EmoBed: Strengthening Monomodal Emotion Recognition via Training with Crossmodal Emotion Embeddings

Jing Han, *Student Member, IEEE,* Zixing Zhang, *Member, IEEE,* Zhao Ren,
and Björn Schuller, *Fellow, IEEE*

Abstract—Despite remarkable advances in emotion recognition, they are severely restrained from either the essentially limited property of the employed single modality, or the synchronous presence of all involved multiple modalities. Motivated by this, we propose a novel crossmodal emotion embedding framework called EmoBed, which aims to leverage the knowledge from other auxiliary modalities to improve the performance of an emotion recognition system at hand. The framework generally includes two main learning components, i. e., joint multimodal training and crossmodal training. Both of them tend to explore the underlying semantic emotion information but with a shared recognition network or with a shared emotion embedding space, respectively. In doing this, the enhanced system trained with this approach can efficiently make use of the complementary information from other modalities. Nevertheless, the presence of these auxiliary modalities is not demanded during inference. To empirically investigate the effectiveness and robustness of the proposed framework, we perform extensive experiments on the two benchmark databases RECOLA and OMG-Emotion for the tasks of dimensional emotion regression and categorical emotion classification, respectively. The obtained results show that the proposed framework significantly outperforms related baselines in monomodal inference, and are also competitive or superior to the recently reported systems, which emphasises the importance of the proposed crossmodal learning for emotion recognition.

Index Terms—Crossmodal learning, emotion recognition, emotion embedding, joint training, triplet loss

✦

1 INTRODUCTION

$\mathbf{A}$UTOMATIC emotion recognition endows machines with the capability of natural and empathic communication with humans. Thus, it is considered to be essential to sustain long-term human–machine interactions and is even critical in shifting the present artificial intelligence into the next generation enhanced with emotional characteristics [1]–[4]. Over the past decades, significant advances have been made to improve the accuracy and robustness of emotion recognition systems in either a monomodal or multimodal scenario.

The *monomodal* emotion recognition systems normally independently explore the prominent features for the emotions of interest, from one specific modality, such as audio, video, image, text, or physiology [5]–[8]. With the advent of deep learning, such modality-specific systems have continually achieved promising performance [9]–[11]. In contrast, the *multimodal* systems tend to jointly utilise several modalities, with an aim to take advantage of complementary or supplementary information from different media cues [2], [12], [13]. Benefited from this, such systems have been consistently evaluated to be superior to the monomodal systems in numerous previous works [14]–[16].

Albeit the notable advantages, in the evaluation phase, most of the multimodal emotion recognition systems re-

- *J. Han, Z. Ren, and B. Schuller are with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany. E-mail: {jing.han, zhao.ren}@informatik.uni-augsburg.de*
- *Z. Zhang and B. Schuller are with Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, UK. E-mail: {zixing.zhang, bjoern.schuller}@imperial.ac.uk*

quire the synchronous presence of all modalities that are employed in the previous training phase [12]–[16]. This severely impedes their application in real life, since it is a common case that information from some particular modalities is missing. For example, a camera could be not always fixed in front of a user, or could not work in darkness, which results in invalid or missing visual signals. Likewise, a user could be silent although she/he is emotional, leading to the missing of audio data. The absence of any involved modalities often leads to corruption or performance degradation of a pre-trained multimodal system [2].

A straightforward way to address this issue often makes use of the integration of an additional component, such as a voice activity detector and a face detector, in front of the multimodal recognition systems [2]. Once the absence of a particular modality is detected, the prediction process can be automatically re-directed to another system that is trained via an accordingly reduced number of modalities. Nevertheless, such a system is normally inferior to the system with all modalities as aforementioned.

To embrace the advantages and avoid the shortages of both systems, in this article, we propose a novel *crossmodal Emotion emBedding* framework, namely *EmoBed*. The underlying idea of this framework is to transfer the knowledge from other auxiliary modalities to a target modality, in order to enhance the performance of a monomodal emotion recognition model. Basically, it consists of two main processes: the *joint multimodal training* process and the *crossmodal training* process. The former process utilises the data from multiple modalities to jointly train a shared network, with an assumption that the knowledge from different modalities

could be implicitly transferred to or fused by the network. Meanwhile, the latter process utilises a triplet loss [17], [18] to minimise the distance of intra-class representations while maximising the inter-class ones, regardless of their corresponding modality types. In doing this, it forces the extracted high-level representations cross modalities shared a similar space, where the intra-class representations have a close distance while the inter-class ones have a long distance.

This framework holds two main advantages compared with traditional multimodal emotion recognition systems [12]–[16]. Firstly, it only requires the data from auxiliary modalities in the training stage, where the knowledge is supposed to be transferred to the target modality. In the inference stage, the auxiliary modalities are not needed anymore. Therefore, it overcomes the synchronous presence problem of traditional multimodal systems. Secondly, when training the network, it does not demand a paired data. That is, the signals from different modalities are unnecessarily time-aligned. Thus, the data from heterogeneous corpora (i. e., modality mismatch) can be used for training our framework. This advantage largely releases the signal-alignment requirement of traditional multimodal systems.

Furthermore, our work is partially inspired by the multi-task learning paradigm, where multiple tasks are jointly trained with a shared network and several task-specific networks. It has been repeatedly demonstrated including in affective computing that such a learning process can lead to a better generalisation of the representations learnt from the shared networks [19]–[23]. Similarly, in this work, we assume that multiple modalities could also benefit for training a monomodal framework, through optimising the parameters of a shared network.

Overall, the major contributions of this work include: (i) We propose a novel learning framework – EmoBed – to explore knowledge from auxiliary modalities for an emotion recognition system; (ii) we jointly train the network with the heterogeneous data, which, however, is unnecessary to be paired; (iii) we extract modality-invariant emotion embeddings in a latent space via a triplet loss. Although triplet loss has been implemented in the emotion recognition literature [24], [25], it was merely utilised to distil discriminative representations within speech signals, which differs from the proposed work that aims to distil the modality-invariant emotion embeddings; finally, (iv) we comprehensively investigate the effectiveness and robustness of the model for both dimensional continuous emotion regression and categorical discrete emotion classification. Note that, for the sake of clarification, we define the traditional emotion recognition system without any crossmodal training as a *classic* emotion recognition system, whereas the system enhanced by our proposed crossmodal technologies as an *enhanced* emotion recognition system. Despite the fact that our proposed approaches can be used for more than one modality, for the sake of simplicity, in the present article, we mainly focus on the visual and audio modalities for emotion recognition, as cameras and microphones are pervasive in the world. Thus, the audio-only or video-only based system is assumed to be monomodal system.

The remainder of this article is structured as follows. First, we brief related works for crossmodal and multimodal emotion recognition systems in Sections 2. After that, in Section 3, we elaborately describe the proposed EmoBed framework in detail. Then, in Section 4 and 5, we perform comprehensive experiments on two audiovisual emotional databases, to assess the performance of the proposed approach in both emotion regression and classification tasks. Finally, in Section 6, we draw conclusions and point out some promising related research directions.

2 RELATED WORK

In this section, we summarise the most related and recently reported works on the crossmodal training and multimodal emotion recognition systems for emotion recognition, respectively.

2.1 Crossmodal Training

Recently, there have been an increasing number of studies to investigate approaches to transfer knowledge across domains or modalities [26]–[29]. One notable relevant work is SoundNet [28] for large-scale natural sound representation learning under a student-teacher training procedure. In particular, a teacher vision network can guide the student sound network to recognise acoustic events in an unsupervised learning manner. During training, KL-divergence between the posterior probabilities of the teacher and student networks is minimised, and thus, knowledge is transferred from well established visual recognition models into audio ones. Of more relevance is a recent work by Albanie et al. [30], where the authors proposed to produce aligned embeddings for speech emotion recognition, by distilling the knowledge of a strong teacher network for facial emotion recognition.

