Privileged Knowledge Distillation for Dimensional Emotion Recognition in the Wild

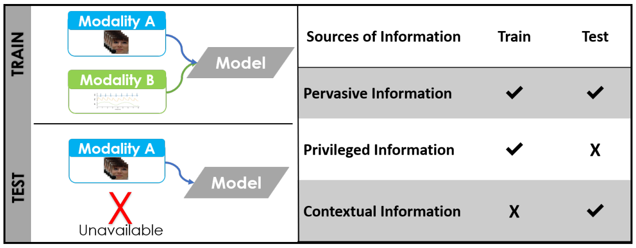
Muhammad Haseeb Aslam [1], Muhammad Osama Zeeshan[1], Marco Pedersoli[1],
Alessandro L. Koerich[1], Simon Bacon[2], Eric Granger[1]
[1] LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
[2]Dept. of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, Canada
muhammad-haseeb.aslam.1@ens.etsmtl.ca, eric.granger@etsmtl.ca

Abstract

Automated emotion recognition (AER) has a growing number of applications, ranging from behavior analysis in assistive robotics and e-learning to depression and pain estimation healthcare. Systems for multimodal AER typically outperform unimodal approaches due to the complementary and redundant semantic information across modalities like visual, audio, language, physiological, etc. However, in practice, only a subset of these modalities is available at inference time, and using multiple modalities increases systems complexity. This paper focuses on video-based AER and aims to enhance the accuracy of unimodal systems by leveraging the Learning Under Privileged Information (LUPI) paradigm with information from multiple modalities. Without loss of generality, this study considers the audio modality as privileged information (only available during training), and introduces a new multimodal to unimodal privileged knowledge distillation (PKD). The teacher network is comprised of a multimodal AER architecture that can process audio-visual information and distills the learned knowledge to a unimodal visual student network. We validate our proposed multimodal PKD method on the challenging RECOLA and Affwild2 datasets for video-based AER, using weak and strong baseline AER architectures, as well as joint cross-attention fusion methods. The proposed method increases the absolute average concordance correlation coefficient accuracy by 8% on the RECOLA dataset, and by 2% on the arousal dimension of the Affwild2 dataset. The code available at multimodal-pkd.

1. Introduction

Automatic emotion recognition (AER) is a challenging problem due to the complex and diverse nature of expressions across individuals and cultures. AER in the wild is a challenging task owing to the fact that real-world emotion portrayal is more subtle than in lab-controlled environments [4, 13]. In lab-controlled environments, emotions are acted or elicited through designed mechanisms, leading to emotion representations that are vivid in nature, and making them easier to distinguish [6]. In real-world environments, emotions are often captured across various modalities, such as facial, vocal, verbal, textual, and body language. The affective state of a person is also reflected in physiological signals like Electroencephalogram (EEG), Electrocardiogram (ECG), Electrodermal Activity (EDA), Heart Rate Variability (HRV), etc [7]. These modalities contain complementary and redundant information about an individual's affective state.



Sources of Information	Train	Test
Pervasive Information	✓	✓
Privileged Information	✓	X
Contextual Information	X	✓

Figure 1. Given a specific real-world video-based AER application, different types of information are available during training and testing. Prevalent information is available at both train and test times (see modality A), while contextual information is only available at test time, and privileged information is only available at train time. Modality B (green) is unavailable at test time, and hence this information is regarded as a privileged modality.

A multimodal analysis is typically considered as a strategy to improve the robustness of AER predictions [3] because: (1) some modalities may be more informative at certain times than others, (2) the complementary information across the modalities can only be captured through multimodal analysis, (3) some information may be corrupted because of noise, occlusion, etc. so the remaining modalities can compensate, and (4) some emotions are far more subtle than others for a given task, and it is only through multimodal analysis that emotion can be accurately predicted. Several models have been proposed in the literature to represent human affective states, including the categorical model in which emotions are categorized as happy, sad, angry, disgusted, fearful, surprised, etc. [8], or the dimensional model, which poses AER as a regression problem. In this categorization, affective dimensions are denoted as continuous values, e.g., valence (positiveness or negativeness), dominance (strong or weak), expectation (anticipation), or arousal (active or passive) [9]. The circumplex model of affect proposed by Russell [26] is the most popular way of expressing continuous emotion representation. This paper focuses on developing cost-effective deep learning (DL) models for video-based AER in the valence-arousal space.

Multimodal AER systems can improve predictive accuracy, although they face challenges like missing modalities and dynamic modality contributions, and require a higher computational complexity. In the context of video-based AER, some modalities (e.g., physiological signals) may not easily be acquired at test time in the operational environment. However, such information can be available during the data-collection phase and may allow for improving the generalization capacity of AER models. Figure 1 shows different sources of information (pervasive, privileged, and contextual), and their availability for video-based AER applications at training and test times.

