From the Lab to the Wild: Affect Modeling via Privileged Information

Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, *Member, IEEE,* and Georgios N. Yannakakis, *Member, IEEE*

Abstract—How can we reliably transfer affect models trained in controlled laboratory conditions (*in-vitro*) to uncontrolled real-world settings (*in-vivo*)? The information gap between in-vitro and in-vivo applications defines a core challenge of affective computing. This gap is caused by limitations related to affect sensing including intrusiveness, hardware malfunctions and availability of sensors. As a response to these limitations, we introduce the concept of privileged information for operating affect models in real-world scenarios (in the wild). Privileged information enables affect models to be trained across multiple modalities available in a lab, and ignore, without significant performance drops, those modalities that are not available when they operate in the wild. Our approach is tested in two multimodal affect databases one of which is designed for testing models of affect in the wild. By training our affect models using all modalities and then using solely raw footage frames for testing the models, we reach the performance of models that fuse all available modalities for both training and testing. The results are robust across both classification and regression affect modeling tasks which are dominant paradigms in affective computing. Our findings make a decisive step towards realizing affect interaction in the wild.

Index Terms—privileged information, machine learning, affect modelling, valence, arousal, physiology, pixels

1 Introduction

Despite the recent advances in Affective Computing (AC), largely based on today's powerful deep learning algorithms [1], [2], [3], [4], [5], [6], affect modeling approaches are still struggling to reliably transfer models trained on data collected in the laboratory (*in-vitro*) to real-world settings (*in-vivo*), that is, in the wild. Sensing affect in controlled laboratory conditions results in high-quality data characterized by precise multimodal measurements. On the contrary, the quality of affect measurements in uncontrolled, real-world settings is limited by a number of environmental and experimental conditions. Consequently, the information gap between in-vitro and in-vivo affect modeling limits the transferability of affect models to real-world applications.

Compared to in-vitro, the quality of in-vivo affect measurements is limited by several factors. First, the environment affects the sensing equipment (e.g. cameras, microphones, physiology sensors). Therefore, the acquired measurements are likely to be either corrupted by experimental noise due to sensors' failures or biased due to environmental conditions under uncontrolled settings. As a result the presence of noise and bias deteriorates the performance of affect models. Second, exploiting multimodal information to model affect is the norm in current AC approaches. However, as AC occurs in-vivo, sensors for capturing multimodal information are unavailable (e.g. electroencephalogram measurements). We cannot assume—and arguably we should not assume—that users engaged in affect interaction will have access to a plethora of specialized sensors for measuring affect in their houses, their cars [7], [8], or public spaces including museums [9], hospitals and rehabilitation centers [10]. Interestingly, such real-world settings define some of the most popular application domains of AC. Finally, capturing information about users in real-world settings comes with a cost in terms of intrusiveness (e.g. requiring users to install and use sensors properly) and privacy (e.g., access to sensitive information via smartphone's webcam and microphone).

This study aims to overcome the current limitations of affect sensing in the wild by introducing the notion of privileged information to affect modeling. In particular, the Learning Using Privileged Information (LUPI) paradigm [11], [12] is best suited for tasks that present different amounts of information available during the models' training and testing phases. In the context of AC privileged information is of utmost importance when there is an information gap between the development and the deployment of an affect model. Our hypothesis is that the LUPI paradigm can be beneficial for mitigating the limitations of affect sensing in the wild, leading to affect models that achieve similar performance in-vitro and in-vivo conditions. We test our hypothesis using the popular RECOLA and SEWA affect databases [13], [14]. These databases consist of multimodal user information annotated across two affect dimensions; arousal and valence. User measurements include facial images, visual and audio features, and electrodermal activity (EDA) and electrocardiogram (ECG) biosignals. They also include continuous arousal and valence traces performed by six annotators. We consider the information from facial images available both in-vitro and in-vivo as capturing facial images requires no special equipment, just conventional cameras. On the contrary, we consider all other modalities to be available only in-vitro since capturing this

- *Konstantinos Makantasis is with the Department of Artificial Intelligence, Univerisity of Malta, Msida, Malta, MSD2080.*
- *Kosmas Pinitas, Antonios Liapis and Georgios N. Yannakakis are with the Institute of Digital Games, Univerisity of Malta, Msida, Malta, MSD2080.*

Manuscript received Month Day, Year; revised Month Day, Year.

information requires specialized laboratory equipment or video and audio processing algorithms.

Similarly to [6], [15], our models of affect rely on raw images both for training and testing phases. Additional knowledge from privileged information is injected into the models during training. During testing, however, the developed models make predictions using only raw images without dependence on privileged information, as would be the case of an in-vivo setting. Our experimental outcomes suggest that exploiting privileged information yields models that perform consistently better than models trained only with images. More importantly, in many cases, models trained under the LUPI paradigm perform equally well as the models that base their predictions on fusing all available modalities (images and privileged information). The results obtained are consistent across the two affect dimensions of arousal and valence and robust across the two most dominant learning paradigms used in AC: classification and regression. This suggests that our approach is robust for training models that can perform in the wild without extensive hyperparameter tuning. Our findings indicate that privileged information is a critical milestone for bridging the gap between in-vitro and in-vivo affect modeling and realising AC in the wild. Models trained on information only available in a lab setting can perform equally well when that information is unavailable or distorted in the wild. Following LUPI, models of affect gain on robustness, unobtrusiveness, accessibility and practicality.

This paper builds upon and extends significantly our earlier work [16] in several ways. First, our study in [16] applies the LUPI paradigm to affect classification. The present study goes one step further and applies the LUPI paradigm to affect modelling problems treated as both classification and regression, covering the vast majority of affect modeling tasks. Second, in [16] privileged information is injected into affect models at the output layer. This study proposes a general methodology for injecting privileged information at any hidden layer of deep learning-based affect models. Therefore, the proposed methodology can derive affect models that exploit privileged information irrespective of the models' architecture choices. Third, in this study, we go beyond arousal prediction within the domain of digital games [6], [15], [16], and evaluate the proposed scheme in two popular publicly available affect databases across two affect dimensions: valence and arousal. We should also mention that one of these databases has been developed for testing the performance of affect models in the wild. The evaluation results suggest that exploiting privileged information can boost the performance of affect models that operate in the wild.

2 Related Work

This section covers the related areas of pixel-based affect modeling, multimodal affect modeling based on audiovisual information and physiology measurements, affect modeling in the wild, and affect modelling under missing data or modalities.

2.1 Pixel-based and Multimodal Affect Modeling

Due to the richness of information encoded in videos and images, eliciting and modeling emotion via visual cues has been a core interest in affective computing [17]. Before the deep learning era, the dominant approach for representing visual content was based on the construction of ad-hoc handcrafted features. Along this line, sophisticated visual descriptors, such as scale-invariant feature transform [18], histogram of oriented gradients [19] and linear binary patterns [20], have been widely used to produce high-level representations of facial patterns used for recognizing emotions [21], [22], [23], [24]. Emotion recognition approaches based on handcrafted visual features are characterized by low computational and memory requirements, and thus, are still being studied for use in real-time embedded systems [25]. In the last decade, convolutional neural networks (CNN) led to a breakthrough in computer vision and visual information processing. During their training, CNNs automatically produce high-level representation of visual information, eliminating the need for handcrafted features construction. Baveye *et al.* [26] proposed one of the first approaches using CNNs for predicting dimensional affective scores from videos. However, the limited number of training data prevented the learning of data-hungry CNN models.