Our approach differs from the one proposed by Albanie et al. [30] mainly in two key aspects. Firstly, although the student network is learnt without access to any form of labelled audio, the main limitation of the approach is that it requires a well-trained complex teacher network, which arguably has not yet been developed in emotion recognition. In our method, however, the network is trained from scratch. Second, the results reported in the work [30] indicate that the performance of the student network falls short of the performance of the teacher network, and cannot compete with a fully supervised network. In contrast, our primary focus goes to provide additional knowledge from a second modality to assist the targeted modality. As a consequence, the performance of our approach is expected to be superior to a fully supervised monomodal network.

2.2 Multimodal Emotion Recognition

As mentioned in Section 1, multimodal fusion approaches have been widely conducted to exploit complementary information, to improve the performance and robustness of emotion recognition systems measurably [14], [31]. To this aim, various data fusion strategies have been extensively utilised, and they normally can be classified into three categories, namely, feature-level fusion, decision-level, and model-level fusion [2]. Typically, feature-level fusion (also known as early fusion) straightforwardly concatenates audio and visual features into one combined feature vector,
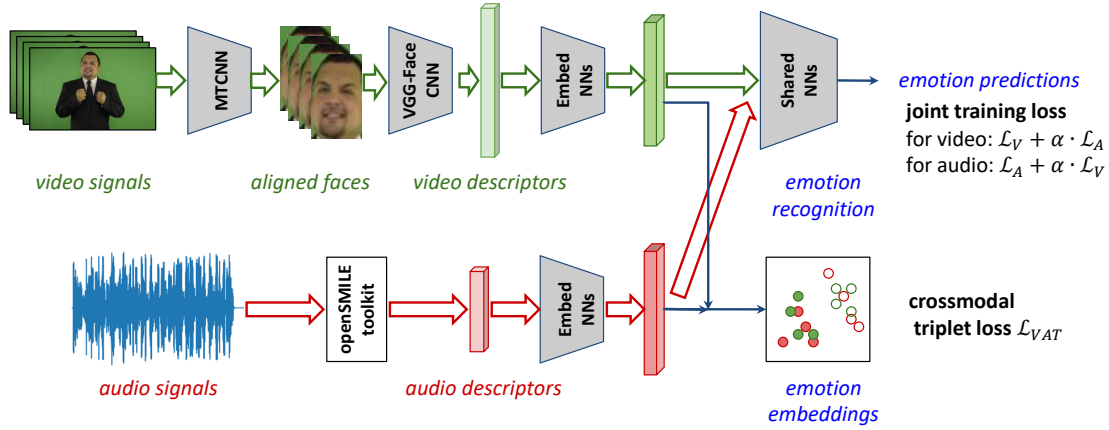
Fig. 1. The proposed crossmodal Emotion emBedding (EmoBed) framework for monomodal emotion recognition.

which is then used as the input for modelling [32], [33]. In contrast, decision-level fusion (also known as late fusion) combines the predictions, rather than the features, from the modality-specific models for a final decision by the use of certain suitable criteria [32], [34]. In addition, model-level fusion fuses the intermediate representations instead, and in this manner learns to model the potential hidden correlations among features [14], [35], [36]. Compared with the former two fusion strategies, model-level fusion is supposedly more bio-inspired from the cognitive perspective [37]. Overall, these three fusion strategies are practical and helpful, to different extents, for audiovisual emotion recognition.

However, these approaches focus on multimodal scenarios and rely heavily on the existence of signals from all sensors. In contrast to these works, we exploit the hidden correlation of multiple modalities in an implicit fusion manner, and thus it later can be implemented in a more flexible setting, as information from auxiliary modalities is not required during inference.

3 CROSSMODAL EMOTION EMBEDDING SYSTEM

In this work, we aim to attain a shared embedding space to explore the latent correlation between audio and video signals for monomodal emotion recognition, and the training stage of the proposed EmoBed framework is depicted in Fig. 1. Typically, after extracting audio and video descriptors via several standard and essential processing steps, we jointly train two modality-specific networks to project these multimodal descriptors into a common space, the representations of which can then be applied to predict emotions.

Mathematically, the two embedding functions f_A : $\mathbb{R}^M \to \mathbb{R}^E$ and f_V : $\mathbb{R}^N \to \mathbb{R}^E$, aim at mapping audio inputs in $\mathbb{R}^M$ and visual inputs in $\mathbb{R}^N$ onto embedded representations accordingly in a shared coordinate space $\mathbb{R}^E$. To learn such embedding functions, we first introduce joint training with audiovisual data in Section 3.1. Moreover, to learn useful semantic representations, we further employ crossmodal triplet loss in the learning process in Section 3.2. Lastly, the proposed EmoBed framework is given in Section 3.3, by integrating the merits of the joint training and the crossmodal training.

3.1 Joint Training with Audiovisual Data

In this subsection, we demonstrate how to learn a common embedding space for monomodal emotion recognition with joint training loss using audiovisual emotional data. Before that, we first briefly differentiate our joint training from other related structures in the following paragraphs, as depicted in Fig. 2.

Three conventional multimodal fusion paradigms are demonstrated in Fig. 2(a)-(c). Note that, although information of multiple modalities is fused at different levels, they all contribute to multimodal emotion recognition systems. Concretely, given $\mathbf{x}_{(\cdot)}$, $\mathbf{e}_{(\cdot)}$, and $y_{(\cdot)}$ denoting the monomodal input feature, the learnt hidden-layer representation, and the output prediction for audio A and video V, respectively, the combination of audio and video knowledge is in forms of $[\mathbf{x}_A; \mathbf{x}_V]$ for feature-level fusion, weighted averaging based on y_A and y_V for decision-level fusion, or $[\mathbf{e}_A; \mathbf{e}_V]$ for model-level fusion. It should be noted that, these models can be utilised, if and only if both $\mathbf{x}_A$ and $\mathbf{x}_V$ are available as inputs of the model, and there is no need for $\mathbf{e}_A$ and $\mathbf{e}_V$ to be of the same dimensions. In the proposed joint training model, albeit the constraint of the existing of both $\mathbf{x}_A$ and $\mathbf{x}_V$ remains during training, the model can then be applied under a monomodal setting.

Furthermore, our model is also dissimilar to multi-task learning, which is illustrated in Fig. 2(d). In multi-task learning, during the training phase, an auxiliary task benefits the main task by updating the parameters in the shared front-end feature-learning network. In contrast, in the proposed model, we assume inputs of an auxiliary modality can help improve the emotion prediction of the main modality, by optimising the parameters of the shared back-end predicting network.