In this paper, we seek to enhance the performance of unimodal AER models by leveraging privileged information. In machine learning (ML), privileged information refers to additional information that is only available to the model for training. A multimodal audio-visual teacher is trained with privileged information, and then this knowledge is distilled to a unimodal visual-only student model. Without loss of generality, we consider vocal expressions as privileged information (an audio modality) and seek to develop a unimodal AER model that predicts the arousal and valence values of a person appearing in a video, based only on his facial expressions (a visual modality). Furthermore, we propose a joint training mechanism for the student model with adaptive weighting to minimize the negative transfer of samples when the teacher model's predictions are incorrect.

The main contributions of this article are summarized as follows. (i) A multimodal to unimodal privileged knowledge distillation (PKD) mechanism with adaptive weighting is proposed to enhance the performance of unimodal AER systems. The proposed system is a model and modality agnostic approach that enables using diverse sources as privileged information (e.g., physiological) only at training time. (ii) Our empirical experiments on the challenging RECOLA and Affwild2 datasets for video-based AER show the merits of our multimodal PKD method when using audio as privileged information. Results indicate that our approach significantly improves the performance of different unimodal AER video-based systems without the complexity of multimodal audio-visual (A-V) systems at test time.

2. Related Work

Audio-Visual (A-V) Fusion: The seminal work in multimodal DL was proposed by Ngiam et al. [16] in which A and V modalities were first separately encoded and then fused using restricted Boltzmann machines and autoencoders. This work opened new pathways to use DL for multimodal analysis. Tzirakis et al. [28] proposed a DL model for A-V fusion, where A features were obtained using a 1D convolutional neural network (CNN), and visual features were processed by a ResNet-50 visual backbone. The two feature vectors were fused and fed into a long short-term memory network (LSTM) for temporal modeling. Rajasekhar et al. [22] proposed a joint cross-attention mechanism for audio-visual fusion, where separated backbones were trained for A and V modalities. The A backbone was a ResNet-18 trained on spectrograms, and the visual features were extracted from a pre-trained inflated 3D CNN (I3D). The extracted features were then fed into the joint cross-attention fusion block, where correlation-based attention weights were calculated to weigh the A and V modalities dynamically.

Multimodal fusion methods are often more effective and robust than unimodal AER systems but are more complex. Usually, the inference time of multimodal AER systems is larger than unimodal AER systems. In the case of feature-level fusion systems, they require pre-processing of several modalities, separated feature extraction backbones, as well as a fusion mechanism. Furthermore, they also require all modalities at test time to maintain inference accuracy. In the case of decision-level fusion, it is easier to incorporate multiple modalities, but such systems fail to capture complementary information among modalities. Techniques such as attention bottleneck in transformers [15] have been proposed to reduce the computational complexity of multimodal systems by minimizing the pairwise attention cost. However, such methods also require all the modalities to be present at test time.

Unlike these methods, we propose a knowledge distillation (KD) mechanism in which the teacher model is computationally expensive and requires the modalities to be present at training time. This model leverages this additional supervision to train a cost-effective student model requiring only one modality at the inference time.

Learning Using Privileged Information (LUPI): Vapnik and Vashist [30] introduced the concept of privileged information, where using additional information only present at training time can outperform the traditional ML paradigm. Since then, this concept has been explored in many applications, like action recognition and person re-identification [27, 31]. In multimodal systems, there is often a concern about the availability of certain modalities. For example, in the case of RGB-D data, it is easier to obtain depth sensing for training data, but in real-life environments, we mostly deal with RGB data. Similarly, in the AER domain, certain modalities (e.g., physiological modalities, like ECG and EDA) may be completely absent at test time or partially missing. A-V modalities can be partially missing due to occlusion, user-initiated muting, transmission/recording errors, etc. Zhao et al. [35] proposed a privileged KD mechanism for online action detection. Their study proposed the idea of using the future frames as privileged information only at training time. However, only partial hidden features of the student model were updated through KD loss because of the teacher-student gap. Such approaches are not well explored in the AER domain. We present a cost-effective multimodal to unimodal privileged KD (PKD) approach and empirically show that PI can improve the performance of unimodal AER systems.

Knowledge Distillation (KD): KD was first proposed by Hinton et al. [11] as a model compression technique based on logit information. Most KD methods in AER focus on cross-modal KD, where the information learned from one modality is distilled into another. Albanie et al. [1] introduced AER from speech, using a teacher-student method that learns embeddings for speech classification in a completely unsupervised manner. The student model dealing with single input modality (A) is trained to reproduce the features of the "teacher" model, which has been trained in a supervised manner on a second modality (V). Deng et al. [5] proposed a multi-task teacher network for three AER tasks (action unit detection, categorical emotions, and valence-arousal prediction), and the output of this network was considered as soft labels for three student models. The student models were trained using ground truth and soft labels. Zhang et al. [33] proposed a visual-to-EEG cross-modal KD mechanism where the stronger visual modality enhances the weaker EEG modality. Knowledge was distilled from the visual teacher to the EEG student using a weighted sum of L1 and concordance correlation coefficient (CCC) loss. All these methods focus on transferring knowledge learned from one modality to another. In contrast, the approach proposed in this paper relies on KD to train accurate unimodal AER systems using a multimodal teacher with additional supervision.

