

Privileged Contrastive Pretraining for Multimodal Affect Modelling

Kosmas Pinitas
Institute of Digital Games
University of Malta
Msida, Malta
kosmas.pinitas@um.edu.mt

Konstantinos Makantasis
Department of Artificial Intelligence
University of Malta
Msida, Malta
konstantinos.makantasis@um.edu.mt

Georgios Yannakakis
Institute of Digital Games
University of Malta
Msida, Malta
georgios.yannakakis@um.edu.mt

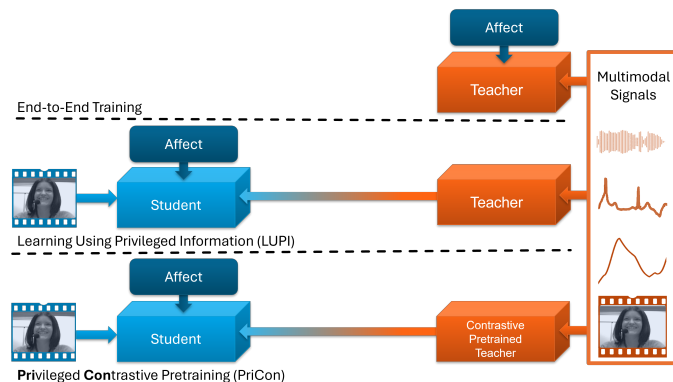


Figure 1: Visualising the proposed Privileged Contrastive Pretraining (PriCon) framework. The top figure illustrates the traditional end-to-end training of an affect model based on a set multimodal signals. The middle figure illustrates a LUPI paradigm for affect modelling, by which the student model learns affect labels from modalities available in the wild (e.g., visual frames in this example) while the teacher model uses additional multimodal signals available in the lab (e.g., audio and physiological signals in this example) during training. The bottom figure visualises the high-level concept of PriCon that trains a teacher model via SCL prior to transferring knowledge to the student via LUPI.

Abstract

Affective Computing (AC) has made significant progress with the advent of deep learning, yet a persistent challenge remains: the reliable transfer of affective models from controlled laboratory settings (*in-vitro*) to uncontrolled real-world environments (*in-vivo*). To address this challenge we introduce the *Privileged Contrastive Pretraining* (PriCon) framework according to which models are first pretrained via supervised contrastive learning (SCL) and then act as teacher models within a Learning Using Privileged Information (LUPI) framework. PriCon both leverages privileged information during training and enhances the robustness of derived affect models via SCL. Experiments conducted on two benchmark affective corpora, RECOLA and AGAIN, demonstrate that models trained using PriCon consistently outperform LUPI and end to end models. Remarkably, in many cases, PriCon models achieve performance comparable to models trained with access to all modalities during both training and testing. The findings underscore the potential of

PriCon as a paradigm towards further bridging the gap between *in-vitro* and *in-vivo* affective modelling, offering a scalable and practical solution for real-world applications.

CCS Concepts

• **Computing methodologies** → *Artificial intelligence; Neural networks*; • **Human-centered computing** → *Human computer interaction (HCI)*.

Keywords

affective computing; arousal; valence; privileged information; representation learning

ACM Reference Format:

Kosmas Pinitas, Konstantinos Makantasis, and Georgios Yannakakis. 2025. Privileged Contrastive Pretraining for Multimodal Affect Modelling. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25)*, October 13–17, 2025, Canberra, ACT, Australia. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3716553.3750766>

1 Introduction

Affective Computing (AC) has made significant leaps forward in recent years due to advances in deep learning [2, 9, 17, 35–37], which directly enhanced the predictive capacity of affect models. A major AC challenge remains, however: transferring models trained in controlled (*in-vitro*) environments to real-world (*in-vivo*) settings.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI '25, Canberra, ACT, Australia

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1499-3/25/10

<https://doi.org/10.1145/3716553.3750766>

Although laboratory conditions enable precise data collection, real-world scenarios introduce noise, privacy concerns, and hardware constraints, widening the *in-vitro* to *in-vivo* gap.

As a response to the above challenges this paper introduces a unified framework that combines the advantages of Supervised Contrastive Learning (SCL) and the Learning Using Privileged Information (LUPI) paradigm [38, 40]. The introduced *Privileged Contrastive Pretraining* (PriCon) framework both exploits additional modalities during training as privileged information for the model and improves the generalisability of derived affect models deployed in *in-vivo* scenarios (see Figure 1). While LUPI allows PriCon to leverage richer data during training than what is available at test time, SCL enhances the quality of learned representations by encouraging class-wise separation in the feature space. We integrate these two components within PriCon to address two key hypotheses. First, LUPI can enhance affect detection under real-world (*in-vivo*) conditions, producing performances comparable to *in-vitro* models (**H1**). Second, PriCon further enhances the effectiveness of *in-vivo* LUPI models (**H2**). Both hypotheses are evaluated on the RECOLA [29] and AGAIN [21] datasets, across different affective dimensions and modalities.

For testing H1, we incorporate fine-grained image features and fused data (i.e., image frames and features) as privileged information during *in-vitro* training, while using only frames for *in-vivo* testing. This approach reflects real-world limitations, where additional modalities (e.g., physiological signals) are impractical due to privacy and hardware constraints. For testing H2, PriCon is applied to *in-vitro* models, refining their representations and improving their transferability to *in-vivo* settings. Our results show that privileged information significantly enhances model performance (validating H1), while training privileged models via PriCon further improves *in-vivo* performance (validating H2). Notably, PriCon trained models achieve performances comparable to those obtained from models trained on all modalities.

This paper makes several key contributions towards advancing affect modelling in real-world settings. First, we evaluate the LUPI paradigm for affect modelling in games and dyadic interactions, demonstrating its potential in dynamic environments. Second, we introduce the PriCon framework that boosts the robustness and transferability of *in-vitro* LUPI models via SCL. Third, we validate the approach introduced on both RECOLA and AGAIN datasets, showing its robustness across affective dimensions (arousal and valence) and tasks (games vs. dyadic interactions). These findings underscore the benefits of *privileged contrastive pretraining* in creating more generalisable affective models for real-world applications.