Let us denote an audio feature vector as $\mathbf{x}_A \in \mathbb{R}^M$ and its corresponding visual feature vector as $\mathbf{x}_V \in \mathbb{R}^N$, where M and N are the dimensions of the audio and visual vectors, respectively. As depicted in Fig. 2(e), $\mathbf{x}_A$ and $\mathbf{x}_V$ are fed into two modality-specific subnetworks, the process of which can be formulated as follows:

$$\mathbf{e}_A = f_A(\mathbf{x}_A), \ \mathbf{e}_V = f_V(\mathbf{x}_V), \tag{1}$$

where the function $f_A(\cdot)$: $\mathbb{R}^M \to \mathbb{R}^E$ and the function

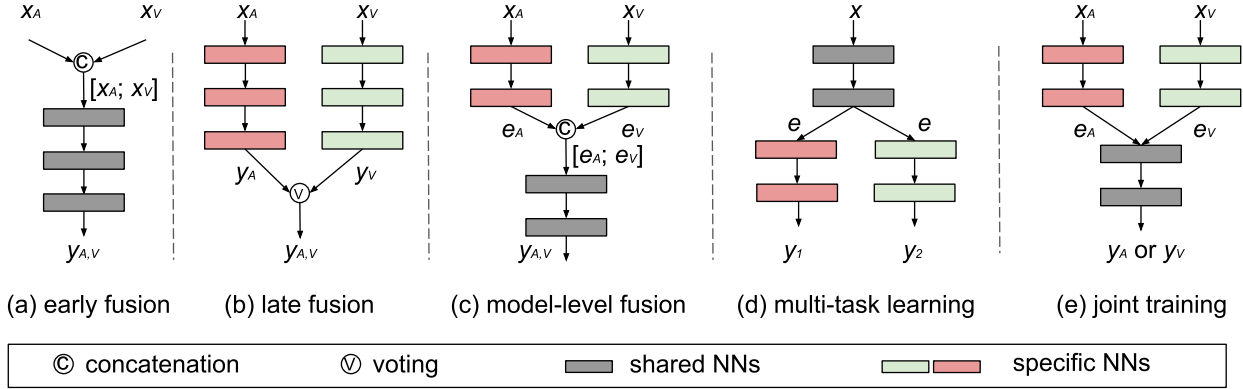Fig. 2. Structure comparison among the proposed joint audiovisual training (e), and other related multimodal learning frameworks (i.e., early fusion (a), late fusion (b), model-level fusion (c)), and multi-task learning (d).

$f_V(\cdot) : \mathbb{R}^N \to \mathbb{R}^E$ map each input of different modalities into the same subspace, resulting in corresponding E-dimensional representations $\mathbf{e}_A$ and $\mathbf{e}_V$. After that, the following shared layers are applied to estimate the final predictions, and this process can be formulated as follows:

$$y_A = f(\mathbf{e}_A), \; y_V = f(\mathbf{e}_V), \tag{2}$$

where the function $f(\cdot) : \mathbb{R}^E \to \mathbb{R}$ estimates final predictions y_A and y_V, separately.

To efficiently aggregate the advantages of different modalities for monomodal emotion recognition (i.e., speech emotion recognition or facial emotion recognition), the model is trained with a set of audiovisual features $\{(\mathbf{x}_A, \mathbf{x}_V)\}$. When the model is applied for speech emotion recognition, the joint loss function $\mathcal{J}(\theta)$ is calculated by:

$$\mathcal{J}(\theta) = \mathcal{L}_A + \alpha \cdot \mathcal{L}_V, \tag{3}$$

where θ denotes the network parameters to be optimised, $\mathcal{L}_A$ and $\mathcal{L}_V$ stand for the loss of audio and video data, respectively, and α denotes the weight of the video prediction loss to regulate its contribution to $\mathcal{J}(\theta)$. The term $\alpha \cdot \mathcal{L}_V$ enforces the optimisation to take the auxiliary modality information into account. Similarly, for facial emotion recognition, the joint loss function in Eq. (3) is altered into

$$\mathcal{J}(\theta) = \mathcal{L}_V + \alpha \cdot \mathcal{L}_A. \tag{4}$$

Moreover, the value of α is optimised on the development set, by achieving the best performance for the selected modality.

3.2 Crossmodal Triplet

The primary focus of this subsection is to learn emotion-discriminative embeddings using crossmodal data via a triplet loss function. In general, triplet loss forces to project the original descriptors into a latent space where instances with similar semantics are pulled together while instances with dissimilar semantics are pushed away. Consequently, the similarity of instances with the same semantic information is preserved in the learnt representations. While the triplet constraint has been investigated intensively for images and text [38], [39], they are now starting to gain traction in audio and video studies [24], [40]. Motivated by these successes, we propose a crossmodal triplet framework by adopting crossmodal triplet loss to supervise the learning process.

To form a triplet $tri = \{\mathbf{e}, \mathbf{e}^+, \mathbf{e}^-\}$, for the embedding of a given instance $\mathbf{e}$, we select embeddings of another two instances, i.e., $\mathbf{e}^+$ and $\mathbf{e}^-$. In particular, $\{\mathbf{e}, \mathbf{e}^+\}$ (denoted as a *positive pair*) are embeddings from the same class, with $\{\mathbf{e}, \mathbf{e}^-\}$ (denoted as a *negative pair*) belonging to different classes. To this end, we first calculate the semantic similarity of paired instances over a batch of embeddings. Typically, given a pair of embeddings $\{\mathbf{e}, \mathbf{e}'\}$, their similarity D can be computed as follows:

$$D_{\mathbf{e},\mathbf{e}'} = \|\mathbf{e} - \mathbf{e}'\|_2, \tag{5}$$

where $\|\cdot\|_2$ denotes the Euclidean distance between two embeddings in the pair. Therefore, a pairwise Euclidean distance matrix can be generated by computing the distance between all embeddings. Note that, the diagonal of the obtained matrix is null, as the distance between one embedding and itself is zero.

Subsequently, for each $\mathbf{e}$, we explore the distance matrix to find $\mathbf{e}^+$ and $\mathbf{e}^-$ from the batch, to form the hardest positive pair $\{\mathbf{e}, \mathbf{e}^+\}$ and the hardest negative pair $\{\mathbf{e}, \mathbf{e}^-\}$, by considering their emotion labels accordingly. Particularly, $\mathbf{e}^+$ is one embedding which has the maximum distance from $\mathbf{e}$, among all embeddings having the same label as $\mathbf{e}$. Conversely, $\mathbf{e}^-$ is another embedding which has the minimum distance from $\mathbf{e}$, among all embeddings with different labels from $\mathbf{e}$.

Once $\mathbf{e}^+$ and $\mathbf{e}^-$ are obtained and applied together with $\mathbf{e}$ to construct the hard triplet for each embedding, the triplet loss constraint $\mathcal{L}_T$ can be estimated over all hard triplets by

$$\mathcal{L}_T = \sum (D_{\mathbf{e}_i, \mathbf{e}_i^+} - D_{\mathbf{e}_i, \mathbf{e}_i^-}). \tag{6}$$

Note that, to compute the crossmodal triplet loss $\mathcal{L}_{VAT}$, we first combine the audio embeddings $\mathbf{e}_A$ and the video embeddings $\mathbf{e}_V$, to form a double-sized batch of embeddings in the form of $\{\mathbf{e}_A; \mathbf{e}_V\}$. Then, a pairwise Euclidean distance matrix is obtained by computing the distance between all paired embeddings. Afterwards, for each embedding (either audio or video), another two embeddings are chosen from the same batch, to form a hard triplet. It is worth mentioning that, when generating the hardest positive or negative pair, we take both the intermodal and

intramodal similarity into consideration. In this manner, the learning process forces the model to narrow the distribution gap of embeddings from different modalities, and to keep the specific emotional semantics intact in the meantime.

Supervised by the crossmodal triplet loss $\mathcal{L}_{VAT}$, the model is forced to minimise the optimisation objective $\mathcal{J}(\theta)$, which can be formulated as:

$$\mathcal{J}(\theta) = \mathcal{L}_{mon} + \beta \cdot \mathcal{L}_{VAT}, \tag{7}$$

where $\mathcal{L}_{mon}$ denotes the conventional monomodal discriminative loss, which is, in our case, $\mathcal{L}_V$ for facial emotion recognition or $\mathcal{L}_A$ for speech.