3. Proposed Method

To improve the accuracy of unimodal AER systems, we propose leveraging the LUPI paradigm with information from multiple sources, and a multimodal privileged KD (PKD) mechanism. As shown in Figure 2, a joint audio-visual feature representation is learned such that semantically similar samples from multiple modalities are closer to each other than dissimilar ones. This is made possible by the additional A modality that is available during model training. For a given video sequence, the PKD mechanism seeks to learn student model embeddings that approach the joint feature embeddings of the teacher model. The student network is jointly optimized using the ground truth and the teacher's multimodal embeddings.

3.1. Teacher-Student Models

A-V information is input to the multimodal teacher model, and separated backbones are used to extract modality-specific features. These A and V features are then input to a fusion module that learns a joint multimodal representation, and this knowledge is distilled into the unimodal student model that does not have access to the privileged modality.

Since knowledge is distilled from the multimodal joint representation space to the unimodal model, the method to obtain joint representations holds crucial importance. To understand the effect of the fusion model on the proposed system, we validate (in Section 4) using: (1) feature concatenation with a fully-connected layer, and (2) a state-of-the-art joint-cross attention fusion technique for AER [22]. The fully-connected layer for concatenation is selected to ensure that the performance gained in the subsequent steps results from the proposed PKD technique. Feature vectors extracted from the A and V backbones are concatenated and then processed through multiple non-linear transformations. The effectiveness of KD methods is mainly affected by the gap in capacity between the student and teacher models. Our method distills knowledge from the A-V teacher model to the V student model. We use the same V networks as the teacher model to minimize this gap in capacity.

3.2. Multimodal Privileged Knowledge Distillation

Knowledge is distilled from the multimodal A-V teacher model to the unimodal visual student network. In the multimodal teacher network, after the feature vectors are fused, and a non-linear transformation is applied, both the A and V features are projected in a subspace where the semantically similar samples are projected closer to each other than the dissimilar ones. This knowledge is then distilled to the unimodal student network by minimizing the distance between the feature embeddings of the teacher model after fusion, and the student model. This distance is minimized by calculating the cosine similarity between the two feature vectors. The cosine loss is jointly minimized along with the CCC loss based on the ground truth.

Let $\mathbf{X}_i^V$ be the modality-specific representation of the V

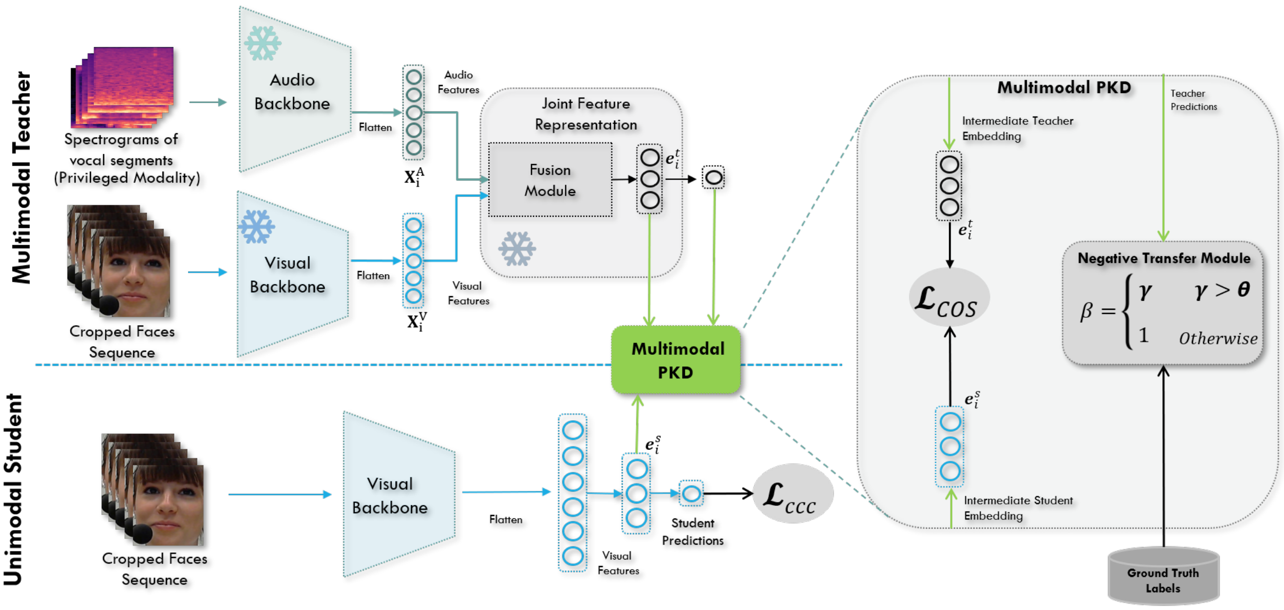


Figure 2. Illustration of the proposed multimodal PKD mechanism with the A and V backbones and fusion (see Sec. 4.2.1). The multimodal teacher model (top) is trained on both A and V modalities. The student model (bottom) only inputs the V modality and learns jointly using the ground truth and multimodal teacher model. The PKD module (right) minimizes the distance between intermediate representations of the teacher and student.