2 Related Work

This section reviews related work on the two key learning paradigms featured in PriCon. Specifically, we survey learning using privileged information and contrastive learning under the lens of affect modelling.

2.1 LUPI for Affect Modelling

While both distillation and LUPI involve a teacher-student framework, they serve different purposes. Traditional knowledge distillation focuses on compressing knowledge by training a student

model to mimic the outputs of a more complex teacher model. For instance, Sun et al. [33] proposed a novel technique for micro-expression detection that distills knowledge from a pre-trained deep teacher neural network to a shallow student neural network via a teacher-student correlative framework enabling the student to capture fine-grained details necessary for accurate micro-expression recognition. Similarly, Liu et al. [12] developed a cross-modal consistency modelling-based knowledge distillation framework for image-text sentiment classification of social media data. By aligning the information between images and text, the student model can effectively capture the sentiment expressed in multimodal social media posts.

Several affective computing problems have asymmetric train and test input distributions. This discrepancy means that the training phase might have access to richer or more diverse data than what is available during testing. Hence, it is not surprising that the LUPI paradigm has started to become popular for AC research. LUPI addresses this asymmetry by allowing the use of additional information during training that is not available during testing. Makantasis et al. [16] introduced a ranking model that treats additional training information as privileged information to rank affect states. This approach leverages richer data available during training, such as audio or contextual information, to improve the ranking model's accuracy in predicting affective states. In another study [18], the same authors predicted arousal from gameplay footage while treating telemetry and heart rate as privileged information. Using detailed physiological and telemetry data during training, the model could better capture the nuances of player arousal based solely on gameplay footage during testing. Zhang et al. [46] proposed a LUPI method to distil EEG representations via capsule-based architectures for both classification (positive, negative, neutral emotion) and regression (vigilance) tasks. The capsule-based architectures captured spatial hierarchies in the data, making the distilled knowledge more effective for the student model.

This work extends LUPI-based affect modelling by applying it to affect classification, a previously unexplored area. Unlike prior studies that focus on ranking affect states or knowledge distillation for efficiency, our approach leverages multimodal privileged information to improve the discriminative power of the unimodal student model, enabling more accurate classification of affective states. Additionally, we enhance model robustness by pretraining the teacher models via SCL, improving generalisation.

2.2 Contrastive Learning for Affect Modelling

Contrastive learning techniques are among the most widely applied methods for learning representation. Notably, Li et al. [10] investigated the impact of unsupervised representation learning for speech emotion recognition. Their study demonstrated that the proposed contrastive predictive coding method based on InfoNCE produced representations that achieved state-of-the-art performance across the activation, valence, and dominance dimensions.

Building on these advances, Mai et al. [15] introduced a novel hybrid contrastive learning framework. This framework integrates intra-modal, inter-modal, and semi-contrastive learning to enable the model to explore cross-modal interactions and preserve inter-class relationships, thereby reducing the modality gap. Their approach demonstrated significant improvements in capturing the

complex relationships between different modalities, such as audio and visual data, in emotion recognition tasks. Pinitas et al. [24] employed supervised contrastive learning on fine-grained multimodal features to develop robust arousal-infused representations. Their work underscored the importance of incorporating supervised learning signals to enhance the quality of representations, leading to higher arousal classification accuracy than the end-to-end alternatives. Yang et al. [43] proposed an innovative low-dimensional supervised cluster-level contrastive learning method. This approach reduces the high-dimensional SCL space to a more manageable three-dimensional affect representation. By efficiently compressing the feature space, their method facilitates more effective and computationally efficient emotion recognition while maintaining high performance. Recently, Pinitas et al. [25] introduced a novel few-shot representation learning framework that fine-tuned self-supervised models for engagement state prediction across FPS games outperforming end-to-end baselines.

Unlike the aforementioned works that focus on unsupervised, hybrid, or cluster-level contrastive learning for affect modelling, this paper integrates supervised contrastive learning within the LUPI paradigm with the aim to improve in-vitro to in-vivo generalisation. Additionally, the focus on affect classification in games and dyadic interactions, rather than general emotion recognition, sets this study apart by tackling dynamic and interactive environments where real-time adaptability is crucial and very much needed.

3 Methodology

This section first describes the use of privileged information and supervised contrastive learning in a concise way and then moves on to outline the model architectures employed.

3.1 Learning Using Privileged Information

Learning using privileged information [38, 40] addresses problems characterised by an asymmetric distribution of information between training and test time. LUPI provides the means to *transfer knowledge* from all available modalities to a machine learning model that makes predictions using only a subset of these modalities [13, 31]. As far as the RECOLA database is concerned, we treat as privileged the information that corresponds to physiology and audiovisual features provided by the database creators [29]. For AGAIN, we consider fine-grained gameplay features as privileged information [21]. Our choice is justified by the fact that capturing physiology requires specialised sensors, while constructing physiology and audiovisual features implies the employment of specific software algorithms. In the same vein, information about the state of the game (e.g., number and actions of enemies) requires access to the game engine itself. On the contrary, information that comes from raw footage frames is considered prevalent information as it can be captured using conventional cameras that can be available at both training and test times.

This study explores the use of privileged information with neural network-based affect models. Following [6, 13, 39], we represent privileged knowledge—i.e., information available only during training, such as physiological signals or high-resolution feature embeddings—within the output of a neural network that has been trained to make predictions based on either all available modalities

(privileged + standard) or on privileged information alone. This model is called *teacher*. A trained teacher model can transfer knowledge obtained through privileged information to another model called *student*. The transfer of knowledge can be achieved by feeding the model with only those modalities of information that are available in the wild and forcing it during training to balance (via hyperparameter α) between learning the task (cross-entropy loss L_{CE}) and learning prediction distributions that match those of the teacher model (KL divergence loss L_{KL}):

$$L_P = (1 - \alpha)L_{CE}(f(x), y) + \alpha L_{KL}(g(\tilde{x}), f(x)) \quad (1)$$

The resulting optimisation objective L_P is shown in Eq. (1). Where $f(x)$ and $g(\tilde{x})$ are the outputs of the student and teacher models, respectively, while $\tilde{x}$ refers to the privileged modalities. It should be noted that 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.

