimodal		✓			72.80
Unimodal			✓		75.90
Unimodal				✓	76.35
Late Fusion		✓	✓	✓	80.72
Attention Mec.		✓	✓	✓	81.80
Ours		✓	✓		80.32
Ours			✓	✓	**89.50**
Ours		✓	✓	✓	**93.17**

IV. EXPERIMENTS AND RESULTS

A. Datasets and Evaluation Metrics

We used the speech part of **RAVDESS** dataset [71], containing 2880 audio-visual recordings acted by 24 *professional actors* pronouncing two lexically identical statements. Each recording was labeled in terms of one of the eight categorical emotions (anger, happiness, disgust, fear, surprise, sadness, calmness and neutral), while the emotions were expressed with two intensity (normal or strong). RAVDESS is class-balanced except the neutral class, which was elicited 50% less time than the other emotion classes. We adapted two cross-validation settings following the methods [42], [48], [27], [28], [13], [72], [44], [53], [12], [52]. The first setting considers the identities of the actors such that the training (validation) and the corresponding testing *k*-folds have no overlap in terms of actors (shown as *actor-split*=✓ hereafter). The second setting,

TABLE II
RESULTS OF THE PROPOSED AND THE BASELINE METHODS ON CMU-MOSEI [43] IN TERMS OF WEIGHTED ACCURACY (*w*-ACC) AND F1 MEASURE. WOUT/ TEXT STANDS FOR THE EXPERIMENTS WHEN THE TEXT MODALITY IS NOT USED WHILE ALL OTHER MODALITIES ARE USED.

Methods	Happy		Sad		Anger		Surprise		Disgust		Fear		Overall	
	w-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1
Late Fusion	59.71	60.17	54.17	27.97	54.58	34.58	50.01	3.31	54.29	34.10	54.92	22.83	54.60	30.50
Attention Mec.	61.27	61.61	55.80	36.09	54.92	37.06	50.34	5.66	55.84	44.15	57.25	43.71	55.90	38.00
Ours wout/ text	63.96	61.84	50.71	12.41	54.88	26.59	50.30	2.76	58.37	35.44	54.79	27.56	55.50	27.77
Ours	**68.82**	**69.20**	**62.93**	**55.70**	**67.91**	**70.09**	**62.93**	**72.73**	**72.91**	**74.25**	**64.49**	**74.85**	**66.70**	**69.50**

TABLE III
RESULTS OF THE PROPOSED METHOD AND THE SOTA MER METHODS TESTED ON RAVDESS [71]. ATT STANDS FOR ATTENTION MECHANISM.

Methods	Actor Split	Feature Learning	ACC (%)
Human performance [71]	-	-	80.00
Ghaleb et al. [42]	✓	Supervised	67.70
Ghaleb et al. [48]	✓	Supervised	69.40
Ghaleb et al. [48] (w/ATT)	✓	Supervised	76.30
Radoi et al. [27]	✓	Supervised	78.70
Ours	✓	Unsupervised	78.54
Beard et al. [44]		Supervised	58.30
Song et al. [28]		Supervised	90.00
Tiwari et al. [52]		Supervised	93.30
Ours		Unsupervised	93.17

instead, applies standard k-fold cross-validation (i.e., *actor-split*=✗). In both settings, k was taken as 10 and the reported results are in terms of accuracy (ACC), which is averaged over the 10-folds, supplying fair comparisons with the MER SOTA [42], [48], [27], [28], [44], [53], [52]. As the same statements are being repeated by the actors in RAVDESS dataset [71], the proposed method (as well as the SOTA) are based only on visual and acoustic modalities.

The **CMU-MOSEI** [43] is the largest multimodal in-the-wild dataset in the MER domain. It consists of more than 23K utterances, belonging to more than 1000 speakers, collected from YouTube videos. Each utterance is labeled with six emotions: happiness, sadness, anger, fear, disgust, and surprise with a [0,3] Likert scale for the presence of each emotion class. Following [43], [21], [41], [29], [26], [46], [14], [15], [16], [17], the emotions were treated as either present or not present (i.e., binary classification), while more than one emotion can be present at the same time, making the task a multi-label problem. There exist ($\approx$ 3000) not-correctly aligned sequences across the modalities. As our approach does not require strict data alignment, we used all sequences as supplied in CMU-MOSEI SDK [73]. In other words, we did not apply any data cleaning, e.g., as in [21]. We also used the recommended dataset split and the evaluation metrics in [43], namely weighted accuracy [74] (*w*-ACC) and F1-measure.