modality, and $\mathbf{X}_i^A$ be the modality-specific representation of the A backbone, where i indicated the input frames. Both $\mathbf{X}_i^V$ and $\mathbf{X}_i^A$ are combined in the fusion module to obtain the joint representation $\mathbf{X}_i^C$. Non-linear transformations are applied to $\mathbf{X}_i^C$ to project it into the joint A-V representation space. Let $\mathbf{e}_i^t$ be the teacher embedding for the i_{th} frame. Similarly, the same frame sequence is fed to the visual student network, and the n-dimensional embedding from the V student network is also obtained during student training. Let $\mathbf{e}_i^s$ be the intermediate embedding of the student network. The cosine loss is calculated between the intermediate multimodal teacher embedding $\mathbf{e}_i^t$, and intermediate student embedding $\mathbf{e}_i^s$ using:

$$\mathcal{L}_{\text{cos}} = 1 - \sum_{i=1}^{N} \frac{\mathbf{e}_i^t \cdot \mathbf{e}_i^s}{\|\mathbf{e}_i^t\| \cdot \|\mathbf{e}_i^s\|} \tag{1}$$

where $\mathbf{e}_i^t \cdot \mathbf{e}_i^s$ represents the dot product of the intermediate teacher and student embedding vector for the i_{th} sample, obtained by multiplying the corresponding components of vectors, and adding up the products, and $\|\mathbf{e}_i^t\| \cdot \|\mathbf{e}_i^s\|$ represents the product of the magnitude of two embeddings.

The CCC loss is calculated from the predictions of the unimodal student network and the corresponding ground truth as:

$$\mathcal{L}_{\text{ccc}} = 1 - \rho_c \tag{2}$$

where ρ_c (CCC measure), is given by:

$$\rho_c = \frac{2.\rho \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3}$$

where μ_x and μ_y are the means of the two variables — predictions and ground-truth, respectively, and σ_x^2 and σ_y^2 are the corresponding variances. ρ is the correlation coefficient between the two variables — predictions and ground truth.

The student network is jointly trained using the loss from the distillation ($\mathcal{L}_{\text{cos}}$) as well as the loss from the ground truth ($\mathcal{L}_{\text{ccc}}$) as:

$$\mathcal{L}_{\text{joint}} = \alpha \cdot \mathcal{L}_{\text{cos}} + (1 - \alpha) \cdot \mathcal{L}_{\text{ccc}} \tag{4}$$

where α controls the weighting of each loss term, which can be tuned as a hyperparameter or learnable weight.

3.3. Negative Transfer

Given access to the privileged modality, the multimodal teacher may provide accurate predictions. However, the teacher model may output inaccurate predictions for some samples, and it is undesirable to distill this knowledge into the student model. To avoid this negative transfer, we propose an adaptive weighting approach. During the training of the student, a frame sequence is fed to the teacher model to obtain the joint intermediate representations. The CCC is also computed from the predictions of the multimodal teacher model for that frame sequence. Then, this information is used to dynamically weigh the KD loss term ($\mathcal{L}_{\text{cos}}$)

in Eq. (4).

$$\gamma = 1 - \rho_c^{teacher} \tag{5}$$

$$\beta = \begin{cases} \gamma, & \text{if } \gamma \leq \theta \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

where θ is a threshold for negative transfer. If the value of γ is greater than θ, i.e., the teacher prediction is "wrong" beyond the threshold, $\mathcal{L}_{\text{cos}}$ is set to zero for that sequence. After incorporating the negative transfer module, Eq. (4) can be rewritten as:

$$\mathcal{L}_{\text{joint}} = \alpha \cdot \mathcal{L}_{\text{cos}} + \beta \cdot \mathcal{L}_{\text{ccc}} \tag{7}$$

In this way, the student model only learns from the teacher if its predictions are accurate, and KD is avoided if the teacher's predictions are "wrong" beyond the threshold. This method implicitly incorporates the ground truth information in the KD loss.