3.3 Crossmodal Emotion Embedding

In this subsection, we illustrate the proposed EmoBed framework, which integrates the triplet constraint (see Section 3.2) into the joint training approach (see Section 3.1). The overall training process is demonstrated in Fig. 1.

Generally, after extracting monomodal descriptors from standard pre-processing procedures, embedding functions $f_A(\cdot)$ and $f_V(\cdot)$ are estimated by two embedded neural networks, respectively, which project audio and video descriptors into a common latent space. Subsequently, the audio and visual embeddings are fed into a shared emotion recognition neural network, which is trained via a joint training loss. Concurrently, the training process is supervised by the triplet loss of the audio and visual embeddings.

Mathematically, when applying the EmoBed framework for audio emotion recognition, the objective function can be formatted as:

$$\mathcal{J}(\theta) = \mathcal{L}_A + \alpha \cdot \mathcal{L}_V + \beta \cdot \mathcal{L}_{VAT} + \lambda \cdot \mathcal{R}(\theta), \tag{8}$$

where $\mathcal{L}_A$ and $\mathcal{L}_V$ represent the discriminative loss function of audio and visual data, respectively, while $\mathcal{L}_{VAT}$ represents the triplet loss function of both audio and visual data. Moreover, the hyperparameters α and β are introduced to weight the contribution of the video data and the triplet loss. Furthermore, λ is applied to control the importance of the regularisation term $\mathcal{R}(\theta)$ (L2). Similarly, when training the EmoBed framework for facial emotion recognition, the objective function in Eq. (8) can be modified by exchanging $\mathcal{L}_A$ and $\mathcal{L}_V$.

After the model has finished training, the components associated with the auxiliary modality can be discarded, while the rest is retained and utilised to recognise emotions.

4 EXPERIMENTAL IMPLEMENTATION

To evaluate our approach comprehensively, we conducted extensive experiments on two multimodal emotional datasets for two tasks, respectively. Specifically, the RECOLA dataset was chosen for *dimensional emotion regression*, whereas the OMG-Emotion dataset for *categorical emotion classification*. In this section, we briefly introduce the two datasets. Then, we describe the experimental setup in detail for the sake of experiment replication, and also the evaluation metrics for a performance comparison.

4.1 Evaluated Databases and Features

4.1.1 RECOLA

This database is widely used for audiovisual dimensional emotion recognition, and a standard database which was previously applied in the AVEC since 2015 [41], [42]. It contains audiovisual recordings of spontaneous and natural interactions from 27 French-speaking participants in order to investigate socio-affective behaviours in the context of remote collaborative tasks. Moreover, time- and value-continuous dimensional emotion annotations in terms of *arousal* and *valence* are given with a constant frame rate of 40 ms for the first five minutes of each recording, by averaging all six annotators and meanwhile taking the inter-evaluator agreement into consideration [43]. The dataset is further equally divided into three disjoint parts, by balancing the gender, age, and mother tongue of the participants. Therefore, each part consists of nine unique recordings, resulting in 67.5 k segments in total for each part (training, development, or test).

When conducting experiments on RECOLA, we employed the same acoustic and visual features as the features utilised to compute the AVEC 2015 and 2016 baselines [41], [44], for a fair comparison with other methods. Specifically, as for the acoustic features, the extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS* [45]) was extracted on all audio recordings by our open-source openSMILE toolkit [46]. This resulted in a set of 88 acoustic features per segment.

In relation to the visual features, we utilised both the *appearance* and *geometric* standard features of the AVEC challenges. That is, we investigated handcrafted video features rather than learnt features from a pre-trained VGG-Face net as shown in Fig. 1, as the inputs of visual embedding nets. This is for a fair comparison with other methods. Similar to the acoustic features, the arithmetic mean and the standard derivation were computed over the sequential handcrafted visual features of each frame using a sliding window of 8 s with a step size of 40 ms. This process led to 168 appearance and 632 geometric visual features.

4.1.2 OMG-Emotion

In our experiments, the One-Minute Gradual-Emotional (OMG-Emotion) Behavior dataset [47] was employed for categorical emotion classification. The OMG-Emotion dataset is composed of 567 emotional monologue videos collected from Youtube, with an average length of one minute. These videos were then divided into utterance-level clips, and annotated by at least five annotators [47]. Seven categorical emotions were considered, i.e., *neutral, happiness, sadness, anger, surprise, fear,* and *disgust.* Majority voting was then applied to compute the gold standard based on all annotations of the same segment. Moreover, the dataset is split into training, development, and test sets, resulting in 2 440, 617, and 2 229 segments for each partition, respectively. Note that, in this work, we performed experiments and reported performances only on the development set, as labels of the test set are not yet accessible.

To extract acoustic features on the OMG-Emotion dataset, we used the eGeMAPS feature set [45], resulting

in 88 features for each utterance. For visual descriptors, firstly, a multi-task cascaded CNN [48] was applied for face detection and alignment on each frame. After that, frame-level intermediate deep features of size 4 096 were extracted from the "fc-7" layer of the VGG-Face model [49], which was pre-trained on a large number of facial images. Lastly, average pooling was conducted on all frames of the same clip to deliver the final utterance-level video descriptors.

4.2 Experimental Setup

The proposed EmoBed networks were implemented with GRU-RNNs. Compared with LSTM units, the employed GRUs have fewer parameters owing to the lack of separate memory cells and output gates, which results in a faster training process and a less-training-data demand for achieving a good generalisation [50]. Moreover, many empirical evaluations have shown that GRUs perform as competitively as LSTM units [51].

For the RECOLA experiments, we fixed the number of hidden layers for the modality-specific subnetworks (i. e., for audio and video) and the modality-shared subnetwork to be two, respectively. Each hidden layer has 120 hidden nodes. To train the network, we utilised the Adam optimisation algorithm with an initial learning rate of 0.001. Moreover, we employed a regularisation term (L2) with a weight decay of 10^{-4}, to improve the model generality. Furthermore, to facilitate the training process, we set the mini-batch size to be four. Additionally, an online standardisation was always applied to the input data by using the means and variations of the training set. All these settings were empirically recommended by our previous work on the RECOLA database after a grid search evaluation strategy [15].

For the OMG-Emotion experiments, we kept in line with the network and the training hyper-parameters, but used one hidden layer as the modality-specific or the modality-shared subnetworks due to the limited size of the OMG-Emotion dataset, and used 64 as the mini-batch size.

When training the network in a crossmodal scenario, we randomly chose the audio and video data, rather than the aligned data pair across audio and video, as the mini-batch. The advantage of this method is that, it does not require the synchronous presence of both modalities in the training phase. This means that we can principally mix the audio-only and video-only databases to complete the network training process.

Additionally, when continuously assessing emotions, the annotators have to take time to perceive acoustic events, understand them, and report the emotional states [52]. To address this annotation delay problem, we took a widely used explicit compensation approach. That is, we shifted the gold standard back in time (i. e., 2.4 s) with respect to the features for all modalities and tasks [41], with an assumption that the delay is invariant with annotators, annotator states, modalities, and tasks.

To refine the obtained predictions for continuous emotion recognition, we further performed a chain of post-processing, including median filtering, centring, scaling, and time-shifting, as suggested by Valstar et al. [41]. The filtering window size W (from 0.12 s to 0.44 s at a rate of 0.08 s) and the time-shifting delay D (from 0.04 s to 0.60 s at a step of 0.04 s) were optimised using a grid search method. All the post-processing parameters were optimised on the development set and then applied to the test set.