B. Comparisons with the Baseline Methods

We compare the proposed approach with the following baseline methods. These baselines are all *supervised* such that cross-entropy and binary cross-entropy losses were used for RAVDESS [71] and CMU-MOSEI [43], respectively. The corresponding results are given in Tables I and II.

Unimodal Learning. Each modality was trained with its associated backbone (described in Sec. III-A) followed by two fully connected (fc) layers with a ReLU activation function. The best results were obtained with the following parameter settings. For acoustic data, the learning rate was initialized with 0.001 and decreased by multiplying it with 0.9 at every 10 epochs. The batch size was 32 and number of epoch was 100. For facial images, the learning rate was 0.01, number of epoch was 150 and the momentum was 0.9. For facial landmarks, the learning rate was 0.001, momentum was 0.9 and the number of epochs was set as the proposed method with patience parameter.

Late Fusion. Recall that late fusion was applied by several SOTA methods, e.g., [27], [28], [26]. Given the modalities and the backbones described, we concatenated the feature embeddings of each modality, and fed them to a shallow network composed of two fc layers with a ReLU activation function. The batch size was taken as 32, the number of epochs was set by the patience parameter, the learning rate and momentum were taken as 0.001 and 0.9, respectively.

Attention Mechanism. As mention in Sec. II, attention mechanism has been frequently applied in MER, hence we adapted it as a baseline too. We first concatenated the feature embeddings obtained from each modality (512 features extracted from each backbone as in our method) and then applied the multi-head attention mechanism of [60]. The batch size was 64, the learning rate was 0.001, and the number of epochs was set to 2000 with the patience parameter described in Sec. III-D. The same scheduler as the proposed method was used.

As seen in Table I, our unsupervised feature learning method outperforms all of the supervised baselines when acoustic and facial landmarks are involved. It is notable that, in the visual domain, the facial landmarks are more effective than the facial images. Out of all baseline methods, late fusion and attention mechanism surpass the unimodal setups, while attention mechanism achieves slightly better results than the late fusion. Overall, all methods perform better in the *actor-split*=✗ setting compared to their *actor-split*=✓ counterpart. This is perhaps as a result of having more training data in the *actor-split*=✗ setting. With reference to Table I, we have further investigated the contribution of used modalities with respect to different emotions by inspecting the confusion matrices. Our observation is that there is no particular modality or a pair of modality which performs better for a specific

TABLE IV
PERFORMANCE COMPARISONS AMONG THE PROPOSED METHOD AND THE SOTA MER METHODS TESTED ON CMU-MOSEI [43] DATASET. THE RESULTS THAT OUR METHOD SURPASSES ARE GIVEN IN YELLOW.

Methods	Happy		Sad		Anger		Surprise		Disgust		Fear		Overall	
	w-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1	*w*-ACC	F1
Unsupervised Feature Learning Methods														
CAE-LR [25]	64.70	65.60	53.20	55.60	61.80	61.90	57.10	70.70	69.00	70.10	60.40	69.20	61.03	65.52
Ours	68.82	69.20	62.93	55.70	67.91	70.09	62.93	72.73	72.91	74.25	64.49	74.85	66.70	69.50
Fully Supervised Methods														
MESM [21]	64.10	72.30	63.00	46.60	66.80	49.30	65.70	27.20	75.60	56.40	65.80	28.90	66.80	46.80
Zhang et al. [15]	71.70	–	64.30	–	66.60	–	62.30	–	72.50	–	64.60	–	67.00	–
FE2E [21]	65.40	72.60	65.20	49.00	67.00	49.60	66.70	29.10	77.70	57.10	63.80	26.80	67.60	47.40
Graph-MFN [43]	66.30	66.30	60.40	66.90	62.60	72.80	53.70	85.50	69.10	76.60	62.00	89.90	62.35	76.33
Delbrouck et al. [41]	–	64.00	–	67.90	–	74.70	–	86.10	–	83.60	–	84.00	–	76.72
Huynh et al. [51]	62.70	63.00	54.40	69.70	59.60	74.30	50.60	85.70	66.00	81.30	52.90	86.40	57.70	76.73
Khare et al. [49]	68.10	68.20	64.30	72.40	67.30	74.80	65.10	87.70	73.60	82.40	63.00	86.60	66.90	78.68
CIA [46]	51.90	71.30	61.80	72.90	67.40	74.70	58.20	86.00	74.10	81.80	63.90	87.80	62.88	79.08
Tsai et al. [14]	71.00	71.00	75.00	72.10	78.30	75.00	90.50	86.10	83.00	82.50	91.70	87.80	81.58	79.08
Wen et al. [16]	72.50	72.60	75.60	70.70	77.10	74.90	90.60	86.10	85.00	83.20	91.70	87.80	82.08	79.22
Shenoy et al. [26]	70.00	68.40	76.10	74.50	83.10	80.90	87.40	84.00	90.30	87.30	89.70	87.00	82.77	80.35
M3ER [29]	–	78.00	–	87.30	–	81.60	–	93.20	–	84.40	–	91.80	–	86.05