The development of medium- and large-scale affect corpora [27], [28] established the use of deep learning in affect modeling [29]. Breuer and Kimmer [30] demonstrated the capacity of CNNs to jointly learn various facial expression recognition tasks. Jung *et al.* [31] boosted the performance of facial-based affect models by exploiting high-level spatiotemporal representations of facial action points produced by CNNs. Ng *et al.* [32] proposed transfer of learning across CNNs for emotion recognition through visual cues. Based on the hypothesis that gameplay footage embeds players' affect, studies in [6], [15] used CNNs to map raw gameplay footage to players' arousal. Finally, Martinez *et al.* [29] presented the first application of CNN models for detecting affect via physiological signals such as skin conductance.

Along with visual information, additional modalities such as audio and physiology measurements have been used for modeling affect. The hypothesis that additional modalities can reveal different aspects of emotions triggered the collection of multimodal information datasets such as the RECOLA database [13] used in this study, the DEAP [33], AMIGOS [34] and SEMAINE [35] datasets. The dominant approach for processing multimodal information is based on fusing the different modalities into a common representation, which is then used as input to machine learning models. Indicatively, Tzirakis *et al.* [36] propose a CNN and a deep residual network to combine auditory and visual information for emotion recognition. Ranganathan *et al.* [37] used deep belief networks to generate multimodal features from face, body gesture, voice and physiology measurements in an unsupervised manner. Siriwardhana *et al.* [38] used self-supervised learning to fuse text, audio and visual information along with transformers for recognizing affect. Seng and Ang [39] presented emotion modeling techniques based on mulitmodal unstructured big-data, while Abdullah *et al.* [40] presented a survey on the application of

deep learning models to multimodal emotion recognition.

The methods discussed above model affect based on visual only or multimodal information captured in well-defined and controlled laboratory conditions. Collecting, however, data in a laboratory requires specialized hardware and software that might not be available in the wild, and yields noise-free and unbiased datasets. Both of these characteristics limit the application of affect models trained with laboratory-generated data to real-world settings. This paper aims to take affective modeling outside of a laboratory's closed boundaries by building models able to predict affect using information that is available in the wild, and at the same time, exploit knowledge through privileged laboratory-generated information.

2.2 Affect Modeling In The Wild

Affect modeling in the wild focuses on developing models able to analyze the emotional state of humans in real-life scenarios that entail uncontrolled conditions. To mitigate the problem of noisy, distorted and biased data, large databases [14], [28], [41], [42], [43], [44] that simulate human emotions in the wild are necessary [45]. The availability of large affect corpora have enabled powerful deep learning models that achieve state-of-the-art affect modeling results.

AffWildNet proposed by Kollias *et al.* [46], [47] achieved the best performance in the *Aff-Wild* challenge [28] by combining CNNs and recurrent neural networks to accurately capture face dynamics. The EmoFAN deep learning model [1] jointly predicted discrete emotional states and continuous affect dimensions by building upon the face alignment network proposed in [48], thereby achieving the best performance on the AfewVA dataset [49]. Aspandi *et al.* [50] estimated affect in the wild by exploiting adversarial neural networks that build high-level representations of audiovisual information. Parthasarathy and Sundaram [51] demonstrated that multimodal deep learning affect models can significantly improve affect detection in the wild. They use multimodal transformers to capture and exploit temporal dynamics of audiovisual information towards detecting affect states. Finally, Kollias and Zafeiriou [52] proposed a unified framework for affect modeling in the wild that considers facial expressions and categorical affect, facial action units, and dimensional affect representations.

Although the studies listed above model affect in the wild, they all require large affect corpora to reduce the impact of noise and bias on the performance of affect models. Instead, this paper relies on the use of privileged information to effectively train and reliably transfer models of affect from controlled laboratory conditions to real-world settings. Rather than use training data captured in the wild, we use high quality laboratory measurements as privileged information for assisting the training of the models, which can then be applied and operate in the wild. Although exploiting privileged information has been proposed in [16] for modeling players' arousal within the domain of digital games, this study extends the aforementioned work by applying the LUPI paradigm to two popular affect datasets beyond games for modeling both arousal and valence.

2.3 Affect Modelling under Missing Data/Modalities

Affect modelling in the wild may suffer from missing or corrupted data due to unforeseen sensor malfunctions. The study in [53] proposes a semi-supervised multi-view model to address the problem of missing modalities. Their approach treats a missing modality as a latent variable which is integrated out during inference. The works in [54], [55] formulate emotion recognition as a multitask learning problem and leverage several classifiers under all combinations of different modalities to avoid the missing data/modality problem. The authors in [56] propose the learning of joint multimodal representations able to predict the representations of any missing modality. They apply their methodology for trimodal emotion recognition using visual, acoustic and textual information. The study in [57] investigates the performance of state-of-the-art transformers for bimodal emotion recognition problems where one of two modalities is missing. The authors of [58] propose a framework based on iterative data augmentations to address the problem of multimodal emotion recognition in conversation tasks with missing modalities. The aforementioned studies attempt to recover information about the missing data or modalities and exploit that information during a model's training or inference. Our approach, instead, does not target the problem of missing data or modalities. It is based on privileged information that is able to transfer knowledge encoded in models trained using a large set of information modalities to models trained on a smaller set of modalities.

Another approach to deal with missing modalities is based on the idea of dynamic fusion. The works in [59], [60] automatically estimate the importance of each modality during training and weigh it accordingly. During testing they exploit the learnt weighting scheme to fuse different modalities based on given inputs. These approaches are different than ours in the sense that our model is trained on several modalities and during testing it operates on a subset of them. Our aim is to introduce to a model knowledge from modalities that are not available during testing and not to fuse several modalities given the current inputs.

Finally, the study in [61] proposes to use separate networks per available modality and then to enforce them to collaborate and learn common semantics across modalities. During testing only the networks that correspond to the available modalities are used. This approach is also different than ours. First, it focuses on correlation matrices of the latent representations to learn common semantics. Second, it employs several unimodal classifiers and a knowledge transfer scheme from all classifiers to the one that will operate with missing modalities. Our approach uses just two classifers, one teacher and one student, and it is not based on correlation-based knowledge transfer. Instead, it uses the knowledge from the teacher to guide, via modifying the loss function, the learning of the student.

3 Use Cases and Data Preprocessing

This section presents the datasets used for experimentally validating our proposed methodology and the data preprocessing steps.

3.1 Datasets

To test the impact of privileged information on affect modeling and investigate the degree to which it can transfer knowledge from in-vitro experiments to in-vivo AC applications we use two multimodal databases: RECOLA [13] and SEWA [14].

RECOLA consists of audio, visual and physiological (EDA and ECG) recordings of online dyadic interactions between 34 participants. Since RECOLA has been used for audiovisual emotion recognition challenges, the creators split the database into two parts; data form 23 participants used for training and developing models of affect are publicly available, while the rest serve for evaluating the performance of the developed models and are not (and will not be) made publicly available. Six assistants (three males and three females) annotated the collected data in terms of arousal and valence. The annotations are continuous, bounded in the range of $[\ 1; 1]$ and provided at 25Hz.

SEWA contains recordings of 398 volunteers watching various advertisements and discussing them via a video-chat software. Volunteers' behaviour captured in completely unconstrained, real-world environments using webcams and microphones. Five raters provided continuous valence and arousal labels for the audio-visual recordings at 66Hz. These annotations were combined to a single ground truth that maximally correlates annotations from all raters. Finally, the ground-truth was normalised in $[0; 1]$. In this study, we use the SEWA basic dataset, which consists of 538 short (10-30 second long) segments cut from the full video-chat recordings. For the rest of the paper, we refer to the SEWA basic dataset as SEWA.