3.2 Privileged Contrastive Pretraining

A central component of the PriCon framework is the pretraining of teacher models using SCL, which lays the foundation for effective privileged supervision. By leveraging class label information, SCL guides the formation of semantically meaningful sample pairs, resulting in highly discriminative representations. This results in an embedding space where similar instances are clustered together, and dissimilar ones are pushed apart, enhancing the model's capacity to generalise.

In the PriCon framework, this teacher pretraining stage is not merely auxiliary but is instrumental in learning rich, class-consistent representation structure. Specifically, for each input x_i (referred to as the *anchor*), SCL constructs a set of *positives* P_i , which includes all other instances in the batch sharing the same class label, and contrasts them against *negatives*—instances from different classes. This structured approach enhances the teacher's ability to model fine-grained intra-class similarities and inter-class separability, which is later transferred to the student model through the LUPI paradigm.

Compared to traditional supervised learning, which typically optimises for loss minimisation without explicitly shaping the latent space, SCL provides a more expressive training signal. By encouraging consistent semantic alignment in the representation space, it equips the teacher with a more nuanced understanding of class distributions—an essential property for the downstream knowledge transfer in PriCon. Formally, the supervised contrastive loss L_{SC} is defined as:

$$L_{SC} = \sum_{i \in I} \frac{-1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(r_i \cdot r_p / \tau)}{\sum_{a \in A_i} \exp(r_i \cdot r_a / \tau)} \quad (2)$$

where I is a set that includes all samples and P_i is the set that includes only the samples that are assigned to the same class as i . A_i is a set that contains any element of set I besides element i . With r_i , r_p and r_a we denote the latent representations of the model for the samples i , p and a , respectively. Finally, τ stands for a non-negative temperature hyperparameter that controls the sharpness of the similarity distribution. It should be noted that i , p and a correspond to the index of the current sample, a sample positive

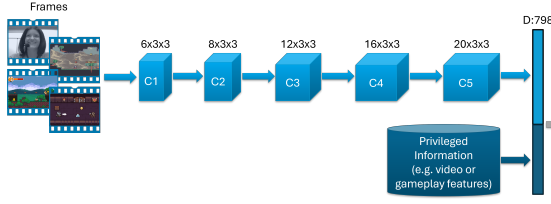


Figure 2: Illustration of the affect models. ‘C’ and ‘D’ denote convolutional and dense layers, respectively. The student model (light blue) and privileged teacher (dark blue) are fused in the fusion teacher via the gray module.

to the current sample, and a sample different from the current one, respectively.

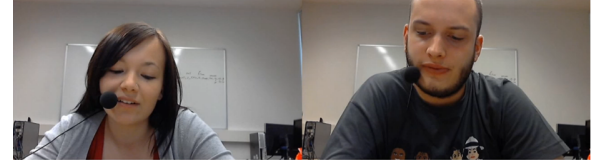
This formulation ensures that representations of samples from the same class are tightly clustered by maximising $r_i \cdot r_p$, while dissimilar samples are discouraged from occupying nearby regions in the latent space via the denominator term. In the context of PriCon, SCL serves as a powerful mechanism to encode privileged information into the teacher model, thus laying a strong representational foundation for effective student training in the subsequent knowledge transfer phase.

3.3 Model Architectures

In this section, we first present the main components used in LUPI, namely student and teacher models. Then, we present the baseline architectures that we compare against our method for assessing the effectiveness of the obtained models. Both student and teacher architectures are illustrated in Figure 2.

3.3.1 Student Model. The student neural network S is a predictive model that learns to estimate the target outcomes by utilising the information from frames. During the training phase, the student model leverages knowledge distilled from a teacher model T , which has access to an additional set of modalities (i.e., privileged information). This privileged information, available only during training, enhances the learning process by providing auxiliary guidance. The goal of the student model is to effectively generalise to new, unseen data, where privileged information is not available, using only the primary feature set. In this work S is a Convolutional Neural Network (CNN) comprising five convolutional layers of 6, 8, 12, 16 and 20 filters, respectively, followed by a dense layer. The first four convolutional layers are configured with a stride parameter of 2, while the fifth convolutional layer uses a stride of 1. Each convolutional layer has a kernel size of dimensions 3×3 and employs the ReLU activation function. Finally, a dense layer with 768 neurons, also activated using ReLU, is applied to produce the encoded frame representation. The output of this layer is fed to the decision layer, which is a simple 2-neuron dense SoftMax-activated layer. A dropout of 0.1 is applied before the last layer.

3.3.2 Teacher Models. The teacher, T , is a predictive model that is trained with access to both the primary feature set (i.e., prevalent information) and an additional set of privileged information. The teacher model is typically designed to achieve high performance by leveraging this richer information. Once trained, the teacher model’s knowledge is distilled and transferred to the student model



(a) RECOLA



(b) Platformer Games (AGAIN)

Figure 3: Sample frames from the datasets. (a) RECOLA: dyadic interactions. (b) AGAIN: three platformer games: *Endless*, *Pirates!*, and *Run'N'Gun!* (from left to right).

S , guiding it to improve its predictions based on the primary feature set alone. The teacher model plays a crucial role in the LUPI framework by acting as a source of enhanced supervision during training. In this paper, we consider two teacher architectures. **Privileged Teacher:** The privileged teacher T_p considers only feature (privileged) information. Consequently, it is a simple ANN consisting of two layers. The first layer has 30 neurons employing the logistic activation function. The output of this layer is fed to the decision layer which is a simple 2-neuron dense SoftMax-activated layer. Once again a dropout of 0.1 is applied before the last layer. **Fusion Teacher:** The fusion teacher T_f has access to both privileged and prevalent information (i.e., features and frames). In particular, it fuses the penultimate layers of the frame and feature models described above. To account for the difference in dimensionality, an additional ReLU-activated layer of 30 neurons is added to the frame encoder to match the dimensions of the feature encoder. In this case the feature encoder is a single 30-neuron layer activated by ReLU. The last layer of the fusion encoder is a linear layer of 60 neurons activated by ReLU. The purpose of this last layer is to fuse information from both frames and features in order to yield the final representation. The output of the fusion layer is fed to a 2-neuron dense layer which is activated by SoftMax. A dropout of 0.1 is applied before the output layer.