4.3 Evaluation Metrics

To evaluate the performance of the continuous emotion regression model, we took the *Concordance Correlation Coefficient* (CCC), which was officially recommended by the AVEC 2015-2018 challenges [41]. The CCC is defined by:

$$r_c = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \qquad (9)$$

where r represents *Pearson's correlation coefficient* (PCC) between two time series (e. g., prediction and gold-standard), μ_x and μ_y denote the mean of each time series, and σ_x^2 and σ_y^2 stand for the corresponding variances. Compared with PCC, the CCC considers not only the shape similarity between the two series but also the value precision. This is especially relevant for estimating the performance of time-continuous emotion prediction models, as both the trends as well as absolute prediction values are relevant for describing the performance of a model. The CCC metric falls into the range of [-1, 1], where $+1$ represents perfect concordance, -1 total discordance, and 0 no concordance at all.

As to the discrete emotion classification tasks, we chose F1 as the evaluation metric, mainly due to the fact that i) it can provide an overview performance in a multi-class setting as it is calculated by the harmonic mean of unweighted precision and recall; ii) it was employed by the OMG-Emotion challenge in 2018 [47]. In general, a higher CCC or F1 indicates a better prediction performance.

Further, to evaluate the statistical significance of performance improvement, unless stated otherwise, we undertook a *Fisher r-to-z transformation* [53] for dimensional continuous emotion regression and a one-tailed z-test for categorical discrete emotion classification. Only if the p-value was lower than .05, a significant difference was triggered.

5 EXPERIMENTAL RESULTS AND DISCUSSIONS

For the sake of fair performance comparison, in this section, we report on conducted experiments on two emotional databases (i. e., RECOLA and OMG-Emotion) with their corresponding standard testbeds of the AVEC 2016 [41] and OMG-Emotion challenges 2018 [47].

5.1 Results on RECOLA

As suggested in [41], for these experiments, we took two visual feature sets (i. e., appearance and geometric) and one acoustic feature set (i. e., eGeMAPS), as aforementioned in Section 4.1.1. This leads to four experimental scenarios: audio (+video-app.), video-app. (+audio), audio (+video-geo.), and video-geo. (+audio), where the modalities shown in the parentheses indicate the employed auxiliary modalities for the corresponding single modalities. In addition, we evaluated the systems on both dimensional *arousal* and *valence* regressions as well. For the classic monomodal systems, we independently performed the training process on either

TABLE 1
Performance comparison in terms of CCC for the **arousal** prediction among the proposed EmoBed systems, related baselines, and other state-of-the-art systems. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions of the RECOLA database. Three feature sets (audio-eGeMAPS, video-appearance, and video-geometric) were employed to evaluate all approaches. Four monomodal scenarios are considered: audio (+video-app.), video-app. (+audio), audio (+video-geo.), video-geo. (+audio), where the modalities in the parentheses are the employed auxiliary modalities. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "⋆" symbol. MTL: multi-talk learning; DDAT: dynamic difficulty awareness training; RE: reconstruction error; PU: perception uncertainty.

CCC	audio (+video-app.)		video-app. (+audio)		audio (+video-geo.)		video-geo. (+audio)	
our frameworks	dev	test	dev	test	dev	test	dev	test
classic monomodal	.766	.605	.512	.411	.766	.605	.499	.399
joint audiovisual training	.769	.611	.520	.401	.769	.611	.515	.413
crossmodal triplet training	.795	.633	.541	.465	.794	.632	.512	.397
EmoBed	.792⋆	.644⋆	**.549⋆**	**.475⋆**	.793⋆	.639⋆	.527⋆	**.417⋆**
	audio-only		video-app.-only		audio-only		video-geo.-only	
state of the art	dev	test	dev	test	dev	test	dev	test
SVR + offset, 2016 [41]	.796	.648	.483	.343	.796	.648	.379	.272
Strength modelling, 2017 [14]	.755	.666	.350	.196	.755	.666	—	—
End-to-end, 2017 [54]	.786	**.715**	.371	.435	.786	**.715**	—	—
MTL (RE), 2017 [55]	.788	.629	.512	.425	.788	.629	.502	.324
MTL (PU), 2017 [21]	.803	.654	.502	.406	.803	.654	.508	.327
Curriculum learning, 2018 [56]	.687	.591	.417	.343	.687	.591	.394	.267
DDAT (RE), 2018 [15]	.807	.694	.539	.437	.807	.694	**.544**	.400
DDAT (PU), 2018 [15]	**.811**	.664	.518	.438	**.811**	.664	.513	.397

TABLE 2
Performance comparison in terms of CCC for the **valence** prediction among the proposed EmoBed systems, related baselines, and other state-of-the-art systems. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions of the RECOLA database. Three feature sets (audio-eGeMAPS, video-appearance, and video-geometric) were employed to evaluate all approaches. Four monomodal scenarios are considered: audio (+video-app.), video-app. (+audio), audio (+video-geo.), video-geo. (+audio), where the modalities in the parentheses are the employed auxiliary modalities. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "⋆" symbol. MTL: multi-talk learning; DDAT: dynamic difficulty awareness training; RE: reconstruction error; PU: perception uncertainty.

CCC	audio (+video-app.)		video-app. (+audio)		audio (+video-geo.)		video-geo. (+audio)	
our frameworks	dev	test	dev	test	dev	test	dev	test
classic monomodal	.504	.381	.545	.525	.504	.381	.619	**.529**
joint audiovisual training	.505	.393	.546	.511	.513	.395	.622	.527
crossmodal triplet training	.515	.404	.564	.517	.512	.405	.636	.517
EmoBed	.514⋆	**.434⋆**	.564⋆	.516	**.521⋆**	**.439⋆**	**.645⋆**	.520
	audio-only		video-app.-only		audio-only		video-geo.-only	
state of the art	dev	test	dev	test	dev	test	dev	test
SVR + offset, 2016 [41]	.455	.375	.474	.486	.455	.375	.612	.507
Strength modelling, 2017 [14]	.476	.364	.592	.464	.476	.364	—	—
End-to-end, 2017 [54]	.428	.369	**.637**	**.620**	.428	.369	—	—
MTL (RE), 2017 [55]	**.519**	.331	.529	.366	.519	.331	.632	.488
MTL (PU), 2017 [21]	.506	.416	.468	.418	.506	.416	.643	.452
Curriculum learning, 2018 [56]	.159	.174	.446	.419	.159	.174	.300	.269
DDAT (RE), 2018 [15]	.508	.422	.528	.457	.508	.422	.639	.471
DDAT (PU), 2018 [15]	.498	.407	.514	.431	.498	.407	.632	.501

audio or video data. That was achieved by setting α to be 0.0 in Eq. (3) and Eq. (4), respectively.

The obtained results for arousal and valence predictions are presented in Tables 1 and 2, respectively. From the two tables, it can be seen that our classic monomodal systems outperform the challenge benchmarks that used the 'SVR + offset' approach [41] in most cases. One exception is the arousal prediction with audio signals, which probably is attributed to the fact that a fixed network structure, rather than the optimised one on the arousal prediction, was employed in our experiments (see Section 4.2). These results further confirm that GRU-RNNs hold the powerful capability to capture long-term context dependence.

Furthermore, when jointly training audio and video data (see Section 3.1), one may notice that the corresponding monomodal systems (based on either audio-eGeMAPS, video-appearance, or video-geometric) can deliver higher CCCs compared with the classic monomodal systems, in seven out of eight cases for the arousal prediction and six out of eight cases for the valence prediction, respectively. This observation implies that such a joint training process can somewhat transfer shared semantic information from other heterogeneous data to the target modality thanks to the implementations of i) a shared subnetwork and ii) a

multi-task learning framework.