Along with the raw recordings, the RECOLA database creators provide features that describe each information modality. For audio information in RECOLA, besides raw audio files, probability of voice activity detection and eGeMAPS acoustic features [62] are also provided. Features describing the statistical properties of EDA and ECG are also given. Finally, visual information is described by raw $1080 \quad 720$ video frames, probability of face detection, optical flow, and detection and movement of 15 emotion-related facial action units. In SEWA, we created 65 OpenSMILE [63] acoustic features from the audio recordings. For RECOLA, we consider the video frames as *pixel information* and the remaining metrics as *visual features* which require complex software to process that may not be available in the wild.

3.2 Data Preprocessing

This study aims to produce models of affect that predict arousal and valence based on different information modalities. We split the interaction session of each participant using overlapping windows. The sliding step and the length of the windows are hyperparameters we consider. In this study, we conduct experiments for a sliding step of 400ms and window lengths of 1, 2 and 3 seconds. Using a fixed sliding step and overlapping time windows, the dataset size (number of time windows) is not affected by the windows' length. By varying the windows' length, the amount of temporal information encoded in each window changes affecting both audiovisual information and physiology features.

TABLE 1
Information modalities in RECOLA and SEWA within each time window along with their dimension. We treat only pixel information as non-privileged.

Modality – RECOLA	**Dimension**
Pixel Information	$320 \times 180 \times 5$ per second
Audio Features (e.g. eGeMAPS, voice activity)	131
Visual Features (e.g. facial action units)	41
Electrocardiogram (ECG) Features	54
Electrodermal activity (EDA) Features	63

Modality – SEWA	**Dimension**
Pixel Information	$320 \times 180 \times 5$ per second
Audio Features (e.g. OpenSMILE features)	65

After splitting the multimodal dataset across time windows, the information associated with each window corresponds to a sequence of raw footage frames concatenated along the channels dimension and the mean values of audiovisual and physiology features. To reduce the computational cost, we use grayscale footage frames resized to $320 \quad 180$ pixels and frame skipping of five frames. Regarding RECOLA, as the data annotation is conducted by six assistants, we use the median annotation values per time instance to mitigate annotators' disagreement [64]. For SEWA, we use the audiovisual annotation values provided with the dataset. Then, the arousal and valence ground truth labels for each window correspond to the mean annotation values within the window's duration [6], [65]. Table 1 summarizes the information modalities that describe each of the time windows.

4 Affect Modeling Using Privileged Information

In this section we detail the *Learning Using Privileged Information* paradigm [11] for building models of affect capable of generalizing in the wild, as well as the architecture of the employed machine learning models.

4.1 Learning Using Privileged Information

Learning Using Privileged Information (LUPI) [11], [12] addresses problems characterized by an asymmetric distribution of information between training and test time; specifically, additional information is given about the training data, which is not available at test time. This setting is prevalent in affective computing. A plethora of different information modalities can be captured in controlled laboratory conditions using specialized hardware and software. In the wild, however, it is impossible to capture the same modalities due to sensors' cost, noisy environments, and invasive capturing procedures. LUPI provides the means to *transfer knowledge* from all the available modalities to a machine learning model that makes predictions using only a subset of these modalities [66], [67]. In other words, LUPI allows an affect model to be trained exploiting knowledge that comes from all the modalities captured in a laboratory setting or via specialized software and hardware. During test time, however, the same model makes predictions using

only those modalities that are available in the wild. The information that is not available during test time is called *privileged information*.

As far as the RECOLA database is concerned, we treat as privileged the information that corresponds to physiology and ausiovisual features provided by the database creators (see Table 1). For SEWA, we consider audio features as privileged. Our choice is justified by the fact that capturing physiology requires specialized sensors, while constructing physiology and audiovisual features implies the employment of specific software algorithms. On the contrary, we consider information that comes from raw footage frames as non-privileged (captured using conventional cameras) being available both at training and test times.

Below, we describe transferring knowledge from privileged information to a machine learning model. At this point, we should clarify that transferring knowledge using LUPI is different from the transfer of learning techniques used in deep learning [32]. Transfer of learning targets small-sample setting problems by finetuning a model trained for a specific task such that it performs well in a similar task. On the contrary, using LUPI focuses on problems with asymmetric distribution of training/testing information and trains the models from scratch.

This study explores the use of privileged information with neural network-based models of affect. Before transferring knowledge that comes from privileged information, we first have to represent it appropriately. Following [67], [68], [69], we represent that knowledge within the latent and the output representations of a neural network that has been trained and makes predictions based on all available modalities or on privileged information only. This model is called *teacher*. Having a teacher model trained, we can transfer knowledge from privileged information to another model called *student*. The transfer of knowledge can be achieved by feeding the model with only those modalities on information that is available in the wild and force it during training to balance between the learning task's loss and learning latent representations that match those of the teacher model. After training, the student model makes predictions based only on the information that is available in the wild, without any dependence on the teacher model or privileged information.

To be more rigorous, let us denote as $\boldsymbol{x}$ and as $\boldsymbol{x}$ the information describing a specific sample fed to a student and a teacher model, respectively, and as y the sample's affect ground truth label. By denoting as $S_l(\boldsymbol{x})$ 2 $\mathbb{R}^d$ and as $T_k(\boldsymbol{x})$ 2 $\mathbb{R}^d$, respectively, the latent representations at the l-th layer and the k-th layer of the student and the teacher models, then the loss that the student is minimizing during training can be defined as:

$$L_{st} = (1 \quad) L(S_o(\boldsymbol{x}); y) + \quad D(S_l(\boldsymbol{x}); T_k(\boldsymbol{x})) \quad (1)$$

where 2 $[0; 1]$, L stands for the learning task's loss (e.g. mean squared error for regression tasks or cross-entropy loss for classification tasks), $S_o(\boldsymbol{x})$ is the output of the student model, and D is a distance metric penalizing deviations between student's and teacher's latent representations.

In the present study, we focus on regression and classification tasks since these are the two most dominant paradigms in affect modeling. In the case of classification, we develop models that predict high vs low arousal/valence while in regression, our models aim to predict the continuous ground truth label of affect.

For classification problems the student loss in Eq. (1) is defined as:

$$L_{st} = (1 \quad) L_{CE}(S_o(\boldsymbol{x}); y) + \quad L_{KL}(S_o(\boldsymbol{x}); T_o(\boldsymbol{x})) \quad (2)$$

where $T_o(\boldsymbol{x})$ are the probabilistic predictions of the teacher model, L_{CE} is the cross-entropy loss, and L_{KL} is the Kullback-Leibler divergence between student's and teacher's probabilistic predictions. Kullback-Leibler divergence is a statistical distance measuring the difference between two distributions—in our case, the probability distributions over the available classes for the student and teacher models—and it has been used within the knowledge distillation paradigm for transferring knowledge between different models [68]. By examining the relation in Eq. (2), it can be made clear that in the case of classification, the student has to minimize the classification loss and at the same time follow the probabilistic predictions of the teacher.

For regression problems, the above procedure can not be followed. The output of regression models is a real value and not a probability distribution, and therefore Kullback-Leibler divergence cannot be used. In addition, under a regression setting, requiring from the student model to follow a teacher's predictions will force the student's output away from the desired ground truth labels. For all those reasons, we inject knowledge about privileged information at the penultimate layer of the student model [69]. To achieve that, we force the output of the penultimate layer of the student model to be close to the corresponding output of the teacher model by defining the student loss in Eq. (1) as:

$$L_{st} = (1 \quad) L_{MSE}(S_o(\boldsymbol{x}); y) + \quad L_{CS}(S_p(\boldsymbol{x}); T_p(\boldsymbol{x})) \quad (3)$$

where L_{MSE} stands for the mean square error, $S_p(\boldsymbol{x})$ and $T_p(\boldsymbol{x})$ the output of the penultimate layer of the student and teacher models, and L_{CS} the cosine similarity. The relation in Eq. (3) indicates that the student minimizes the regression loss and at the same time it tries to match the representation at its penultimate layer to the representation produced by the teacher. By forcing the latent representations of the student to match the latent representations of the teacher, knowledge about privileged information can be injected at *any* layer of a neural network-based affect model. Therefore, the proposed approach can be applied to any deep learning affect model irrespective of the architecture design choices.