3.3.3 Baselines. There are two baseline architectures employed in this paper corresponding to student models that do not consider privileged information at all. The first architecture, used for comparison with the conventional LUPI framework, performs end-to-end affect classification from frames, bypassing the representation learning process. The second baseline, used for comparison with the PriCon framework, first trains the latent dimension of the student model via SCL and then probes the frozen learned embeddings with a SoftMax activated layer as the one used in all models of this work.

4 Datasets

This section presents the datasets and preprocessing steps used to evaluate the performance of the proposed approach and test our hypotheses.

4.1 RECOLA Database

The RECOLA dataset (Figure 3a) contains 9.5 hours of synchronised multimodal data from 46 French-speaking participants, including audio, visual, and physiological signals (electrocardiogram and electrodermal activity). It features annotations for arousal and valence provided by six experts (third-person annotation). The dataset includes raw footage (frames) along with fine-grained feature sets: 40 visual features (e.g., facial action units, head pose, optical flow), 130 audio features (e.g., voice intensity, pitch, mel-frequency spectral coefficients), and 116 physiological features extracted from cardiac and electrodermal activity. Raw data is also available for further processing. Of the 46 participants, 34 consented to data sharing, 23 of which are included in the public release.

4.1.1 Processing RECOLA. As mentioned earlier, the RECOLA database offers annotation traces for both arousal and valence. The same preprocessing approach is followed for both affective dimensions. Specifically, we segment each participant’s session (features and frames) into overlapping time windows using a sliding step of 400 ms and window lengths of 1, 2, and 3 s. These hyperparameters—the sliding step and window length—influence the size of the dataset and the temporal granularity of information contained in each window. Since the features and annotations are already synchronised, there is no need to account for the reaction time between the stimulus and the emotional response. After segmentation, each time window contains a sequence of feature vectors and a sequence of frames. To reduce computational complexity, we compute the average value for each feature within the window, representing the time window with a single feature vector. This ensures that the dimensionality of the feature vector is independent of the window length. For frame sequences, we retain 5 greyscale frames with dimensions 224×224 per second. For example, the input for a 3 s time window consists of a single feature vector (via averaging) and a frame tensor with dimensions $15 \times 224 \times 224$.

When it comes to affect annotation, we use the median annotation values per time window in order to derive a single trace segment—one per affect dimension—that mitigates inter-annotator disagreement [4]. The median affect trances are then averaged across time windows resulting in two scalar values per time window (g_a for arousal and g_v for valence). To produce the corresponding class labels we binarise the corresponding affect values. In RECOLA, the binarisation criterion is determined by the median ground truth value of the entire set of affect annotation traces ($\tilde{g}_s$) and a threshold ϵ . Specifically, a time window i is labelled as “high” if $g_{s_i} > \tilde{g}_s + \epsilon$ and as “low” if $g_{s_i} < \tilde{g}_s - \epsilon$. The median is used for RECOLA because each video stimulus has been annotated by the same six expert annotators, making the median a robust measure of central tendency. For RECOLA, ϵ is set to 0.1, ensuring a precise exclusion of ambiguous annotations.

4.2 AGAIN Dataset: Platformer Games

For the experiments reported in this paper, we focus on the three games of the platformer genre featured in the AGAIN dataset (Figure 3b) as they offer sufficiently diverse gameplay properties without in need of excessive computation for experimental validation. The three games examined include **Endless**, an infinite runner where players must avoid obstacles while automatically moving

ever rightward, **Pirates!**, a jumping platformer similar to *Super Mario Bros* (Nintendo, 1985), and **Run’N’Gun!**, a more complex game which requires players to move while aiming and shooting at enemies. All games have arcade-style controls of varying complexity, with Run’N’Gun! being the most complex of the three. Each game assigns a score to the player depending on their in-game performance. For the purpose of controlling the data collection process, we limit gameplay duration for all games to two minutes. The AGAIN-Platformers dataset consists of 120 participants that played and annotated their own gameplay in terms of arousal. Apart from raw gameplay footage the creators of AGAIN have also provided a high-level feature set for each game. The general feature set includes 14 features that describe player actions, environmental elements, and game events. Moreover, the dataset includes a set of game-specific extracted features for each game (33 for Endless!, 39 for Pirates! and 47 for Run’N’Gun!). The game-specific features correspond to player status (e.g., player health), gameplay events (e.g., bot aims at player), bot status and the proximal and general game context (e.g., bot-player distance and pickups visible).

4.2.1 Processing AGAIN. Each game in the AGAIN dataset provides self-reported arousal annotations, which are used in this work with a consistent preprocessing approach applied separately to each of the three games. Similarly to RECOLA, we segment each participant’s session into overlapping time windows using a sliding step of 500 ms and window lengths of 1, 2, and 3 s. Unlike the RECOLA dataset, synchronisation is required for AGAIN. To account for the reaction time between stimulus and emotional response, the arousal annotations are shifted backward by 1 s [22, 23]. Furthermore, the trace of each participant is normalised within $[0, 1]$ due to the unbounded nature of the annotations in this dataset.

After preprocessing, each time window contains a sequence of feature vectors and a sequence of frames. To reduce computational load, we compute the average value of each feature within the window, resulting in a single feature vector that represents the entire window. For frame sequences, we retain 5 greyscale frames with dimensions 224×224 per second. In the case of arousal annotation, there is no need to mitigate the disagreement between annotators since each participant provides a single trace per game per session (i.e., self-reporting annotations). For each session, we compute the arousal score g_a as the mean value of this trace, yielding a scalar arousal score per session. This preprocessing pipeline is repeated independently for each of the three games in the dataset. For the AGAIN dataset, the binarisation criterion is based on the mean value of the annotation traces ($\tilde{g}_s$) for each session, as it allows us to anchor binarisation relative to each participant’s own reporting baseline, rather than applying a fixed global threshold that may not generalise across participants or game contexts [19, 28]. For AGAIN, ϵ is set to 0.2, accommodating the variability in self-reported annotations. These thresholds were empirically validated to ensure a balanced exclusion of ambiguous annotations while maintaining sufficient data [17].