4. Experimental Methodology

4.1. Datasets and Evaluation Measures

RECOLA: The Remote Collaborative and Affective (RECOLA) [25] dataset comprises 9.5 hours of multimodal recordings of participants performing a collaborative task during a video conference. There are 46 participants in total. Apart from the A-V data, the RECOLA dataset also includes the Electrocardiogram (ECG) and Electrodermal Activity (EDA) signals. The dataset is split into training, development (or validation), and test sets with nine videos in each. Each video is a continuous recording of 5 minutes which is annotated for valence, arousal, and liking. The annotation frequency is 25 Hz. This dataset has been widely used in the affecting computing community because of its uncontrolled setting. It has been used in multiple affective computing challenges like Audio Visual Emotion Challenge (AVEC) 2015 [2, 24], 2016 [29], etc. The training and validation sets are publicly available, but not the test set annotations. Since the discontinuation of the AVEC challenge, many studies have reported their results on the validation set. Therefore, we also report the results of the proposed method on the validation dataset.

Affwild2 Dataset: The Affwild2 dataset is a collection of 564 in-the-wild videos gathered from YouTube [14]. The dataset is annotated for three behavior analysis tasks, including continuous values of valence and arousal ranging from -1 to +1, categorical expressions, and action unit detection. The valence/arousal set is used in this study. This is among the most extensive datasets available for affective behavior analysis in the wild, having huge diversity in capture conditions, ages, genders, ethnicities, etc., making it a challenging dataset in terms of generalization. The dataset is divided into training, validation, and test subsets having 351, 71, and 152 videos, respectively. Since the test set annotations are not publicly available, we use the validation set to report our results.

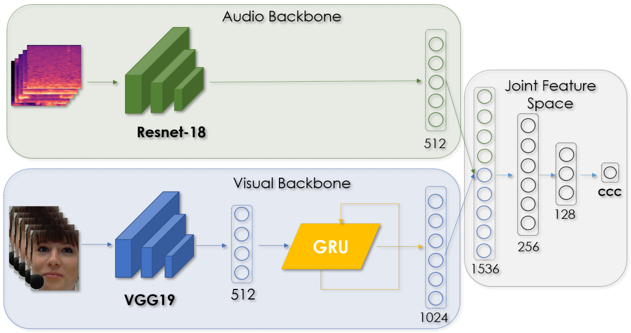


Figure 3. The multimodal teacher model for the RECOLA dataset. The A model is a ResNet-18 (top) trained on spectrograms. The visual model (bottom) is a VGG19 pre-trained on the FER2013 dataset, followed by a GRU.

Evaluation Measure: Since the annotations are continuous values for both RECOLA and Affwild2 datasets, we use the CCC, given in Eq. (3), to measure of agreement between predicted values and the ground truth.

4.2. Implementation Details

4.2.1 Weak Backbones and Fusion on RECOLA

The proposed method works on raw videos rather than features provided with the dataset. In multimodal computing, proper synchronization of modalities is crucial and requires additional pre-processing steps. We have used MTCNN to detect faces from the videos [32] for the V modality. The frames are extracted at 25 fps to match the annotation frequency. The extracted faces are then resized to 112×112 pixels for standardization. To cater to the missing frames, we have used two strategies: i) black frames; ii) frame retention (see Sec. 5.1.1). The V backbone is a VGG19 network, pre-trained in the FER2013 dataset, followed by a one-layer GRU with a 512-dimensional input vector, and a 1024-dimensional output vector. The learning rate for the VGG19 and the ResNet-18 was set to 10^{-6} and 10^{-5}, respectively. In addition, the batch size was set to eight for all experiments. For the A modality, the speech signal is first extracted from the videos, converted to a mono channel, and resampled to 16 kHz. Then, we divided the speech signal into 1-second clips to properly synchronize the two modalities. These clips correspond to 25 frames in the V modality and 25 annotation values. Finally, a short-time Fourier transform is used to obtain spectrograms of resolution 640×480 pixels. The spectrograms are then converted to a log-Mel scale, followed by z-score normalization. The spectrograms are then fed to the ResNet-18. Adam optimizer updates the weights. The batch size is set as 8 for the A backbone as well. The fully connected layers are removed, and a 512-dimensional feature vector is obtained by flattening the last convolutional layer.

For the A-V fusion, the 512-dimensional feature vector is concatenated with the 1024-dimensional output of the GRU from the V backbone, and a 1536-dimensional fused feature vector is obtained, which and processed through multiple non-linear transformations to project both modalities in a common subspace. Figure 3 shows the proposed multimodal teacher model for the weak setting. For the unimodal V student network, the same setting is used for the V backbone in the multimodal teacher network. The proposed approach was implemented using PyTorch [20] and experiments were carried out on servers with Nvidia A100 40GB GPUs.

4.2.2 Strong Backbones and SOTA Fusion on the Affwild2 dataset

For the Affwild2 dataset, we employ the network proposed by Rajasekhar et al. [21]. Cropped-aligned face images included in the dataset were used for the V modality. The V feature extractor is an R3D network with LSTM. The batch size for the V modality is 8, and the network is training using a 10^{-3} learning rate. For the A modality, the extracted audio is segmented into multiple short segments corresponding to 256 frames in the V modality. Discrete Fourier Transform (DFT) is used to obtain spectrograms of resolution 64×107 pixels. The audio feature extractor is a ResNet-18 model which is trained from scratch. The batch size is set to 64, and the network is trained using an initial learning rate of 10^{-3}. The audio and visual feature vectors are fed into the joint cross-attention module for the A-V fusion. The dimension of the A-V feature vector is 1024. The weights in the cross-attention module are updated using the Adam optimizer.