Both Eq. (2) and Eq. (3) rely heavily on the parameter, which determines the impact of the privileged information on the training of the student model. By increasing the value of , we force the student model to weigh more the knowledge coming from the teacher models and pay less attention to ground truth labels. When $= 0$ the student considers only ground truth labels without exploiting privileged information. On the contrary, when $= 1$ the student follows the teacher disregarding any information from the ground truth labels.

In the presented case study, we employ two teacher models. The first model is trained using only privileged information, and the second using all available modalities (privileged information plus pixel information). The student model is trained using raw interaction footage frames and

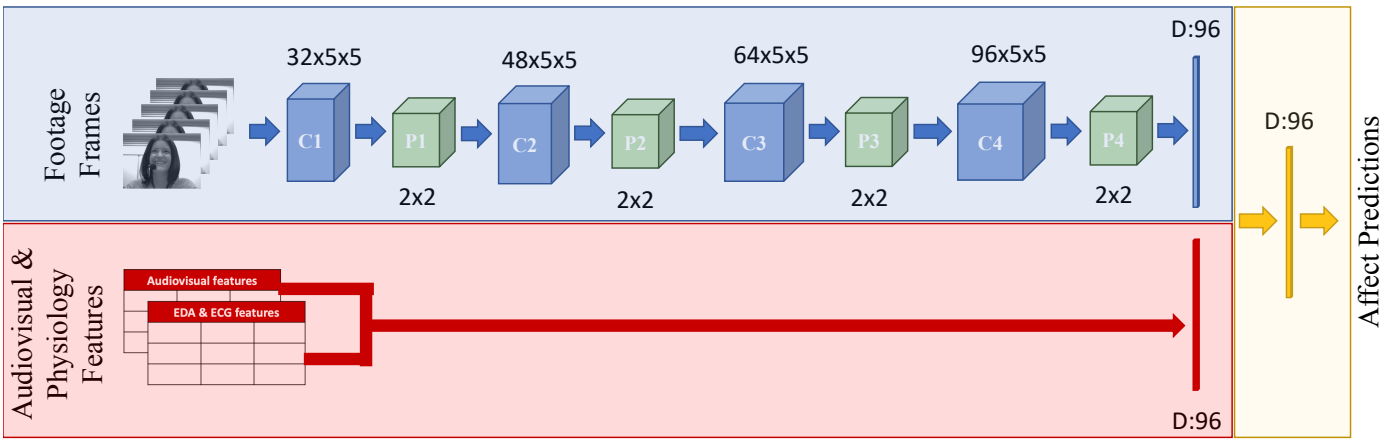

Fig. 1. Architecture of the employed models of affect. The convolutional, max-pooling and dense layers are demoted by "C", "P" and "D" respectively. The blue-shaded stream corresponds to the PixelNet and student models and the red-shaded to the PrivNet. The FusionNet combines the blue and red stream with the yellow-shaded module that fuses the information from the different modalities.

TABLE 2
Information modalities used for training and testing the different models. Details of each modality are found in Table 1.

RECOLA	Pixel		Audio		Visual		ECG		EDA	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
PixelNet	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
FusionNet	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PrivNet	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
StudentNet	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗

SEWA	Pixel		Audio		Visual		ECG		EDA	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
PixelNet	✓	✓	✗	✗	-	-	-	-	-	-
FusionNet	✓	✓	✓	✓	-	-	-	-	-	-
PrivNet	✗	✗	✓	✓	-	-	-	-	-	-
StudentNet	✓	✓	✓	✗	-	-	-	-	-	-

teacher's knowledge. We should emphasize that after training the student makes predictions using information *solely* from raw frames.

4.2 Machine Learning Models of Affect

The student model—we name it *StudentNet*—used in this study is a 2D CNN with four convolutional layers that receives as input a sequence of grayscale footage frames concatenated along the channels dimension. The first two convolutional layers consist of 32 and 48 learnable kernels of dimension 5 x 5 and stride equal to 2. The third and fourth convolutional layers consists of 64 and 96 learnable kernels of dimension 3 x 3 and stride 1. A 2 2 max-pooling layer follows each of the convolutional layers. The last convolutional layer's output is fed to a dense layer with 96 hidden neurons and then is propagated to the output layer. At the penultimate layer of the student model we also use dropout with 10% probability.

As mentioned above, we use two teacher models; *PrivNet* trained with only privileged information, and *FusionNet* trained with all information modalities (see Table 2). PrivNet is a fully connected feed forward neural network with one hidden layer with 96 neurons. FusionNet is a two-stream network: the stream that processes the footage frames has the same architecture as the student model, while the stream that processes the privileged information is similar to PrivNet. The outputs of the two streams are concatenated and pass through a dense layer with 96 neurons before they are fed to the output layer.

To evaluate the impact of privileged information on the affect model's performance, we build a baseline model: a CNN trained and making predictions by exploiting only pixel information of the footage frames. We name this model *PixelNet* following the reported benefits of modeling affect solely from pixels [6], [15]. The architecture of PixelNet is the same as the architecture of the student model.

For all models, we use the Adam optimizer [70] with learning rate 0:001, batches of size 256 to train the models, and ReLU as activation function for all models' layers. Table 2 presents the modalities used for training and testing the performance of the employed models. Figure 1 presents the architecture of all employed models. The blue-shaded stream corresponds to PixelNet and StudentNet and the red-shaded to PrivNet. FusionNet consists of both blue and red streams, and it employs the yellow-shaded module for fusing the information from the different modalities before the output layer.

It should be noted that this study does not aim to produce state-of-the-art results. Therefore, we did not conduct any architecture search, including the investigation of different fusion strategies, for deriving the best possible learners for the problems at hand. This study aims, first, to rigorously formalise a methodology for exploiting privileged information in affect modelling and under different learning paradigms, and, second, to showcase the benefits of infusing privileged information into affect models.

5 Results

This section presents the framework for evaluating the impact of privileged information on affect modeling and the experimental results obtained.

TABLE 3
Data points for the two datasets, per time window and for classification (Class.) or regression tasks.

Datapoints	RECOLA			SEWA		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Arousal Class.	13752	13612	13444	20879	19590	18363
Valence Class.	10827	10750	10630	20891	19625	18512
Regression	16750	16692	16635	25858	22468	21150

5.1 Evaluation Framework

As mentioned earlier, we treat affect modeling both as a regression and as a classification task. In the former our models attempt to predict continuous ground truth affect annotations. In the latter, however, the models learn to map user information to low and high arousal/valence classes. Therefore in classification, we use a split criterion value t that determines the class (low vs high) of each data point. We set the median values as our split criterion ($t = 0.07$ and $t = 0.05$ for arousal and valence, respectively) in an attempt to create balanced datasets for both affect dimensions. Having defined parameter t, we assign the examples whose annotation values is larger than $t +$ to the high arousal/valence class, and the examples with annotation values smaller than t to the low arousal/valence class. The parameter determines a region around the class-splitting value within which annotation values are treated as uncertain and ignored during affect classification to avoid unstable classifiers due to trivial differences in their inputs [6], [15]. Based on the successful findings of [16], we set $= 0.1$. The procedure described above results in datasets of different sizes, listed in Table 3; since regression does not use an uncertainty threshold, the entire dataset is used for both arousal and valence regression.