emotion class(es).

Given the better performances of late fusion and attention mechanism compared to unimodal learning in Table I, we inherited them to test on CMU-MOSEI dataset [43] when four modalities (text, facial images, acoustic and facial landmarks) are used. Additionally, in order to investigate the contribution of the *text modality*, we compare the results of the proposed method with the performance of the proposed method when the text is discarded (shown as wout/ text). The corresponding results can be seen in Table II. Our method outperforms the baselines for all emotion classes (especially for surprise) as well as on average (see Table II). Also, the performances of our method do not fluctuate across different emotion classes, meaning that our method generalize better than the baseline methods. In overall there exist a drop of 11.2% and 41.73% for w-ACC and F1-measure, respectively, when the text modality is discarded from the pipeline of the proposed method, showing the positive contribution of the text modality.

C. Comparisons with the State-of-the-art Methods

We compare our approach with several SOTA MER methods. Concerning RAVDESS [71], the performances are given in Table III. The fact that "human performance" is not 100% presents the difficulty of MER task. It is remarkable that our approach surpasses several supervised competitors: [42], [48], [44], [28] with a margin of 2-35% despite working in a more difficult (unsupervised) setting. It also performs on par with supervised approaches: [27], [52]. The results for CMU-MOSEI [43] are given in Table IV. There exist a very recent unsupervised feature learning approach (namely CAE-LR [25]) tested on CMU-MOSEI [43] for multimodal sentiment analysis. CAE-LR [25] achieved the best results for multimodal sentiment analysis compared to other unsupervised counterparts. Motivated by this, we adapted the authors' code for MER. Instead of applying Logistic Regression, we performed Linear Evaluation [58], which is the common protocol for unsupervised learning if the downstream task is classification (notice that we apply it for the proposed method as well, i.e., the prediction layer). For all emotion classes and on overall, our method achieves much better results than CAE-LR [25], showing the effectiveness of the contrastive loss in multimodal setting compared to convolutional autoencoders. It is worth noting that, on average, our method is better than several fully supervised techniques: MESM [21], FE2E [21], Graph-MFN [43], [51], CIA [46]. Considering that these methods integrate relatively complex supervised techniques; attention mechanisms, transformers, graphs, the better performance of our method is very promising.

V. CONCLUSION

We presented an unsupervised multimodal feature learning approach, which was tested on discrete emotion recognition. Our method is a pioneer in the MER litreature, being based on pairwise contrastive learning. Experiments show that the performance of our approach is better than the supervised baselines and unsupervised counterpart, while being competitive to several complex supervised SOTA and even surpassing a few. Being an unsupervised feature learning method, the proposed approach is transferable to other domains without retraining (not even tuning) the representation model itself.

The proposed method keeps the modality pairings the same for all data (i.e., emotions) and the way we learn the features gives equal importance to each modality. An alternative could be having different modality pairings for different emotion classes. This will be further investigated as future work.

ACKNOWLEDGMENT

This work was supported by the EU H2020 SPRING project (No. 871245) and by Fondazione VRT.

REFERENCES

[1] "SPRING: Socially pertinent robots in gerontological healthcare," https://spring-h2020.eu, accessed on 2021-12-12.

[2] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," in *In INTERSPEECH*, 2006.

[3] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion recognition for healthcare surveillance systems using neural networks: A survey," 2021.

[4] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion recognition in e-learning systems," in *In ICMCS*, 2018, pp. 1–6.