5. Results and Discussion

5.1. Results on RECOLA data

Visual Student Network after PKD: Table 1 shows the average CCC of valence and arousal using the unimodal visual student model obtained with the PKD mechanism on the AER system with weak baselines and fusion and on the RECOLA dataset. Figure 4 shows the arousal and valence values predicted over time by models against the ground truth values. The multimodal teacher model with frame retention (FR) has access to A and V modalities at both training and test time and serves as the upper bound for our system. The unimodal (visual-only) network is a VGG19 followed by a one-layer GRU. Such a model serves as the lower bound for our system, and it is only trained using the V modality, and the same is used at test time as well. The unimodal student model with PKD learns jointly from the ground truth and the superior multimodal teacher model. The model takes only input data from the V modality and predicts the valence and arousal values. This shows that using the proposed PKD mechanism, the performance of the visual-only model is increased from 0.342 to 0.428 for valence and 0.457 to 0.531 for arousal. Thus, PKD is achieved where audio modality is only used at the training time and increases the performance of the visual-only model by 8%.

Method	Valence	Arousal
Multimodal Teacher	0.472	0.586
Unimodal (visual-only)	0.342	0.457
Student w/ PKD (ours)	0.428	0.531

Table 1. Average CCC of valence and arousal evaluated on the RECOLA development set.

Comparison with unimodal visual methods: Table 2 compares the unimodal visual student model (VGG19) with PKD to unimodal visual-only SOTA methods, the AER system with weak baselines and fusion, on the RECOLA dataset. For a fair comparison, we only compare with the visual-only methods that utilize raw videos instead of feature sets (appearance/geometric) provided by the AVEC challenge organizers [24, 29] because those methods do not suffer from the missing (visual) modality problem. It is important for effective AER in a real-world environment to develop methods that work on raw videos that are recorded in-the-wild conditions. The proposed method ranks lower than [23, 28] because computationally expensive visual backbones like ResNet-50 and I3D backbones are used in those methods.

Method	V Network	Valence	Arousal
Ortega et al. [19]	Custom CNN	0.25	0.35
Tzirakis et al. [28]	ResNet-50	0.62	0.43
Praveen et al. [23]	I3D	0.64	0.58
Unimodal (visual-only)	VGG19	0.34	0.46
Student w/ PKD (ours)		0.43	0.53

Table 2. Comparison of the average CCC for valence and arousal of visual-only methods on the RECOLA development set.

Computational Complexity: In this paper, the privileged modality is effectively utilized in order to enhance the accuracy of the unimodal AER model. However, the effectiveness of our proposed unimodal visual student model with

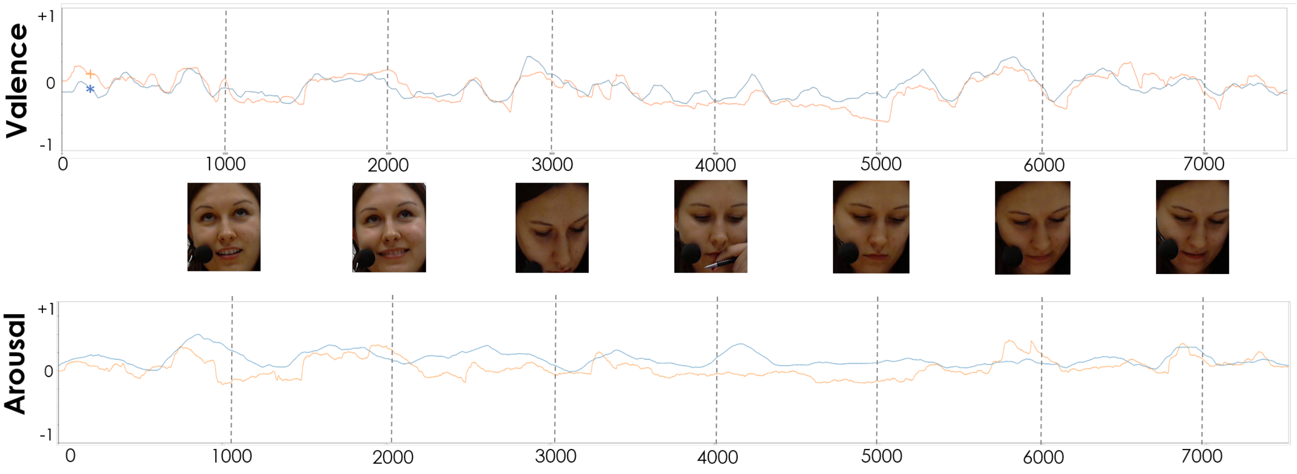


Figure 4. Graphs of predicted CCC* values (blue) and ground truth^{+} (orange) for valence and arousal using the unimodal visual student model with PKD on a RECOLA video. The visualization comprises 7500 frames; the facial frame for every 1000th frame is shown.