5 Results

This section presents the framework for evaluating the impact of contrastive teacher pretraining for privileged information on affect modelling and the experimental results obtained.

5.1 Evaluation Framework

As mentioned earlier, we treat affect as a classification task. Hence, our models attempt to predict high and low affect (arousal or valence) states. We adopt a 5-fold cross-validation scheme to evaluate the models' performance across all experiments reported in the paper. When splitting a dataset, we make sure that data from the same participant is either in the training or the test set, not both. Before any train-test split we hold out 10% of the participants randomly as our validation set. This set is used for stopping training early so as to avoid model overfitting; i.e., training stops after 5 epochs without loss improvement on the validation set. The training, validation and test sets are the same for all models reported. We report models' performance in terms of binary classification accuracy, since after class splitting, the datasets are balanced. A key component of the LUPI training objective is the hyperparameter α , which controls the teacher's influence. We perform an exhaustive search over candidate values $\alpha \in \{0.25, 0.5, 0.75, 1.0\}$ using the validation set and adopt the value that yields the best performance for subsequent experiments. All reported results are statistically validated via pairwise t-tests ($p < 0.05$).

In this work, student models make predictions using solely information that is available in the wild; in our experiments that is the raw footage frames. The student models trained via LUPI are denoted as S_p and S_f and have access to privileged information through the Privileged Teacher and the Fusion Teacher, respectively. We compare the student's performance against the performance achieved by the Fusion Teacher (T_f) that uses all modalities (i.e., features & frames) for training and testing, and the Privileged Teacher (T_p) that makes predictions using only privileged information. E refers to models that do not utilise privileged information during training. In Tables 1 and 2 the LUPI columns refer to the conventional approach where the teachers have been trained in an end-to-end fashion in this case E is also trained in the same manner. The PC column refers to LUPI with teachers that have been pretrained via SCL (proposed PriCon framework). The accuracy of the T_p , T_f and E models in the PC setting is obtained by applying a linear probe on top of the learned representations.

5.2 The Importance of Privileged Information

To test **H1**—which posits that LUPI can enhance affect detection under real-world (in-vivo) conditions, producing performances comparable to in-vitro models—we evaluate and compare the performance of baseline, teacher, and student models within the LUPI framework. Table 1 reports the average 5-fold validation accuracy for high–low arousal and valence classification tasks on the RECOLA dataset. The results are analysed across three different time windows (1s, 2s, 3s). The models evaluated include baseline models, teacher models (Fusion and Privileged Teachers), and student models trained via the LUPI framework (see LUPI column).

Baseline models (E) consistently achieve low accuracy scores across all six experimental settings, indicating that models relying solely on raw frame data are unable to achieve competitive performance. This highlights the limitations of training models without incorporating privileged information or additional learning signals. The teacher models—namely the Fusion Teacher (T_f) and the Privileged Teacher (T_p)—yield significantly higher accuracy

Table 1: RECOLA: Average accuracy prediction for arousal and valence. E corresponds to baseline models. S_p and S_f correspond to student models that have been trained under the privileged (T_p) and fusion (T_f) teacher, respectively. LUPI and PC denote, respectively, vanilla learning via privileged information and our introduced PriCon framework. Bold values correspond to the student model with the highest accuracy and underlined values correspond to models whose performance is statistically on par ($p \geq 0.05$) with that model.

Arousal		1 second		2 seconds		3 seconds	
Model		LUPI	PC	LUPI	PC	LUPI	PC
E		59.41	60.60	55.73	57.86	58.22	60.85
S_p		63.61	65.27	60.68	60.64	60.96	63.07
S_f		60.75	63.74	61.20	62.31	61.62	62.31
T_p		69.24	71.41	71.59	73.58	73.11	75.45
T_f		63.04	<u>64.60</u>	66.73	68.86	66.22	69.49
Valence		1 second		2 seconds		3 seconds	
Model		LUPI	PC	LUPI	PC	LUPI	PC
E		<u>60.84</u>	<u>60.19</u>	<u>60.60</u>	<u>60.40</u>	62.36	61.34
S_p		<u>59.94</u>	62.26	<u>60.07</u>	<u>59.82</u>	64.90	62.42
S_f		<u>61.66</u>	<u>61.58</u>	60.78	<u>60.09</u>	62.51	59.61
T_p		<u>60.95</u>	<u>62.10</u>	56.51	<u>59.48</u>	<u>65.15</u>	<u>62.80</u>
T_f		<u>62.74</u>	<u>62.38</u>	<u>61.26</u>	<u>61.99</u>	<u>64.74</u>	<u>64.21</u>

Table 2: AGAIN: Average accuracy prediction for arousal across three platformer games of the dataset. Bold values correspond to the student model with the highest accuracy and underlined values correspond to models whose performance is statistically on par ($p \geq 0.05$) with that model.