[5] P. Paikrao, A. Mukherjee, D. K. Jain, P. Chatterjee, and W. Alnumay, "Smart emotion recognition framework: A secured iovt perspective," *IEEE Consumer Electronics Magazine*, pp. 1–1, 2021.

[6] L. R. Demenescu, K. A. Mathiak, and K. Mathiak, "Age- and gender-related variations of emotion recognition in pseudowords and faces," *Experimental Aging Research*, vol. 40, no. 2, pp. 187–207, 2014.

[7] S. Olderbak, O. Wilhelm, A. Hildebrandt, and J. Quoidbach, "Sex differences in facial emotion perception ability across the lifespan," *Cognition and Emotion*, vol. 33, no. 3, pp. 579–588, 2019.

[8] J. Engelmann and M. Pogosyan, "Emotion perception across cultures: the role of cognitive mechanisms," *Frontiers in Psychology*, vol. 4, p. 118, 2013.

[9] D. Furnes, H. Berg, R. M. Mitchell, and S. Paulmann, "Exploring the effects of personality traits on the perception of emotions from prosody," *Frontiers in Psychology*, vol. 10, p. 184, 2019.

[10] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, p. 423–443, 2019.

[11] C. Beyan, S. Karumuri, and G. Volpe, "Modeling multiple temporal scales of full-body movements for," *IEEE Trans. on Affective Computing*, pp. 1–1, 2021.

[12] M. Abdullah, M. Ahmad, and D. Han, "Facial expression recognition in videos: An cnn-lstm based model for video classification," in *In ICEIC*, 2020, pp. 1–3.

[13] A. Shirian, S. Tripathi, and T. Guha, "Dynamic emotion modeling with learnable graphs and graph inception network," *IEEE Trans. on Multimedia*, pp. 1–1, 2021.

[14] Y.-H. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *In EMNLP*, 2020, pp. 1823–1833.

[15] D. Zhang, L. Wu, S. Li, Q. Zhu, and G. Zhou, "Multi-modal language analysis with hierarchical interaction-level and selection-level attentions," in *In IEEE ICME*, 2019, pp. 724–729.

[16] H. Wen, S. You, and Y. Fu, "Cross-modal dynamic convolution for multi-modal emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 78, 2021.

[17] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61 672–61 686, 2020.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *In ICML*. PMLR, 2020, pp. 1597–1607.

[19] N. Rai, E. Adeli, K.-H. Lee, A. Gaidon, and J. C. Niebles, "Cocon: Cooperative-contrastive learning," in *In IEEE/CVF CVPR Workshops*, June 2021, pp. 3384–3393.

[20] L. Wang, K. Kawakami, and A. van den Oord, "Contrastive predictive coding of audio with an adversary," in *In INTERSPEECH*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 826–830.

[21] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal end-to-end sparse model for emotion recognition," in *In the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5305–5316.

[22] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Information Sciences*, vol. 569, pp. 90–109, 2021.

[23] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *In NeurIPS*, vol. 33, 2020, pp. 22 243–22 255.

[24] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *In IEEE ICASSP*, 2021, pp. 6329–6333.

[25] P. Koromilas and T. Giannakopoulos, "Unsupervised multimodal language representations using convolutional autoencoders," *ArXiv*, vol. abs/2110.03007, 2021.

[26] A. Shenoy and A. Sardana, "Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020, pp. 19–28.

[27] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, 2021.

[28] Y. Song, Y. Cai, and L. Tan, "Video-audio emotion recognition based on feature fusion deep learning method," in *IEEE Int. Midwest Symposium on Circuits and Systems (MWSCAS)*, 2021, pp. 611–616.

[29] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *In AAAI*. AAAI Press, 2020, pp. 1359–1367.

[30] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *In IEEE ICASSP*, 2019, pp. 7390–7394.

[31] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Transformer based unsupervised pre-training for acoustic representation learning," in *In IEEE ICASSP*, 2021, pp. 6933–6937.

[32] Y. Xiao, D. Wang, and L. Hou, "Unsupervised emotion recognition algorithm based on improved deep belief model in combination with probabilistic linear discriminant analysis," *Personal Ubiquitous Comput.*, vol. 23, no. 3–4, p. 553–562, jul 2019.