To evaluate the models' performance we follow a 5-fold cross-validation scheme. When splitting the dataset, we do not include data from the same participant in both training and test sets. We also use 10% of the training data as a validation set to activate early stopping criteria and avoid model overfitting; training stops after 10 epochs without loss improvement on the validation set. All employed models of affect are evaluated using precisely the same data, i.e., the training, the validation and the test sets are the same for all models. Finally, for affect classification we report models' performance in terms of binary classification accuracy. When we treat affect modeling as a regression task, we evaluate models in terms of Pearson's Correlation Coefficient (PCC) [71] and Concordance Correlation Coefficient (CCC) [72] since these metrics are widely used to quantify the performance of affect models (CCC is also used in AVEC [73] challenges to evaluate models performance on the RECOLA dataset). PCC is used to linearly correlate two variables—in our case the predicted and ground truth affect labels—while CCC is used to measure the agreement (reliability) between the predicted and ground truth labels.

5.2 Teacher's Impact on Student's Performance

We start our analysis by investigating the impact of the teacher on the performance of the student model as determined by parameter in Eq. (1). We use PrivNet and FusionNet as our teacher models. Initially we train these models for modeling affect both as a classification and as a regression task. We train the student models using five values for parameter : $2\ \{0, 0.25, 0.5, 0.75, 1\}$. With $= 0$ the student considers only ground truth labels without exploiting privileged information. On the contrary, with $= 1$ the student follows the teacher disregarding any information from the ground truth labels. In this study we aim to introduce the concept of privileged information in AC. Therefore, instead of using cross validation to estimate the optimal value for , we choose to use a set of predefined values for that parameter. We believe that presenting the performance of student models for different preset values of provides better insights to the models' behaviour. Table 4 presents the results of this investigation. Majority class represents a classifier that always predicts the class which is the most frequently encountered in the training set.

TABLE 4
The effect of parameter on students' average binary classification accuracy (%) when the PrivNet (top) and the FusionNet (bottom) models are used as teachers. Bold values indicate the highest classification accuracy achieved across all different values of .

RECOLA

PrivNet Teacher	Arousal			Valence		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Majority Class	50.28	50.39	50.82	60.75	60.46	60.99
Student (= 0)	60.78	57.34	61.23	62.32	62.50	65.48
Student (= 0.25)	**64.04**	60.29	**62.36**	58.08	**64.10**	63.72
Student (= 0.5)	60.12	59.21	59.52	**62.69**	63.04	63.14
Student (= 0.75)	61.87	**64.83**	59.01	62.39	62.09	63.99
Student (= 1)	60.25	60.95	60.84	62.28	63.28	**65.55**

FusionNet Teacher	Arousal			Valence		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Majority Class	50.28	50.39	50.82	60.75	60.46	60.99
Student (= 0)	60.78	57.34	61.23	**62.32**	**62.50**	**65.48**
Student (= 0.25)	60.14	59.90	**61.34**	60.80	61.69	60.84
Student (= 0.5)	**61.06**	**61.78**	61.06	61.91	62.05	63.06
Student (= 0.75)	59.82	59.28	56.27	62.22	60.12	64.04
Student (= 1)	60.18	59.78	58.18	58.56	58.73	61.80

SEWA

PrivNet Teacher	Arousal			Valence		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Majority Class	58.49	57.16	56.36	52.28	50.90	50.90
Student (= 0)	**66.05**	65.93	66.85	**65.86**	**66.17**	66.25
Student (= 0.25)	65.73	**66.83**	**69.10**	65.79	65.46	67.07
Student (= 0.5)	63.64	66.26	68.13	65.41	61.43	**67.57**
Student (= 0.75)	61.63	64.26	63.40	59.10	63.85	63.32
Student (= 1)	60.40	60.72	62.16	55.85	60.44	61.71

FusionNet Teacher	Arousal			Valence		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Majority Class	58.49	57.16	56.36	52.28	50.90	50.90
Student (= 0)	66.05	65.93	66.85	**65.86**	66.17	66.25
Student (= 0.25)	64.21	**68.58**	69.14	65.40	64.50	**66.55**
Student (= 0.5)	**66.62**	66.99	67.97	63.52	65.86	66.54
Student (= 0.75)	66.49	66.83	68.23	63.52	**66.73**	65.26
Student (= 1)	65.40	65.39	**70.77**	64.94	64.49	66.04

For RECOLA, PrivNet appears to be a more powerful teacher than FusionNet. PrivNet exploits handcrafted audiovisual and physiological features, which can better capture arousal and valence [13]. On the contrary, FusionNet, which uses all available modalities, seems to provide less informative predictions to the student models. This

TABLE 5
The effect of parameter on students' average performance in terms of PCC/CCC when the PrivNet (top) and the FusionNet (bottom) models are used as teachers. Bold values indicate the highest PCC and CCC performance achieved across all different values of .

RECOLA

PrivNet Teacher	**Arousal**			**Valence**		
	1 second	2 seconds	3 seconds	1 second	2 seconds	3 seconds
Student (= 0)	0.165 / 0.100	0.184 / 0.095	0.230 / 0.166	**0.286 / 0.166**	**0.214 / 0.133**	0.190 / 0.138
Student (= 0:25)	0.174 / 0.114	0.172 / 0.118	**0.298 / 0.218**	0.209 / 0.120	0.151 / 0.065	**0.249 / 0.159**
Student (= 0:5)	**0.239 / 0.187**	0.251 / 0.159	0.200 / 0.137	0.254 / 0.128	0.048 / 0.032	0.196 / 0.115
Student (= 0:75)	0.225 / 0.118	**0.280 / 0.179**	0.201 / 0.132	0.194 / 0.105	0.053 / 0.047	0.179 / 0.097
Student (= 1)	-0.125 / -0.017	-0.086 / -0.023	0.114 / 0.001	-0.088 / 0.000	-0.168 / -0.150	-0.058 / -0.002

FusionNet Teacher	**Arousal**			**Valence**		
	1 second	2 seconds	3 seconds	1 second	2 seconds	3 seconds
Student (= 0)	0.165 / 0.100	0.184 / 0.095	0.230 / 0.166	0.286 / 0.166	**0.214 / 0.133**	0.190 / 0.138
Student (= 0:25)	0.171 / 0.114	0.215 / 0.160	0.280 / 0.214	0.175 / 0.101	0.161 / 0.111	0.257 / 0.160
Student (= 0:5)	**0.263 / 0.169**	0.196 / 0.111	**0.323 / 0.235**	0.192 / 0.112	0.172 / 0.111	0.279 / 0.162
Student (= 0:75)	0.163 / 0.113	**0.227 / 0.160**	0.193 / 0.148	**0.297 / 0.195**	0.065 / 0.031	**0.281 / 0.175**
Student (= 1)	0.080 / 0.013	-0.147 / -0.009	0.068 / 0.001	-0.044 / -0.001	-0.151 / -0.007	-0.092 / -0.002

SEWA

PrivNet Teacher	**Arousal**			**Valence**		
	1 second	2 seconds	3 seconds	1 second	2 seconds	3 seconds
Student (= 0)	0.301 / 0.249	**0.412 / 0.374**	0.270 / 0.171	0.577 / 0.516	0.595 / 0.564	0.536 / 0.479
Student (= 0:25)	**0.399 / 0.361**	0.393 / 0.365	**0.391 / 0.332**	0.589 / 0.566	0.611 / 0.587	0.499 / 0.439
Student (= 0:5)	0.375 / 0.333	0.391 / 0.368	0.330 / 0.297	**0.601 / 0.580**	0.625 / 0.605	**0.609 / 0.585**
Student (= 0:75)	0.333 / 0.279	0.386 / 0.319	0.361 / 0.343	0.591 / 0.546	**0.640 / 0.623**	0.583 / 0.560
Student (= 1)	-0.077 / -0.022	0.006 / 0.002	-0.132 / -0.029	0.082 / 0.027	0.054 / 0.000	0.115 / 0.023