Model	Parameters	MACs
Multimodal Teacher	39.7M	10.2G
Student w/ PKD	27.8M	4.8G

Table 3. Comparison of the computational complexity (number of parameters and MAC operations) between the multimodal teacher and unimodal student.

PKD can also be analyzed by comparing the computational complexity of the teacher and student models. Table 3 shows the number of parameters and multiply-accumulate (MAC) operations of the multimodal teacher and unimodal student model. The teacher model requires 39.7M trainable parameters, which amounts to 10.2 GMACs. The student model, on the other hand, has 27.8M parameters amounting to 4.8 GMACs. As shown in Figure 2, the student model does not have the audio backbone or the fusion module. The total number of parameters is reduced by 11.9M, and the GMACs are reduced by 50%.

5.1.1 Ablation Studies

To assess the contribution of the negative transfer module and frame retention mechanism, we perform two ablations: i) where the knowledge is distilled from the multimodal teacher to the unimodal student without any regularizer that prohibits KD when the teacher model's predictions are inaccurate, and ii) where black frames are fed to the model instead for frame retention.

Negative Transfer: The negative transfer module (NTM) prohibits the student model from learning from the teacher in sequences where the teacher's predictions are ”wrong” beyond the threshold. Table 4 shows that the performance of the unimodal student declines if the student is allowed to learn from the teacher without any regularization. The average CCC for valence goes from 0.42 to 0.39, and the drop from 0.53 to 0.50 is observed in the arousal dimension. The unimodal visual student w/o NTM is trained using Eq. (4) where the value for α is empirically set as 0.2.

Method	Valence	Arousal
Student w/ PKD, w/o NTM	0.39	0.50
Student w/ PKD, w/ NTM	0.42	0.53

Table 4. The average CCC for valence and arousal on the RECOLA development set using the unimodal visual student model with PKD with and without negative transfer module.

Frame Retention: Videos recorded in-the-wild setting are usually prone to the missing modality problem. This may be due to a multitude of reasons, either the visual or audio stream is corrupted due to transmission or recording error, user-initiated muting or covering, occlusion, etc. In this context, the MTCNN algorithm cannot detect and crop faces where the participant is not facing the camera directly. Traditionally, a black frame is fed instead of the missing visual frame. This strategy is functional and is necessary to maintain synchronization between the modalities. However, the input is considered noise for the system and adversely affects the performance.

An alternative is to retain the last detected visual frame and repeat it for all the frames where no face is detected. It is observed that in the frames where the participant is not facing the camera, the annotation values do not change drastically. This phenomenon motivated us to use the frame retention strategy instead of black frames because if the

Method	Valence	Arousal
Multimodal teacher w/o FR	0.40	0.53
Multimodal teacher w/ FR	0.47	0.56

Table 5. Average CCC for valence and arousal on the RECOLA development set using the multimodal teacher model with and without frame retention (FR).

gold standard annotation values are not deviating drastically from the missing frames, the retained frames would have a higher correlation to the last detected frame. It can be observed from Table 5 that the valence value goes up from 0.40 to 0.47, whereas the average CCC for arousal is only marginally improved. This result supports the notion that the valence dimension is more affected by the visual modality, and it is easier to gauge arousal from the audio modality [28].

5.2. Results on Affwild2 Data

The proposed PKD method was also validated on the Affwild2 dataset. It is observed that the visual student performs well for the arousal dimension, and the absolute CCC value is improved by 2%. However, the performance goes down for the valence dimension. The visual student is unable to learn generalizable embeddings for the valence dimension. This is perhaps due to the large capacity gap between the teacher and student models. When sophisticated fusion mechanisms like joint cross-attention-based feature fusion are applied, the visual-only student model fails to learn feature embeddings close to the multimodal embeddings that the joint cross-attention fusion learned. Another reason may be the difference between the datasets. The Affwild2 dataset is a collection of videos collected from the web, which differ significantly in terms of quality, capture conditions, ethnicity, age, etc., making it a challenging dataset. In Table 6, we compare our visual-student using PKD with SOTA visual-only methods for valence/arousal values in the Affwild2 development dataset.

6. Conclusion

This study proposes a multimodal to unimodal PKD mechanism with adaptive weighting to enhance the performance of unimodal AER systems. The weights of the student model are jointly optimized using the ground truth and cosine loss between student and teacher intermediate embeddings. Without loss of generality, we use audio modality as privileged information. We validate the proposed approach on RECOLA and Affwild datasets. The proposed PKD method improved the performance of the visual-only student network by 8% on the RECOLA dataset. In the Affwild2 dataset, the overall average CCC was improved by 2% for the arousal dimension.