Endless		1 second		2 seconds		3 seconds	
Model		LUPI	PC	LUPI	PC	LUPI	PC
E		65.96	<u>67.83</u>	63.14	65.84	65.52	65.54
S_p		66.71	67.95	67.08	68.21	66.36	69.51
S_f		66.17	<u>67.26</u>	66.07	<u>67.50</u>	66.27	<u>69.29</u>
T_p		71.57	72.65	73.25	74.34	72.77	81.33
T_f		71.16	72.83	73.14	74.66	72.06	78.95
Pirates!		1 second		2 seconds		3 seconds	
Model		LUPI	PC	LUPI	PC	LUPI	PC
E		<u>67.63</u>	<u>67.03</u>	66.22	65.49	<u>66.54</u>	<u>66.57</u>
S_p		<u>66.79</u>	<u>66.89</u>	66.03	<u>67.06</u>	66.42	<u>66.86</u>
S_f		<u>66.60</u>	67.12	<u>66.89</u>	67.25	66.01	67.39
T_p		65.36	<u>67.46</u>	65.50	67.05	65.66	<u>66.72</u>
T_f		65.66	<u>67.16</u>	65.62	<u>67.75</u>	65.65	66.46
Run'N'Gun!		1 second		2 seconds		3 seconds	
Model		LUPI	PC	LUPI	PC	LUPI	PC
E		67.40	70.21	68.11	69.95	68.07	68.77
S_p		70.40	71.32	70.43	71.41	69.73	71.11
S_f		<u>70.66</u>	<u>71.21</u>	69.80	72.05	68.96	<u>70.43</u>
T_p		73.05	74.66	74.15	75.14	73.25	73.68
T_f		<u>71.40</u>	73.43	<u>72.11</u>	73.95	73.07	73.88

values for arousal compared to the baseline models. T_p teachers, which consider only privileged features, consistently achieve the highest accuracy, establishing the upper bound for arousal classification. On the other hand, T_f teachers—trained on both privileged information and frames—achieve competitive results but typically perform worse than T_p .

LUPI-trained student models (S_p, S_f) demonstrate substantial improvements over baseline models. Specifically, S_f , which is trained under the supervision of the Fusion Teacher, performs on par with or even outperforms S_p , which is guided by the Privileged Teacher. This finding suggests that the inclusion of frames in the T_f teacher can allow for better information transfer during the LUPI training process.

While significant improvements are observed for arousal classification, valence classification results remain relatively stable across models, training strategies, and time window lengths. This can be attributed to the inherent challenges of valence prediction, which relies on subtle and ambiguous cues that are less dynamic and harder to capture than arousal. Valence-related expressions often appear visually similar across emotional contexts, making it difficult for models using raw frames to learn discriminative features [1, 20]. Additionally, while privileged information provides strong supervisory signals for arousal, it may not capture sufficient valence-specific information, leading to weaker guidance from the T_p and T_f teachers. This suggests that valence recognition might require additional contextual information, such as semantics or interaction dynamics [32, 42].

Table 2 presents the arousal classification results for the AGAIN dataset, which includes three platformer games: *Run'N'Gun!*, *Pirates!*, and *Endless*. Each game is evaluated for time windows of 1s, 2s, and 3s. In *Run'N'Gun!*, T_p achieves the highest performance, establishing an upper bound. Among the LUPI-trained student models, S_p performs on par with or slightly better S_f , and both improve upon the baseline models (E).

For *Pirates!*, performance remains relatively stable across all time windows, with limited improvements as the temporal context increases. This stability suggests that arousal cues are less dynamic and harder to capture in this game due to subtler changes in play. In this case, both teachers perform on par; the same holds for S_f and S_p among the student models. It is worth noting that here, teacher models perform on par with student models that do not use privileged information. Consequently, there is no information gain to be exploited by the LUPI framework. Finally, in *Endless*, the end-to-end teachers perform on par. Among the LUPI-trained students, S_f and S_p also perform similarly across all time windows, consistently improving over the baseline models E .

The obtained results collectively across two different datasets and 3 different affective dimensions validate **H1**, confirming that the LUPI framework can enhance affect detection under real-world (in-vivo) conditions. LUPI-trained student models consistently outperform baseline models and, in several cases, closely approximate the performance of their privileged teacher counterparts. This demonstrates the potential of privileged supervision to guide the learning of compact, frame-only student models that generalise well in realistic settings where privileged modalities are unavailable at test time.

5.3 Privileged Contrastive Pretraining

To evaluate **H2**, we investigate the impact of PriCon—a framework that integrates SCL within LUPI—on the performance of both teacher and student models. The central question is whether teacher pretraining via SCL leads to more effective student learning and ultimately enhances binary affect classification. To evaluate this, we focus on the results reported in the ‘PC’ columns of Tables 1 and 2, which isolate the influence of PriCon on model performance across different datasets, affective states, and time window configurations.

For the RECOLA dataset (Table 1), PriCon significantly improves model performance across the board. SCL-pretrained teacher models, T_f and T_p , (PC column) not only match but often outperform their end-to-end LUPI counterparts. This indicates that contrastive pretraining enables the learning of more discriminative and generalisable affect representations, which are beneficial regardless of whether the teacher is trained on fused or privileged data. Importantly, these benefits are transferred downstream: student models S_p and S_f consistently outperform the end-to-end baseline E , and in multiple cases even match or exceed the performance of SCL-pretrained E . These results validate the effectiveness of the PriCon framework, allowing student models to benefit from the increased performance of the SCL pretrained teachers. Valence classification results, however, remain relatively stable. While PriCon provides modest gains, the improvements are less substantial due to the subtle and ambiguous nature of valence cues in visual data, as discussed in Section 5.2.

In the AGAIN dataset (Table 2), PriCon continues to show strong and consistent gains for student models. In *Run'N'Gun!*, both S_p and S_f outperform not only the baseline E but also their end-to-end LUPI versions, reinforcing the utility of combining contrastive objectives (for teacher pretraining) with the conventional LUPI paradigm. This demonstrates PriCon’s capacity to bridge the modality gap during student training and produce better generalising models. In *Pirates!*, although the overall gains are smaller, PriCon yields noticeable improvements in 2-second time windows. Here, S_p and S_f significantly outperform E , even when teacher performance remains similar. In *Endless*, PriCon teachers train students (see PC column) that achieve higher accuracy than both their end-to-end counterparts (LUPI column) and the baseline models E . Notably, the performance gap is most pronounced for the 2s and 3s windows, suggesting that temporal context plays a role in enhancing the benefits of SCL teacher pretraining.

In summary, obtained results validate **H2**, demonstrating the effectiveness of the PriCon framework. By integrating SCL for teacher pretraining and LUPI for privileged information distillation, PriCon enables the development of robust affective models with improved generalisation across tasks (affect dimensions), input modalities, and temporal resolutions.