Rather than the joint audiovisual training, when performing the triplet constraint across the audio and video modalities (see Section 3.2), the obtained CCCs (the third lines in Tables 1 and 2) show that the introduced enhanced monomodal systems again generally significantly outperform the classic monomodal systems in most cases (Fisher r-to-z transformation, $p < .05$). This suggests that the implementation of the triplet constraint is helpful to distil emotional discriminative representations, not only in a monomodal scenario [24] but also in a crossmodal scenario as investigated in this article.

Finally, when we simultaneously carried out the crossmodal triplet training as well as the joint audiovisual training processes, it can be seen that the EmoBed systems achieve the best performance in most cases, i. e., six out of eight cases for the arousal regression and five out of eight cases for the valence regression. For example, the obtained CCCs on the test set of the audio-based model are boosted from .605 to .644 and .639 for arousal regression, and from .381 to .434 and .439 for valence regression, when respectively integrating with video-appearance and video-geometric feature sets in the training process. Furthermore, the best CCCs achieved by the video-based models reach to .475 and .417 with appearance and geometric feature sets, respectively, with absolute CCC increases of .064 and .018 compared with the classic monomodal systems for arousal regression. Such an observation, however, cannot be found for the video-based valence regression models on the test set. This exception possibly is attributed to the distribution mismatch between the development and test partitions. Overall, it is concluded that the proposed EmoBed can largely supply additional knowledge from audio signals to alleviate the shortage of video signals, and vice verse.

Meanwhile, as presented in Tables 1 and 2, the EmoBed systems achieve comparable or better performance to other state-of-the-art methods, such as the end-to-end systems [54], curriculum learning systems [56], multi-task learning systems [21], and Dynamic Difficulty Awareness Training (DDAT) systems [15]. Although most of these reported systems utilised a variety of multimodal fusion approaches to boost their final performance, these systems, in the reference stage, demand the simultaneous occurrence of all modalities that appear in the training stage, to the best of our knowledge. It also has to be noted that the best achieved results are delivered by using the end-to-end system for the arousal predictions when using audio signals (i. e., .715 of CCC) and for the valence predictions when using video signals (i. e., .620 of CCC). This observation further confirms that optimising the whole network from the raw signals to the target can benefit the efficient extraction of high-level representations. Therefore, in the future we will further implement the EmoBed system in an end-to-end fashion, which is expected to further enhance the performance of monomodal emotion recognition.

5.2 Results on OMG-Emotion

For our experiments on the OMG-Emotion database, we conducted seven-class categorical emotion classification tasks on audio and visual signals. Table 3 presents the

TABLE 3
Performance of the proposed EmoBed systems, related baselines, and other reported systems in terms of F1 on the development set of the OMG-Emotion dataset. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "⋆" symbol.

F1 [%]	audio	video
other approaches		
SVM [47]	33.0	—
RF [47]	—	37.0
our frameworks		
classic monomodal	36.5	37.9
joint audiovisual training	40.2	42.1
crossmodal triplet training	40.7	41.0
EmoBed	**41.7⋆**	**43.9⋆**

performance of the models in terms of F1 on the development set (note that the annotations of the test set are not publicly available). From the table, we can see that on this database, our classic monomodal models outperform the other methods reported in the literature [47], i. e., Support Vector Machine (SVM) and Random Forest (RF). More specifically, our classic monomodal models yields higher F1 than SVM (36.5 % vs 33.0 %) for audio, and than RF (37.9 % vs 37.0 %) for video.

Additionally, comparing the proposed joint audiovisual training models with the classic monomodal systems, it is noticed that the former approach outperforms the latter one by a large margin, i. e., 40.2 % vs 36.5 % for audio and 42.1 % vs 37.9 % for video. These experimental results again indicate that, the proposed joint audiovisual training approach is plausible to promote performances of monomodal emotion classification. Furthermore, similar results are also obtained when utilising the triplet training approach to distil the salient representations across multiple modalities. Nevertheless, the highest F1s are achieved by means of the EmoBed systems, which deliver 5.2 % and 6.0 % absolute performance gain compared with the classic monomodal systems when using audio or video signals, respectively. All these observations further confirm the findings discovered from the RECOLA database.

5.3 Visualisation of Emotion Embeddings

To investigate how the proposed crossmodal learning framework benefit emotion recognition, we extracted the learnt representations from the classic monomodal systems and the proposed EmoBed systems. Fig. 3 illustrates the distribution of the learnt representations on the development set of the RECOLA database by means of t-Distributed Stochastic Neighbour Embedding (t-SNE). It can be seen that with the classic monomodal systems, the learnt representations can be easily distinguished into three parts by the modalities they stem from, in either arousal (cf. Fig. 3 (a)) or valence (cf. Fig. 3 (c)) prediction. Specifically, the representations learnt from different modalities almost have no overlap albeit they belong to the same emotion states. In stark contrast, the representations extracted from EmoBed systems are visibly clustered together based on their emotional properties (cf. Fig. 3 (b) and (d) for arousal and valence, respectively).
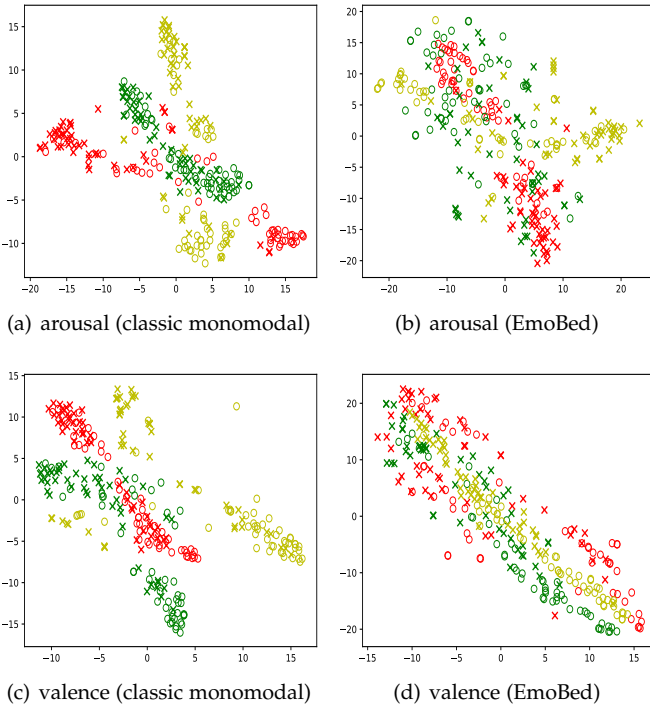
(a) arousal (classic monomodal)   (b) arousal (EmoBed)

(c) valence (classic monomodal)   (d) valence (EmoBed)

Fig. 3. Visualisation of the learnt representations of the development set of the **RECOLA** database when using the proposed EmoBed systems or the classic monomodal systems. Red, green, and yellow markers: representations from audio (eGeMAPS), video (appearance), and video (geometric) modalities; circle and cross markers: high and low arousal/valence.

(a) classic monomodal   (b) EmoBed

Fig. 4. Visualisation of the learnt representations of the development set of the **OMG-Emotion** database when using the proposed EmoBed systems or the classic monomodal systems. Red and green markers: representations from audio and video modalities; circle and cross markers: neutral and sad categories.

Such an observation is even more noticeable on the OMG-Emotion database (cf. Fig. 4). Note that, for the sake of simplicity, we merely chose two emotional categories (i. e., sad and neutral) for visualisation. Likewise, one can find that the representations belonging to the same emotional category share almost the same latent space.