FusionNet Teacher	**Arousal**			**Valence**		
	1 second	2 seconds	3 seconds	1 second	2 seconds	3 seconds
Student (= 0)	0.301 / 0.249	0.412 / 0.374	0.270 / 0.171	0.577 / 0.516	0.595 / 0.564	0.536 / 0.479
Student (= 0:25)	**0.388 / 0.328**	**0.418 / 0.385**	0.401 / 0.383	0.602 / 0.579	**0.621 / 0.608**	**0.611 / 0.586**
Student (= 0:5)	0.351 / 0.322	0.387 / 0.350	**0.414 / 0.382**	0.600 / 0.580	0.612 / 0.598	0.601 / 0.578
Student (= 0:75)	0.363 / 0.330	0.383 / 0.344	0.294 / 0.224	**0.611 / 0.595**	0.621 / 0.605	0.593 / 0.597
Student (= 1)	0.027 / 0.020	-0.148 / -0.020	0.090 / 0.036	-0.102 / -0.045	-0.035 / -0.026	-0.076 / -0.003

suggests, first, that privileged information can be better correlated to binary classification target variables compared to raw pixels' information and, second, that the joint distribution between raw pixels' information and target variables is not similar to the joint distribution between privileged information and target variables. Fusing modalities with dissimilar joint probability distributions will most likely deteriorate the learning model's performance since it increases the model's learning capacity (number of trainable paramters) without improving the quality of information used to model the data. As far as the affect dimensions are concerned, student models can better capture arousal than valence. For 2 seconds time windows, the best relative performance improvement between a student that exploits and the student that disregards privileged information is 13%, while for valence the corresponding improvement is 2.5%. This agrees with state-of-the-art results on RECOLA, which indicates that affect measurements can better capture arousal than valence [74].

For SEWA, student models seem to benefit more from the FusionNet teacher. This implies that for this dataset, raw pixels' information and audio features complement each other. Similar to RECOLA, we observe a larger relative performance improvement for the arousal dimension (i.e. 5.8% with three-second time windows). Interestingly, when PrivNet is used as a teacher, the student that performs best for half of the cases is the one that disregards privileged information. By combining the above two observations, we can conclude that for SEWA, the most informative modality for capturing arousal and valence is that of raw pixels.

Table 5 presents the results of this investigation when we treat affect modelling as a regression task. Regarding the arousal dimension, we observe that student models' performance improves when they exploit teachers' information for both datasets. In this case, however, we see that the time window duration is vital for selecting the most informative teacher. For the RECOLA dataset, FusionNet appears to be a more informative teacher for 1 and 3 seconds time windows, while for SEWA, FusionNet teacher works better than PrivNet for 2 and 3 seconds time windows. Therefore, the temporal dimension highly affects the quality of information carried by each modality. Privileged information does not seem to improve valence modelling using 1 and 2 seconds time windows for RECOLA. In most cases, the student that disregards privileged information, irrespectively from the teacher used, achieves the best performance. For 3 seconds time windows, however, the student that exploits information from the FusionNet teacher achieves 48% (27%) relative performance improvement in terms of PCC (CCC) compared to the student that disregards teachers' information. For SEWA, however, information from a teacher is beneficial for the student models, irrespective of the windows' duration, resulting in a relative performance improvement of 13% for three-second time windows with the FusionNet teacher. Similar to the classification results, FusionNet appears to be a more informative teacher.

Based on the results presented above, we can conclude that transferring knowledge from privileged information to student models can improve their performance for both regression and classification affect modeling tasks. Parameter α significantly affects the student models' performance and it should be appropriately set according to the problem at hand. However, forcing the student to disregard ground truth labels and follow exclusively the teacher seems to negatively affect its performance, especially for regression tasks, where we observe no or negative correlation between affect measurements and target variables.

5.3 The Importance of Privileged Information

In this section we investigate the impact of privileged information on building models of affect that can operate in the wild. As mentioned above, the student models make predictions using solely information that is available in the wild; in our case the raw frames of the interaction footage. We compare the student models' performance against the performance achieved by the FusionNet model that uses all information modalities captured in laboratory environments for training and testing, and PixelNet that makes predictions using the same information modalities as the student models. For the following investigation, we use non-zero α parameter values that yield the most accurate student models based on the sensitivity analysis covered in Section 5.2. Moreover, we repeat the 5-fold cross-validation scheme (see Section 5.1) five times after reshuffling the participants, to increase the robustness of statistical tests for evaluating the significance of our results. For the sake of completeness, we also present the performance of PrivNet that makes predictions using only privileged information.

Figure 2 presents the results of this comparison when affect modeling is treated as a classification task. For all but one scenario (arousal classification on SEWA with 1 second time windows) and for both affect dimensions considered the student model trained using information from teachers performs on par or better than PixelNet. The relative improvement of the best student model over PixelNet is more than 4% in 6 of 12 instances and more than 1% in 8 of 12 instances. Out of these, one instance has over 8% relative improvement in arousal classification using RECOLA data with 2 seconds time windows. Surprisingly, for half of the scenarios, student models perform on par with FusionNet despite the fact that the latter uses more modalities. These results suggest that student models, when appropriately parameterized, can be efficiently applied in the wild for two reasons: first, they perform on par with or better than PixelNet and second, they closely approximate the performance of FusionNet, although they use only the information that is available in the wild. The PrivNet model achieves the highest accuracy for arousal classification using the RECOLA data; this model, however, uses solely privileged information, and, thus, it cannot operate in the wild.

In Figure 3 we compare the employed models when affect modeling is treated as a regression task. The best student model has a relative improvement over PixelNet by more than 4% in 7 out of 12 instances for PCC and in 9 out of 12 instances for CCC. Out of these, one instance has over 45% relative improvement in arousal regression

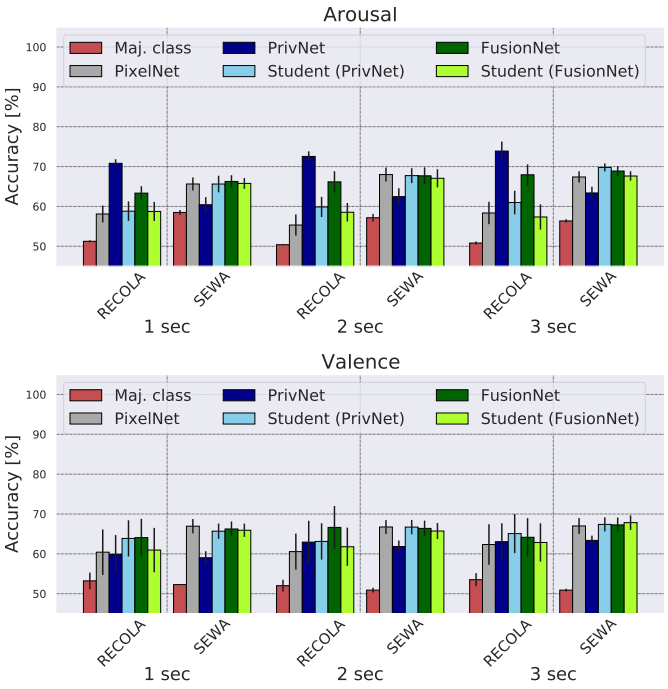

Fig. 2. Average affect classification accuracy of different models, along with the 95% confidence intervals across the five 5-fold cross-validation runs.

using SEWA data with 1 second time windows. Similarly to the classification case, the performance of the student models, when appropriately parameterized, approximates the performance of FusionNet and for 9 out of 24 instances it achieves higher performance.