6 Discussion

The findings presented in this paper underscore the significant potential of *privileged contrastive pretraining* for affect modelling. By focusing on binary classification (high vs. low affective state) our introduced method demonstrated improved robustness and performance across both RECOLA and AGAIN datasets. However,

several methodological limitations and considerations must be acknowledged. Although arousal classification benefits substantially from PriCon, valence results show only marginal gains, indicating that valence remains a harder target for generalisation. Thus future work should explore the integration of user context [34] and MAMBA [11] architectures within the PriCon framework to further improve valence prediction.

The study employed an exhaustive search protocol to fine tune the α hyperparameter, which controls the influence of the teacher in LUPI. While this approach ensured optimal performance for the chosen datasets, it may not generalise well to other datasets or applications. The fixed hyperparameter tuning approach assumes a universal solution, potentially missing dataset-specific variations. Tuning was done for 1-second time windows, assuming the optimal α value is consistent across all windows. While more efficient methods like Bayesian optimisation could improve tuning, their computational demands were deemed impractical for this study. Future work could explore adaptive optimisation techniques such as gradient-based hyperparameter learning [3, 14] for better efficiency and performance across datasets.

Beyond hyperparameter tuning, the manual selection of privileged features represents another limitation of this study. While the selected features were thoughtfully aligned with domain knowledge and task-specific objectives, the ad-hoc approach followed may have overlooked complex, latent structures within the datasets that automated methods could exploit. Attention mechanisms and feature-ranking algorithms [5, 45] could help identify and prioritise the most relevant features, revealing patterns missed by manual selection. However, implementing these methods would require additional resources, potentially diverting attention from the study's main goal of evaluating the proposed framework. While manual selection ensured interpretability, future work could explore automated techniques to enhance scalability and effectiveness.

While the binary classification paradigm is widely used to showcase the robustness of AC frameworks [7, 8, 26, 30], a natural extension would be to explore the potential of PriCon in semi-supervised and self-supervised learning paradigms. Such extensions could unlock the ability to learn powerful general-purpose representations that are better suited for diverse affect modelling tasks, especially in scenarios with limited labelled data. These paradigms would be particularly beneficial for scaling affective computing systems in real-world applications where data labelling is often costly and time-consuming.

Another promising direction involves extending PriCon to ranking paradigms, which are essential for ordinal affect modelling tasks [16, 44]. The incorporation of PriCon into these paradigms could enable models to better exploit privileged information for inferring subtle affective gradients, thereby improving their ability to model subjective human experiences. Furthermore, combining PriCon with few-shot learning strategies [27, 41] could facilitate robust affect modelling in low-data regimes, a common challenge in real-world scenarios. However, these extensions would require significant methodological adaptations. These adaptations—while outside the scope of the current study—represent exciting opportunities for advancing the field and addressing more complex affective computing challenges in future research. Finally, regarding evaluation, this study reports binary classification accuracy since the

datasets are balanced. Nonetheless, future work could explore alternative metrics—such as F1 score, AUC, or mean absolute error—that may offer complementary insights, particularly in the presence of class imbalance or in regression-based affect modelling.

7 Conclusions

In this paper, we introduced a methodology for affect modelling in real-world scenarios by leveraging privileged information and teacher pre-training via SCL. Our central hypothesis posits that privileged information can facilitate the reliable transfer of affect models from controlled environments, where abundant high-quality data is available, to real-world settings. We evaluated this hypothesis in the RECOLA and AGAIN datasets, which include both raw visual data and high-level handcrafted features. We considered all handcrafted features as privileged information that is accessible solely during model training, while also considering raw visual data such as video frames as available during both training and testing. As part of the PriCon framework, we pretrain teacher models via SCL and subsequently transfer their knowledge to student models, which operate without privileged information during testing. This approach assumes that teachers with a higher predictive power can produce more robust student models.

Our findings for arousal and valence prediction highlight the combined advantages of leveraging privileged information and contrastive pretraining. Affect models trained via the PriCon framework not only match but often exceed the performance of their teacher models, while operating solely on frame-based features at test time. Remarkably, in many cases, PriCon models achieve performance comparable to models trained with access to all modalities during both training and testing. The proposed methodology has broad applicability to affective computing tasks involving multi-modal data. It is especially valuable in scenarios where access to certain modalities is limited or unavailable, offering a scalable and effective approach to emotion recognition in the wild. The findings underscore the potential of PriCon as a paradigm towards further bridging the gap between *in-vitro* and *in-vivo* affective modelling, offering a scalable and practical solution for real-world applications.

Safe and Responsible Innovation Statement

The experiments presented in this paper were conducted using publicly available datasets containing frames, fine-grained features, and affect annotations. Any personally identifiable information was anonymised with untraceable IDs, ensuring that such data was not accessible to us. Participants consented to having their faces publicly available. Furthermore, the datasets used do not contain potentially offensive content, and the proposed framework is based on open-source methods, ensuring reproducibility. Finally, to the best of our knowledge, our work does not contribute to the development of deceptive applications or exacerbate existing privacy or discriminatory concerns.

Acknowledgments

This project has received funding from the Malta Council for Science and Technology through the SINO-MALTA Fund 2022, Project OPtiMaL.

References

- [1] Muhammad Haseeb Aslam, Muhammad Osama Zeeshan, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, and Eric Granger. 2023. Privileged knowledge distillation for dimensional emotion recognition in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3338–3347.
- [2] Anthony F. Botelho, Ryan S. Baker, and Neil T. Heffernan. 2017. Improving Sensor-Free Affect Detection Using Deep Learning. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 40–51. doi:10.1007/978-3-319-61425-0_4
- [3] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. 2017. Forward and reverse gradient-based hyperparameter optimization. In *International conference on machine learning*. PMLR, 1165–1173.
- [4] Oliver Grewe, Frederik Nagel, Reinhard Kopiez, and Eckart Altenmüller. 2007. Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion* 7, 4 (2007), 774.
- [5] Ning Gui, Danni Ge, and Ziyin Hu. 2019. AFS: An attention-based mechanism for supervised feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3705–3713.
- [6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *Stat* 1050 (2015).
- [7] Smith K Khare, Victoria Blanes-Vidal, Esmail S Nadimi, and U Rajendra Acharya. 2024. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information fusion* 102 (2024), 102019.
- [8] Gs Shashi Kumar, Niranjana Sampathila, and Roshan Joy Martis. 2023. Classification of human emotional states based on valence-arousal scale using electroencephalogram. *Journal of Medical Signals & Sensors* 13, 2 (2023), 173–182.
- [9] Lin Li, Tiong-Thye Goh, and Dawei Jin. 2020. How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. *Neural Computing and Applications* 32, 9 (May 2020), 4387–4415. doi:10.1007/s00521-018-3865-7
- [10] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. 2021. Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6329–6333.
- [11] Yuheng Liang, Zheyu Wang, Feng Liu, Mingzhou Liu, and Yu Yao. 2025. Mamba: A mamba-based approach for continuous emotion recognition in valence-arousal space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5651–5656.
- [12] Huan Liu, Ke Li, Jianping Fan, Caixia Yan, Tao Qin, and Qinghua Zheng. 2022. Social Image-text Sentiment Classification With Cross-Modal Consistency and Knowledge Distillation. *IEEE Transactions on Affective Computing* (2022).
- [13] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. (2016).
- [14] Dougal Maclaurin, David Duvenaud, and Ryan Adams. 2015. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*. PMLR, 2113–2122.
- [15] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* 14, 3 (2022), 2276–2289.
- [16] Konstantinos Makantasis. 2021. Affranknet+: ranking affect using privileged information. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–8.
- [17] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2021. The Pixels and Sounds of Emotion: General-Purpose Representations of Arousal in Games. *IEEE Trans. on Affective Computing* (2021).
- [18] Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2021. Privileged Information for Modeling Affect In The Wild. In *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*.
- [19] Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, and Georgios N Yannakakis. 2022. The Invariant Ground Truth of Affect. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–8.
- [20] Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, and Georgios N Yannakakis. 2023. From the lab to the wild: Affect modeling via privileged information. *IEEE Transactions on Affective Computing* (2023).
- [21] David Melhart, Antonios Liapis, and Georgios N. Yannakakis. 2021. The Affect Game Annotation (AGAIN) Dataset. *arXiv preprint arXiv:2104.02643* (2021). arXiv:2104.02643 [cs.LG]
- [22] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2021. Towards general models of player experience: A study within genres. In *2021 IEEE Conference on Games (CoG)*. IEEE, 01–08.
- [23] Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2022. RankNEAT: outperforming stochastic gradient search in preference learning tasks. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 1084–1092.
- [24] Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2022. Supervised contrastive learning for affect modelling. In *Proceedings of the International Conference on Multimodal Interaction*. 531–539.
- [25] Kosmas Pinitas, Konstantinos Makantasis, and Georgios N Yannakakis. 2025. Across-game engagement modelling via few-shot learning. In *European Conference on Computer Vision*. Springer, 390–406.
- [26] Kosmas Pinitas, Nemanja Rasajski, Matthew Barthet, Maria Kaselimi, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. [n.d.]. Varying the context to advance affect modelling: A study on game engagement prediction.
- [27] Kosmas Pinitas, Nemanja Rasajski, Konstantinos Makantasis, and Georgios N Yannakakis. 2024. Silhouette Distance Loss for Learning Few-Shot Contrastive Representations. *Proceedings of Machine Learning Research* 1 (2024), 18.
- [28] Kosmas Pinitas, David Renaudie, Mike Thomsen, Matthew Barthet, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2023. Predicting Player Engagement in Tom Clancy's The Division 2: A Multimodal Approach via Pixels and Gamepad Actions. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 488–497.
- [29] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. 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*.
- [30] Roberto Sánchez-Reolid, Francisco López de la Rosa, Daniel Sánchez-Reolid, María T López, and Antonio Fernández-Caballero. 2022. Machine learning techniques for arousal classification from electrodermal activity: A systematic review. *Sensors* 22, 22 (2022), 8886.
- [31] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. 2013. Learning to rank using privileged information. In *Proc. of the IEEE Int. Conf. on computer vision*. 825–832.
- [32] Nicholas E Souter, Ariyana Reddy, Jake Walker, Julián Marino Dávolos, and Elizabeth Jefferies. 2023. How do valence and meaning interact? The contribution of semantic control. *Journal of Neuropsychology* 17, 3 (2023), 521–539.
- [33] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. 2020. Dynamic micro-expression recognition using knowledge distillation. *IEEE Transactions on Affective Computing* 13, 2 (2020), 1037–1043.
- [34] Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. 2019. Modeling user context for valence prediction from narratives. *arXiv preprint arXiv:1905.05701* (2019).
- [35] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 3, 1 (Jan. 2021), 42–50. doi:10.1038/s42256-020-00280-0
- [36] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalios A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. 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)*. 5200–5204. doi:10.1109/ICASSP.2016.7472669
- [37] Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. 2021. End-to-end multimodal affect recognition in real-world environments. *Information Fusion* 68 (April 2021), 46–53. doi:10.1016/j.inffus.2020.10.011
- [38] Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research* 16, 1 (2015), 2023–2049.
- [39] Vladimir Vapnik and Rauf Izmailov. 2017. Knowledge transfer in SVM and neural networks. *Annals of Mathematics and Artificial Intelligence* 81, 1 (2017), 3–19.
- [40] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22, 5–6 (2009), 544–557.
- [41] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.
- [42] Matthias J Wieser and Tobias Brosch. 2012. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in psychology* 3 (2012), 471.
- [43] Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing* (2023).
- [44] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing* 12, 1 (2018), 16–35.
- [45] Hyunjin Yoon, Kiyoung Yang, and Cyrus Shahabi. 2005. Feature subset selection and feature ranking for multivariate time series. *IEEE transactions on knowledge and data engineering* 17, 9 (2005), 1186–1198.
- [46] Guangyi Zhang and Ali Etamad. 2021. Distilling EEG representations via capsules for affective computing. *arXiv preprint arXiv:2105.00104* (2021).