These findings indicate that the representations learnt by the proposed EmoBed are somewhat invariant to the modalities. By making use of the emotion embedding space, the emotional representations extracted from audio and video signals are able to implicitly fuse the knowledge from each other. Thus, the exploitation of mutual information possibly leads to performance improvement for a monomodal system.

5.4 Discussion

To demonstrate the importance of the learning from auxiliary modalities for the monomodal emotion recognition system, we independently investigated the impact of weight change with respect to the counterpart modality, when in a joint audiovisual training process or in a crossmodal triplet training process.

Fig. 5 (a) and Fig. 6 (a) depict the relationship between the obtained CCCs and the weight α (cf. Eq. (3) and (4)) on the RECOLA database. It is noted that the model performance is improved when the weight increases to some values for the video-based arousal regression models (green and cyan lines in Fig. 5 (a)). Similar observations can be made as well for the audio-based valence regression models
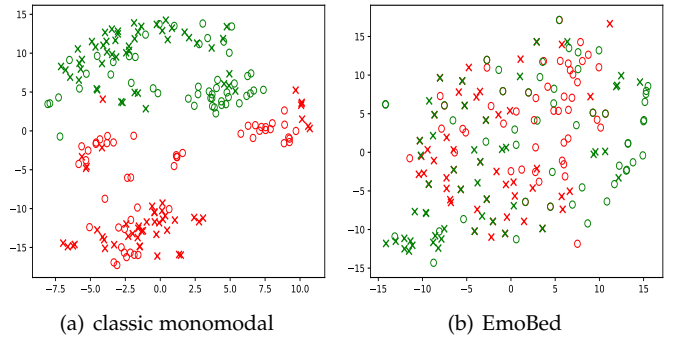
(blue and red lines in Fig. 6 (a)). Therefore, this behaviour again indicates that learning from other modalities indeed can help the enhancement of traditional monomodal systems. Yet, it is also noted that the audio-based arousal and the video-based valence regression models almost remain without obvious performance improvement. This might suggest that transferring the information from the modality with richer knowledge to the one with sparse knowledge is much easier than the other way around, as audio signals often lead to higher CCCs for arousal regression while video signals for valence regression.

Further, Fig. 5 (b) and Fig. 6 (b) illustrate the relationship between the obtained CCCs and the weight β (cf. Eq. (7)) on the RECOLA database. Obviously, it can be seen that the obtained CCCs remarkably grow with the increase of weight β in all cases for arousal (cf. Fig. 5 (b)) and valence (cf. Fig. 6 (b)) regression. Specifically, when $\beta = 1.0$, i. e., the triplet training contributes equally as the traditional emotion regression training, the systems yield the best CCCs in all cases for arousal regression. Nevertheless, the audio- and video-based valence regression systems deliver the best CCCs only when $\beta = 0.4$ and $\beta = 0.8/1.2$, respectively. The lower contribution from triplet loss implies that it might be more difficult to distil the valence-salient representations than the arousal-salient representations by means of triplet training.

Moreover, we conducted a similar investigation on the OMG-Emotion database for categorical emotion classification. Fig. 7 explicitly quantifies the contributions of joint audiovisual training (a) and crossmodal triplet training (b) when in a crossmodal learning framework. For a joint audiovisual training system, when $\alpha = 0.0$, i. e., no contribution from the auxiliary modality, the model is learnt based on only the loss of each modality, separately. When α increases, i. e., the contribution of the auxiliary modality during training increases, the performance of monomodal emotion recognition (audio or video) is improved first, until a point where the contribution of the auxiliary modality might actually penalise the learning objective too much and even harm the learning of the main modality, and thus performances start to decrease. Similar observations can be found for crossmodal triplet training systems.

(a) joint audiovisual training   (b) crossmodal triplet loss

audio (video-app.)   video-app. (audio)
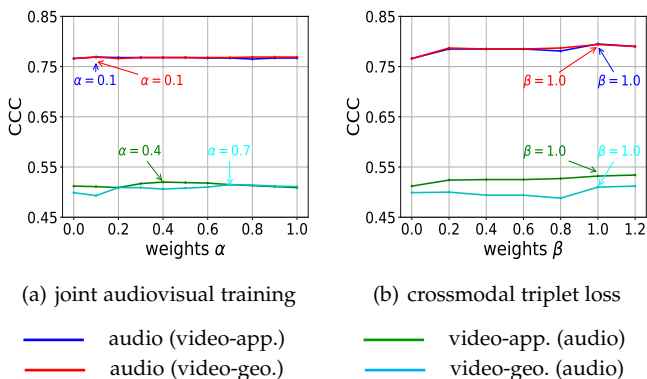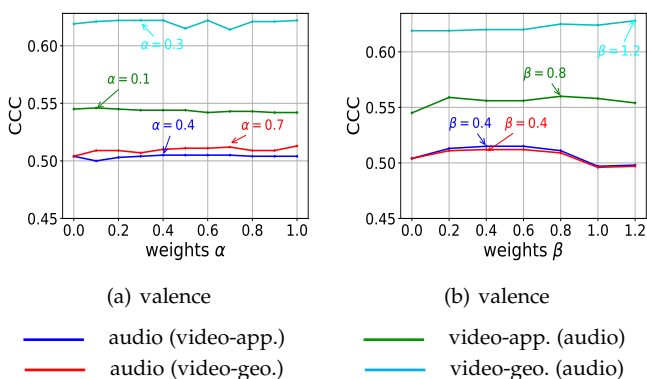
audio (video-geo.)   video-geo. (audio)

Fig. 5. Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems (a), and impact of the crossmodal triplet loss on the *crossmodal triplet training* systems (b), with the **RECOLA** database for **arousal** regression. The best performed α or β is indicated in each case.

(a) valence   (b) valence

audio (video-app.)   video-app. (audio)

audio (video-geo.)   video-geo. (audio)

Fig. 6. Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems (a), and impact of the crossmodal triplet loss on the *crossmodal triplet training* systems (b), with the **RECOLA** database for **valence** regression. The best performed α or β is indicated in each case.

To this end, proper values of the weight α and β need to be identified for the tasks at hand. We can observe from the figure that, the best performance for both audio and video emotion classification is reached when $\alpha = 0.5$ in joint audiovisual training systems; whereas the best performance for audio and video emotion classification is achieved when $\beta = 0.3$ and $\beta = 0.8$, respectively, in crossmodal triplet training systems.

6 CONCLUSION

Different from previous emotion recognition works which have focused on either the traditional monomodal (i. e., modality-specific) systems or multimodal systems, in this article, we proposed an enhanced system through exploring the information across auxiliary modalities. To implement this system with exemplary audio and visual modalities, we, on the one hand, utilised a shared emotion recognition network for both audio and video data, so that the complementary information from an auxiliary modality can be implicitly transferred to the target modality. On the other hand, we applied a triplet constraint over acoustic and visual representations to distil emotional embeddings that are invariant to the modalities. The proposed learning frameworks were systematically evaluated on the two
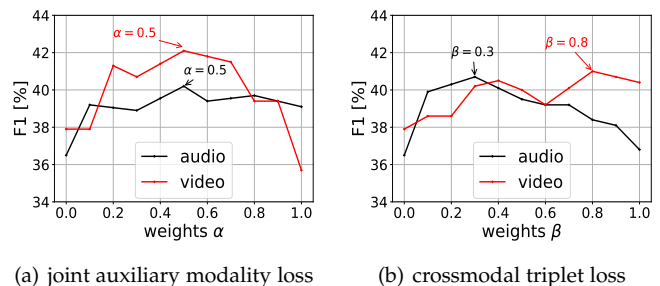

(a) joint auxiliary modality loss   (b) crossmodal triplet loss

Fig. 7. Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems (a), and impact of the crossmodal triplet loss on the *crossmodal triplet training* systems (b), with the **OMG-Emotion** database. The best performed α or β is indicated in each case.

benchmark databases RECOLA and OMG-Emotion. Experimental results have demonstrated that the proposed methods significantly improve the prediction performance of a monomodal system, by fusing an additional modality in the training process.