Method	V Network	Valence	Arousal
Baseline [12]	ResNet-50	0.31	0.17
Zhang et al. [34]	SENet-50	0.28	0.34
He et al. [10]	MobileNet	0.28	0.44
Nguyen et al. [17]	RegNet + GRU	0.43	0.57
Geesung et al. [18]	ResNeXt + SENet	0.51	0.48
MMT (JCA) [22]	I3D	0.67	0.59
Unimodal (visual-only)		0.41	0.51
Student w/ PKD (ours)		0.37	0.53

Table 6. Comparison of the proposed system with visual-only SOTA methods on the Affwild2 validation set.

We conclude that the proposed PKD method works well in the fully-connected layer fusion, where the fully connected layers are stacked together and jointly updated. However, when a more complex fusion mechanism is employed, the capacity gap between the student and teacher model is increased, and the KD efficiency is decreased. More sophisticated KD methods are required to overcome this. One of the limitations of this work is that it requires two-stage training, where, in the first stage, the superior multimodal teacher is trained, and in the second stage, the knowledge is distilled to the unimodal student. Online KD methods should be explored to overcome this limitation. Another limitation is that the student can be only as good as the teacher. Since this method only uses one teacher network, the student may not generalize well. Multiple teacher ensemble or self-distillation methods are possible extensions of this work. In the future, we also intend to use physiological signals as a privileged modality.

Acknowledgements: This work was supported by the Fonds de recherche du Québec – Santé, the Natural Sciences and Engineering Research Council of Canada, and the Canada Foundation for Innovation.

References

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. MM '18, page 292–301, New York, NY, USA, 2018. Association for Computing Machinery. 3

[2] Patrick Cardinal, Najim Dehak, Alessandro Lameiras Koerich, Jahangir Alam, and Patrice Boucher. ETS system for AV+EC 2015 challenge. In *5th International Workshop on Audio/Visual Emotion Challenge*, pages 17–23, 2015. 5

[3] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In Christos Boukis, Aristodemos Pnevmatikakis, and Lazaros Polymenakos, edi-

tors, *Artificial Intelligence and Innovations 2007: from Theory to Applications*, 2007. 1

[4] Wheidima Carneiro de Melo, Eric Granger, and Abdenour Hadid. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE Transactions on Affective Computing*, 13(3):1581–1592, 2022. 1

[5] Didan Deng, Zhaokang Chen, and Bertram E. Shi. Multi-task emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599, 2020. 3

[6] Abhinav Dhall, Roland Goecke, Tom Gedeon, and Nicu Sebe. Emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 10, 03 2016. 1

[7] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Bučinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20:592, 01 2020. 1

[8] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. 2

[9] Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057, 2007. PMID: 18031411. 2

[10] Ruian He, Zhen Xing, Weimin Tan, and Bo Yan. Feature pyramid network for multi-task affective analysis. *ArXiv*, abs/2107.03670, 2021. 8

[11] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. 3

[12] D. Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection and; multi-task learning challenges. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2327–2335, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 8

[13] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, A. Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127:907–929, 2018. 1

[14] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, A. Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127:907–929, 2018. 5

[15] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213, 2021. 2

[16] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress. 2

[17] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial behavior analysis in-the-wild video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2511–2516, 2022. 8

[18] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a past facial image sequence. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3543–3549, 2021. 8

[19] Juan D. S. Ortega, Patrick Cardinal, and Alessandro Lameiras Koerich. Emotion recognition using fusion of audio and video features. *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3847–3852, 2019. 6

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. 6

[21] R Gnana Praveen, Patrick Cardinal, and Eric Granger. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 1–1, 2023. 6

[22] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2485–2494, 2022. 2, 3, 8

[23] R. Gnana Praveen, Eric Granger, and Patrick Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021. 6

[24] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. Av+ ec 2015–the first affect recognition challenge bridging across audio, video, and physiological data. 01 2015. 5, 6

[25] Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013. 5

[26] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 2

[27] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[28] Panagiotis Tzirakis, George Trigeorgis, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, PP, 04 2017. 2, 6, 8

[29] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. AVEC '16, page 3–10, New York, NY, USA, 2016. Association for Computing Machinery. 5, 6

[30] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. Advances in Neural Networks Research: IJCNN2009. 2

[31] Xun Yang, Meng Wang, and Dacheng Tao. Person reidentification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2018. 3

[32] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5

[33] Su Zhang, Chuangao Tang, and Cuntai Guan. Visual-to-eeg cross-modal knowledge distillation for continuous emotion recognition. *Pattern Recognition*, 130:108833, 2022. 3

[34] Zihang Zhang and Jianping Gu. Facial affect recognition in the wild using multi-task learning convolutional network. *ArXiv*, abs/2002.00606, 2020. 8

[35] Peisen Zhao, Lingxi Xie, Jiajie Wang, Ya Zhang, and Qi Tian. Progressive privileged knowledge distillation for online action detection. *Pattern Recognition*, 129:108741, 09 2022. 3