In both Fig. 2 and Fig. 3 we can also observe a large performance gap from PrivNet to the Student model that uses PrivNet as teacher. This suggests that the correlation patterns between privileged information and target variables are different from the correlation patterns between raw pixels' information and target variables. The fact that the student models try to balance between two contradicting objectives—see Eq. (2)—during training likely results in the observed performance gap.

To evaluate the significance of our results, we performed the D' Agostino-Pearson normality test to check whether or not the paired differences of the tested models' performance come from a normal distribution. In the cases where the hypothesis that our data follow a normal distribution can not be rejected, we performed a one-tailed paired t-test to measure the significance of our results. However, when the normality hypothesis didn't hold, we performed the Wilcoxon signed rank test, a non-parametric version of the paired t-test. All tests were performed at a significance level of 0.05. When the RECOLA dataset is used, the student models perform significantly better than the PixelNet for 2 and 3 seconds time windows for arousal classification and for 1-second windows for valence classification. Under the regression paradigm, student models perform significantly better than PixelNet for 2 seconds time windows and 1-second time windows for arousal and valence, respectively, and for both evaluation metrics. For the SEWA dataset, the student model using PrivNet as teacher performs significantly better for 3 seconds time windows for arousal

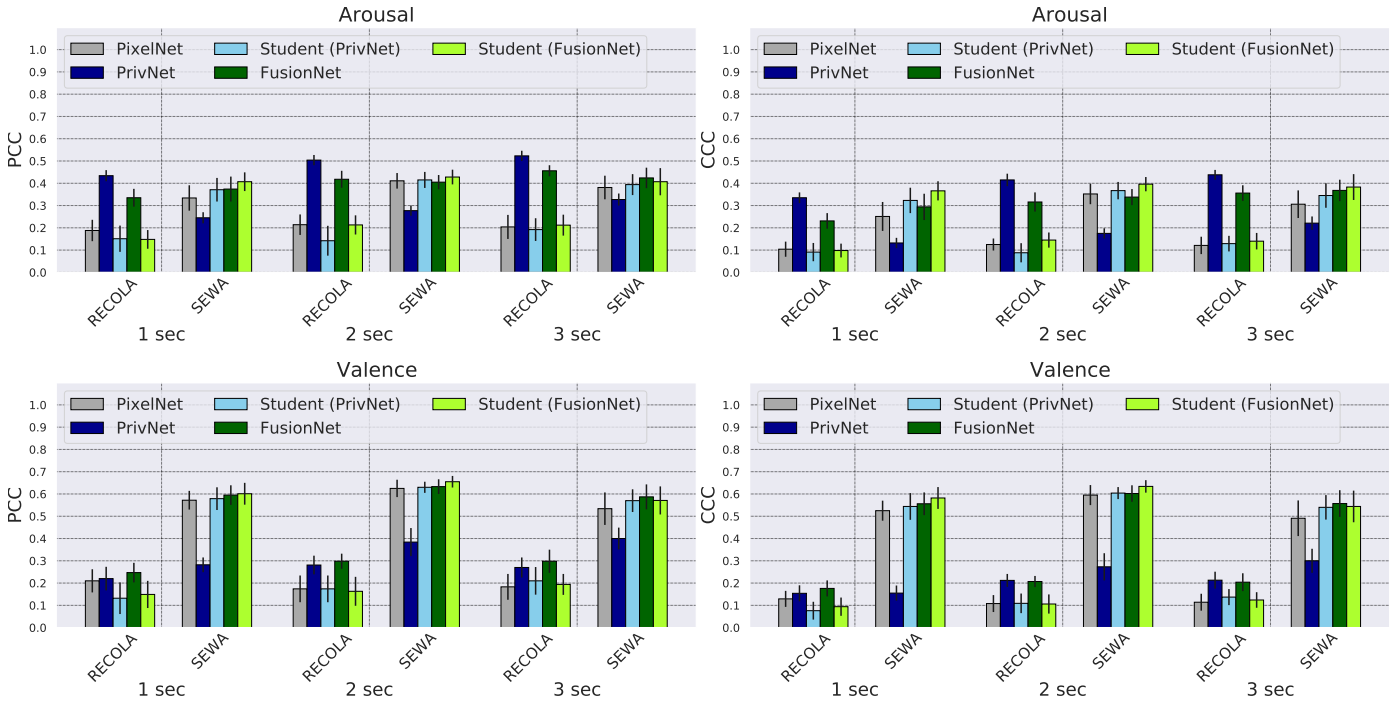

Fig. 3. Comparison of the employed models in terms of average Pearson's Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC) when affect modeling is treated as a regression task along with the 95% confidence intervals across the five 5-fold cross-validation runs.

classification, and the same holds for the student model using FusionNet as teacher for 1 second time windows and valence classification. For arousal regression, student models perform significantly better than PixelNet for 1 second time windows, both for CCC and PCC. For valence regression, however, the student model that uses FusionNet as teacher performs significantly better than PixelNet, both for CCC and PCC, irrespective of the windows' length. In summary, the student model with FusionNet as teacher has significantly higher performance than PixelNet in 6 of 24 instances across datasets, while with PrivNet as teacher the student model has significantly higher performance than PixelNet in 5 of 24 instances. Overall, one or both students outperform significantly PixelNet in 8 of 24 instances, while PixelNet outperforms at least one student in 2 of 24 instances.

Based on the above experiments with two learning paradigms (classification and regression), two affect dimensions and two datasets, we conclude that student models, under the right parameters, achieve average performance values close to, or even higher, than those of FusionNet. We observe a similar behaviour across all scenarios tested, indicating that the LUPI paradigm can provide the means for building accurate models of affect that operate in the wild. Moreover, the student and the PixelNet models use the same kind of information for making predictions. The student, however, appears more robust across different scenarios and achieves on par or higher average performance for all settings.

We should note that this study does not focus on building models that achieve state-of-the-art performance on the employed datasets. For this reason we use simple neural network architectures for our models without any parameter tuning targeting these datasets. Therefore, our results are not directly comparable to the state-of-the-art results reported for RECOLA via the AVEC challenges or for SEWA.

6 Discussion

Testing affect models in the wild comes with costs associated primarily to affect sensing. One would assume that if an affect model has access to fewer modalities during testing in the wild (e.g. due to hardware/software failures or even due to the unavailability of sensors) the result will be detrimental for its accuracy. The results, however, obtained in this and our previous study [16] suggest otherwise. The LUPI paradigm [11], [12] provides the means to mitigate the above limitations of affect modeling in the wild. LUPI produces models that operate in the wild—having access only to a limited set of modalities (in this study raw interaction footage frames)—with small or no actual cost in performance. Our findings suggest that LUPI models can approximate or even surpass the performance of the fusion models that consider all modalities of information during both training and testing. Most importantly via LUPI affect models we gain on accessibility, cost and intrusiveness, bringing affective computing a decisive step closer to real-world applications.

In the presented case study, by examining the performance of PrivNet and PixelNet models for arousal modelling on the RECOLA dataset we see that audiovisual and physiology handcrafted features are more powerful predictors than raw footage frames. This scenario is very common in affective computing, and in general in machine learning, where task-specific handcrafted features are used to improve the performance of learning models [75]. This fact emphasizes the importance of LUPI since student models exploit powerful handcrafted features during training to learn to make accurate predictions with low-level easy

to capture information (without any dependence on the features mentioned above).

While the LUPI paradigm seems to be robust across the modalities and affect dimensions examined in this paper, our hypothesis that privileged information is beneficial for multimodal affect-based interaction needs to be tested further. Even though vanilla convolutional neural networks appear to be performing well in this and earlier studies [6], [16], our plan is to test a number of different deep learning architectures for potentially improving the performance of LUPI models. Moreover, this study embeds the LUPI paradigm in the supervised affect modeling setting. We aim to extend the current methodology for transferring privileged information knowledge to semi-supervised and self-supervised learning paradigms in an attempt to learn powerful general-purpose representations for affect modeling. Another possible extension of this work is the application of LUPI beyond classification and regression to ranking and preference learning paradigms for ordinal affect modeling tasks [76], [77].