Albeit the efficiency, the proposed learning framework could be further developed in the future. For example, in the triplet training process, the annotation uncertainty information could be utilised as a new distance measure between the learnt representations. In addition, it is also worth to train the model by using large-scale heterogeneous datasets from a variety of domains. Moreover, in this work, we independently conducted the feature extraction process before performing the crossmodal learning process. It might be helpful if we further combine the two processes together as an end-to-end framework.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[3] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 406–417, Oct. 2014.

[4] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, Mar. 2016.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[6] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[7] C. Clavel and Z. Callejas, "Sentiment analysis: From opinion mining to human-agent interaction," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 74–93, Jan. 2016.

[8] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Transactions on Affective Computing*, June 2018, 20 pages.

[9] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, Mar. 2017, 14 pages.

[10] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives," *IEEE Computational Intelligence Magazine*, 2018, 13 pages.

[11] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[12] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[13] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.

[14] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-world automatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, vol. 65, pp. 76–86, Sep. 2017.

[15] Z. Zhang, J. Han, and B. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, vol. PP, Sep. 2018, 13 pages.

[16] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, Oct. 2018.

[17] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, no. Mar., pp. 1109–1135, 2010.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 815–823.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 1, pp. 1–29, Mar. 2012.

[20] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, Jan. 2017.

[21] J. Han, Z. Zhang, M. Schmitt, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc. ACM International Conference on Multimedia (MM)*, Mountain View, CA, 2017, pp. 890–897.

[22] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proc. 8th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Seoul, South Korea, 2018, pp. 65–72.

[23] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, Dec. 2018, 14 pages.

[24] J. Han, Z. Zhang, G. Keren, and B. Schuller, "Emotion recognition in speech with latent discriminative representations learning," *Acta Acustica united with Acustica*, vol. 104, no. 5, pp. 737–740, Sep. 2018.

[25] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India, 2018, pp. 3673–3677.

[26] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[27] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[28] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain, 2016, pp. 892–900.

[29] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM International Converence on Multimedia (MM)*, Mountain View, CA, 2017, pp. 154–162.

[30] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. ACM International Conference on Multimedia (MM)*, Seoul, Korea, 2018, pp. 292–301.

[31] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, Jan. 2012.

[32] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. International Conference on Multimodal Interfaces (ICMI)*, State College, PA, 2004, pp. 205–211.

[33] M. Wimmer, B. Schuller, D. Arsic, B. Radig, and G. Rigoll, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, Funchal, Portugal, 2008, pp. 145–151.

[34] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 2462–2465.

[35] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. S. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005, pp. 967–972.

[36] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5005–5009.

[37] S. Toprak, N. Navarro-Guerrero, and S. Wermter, "Evaluating integration strategies for visuo-haptic object recognition," *Cognitive computation*, vol. 10, no. 3, pp. 408–425, June 2018.

[38] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Prof. International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*, Copenhagen, Denmark, 2015, pp. 84–92.

[39] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. International Conference on Information and Knowledge Management (CIKM)*, San Francisco, CA, 2013, pp. 2333–2338.

[40] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 1002–1014, Apr. 2018.

[41] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[42] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftçi, H. Güleç, A. Salah, and M. Pantic, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. 8th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Seoul, South Korea, 2018, pp. 3–13.

[43] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, 2013, pp. 1–8.

[44] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Brisbane, Australia, 2015, pp. 3–8.

[45] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[46] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM International Conference on Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[47] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-emotion behavior dataset," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, Rio, Brazil, 2018, pp. 1408–1412.

[48] 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, Oct. 2016.

[49] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Machine Vision Conference*, Swansea, UK, 2015, pp. 1–12.

[50] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, Doha, Qatar, 2014, pp. 103–111.

[51] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 2342–2350.

[52] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, Apr. 2015.

[53] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Abingdon, UK: Routledge, 2013.

[54] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on End-to-End Speech and Language Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[55] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 2367–2371.

[56] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *arXiv preprint arXiv:1805.10339*, May 2018.

Zixing Zhang (M'15) received his master degree in physical electronics from the Beijing University of Posts and Telecommunications (BUPT), China, in 2010, and his PhD degree in computer engineering from Technical University of Munich (TUM), Germany, in 2015. Currently, he is a research associate with the Department of Computing at the Imperial College London (ICL), UK, since 2017. Before that, he was a postdoctoral researcher at the University of Passau, Germany, from 2015 to 2017. He has (co-)authored more than 80 publications in peer-reviewed books, journals, and conference proceedings to date. His research mainly focuses on deep learning technologies for the speaker-centred state and health computing. He has organised special sessions, such as at the IEEE 7th Affective Computing and Intelligent Interaction (ACII) conference in 2017 and at the 43nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2018. Moreover, he serves as a reviewer for numerous leading-in-their fields journals and conferences, a programme committee member and an area chair for many international conferences, and a leading guest editor of the IEEE Transactions on Emerging Topics in Computational Intelligence.

Zhao Ren received her master degree in Computer Science and Technology from the Northwestern Polytechnical University (NWPU) in the P.R. China, 2017. Currently, she is a EU-researcher and working on her Ph.D. degree at the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, where she is involved with the H2020-MSCA-ITN-ETN project TAPAS, for health-related speech analysis. Her research interests mainly lie in transfer learning, unsupervised learning, and deep learning for the application in health care and wellbeing.

Björn Schuller (M'05-SM'15-F'18) received his diploma in 1999, his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is Professor of Artificial Intelligence in the Department of Computing at the Imperial College London/UK, where he heads GLAM – the Group on Language, Audio & Music, Full Professor and head of the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, and CEO of audEERING. He was previously full professor and head of the Chair of Complex and Intelligent Systems at the University of Passau/Germany. Professor Schuller is President-emeritus of the Association for the Advancement of Affective Computing (AAAC), elected member of the IEEE Speech and Language Processing Technical Committee, Fellow of the IEEE, and Senior Member of the ACM. He (co-)authored 5 books and more than 800 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 23 000 citations (h-index = 73). Schuller is general chair of ACII 2019, co-Program Chair of Interspeech 2019 and ICMI 2019, repeated Area Chair of ICASSP, and former Editor in Chief of the IEEE Transactions on Affective Computing next to a multitude of further Associate and Guest Editor roles and functions in Technical and Organisational Committees.

Jing Han (S'16) received her bachelor degree (2011) in electronic and information engineering from Harbin Engineering University (HEU), China, and her master degree (2014) from Nanyang Technological University, Singapore. She is now working as a doctoral student with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Germany, involved in two EU's Horizon 2020 projects SEWA and RADAR CNS. She reviews regularly for IEEE Transactions on Cybernetics, IEEE Access, and IEEE Signal Processing Letters. Her research interests are related to deep learning for multimodal affective computing and health care. Besides, she co-chaired the 7th Audio/Visual Emotion Challenge (AVEC) and workshop in 2017, and served as a program committee member of the 8th AVEC challenge and workshop in 2018. Moreover, she was awarded student travel grants from IEEE SPS and ISCA to attend ICASSP and INTERSPEECH in 2018.