[33] M. Awiszus, S. Graßhof, F. Kuhnke, and J. Ostermann, "Unsupervised features for facial expression intensity estimation over time," in *CVPR Workshops*, 2018.

[34] G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales, N. Robertson, and Y. Yang, "Deep multi-task learning to recognise subtle facial expressions of mental states," in *In ECCV*, September 2018.

[35] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," *IEEE SISY*, Sep 2021.

[36] M. Seo and M. Kim, "Fusing visual attention cnn and bag of visual words for cross-corpus speech emotion recognition," *Sensors*, vol. 20, no. 19, 2020.

[37] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, p. 145, 2020.

[38] G. Sharma and A. Dhall, *A Survey on Automatic Multimodal Emotion Recognition in the Wild*. Cham: Springer International Publishing, 2021, pp. 35–64.

[39] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," *Affect and emotion in HCI*, pp. 92–103, 2008.

[40] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *In ACM ICMI*, 2013, p. 517–524.

[41] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020.

[42] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *In ACII*, 2019, pp. 552–558.

[43] A. 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 *ACL*, 2018.

[44] R. Beard, R. Das, R. W. M. Ng, P. G. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *In CoNLL*, 2018, pp. 251–259.

[45] W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *In Grand Challenge and Workshop on Human Multimodal Language*, 2018.

[46] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *In EMNLP-IJCNLP*, 2019, pp. 5647–5657.

[47] M. S. Akhtar, D. S. Chauhan, and A. Ekbal, "A deep multi-task contextual attention framework for multi-modal affect analysis," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 3, 2020.

[48] E. Ghaleb, J. Niehues, and S. Asteriadis, "Multimodal attention-mechanism for temporal emotion recognition," in *In IEEE ICIP*, 2020, pp. 251–255.

[49] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 381–388.

[50] X. Chang and W. Skarbek, "Multi-modal residual perceptron network for audio–video emotion recognition," *Sensors*, vol. 21, no. 16, 2021.

[51] V. T. Huynh, H.-J. Yang, G.-S. Lee, and S.-H. Kim, "End-to-end learning for multimodal emotion recognition in video with adaptive loss," *IEEE MultiMedia*, vol. 28, no. 2, pp. 59–66, 2021.

[52] P. Tiwari, H. Rathod, S. Thakkar, and A. D. Darji, "Multimodal emotion recognition using sda-lda algorithm in video clips," *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[53] A. Jaratrotkamjorn and A. Choksuriwong, "Bimodal emotion recognition using deep belief network," in *Int. Computer Science and Engineering Conference*, 2019, pp. 103–109.

[54] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[55] J.-B. Delbrouck, N. Tits, and S. Dupont, "Modulated fusion using transformer for linguistic-acoustic emotion recognition," in *First International Workshop on Natural Language Processing Beyond Text*, 2020.

[56] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[57] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *In ACMMM*. New York, NY, USA: Association for Computing Machinery, 2018, p. 292–301.

[58] G. Paoletti, J. Cavazza, C. Beyan, and A. Del Bue, "Unsupervised Human Action Recognition with Skeletal Graph Laplacian and Self-Supervised Viewpoints Invariance," in *In BMVC*, 2021.

[59] L. Linguo, W. Minsi, N. Bingbing, W. Hang, Y. Jiancheng, and Z. Wenjun, "3d human action representation learning via cross-view consistency pursuit," in *In CVPR*, 2021.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *In NeurIPS*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.

[61] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[62] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *In CVPR*, 2018, pp. 6450–6459.

[63] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[64] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *In ICCV*, 2017.

[65] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *In IJCAI*. AAAI Press, 2018, p. 3634–3640.

[66] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," *In IEEE ICASSP*, pp. 4784–4788, 2018.

[67] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *In the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[68] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[69] A. Radford and J. W. e. a. Kim, "Learning transferable visual models from natural language supervision," in *In ICML*, 2021, pp. 8748–8763.

[70] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," 2021.

[71] 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, pp. 1–35, 05 2018.

[72] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[73] "CMU Multimodal SDK," github.com/A2Zadeh/CMU-MultimodalSDK, accessed on 2021-12-01.

[74] E. Tong, A. Zadeh, C. Jones, and L. Morency, "Combating human trafficking with multimodal deep models," in *In ACL*, R. Barzilay and M. Kan, Eds., 2017, pp. 1547–1556.