7 Conclusions

In this paper we introduce a methodology for building models of affect in the wild exploiting *privileged information*. Our hypothesis is that learning using privileged information can be used to reliably transfer affect models from controlled laboratory to uncontrolled real-world settings. To test our hypothesis we used the RECOLA and SEWA affect databases that include raw visual information, and audio-visual and physiology high-level handcrafted features. We consider all handcrafted features as privileged information (i.e. only available during model training) and assume that raw visual information corresponding to interaction footage frames is available during both training and testing. Under this setting, we treat affect modeling both as a classification and regression task, and develop models for predicting users' arousal and valence.

The core results suggest that affect models trained using privileged information perform equally well or even better than fusion affect models that consider all modalities. Importantly for affective computing research, privileged information affect models do not require access to costly, intrusive or impractical modalities when tested in the wild. Therefore, the findings of the paper bring affective computing one step closer to realising affect interaction in the wild. The proposed methodology for knowledge transfer by following the teacher's predictions and representations has direct applications to any affect modeling task that considers multimodal data and is required to operate in the wild or to make predictions using a subset of the available modalities. Potential applications include (but are not limited to) driver-assisting systems, affective robots, affect-aware recommendation systems, affective games [16], [78] and health applications at home such as stress monitoring and seizure detection.

Acknowledgments

This work has been supported by the European Union's Horizon 2020 research and innovation programme from the TAMED (Grant Agreement No. 101003397) and AI4media projects (Grant Agreement No. 951911). Antonios Liapis was supported by the Malta Council for Science and Technology (MCST) under the FUSION R&I: Research Excellence Programme (Project number: REP-2022-017).

References

[1] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, Jan. 2021.

[2] A. F. Botelho, R. S. Baker, and N. T. Heffernan, "Improving Sensor-Free Affect Detection Using Deep Learning," in *Artificial Intelligence in Education*, ser. Lecture Notes in Computer Science, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Cham: Springer International Publishing, 2017, pp. 40–51.

[3] L. Li, T.-T. Goh, and D. Jin, "How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4387–4415, May 2020.

[4] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5200–5204.

[5] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, Apr. 2021.

[6] K. Makantasis, A. Liapis, and G. N. Yannakakis, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Trans. on Affective Computing*, 2021.

[7] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien, "Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car," *Advances in human-computer interaction*, 2010.

[8] M. Braun, J. Schubert, B. Pfleging, and F. Alt, "Improving driver emotions with affective strategies," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 21, 2019.

[9] J. Kim and D. R. Fesenmaier, "Measuring emotions in real time: Implications for tourism experience design," *Journal of Travel Research*, vol. 54, no. 4, pp. 419–429, 2015.

[10] U. Tripathi, R. Saran, V. Chamola, A. Jolfaei, and A. Chintanpalli, "Advancing remote healthcare using humanoid and affective systems," *IEEE Sensors Journal*, 2021.

[11] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[12] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer." *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2023–2049, 2015.

[13] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. of the IEEE Int. conf. and workshops on automatic face and gesture recognition*, 2013.

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

[15] K. Makantasis, A. Liapis, and G. N. Yannakakis, "From pixels to affect: a study on games and player experience," in *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, 2019.

[16] K. Makantasis, D. Melhart, A. Liapis, and G. N. Yannakakis, "Privileged information for modeling affect in the wild," in *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, 2021.

[17] R. W. Picard, *Affective computing*. MIT press, 2000.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60:2, pp. 91–110, 2004.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Int. Conf. on computer vision and pattern recognition*, vol. 1, 2005, pp. 886–893.

[20] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. of the European Conf. on Computer Vision*, 2004, pp. 469–481.

[21] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, 2018.

[22] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "Emotion recognition from arbitrary view facial images," in *Proc. of the European Conf. on Computer Vision*. Springer, 2010, pp. 490–503.

[23] J. K. J. Julina and T. S. Sharmila, "Facial emotion recognition in videos using hog and lbp," in *Proc. of the IEEE Int. Conf. on Recent Trends on Electronics, Information, Communication & Technology*, 2019, pp. 56–60.

[24] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based hog features," in *Proc. of the IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 2011, pp. 884–888.

[25] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study," in *Proc. of the IEEE conf. on computer vision and pattern recognition workshops*, 2014.

[26] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Proc. of the IEEE Int. Conf. on affective computing and intelligent interaction*, 2015, pp. 77–83.

[27] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 995–1007, 2012.

[28] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: valence and arousal 'in-the-wild' challenge," in *Proc. of the IEEE Conf. on computer vision and pattern recognition workshops*, 2017, pp. 34–41.

[29] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational intelligence magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[30] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[31] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. of the IEEE Int. Conf. on computer vision*, 2015, pp. 2983–2991.

[32] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. of the Int. Conf. on multimodal interaction*, 2015, pp. 443–449.

[33] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[34] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. on Affective Computing*, 2018.

[35] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

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

[37] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision*, 2016.

[38] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.

[39] J. K. P. Seng and K. L.-M. Ang, "Multimodal emotion and sentiment modeling from unstructured big data: Challenges, architecture, & techniques," *IEEE Access*, vol. 7, pp. 90 982–90 998, 2019.

[40] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 02, pp. 52–58, 2021.

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

[42] M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[43] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[44] Z. Zheng, C. Cao, X. Chen, and G. Xu, "Multimodal emotion recognition for one-minute-gradual emotion challenge," *arXiv preprint arXiv:1805.01060*, 2018.

[45] D. Kollias, A. Tagaris, and A. Stafylopatis, "On line emotion detection using retrainable deep neural networks," in *Proc. of the IEEE Symp. Series on Computational Intelligence*, 2016.

[46] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 26–33.

[47] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[48] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2017, pp. 1021–1030.

[49] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[50] D. Aspandi, A. Mallol-Ragolta, B. Schuller, and X. Binefa, "Adversarial-based neural network for affect estimations in the wild," *arXiv preprint arXiv:2002.00883*, 2020.

[51] S. Parthasarathy and S. Sundaram, "Detecting expressions with multimodal transformers," in *Proc. of the IEEE Spoken Language Technology Workshop*, 2021, pp. 636–643.

[52] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

[53] C. Du, C. Du, H. Wang, J. Li, W.-L. Zheng, B.-L. Lu, and H. He, "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 108–116.

[54] M. Pagé Fortin and B. Chaib-Draa, "Multimodal sentiment analysis: A multitask learning approach." in *ICPRAM*, 2019, pp. 368–376.

[55] M. Pagé Fortin and B. Chaib-draa, "Multimodal multitask emotion recognition using images, texts and tags," in *Proceedings of the ACM Workshop on Crossmodal Learning and Application*, 2019, pp. 3–10.

[56] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2608–2618.

[57] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 400–404.

[58] N. Wang, H. Cao, J. Zhao, R. Chen, D. Yan, and J. Zhang, "M2r2: Missing-modality robust emotion recognition framework with iterative data augmentation," *IEEE Transactions on Artificial Intelligence*, 2022.

[59] C. Chen, S. Rosa, Y. Miao, C. X. Lu, W. Wu, A. Markham, and N. Trigoni, "Selective sensor fusion for neural visual-inertial odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 542–10 551.

[60] H. Yang, T. Wang, and L. Yin, "Adaptive multimodal fusion for facial action units recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2982–2990.

[61] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1165–1174.

[62] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Trans. on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[63] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature