frontiers in Physiology

ORIGINAL RESEARCH
published: 01 September 2021
doi: 10.3389/fphys.2021.720464

# Multi-Modal Pain Intensity Assessment Based on Physiological Signals: A Deep Learning Perspective

*Patrick Thiam* [1,2]*, *Heinke Hihn* [2], *Daniel A. Braun* [2†], *Hans A. Kestler* [1†] *and Friedhelm Schwenker* [2†]

[1] *Institute of Medical Systems Biology, Ulm University, Ulm, Germany,* [2] *Institute of Neural Information Processing, Ulm University, Ulm, Germany*

OPEN ACCESS

***Edited by:***
*Rajesh Kumar Tripathy,*
*Birla Institute of Technology and Science, India*

***Reviewed by:***
*Rishi Raj Sharma,*
*Defence Institute of Advanced Technology (DIAT), India*
*Manish Sharma,*
*Institute of Infrastructure, Technology, Research and Management, India*

****Correspondence:***
*Patrick Thiam*
*patrick.thiam@uni-ulm.de*

*† These authors have contributed equally to this work and share senior authorship*

***Specialty section:***
*This article was submitted to Computational Physiology and Medicine,*
*a section of the journal Frontiers in Physiology*

***Received:*** *04 June 2021*
***Accepted:*** *30 July 2021*
***Published:*** *01 September 2021*

***Citation:***
*Thiam P, Hihn H, Braun DA, Kestler HA and Schwenker F (2021) Multi-Modal Pain Intensity Assessment Based on Physiological Signals: A Deep Learning Perspective. Front. Physiol. 12:720464. doi: 10.3389/fphys.2021.720464*

Traditional pain assessment approaches ranging from self-reporting methods, to observational scales, rely on the ability of an individual to accurately assess and successfully report observed or experienced pain episodes. Automatic pain assessment tools are therefore more than desirable in cases where this specific ability is negatively affected by various psycho-physiological dispositions, as well as distinct physical traits such as in the case of professional athletes, who usually have a higher pain tolerance as regular individuals. Hence, several approaches have been proposed during the past decades for the implementation of an autonomous and effective pain assessment system. These approaches range from more conventional supervised and semi-supervised learning techniques applied on a set of carefully hand-designed feature representations, to deep neural networks applied on preprocessed signals. Some of the most prominent advantages of deep neural networks are the ability to automatically learn relevant features, as well as the inherent adaptability of trained deep neural networks to related inference tasks. Yet, some significant drawbacks such as requiring large amounts of data to train deep models and over-fitting remain. Both of these problems are especially relevant in pain intensity assessment, where labeled data is scarce and generalization is of utmost importance. In the following work we address these shortcomings by introducing several novel multi-modal deep learning approaches (characterized by specific supervised, as well as self-supervised learning techniques) for the assessment of pain intensity based on measurable bio-physiological data. While the proposed supervised deep learning approach is able to attain state-of-the-art inference performances, our self-supervised approach is able to significantly improve the data efficiency of the proposed architecture by automatically generating physiological data and simultaneously performing a fine-tuning of the architecture, which has been previously trained on a significantly smaller amount of data.

**Keywords: physiological signals, signal processing, deep neural networks, information fusion, pain intensity assessment**

# 1. INTRODUCTION

The area of research specific to the development of autonomous and objective pain assessment and management systems has been attracting a lot of interest from both medical and engineering research communities lately (Argüello Prada, 2020; Eccleston et al., 2020; Walter et al., 2020). This is due to the fact that an automatic and effective pain assessment system is more than desirable in the context of telemedicine and remote patient monitoring (Schobel et al., 2021), as well as in cases where an individual is unable to accurately assess and successfully report some currently experienced or observed pain episode. The inability to properly assess and effectively report specific pain episodes can be caused by various factors ranging from psychological or cognitive impairments, to physical and cultural predispositions. In such cases, the reliance on self-reporting tools such as the Visual Analogue Scale (VAS) (Hawker et al., 2011) or the Numerical Rating Scale (NRA) (Eckard et al., 2016) would potentially lead to some unsuitable and inadequate pain relief therapy. Meanwhile, suitable information stemming from an autonomous and objective pain assessment system based on measurable behavioral, anatomical and physiological parameters could provide some additional and relevant insight regarding the underlying pain episode, therefore helping to significantly improve both pain assessment and management.

In concordance with the aforementioned increasing interest, as well as technological advances in such areas as sensor systems and data persistence (which enables researchers to proceed with the recording of a diverse set of measurable autonomic parameters using a plethora of advanced sensor systems and wearables), a gradually growing amount of approaches are being proposed for the development of automatic pain assessment systems. Most of these approaches consist of various machine learning methods built upon different types of collected audiovisual and bio-physiological data, that are optimized and subsequently applied in both clinical and experimental settings. Depending on the amount and diversity of sensors used during the data collection phase, several signals have been assessed and evaluated in various settings for the development of pain assessment systems. Some of the most prominently used signals constitute of the audio signal (e.g., paralinguistic vocalizations) (Tsai et al., 2016, 2017; Thiam et al., 2017; Thiam and Schwenker, 2019), the video signal (e.g., facial expressions) (Rodriguez et al., 2017; Werner et al., 2017; Tavakolian and Hadid, 2019; Thiam et al., 2020b), specific bio-physiological signals such as the Electrodermal Activity (EDA), the Electrocardiogram (ECG), the Electromyography (EMG), or the Respiration (RSP) signal (Walter et al., 2014; Campbell et al., 2019; Thiam et al., 2019a), and also bodily expression signals (Dickey et al., 2002; Olugbade et al., 2019; Uddin and Canavan, 2020).

According to the variety of data collected, different types of machine learning approaches have also been proposed and assessed to perform a specific and effective pain assessment task. The proposed approaches range from uni-modal techniques which rely on a single modality (or channel) to perform the underlying inference task, to multi-modal techniques which rely on a set of multiple and diverse modalities to perform the underlying pain assessment task. Typical uni-modal approaches consist of extracting relevant information in the form of a specific feature representation from the underlying modality and subsequently using the feature representation to perform the optimization of a specific inference model (Sharma et al., 2019, 2020). Multi-modal approaches on the other hand, are designed to perform an aggregation of a set of information stemming from multiple and heterogeneous modalities by applying a specific information fusion technique, in order to improve both the performance as well as the robustness of an inference system. Rather than relying on a single channel, an effective and smart combination of complementary information stemming from multiple channels mitigates the drawbacks specific to each single channel, while improving the generalization ability of the optimized inference system in comparison to one based on a single modality (Kächele et al., 2016; Bellmann et al., 2018; Thiam et al., 2018).

In the following work, a multi-modal information aggregation approach based on Deep Denoising Convolutional Auto-Encoders (DDCAEs) is proposed for the assessment of pain intensities based on bio-physiological signals, and subsequently evaluated in terms of classification, regression and data efficiency performances. The proposed approach is characterized by a concurrent and autonomous optimization of the feature representations specific to the involved channels, as well as the simultaneous optimization of a feed-forward neural network performing the underlying inference task. The contribution of the current work is four-fold: first of all, a multi-modal DDCAE architecture (originally proposed in Thiam et al., 2020a) is proposed and described, for the assessment of different levels of pain elicitation based on a set of diverse bio-physiological modalities. Secondly, a gating layer is also proposed and used in combination with the multi-modal DDCAE architecture in order to perform the aggregation of the information stemming from the underlying modalities before being subsequently used to perform a specific inference task. The resulting architecture is consistently evaluated and further extended with an attention mechanism in order to significantly improve the performance of the designed inference system. Next, the resulting model is further extended by introducing a novel Self-Supervised Learning approach to improve the data efficiency of the designed deep architecture. Self-Supervised Learning (SSL) (Jing and Tian, 2020; Jaiswal et al., 2021) is a form of representation learning (Bengio et al., 2013), where the aim is to learn a meaningful representation to improve a final supervised learning task. Our SSL method is based on an information-theoretic approach which enables Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014) to learn a meaningful and compressed data representation. This representation is then used to generate new data which in turn is utilized to perform a final fine-tuning step. To further improve the generalization ability of the deep model, we apply an information-processing constraint on the Auto-Encoders, the fusion gate, and on the classifier. A consistent benchmark of the designed approaches is provided based on both the BioVid Heat Pain Database (Walter et al., 2013) and the SenseEmotion Database (Velana et al., 2017), where we were able to show that our SSL approach produces results that are only marginally lower compared to data augmentation while only using a fraction of the data.

The remainder of the present work is structured as follows. Section 2 provides an overview of some related work involving multi-modal approaches for pain assessment based on bio-physiological signals. This includes conventional as well as deep learning approaches. The proposed approaches, as well as the data used for the assessment of the proposed approaches are described in section 3. The results specific to the performed experiments are depicted and described in section 4. Finally, a discussion of the achieved results, as well as a description of potential future works is provided in section 5, and the work is concluded with an outlook in section 6.

## 2. RELATED WORK

Multi-modal information fusion approaches are designed with the primary goal of significantly improving the performance of an inference model by effectively combining complementary information stemming from a set of diverse modalities (Kittler and Roli, 2000; Kuncheva, 2004; Palm and Schwenker, 2009; Roli, 2009). Conventional information fusion approaches therefore rely on a set of carefully designed hand-crafted features (extracted individually from each input channel) in combination with an information aggregation approach in order to perform the underlying inference task. Hence, the overall performance of the resulting inference model depends on both the relevance of the extracted feature representations (with regards to the underlying inference task), as well as the ability of the designed aggregation approach to effectively combine the information stemming from the resulting heterogeneous set of hand-crafted features. Some of the most prominently used fusion methods consist of early fusion and late fusion approaches.

Early fusion consists of concatenating the extracted feature representations specific to the underlying modalities into a single and high dimensional feature representation, which is subsequently fed into a classification (or regression) model in order to perform the corresponding inference task. The authors in Walter et al. (2014) extract various features from each input channel [EMG, ECG, Skin Conductance Level (SCL)] and perform the classification of several levels of heat-induced pain intensity using early fusion in combination with a Support Vector Machine (SVM) classification model (Abe, 2010). Similarly, the authors in Chu et al. (2014) extract features from similar modalities and use a combination of early fusion and Linear Discriminant Analysis (LDA) (Fisher, 1936) to perform the classification of several levels of electrical pain stimulation. In Chu et al. (2017), the authors also perform an early fusion of a set of features extracted individually from each modality including SCL, ECG, Blood Volume Pulse (BVP), and subsequently selected using genetic algorithms. The classification is subsequently performed using either a SVM, a k-Nearest Neighbor (k-NN) algorithm or a LDA model. The authors in Werner et al. (2014) and Kächele et al. (2015) extract various features from each input channel and perform the classification of several levels of heat-induced pain intensity using early fusion in combination with a Random Forest (RF) classification model (Breiman, 2001). In Ricken et al. (2020), the authors use the same approach in order to perform the classification of different levels of thermal and electrical pain stimuli, based on a similar set of modalities.

Late fusion on the other hand, consists of the combination at a higher level of aggregation of the outputs of a diverse set of inference models trained on various feature representations. In Kessler et al. (2017), the authors designed and assessed an hierarchical fusion architecture consisting of an Artificial Neural Network (ANN) and the Moore-Penrose Pseudoinverse aggregation rule (Schwenker et al., 2006) for the aggregation of several base classifiers' outputs (consisting of RF models), in order to perform the classification of several levels of artificially induced pain elicitation based on several bio-physiological signals including remote Photoplethysmography (rPPG), ECG, and RSP. In Bellmann et al. (2020), the authors propose a dominant channel fusion approach consisting of first identifying the most relevant input channel and using a subsequent combination of the identified most relevant channel and the remaining ones to create an ensemble of classifiers. The final output of the resulting ensemble is computed by applying an average (Mean) aggregation rule. The approach is assessed on several data sets comprising bio-physiological modalities such as EMG, ECG, EDA, and RSP. In Bellmann et al. (2021), a novel late fusion approach consisting of a combination of mixture of experts and stacked generalization approaches is proposed and assessed on different data sets involving the bio-physiological modalities EMG, ECG, and EDA. The authors in Kächele et al. (2016), Thiam and Schwenker (2017), and Werner et al. (2019) use a combination of RF classification models (trained individually on various feature representations), and a Moore-Penrose Pseudoinverse aggregation approach in order to perform the underlying pain related classification tasks. In Lim et al. (2019), the authors propose a bagged ensemble of Deep Belief Networks (DBNs) (Lopes and Ribeiro, 2015) for the assessment of patient's pain level during surgery, using photoplethysmography (PPG). The ensemble of bagged DBNs is also trained on a set of handcrafted features.

Meanwhile, the processes involved in the manual engineering of feature representations and the selection of relevant features for a specific modality are complex and time consuming. Some specific expert knowledge in the area of application is needed in order to ensure that the resulting and final feature representation is relevant for the task at hand. Moreover, since each single feature representation is specific to the corresponding channel and generated independently from the other task-related modalities, finding a suitable information aggregation approach, that effectively combines the complementary information stemming from the channels, can be very tedious. Thus, a growing amount of work has been investigating the application of deep learning approaches with the goal of enabling a system to autonomously learn not only suitable feature representations, but also effective information aggregation parameters, directly from the corresponding and preprocessed raw input signals. The authors in Thiam et al. (2019a) propose a deep neural network for the classification of different levels of nociceptive pain based on ECG, EMG, and EDA signals, characterized by a weighted aggregation layer performing the combination of the outputs of modality specific Convolutional Neural Networks (CNNs) (LeCun et al., 2015). The whole architecture is trained in an end-to-end manner and was able to attain state-of-the-art classification performances on the BioVid Heat Pain Database. In Thiam et al. (2020a), the authors perform a benchmarking of different types of multi-modal DDCAE architectures, each model characterized by a specific joint representation learned simultaneously from various input modalities. In Subramaniam and Dass (2021), the authors propose a hybrid deep learning network consisting of shallow CNNs that extract information from the raw input signals, and the resulting feature representations are subsequently fed to a Long Short-Term Memory (LSTM) recurrent neural network (Hochreiter and Schmidhuber, 1997) that performs the aggregation of the extracted information followed by the classification of different levels of pain elicitation. The approach is also evaluated on The BioVid Heat Pain Database, with the ECG and EDA modalities as input signals.

Self-Supervised Learning has seen some recent attention in the machine learning community. Most algorithms and approaches fall into the category of image classification and generation (Tung et al., 2017) and language modeling (Lan et al., 2020), but the principle is general enough to be applied to a variety of learning problems (Baevski et al., 2020; Ravanelli et al., 2020; Sekar et al., 2020). Generative models [e.g., Generative Adverserial Nets (Goodfellow et al., 2014), Variational Auto-Encoders (Kingma and Welling, 2014), and Boltzmann Machines (Salakhutdinov and Hinton, 2009)] have been recently successfully applied to learn feature representations (Pathak et al., 2016; Donahue et al., 2017; Zhang et al., 2017), as generative models are a natural example of SSL models as the main goal is to find a representation that produces realistic data, e.g., photo-realistic images. However, there has been only little prior work that utilizes SSL in the context of automatic pain assessment and related fields. The work of Tavakolian et al. (2020) proposes a SSL approach to facial recognition for automatic pain assessment. They implement a novel similarity metric and train a Siamese Network on video streams to optimize the metric, distill the network, and then fine-tune the network on pain assessment training data. The authors of Das et al. (2021) introduce an explainable Self-Supervised Representation Learning paradigm to learn temporal facial patterns. They apply their method to predict speech behavior from stuttering adults. Both methods we discussed are specific to their application, whereas our SSL method is general enough to be applied to any learning problem. At the time of publication of this study there were no further SSL applications to automatic pain assessment to the best of our knowledge.

The information-theoretic SSL principle we propose is based on two main ideas: (i) variational Auto-Encoders with adaptive priors and (ii) a encoder-decoder hierarchy. Adaptive VAEs have been first introduced by Hihn et al. (2018) as a way to learn a data dependent prior that can be use to generate new samples efficiently. Although evaluated only on small scale data it has shown promising results, as the introduced information processing constraint enforces abstract latent representations retaining only information that is useful for the task at hand. The idea of using information processing constraints has been investigated extensively in the reinforcement learning community (Houthooft et al., 2016; Galashov et al., 2019; Grau-Moya et al., 2019; Hihn et al., 2019; Leibfried et al., 2019). We extend this idea to the supervised learning setting and combine it further with a encoder-decoder structure used for classification. In Peng et al. (2017), the authors introduce an information-theoretic formalization of a encoder-decoder hierarchy that is based on bounded rationality (Genewein et al., 2015). The authors show how such policies can be learned in an on-line manner. They evaluate their method in a reinforcement learning setting using a simulated humanoid robot platform and show that it is able to learn a meaningful representation of the environment.

The current work aims at improving the performance of a pain assessment model by enabling a specific ANN to perform autonomously and simultaneously both the extraction of feature representations specific to the input modalities, as well as the aggregation of the information stemming from the generated representations. Moreover, a self-learning approach is proposed as an alternative to conventional data augmentation approaches and assessed in the context of multi-modal pain assessment based on bio-physiological signals.

## 3. MATERIALS AND METHODS

Similarly to conventional Auto-Encoders (AEs) (Hinton and Zemel, 1994; Hinton and Salakhutdinov, 2006), a DDCAE consists of an encoder and a decoder. Both encoder and decoder are Convolutional Neural Networks (CNNs), whereby the encoder maps its input into a low dimensional latent space, while the decoder is optimized to reconstruct the encoder's input, based on the computed latent space representation. Moreover, the encoder's input consists of a corrupted signal (which is generated by adding a noisy signal to the clean and unaltered input signal). The parameters of both encoder and decoder are therefore optimized to reduce the reconstruction error between the decoder's output and the unaltered original input signal. The resulting robust bottleneck representation can be subsequently used to train a specific inference model.

### 3.1. Multi-Modal Deep Denoising Convolutional Auto-Encoder

In the current work, an information fusion architecture based on DDCAEs is proposed to perform the aggregation of information stemming from a set of diverse bio-physiological channels in the context of pain assessment. The proposed architecture, which is depicted in **Figure 1**, consists of learning a single latent representation for each input channel, while simultaneously optimizing a gating layer to generate a single weighted representation based on the generated channel specific latent representations. The generated weighted representation is subsequently used to optimize an inference model performing either the classification or the regression task at hand (the

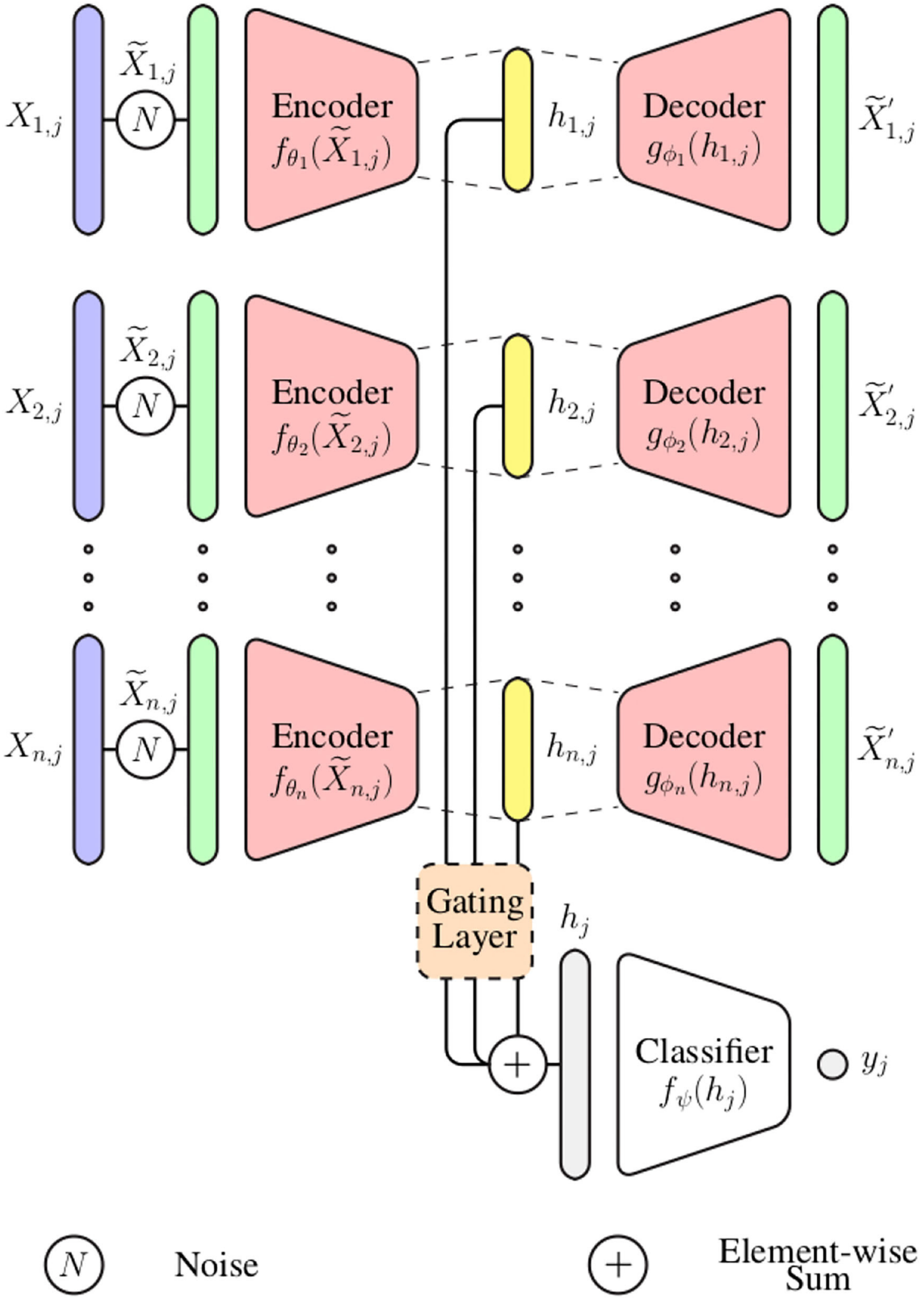
$X_{1,j}$
$\widetilde{X}_{1,j}$
N
Encoder $f_{\theta_1}(\widetilde{X}_{1,j})$
$h_{1,j}$
Decoder $g_{\phi_1}(h_{1,j})$
$\widetilde{X}'_{1,j}$
$X_{2,j}$
$\widetilde{X}_{2,j}$
N
Encoder $f_{\theta_2}(\widetilde{X}_{2,j})$
$h_{2,j}$
Decoder $g_{\phi_2}(h_{2,j})$
$\widetilde{X}'_{2,j}$
$X_{n,j}$
$\widetilde{X}_{n,j}$
N
Encoder $f_{\theta_n}(\widetilde{X}_{n,j})$
$h_{n,j}$
Decoder $g_{\phi_n}(h_{n,j})$
$\widetilde{X}'_{n,j}$
Gating Layer
$h_j$
+
Classifier $f_{\psi}(h_j)$
$y_j$
N Noise
+ Element-wise Sum

**FIGURE 1 |** Multi-modal deep denoising convolutional auto-encoder (DDCAE) architecture.

inference model in this case is a feed-forward neural network). The whole architecture is trained in an end-to-end manner.

In the following paragraph, the parameter $i$ points at the *ith* input channel with $i \in \mathbb{N}$ and $n \in \mathbb{N}$ depicts the total number of input channels ($1 \leq i \leq n$). The parameter $j$ points at the *jth* training sample and $\mathcal{N} \in \mathbb{N}$ depicts the total number of training samples ($1 \leq j \leq \mathcal{N}$). Therefore, the training set specific to the *ith* input channel can be represented as follows: $\{X_{i,j} \in \mathbb{R}^{1\times m}\}_{j=1}^{\mathcal{N}}$ (the parameter $m \in \mathbb{N}$ depicts the dimensionality of the training samples, since each of them consists of a 1-dimensional bio-physiological signal).

For each channel $i \in \mathbb{N}$, a set of noisy input signals $\{\widetilde{X}_{i,j} \in \mathbb{R}^{1\times m}\}_{j=1}^{\mathcal{N}}$ is first generated by altering the original signals $\{X_{i,j} \in \mathbb{R}^{1\times m}\}_{j=1}^{\mathcal{N}}$. Each noisy signal is subsequently fed into the corresponding encoder $f_{\theta_i}$ in order to generate the latent representation $h_{i,j}$:

$$h_{i,j} = f_{\theta_i}(\widetilde{X}_{i,j}) \tag{1}$$

with $\theta_i$ corresponding to the set of trainable parameters of the encoder specific to the *ith* channel. The generated latent representation is further fed into the decoder $g_{\phi_i}$, which generates an output $\widetilde{X}_{i,j}{}'$:

$$\widetilde{X}_{i,j}{}' = g_{\phi_i}(h_{i,j}) \tag{2}$$

Subsequently, a gating layer, which is depicted in **Figure 2**, is used to generate a single weighted representation based on the generated modality specific latent representations $h_{i,j} \in \mathbb{R}^{d_i}$. This approach requires that all latent representations have the same dimensionality: $\forall i \in \{1, 2, \cdots, n\}$, $d_i = \eta \in \mathbb{N}$. Each latent representation $h_{i,j}$ is first normalized by going through a layer with an hyperbolic tangent activation function (*tanh*):

$$\widehat{h}_{i,j} = tanh(\widehat{W}_i h_{i,j} + \widehat{b}_i) \tag{3}$$

where the trainable parameters of the normalization layer consist of $\widehat{W}_i \in \mathbb{R}^{\eta\times\eta}$ and $\widehat{b}_i \in \mathbb{R}^{\eta}$.

The resulting normalized outputs are subsequently concatenated into a single vector $\widehat{h_j} = \left[\widehat{h}_{1,j}, \cdots, \widehat{h}_{n,j}\right]$ ($\widehat{h_j} \in [-1, 1]^{n\cdot\eta}$) and fed into a layer with a softmax activation function, in order to generate the weights specific to each specific feature:

$$w_k = \frac{exp(W_k\widehat{h_j} + b_k)}{\sum_{l=1}^{n\cdot\eta} exp(W_l\widehat{h_j} + b_l)} \tag{4}$$

where the trainable parameters specific to the softmax layer consist of $W_k \in \mathbb{R}^{n\cdot\eta}$ and $b_k \in \mathbb{R}$, with $1 \leq k \leq n\cdot\eta$. The final weighted representation is subsequently generated through a weighted sum of all channel specific latent representation ($h_{i,j}$), using the computed weights ($w = \{w_k\}_{k=1}^{n\cdot\eta}$):

$$h_j = \bigoplus_{i=1}^{n} \widehat{w_i} \odot h_{i,j} \tag{5}$$

where $\widehat{w_i} \in \mathbb{R}^{\eta}$ and $\widehat{w_i} = \left\{w_{(i-1)\cdot\eta+1}, w_{(i-1)\cdot\eta+2}, \cdots, w_{(i-1)\cdot\eta+\eta}\right\}$. Also, $\odot$ denotes the element-wise product, while $\oplus$ denotes the element-wise sum. $h_j \in \mathbb{R}^{\eta}$ is the resulting weighted representation, which is further fed into an inference model $f_\Psi$ to perform either a classification or regression task [$y_j = f_\Psi(h_j)$].

The parameters of the DDCAEs are optimized to minimize the reconstruction error between each decoder's output $\widetilde{X}_{i,j}{}'$ and the original unaltered signal $X_{i,j}$. In the current work, the reconstruction error consists of the mean squared error function:

$$\mathcal{E}_i = \frac{1}{\mathcal{N}}\sum_{j=1}^{\mathcal{N}} \left\| X_{i,j} - \widetilde{X}_{i,j}{}' \right\|_2^2 + \lambda \left\| W_i \right\|_2^2 \tag{6}$$

where $\lambda \left\| W_i \right\|_2^2$ represents a regularization term, with $W_i = \{\theta_i, \phi_i\}$ representing the set of all trainable parameters of the DDCAE specific to the *ith* modality. The parameters of the inference model are optimized accordingly to the task at hand. In the case of a classification task, the corresponding loss function is the categorical cross-entropy loss:

$$\mathcal{L}_{f_\Psi} = -\sum_{cl=1}^{c} y_{cl} log(\widehat{y_{cl}}) \tag{7}$$

where $y_{cl}$ is the ground-truth value of the $cl^{th}$ class and $\widehat{y_{cl}}$ is the corresponding classification output value ($c \in \mathbb{N}$ corresponds to the total number of classes). In the case of a regression task, the corresponding loss function is the mean squared error function:

$$\mathcal{L}_{f_\Psi} = \frac{1}{\mathcal{N}}\sum_{j=1}^{\mathcal{N}} \left\| f_\Psi(h_j) - y_j \right\|_2^2 \tag{8}$$

Since the entire architecture is trained in an end-to-end manner, the entirety of the parameters are optimized by minimizing the following objective function:

$$\mathcal{L} = \sum_{i=1}^{n} \alpha_i \mathcal{E}_i + \alpha_\Psi \mathcal{L}_{f_\Psi} \tag{9}$$

where the parameters $\alpha_i$ and $\alpha_\Psi$ are regularization weights assigned to the loss functions specific to each of the models.

## 3.2. Attention Mechanism

Inspired by the dynamics involved in visual perception (Luck and Ford, 1998), artificial attention mechanisms consist of approaches designed in order to perform the assessment and selection of relevant visual cues, accordingly to the underlying visual task. An intelligent selective processing of specific regions of interest has proven to be very effective and able to significantly improve the performance of deep neural networks for visual computing tasks in areas such as visual image captioning (Chen et al., 2017), object detection (Woo et al., 2018), and image segmentation (Zhao et al., 2020). Inspired by such approaches, the previously described DDCAEs (see section 3.1) are extended

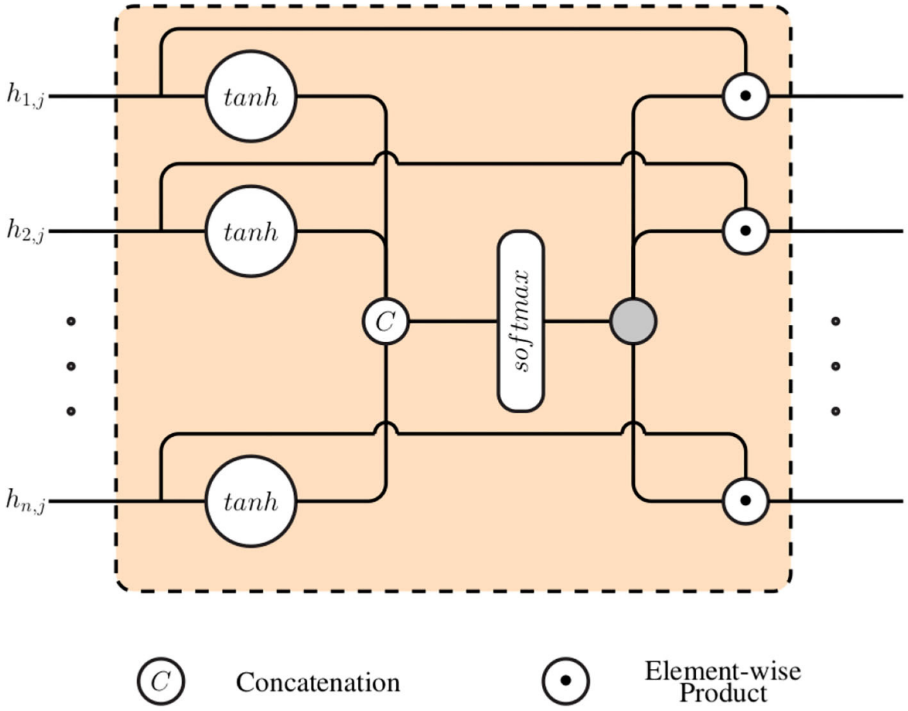

**FIGURE 2 |** Gating layer.

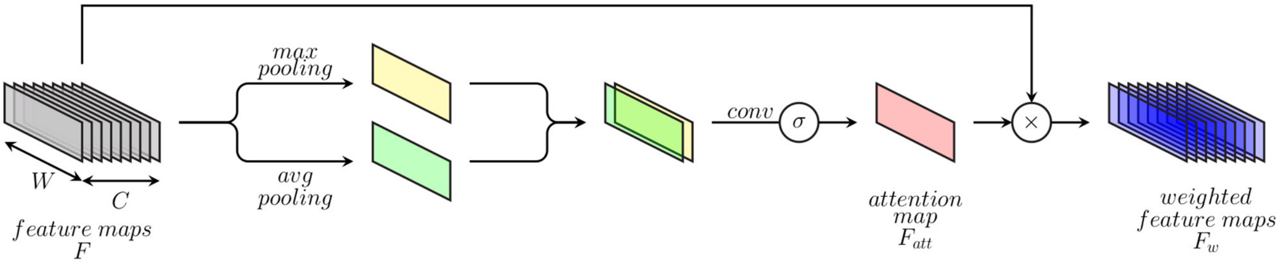

**FIGURE 3 |** Attention mechanism.

with a 1-dimensional attention mechanism, in order to focus on the most relevant feature descriptors accordingly to the task at hand, therefore improving the overall performance of the inference model. The proposed attention mechanism (which is depicted in **Figure 3**) consists of an extension to 1-dimensional feature maps of a spatial attention module for 2-dimensional feature maps originally proposed in Woo et al. (2018). The attention mechanism aims at generating a weighted representation of the optimized feature maps (or feature representations) according to the relevance of each feature descriptor for the specific task at hand.

This is done by generating a specific weighting mask through a channel-wise aggregation of information, followed by a specific convolution operation with a sigmoid activation function, therefore highlighting the relevance of specific regions of interest within the feature maps and weighting these regions accordingly. The attention mechanism consists of first applying average-pooling and max-pooling operations over the channel axis of a set of feature maps stemming from an intermediate convolution layer. The resulting feature maps from both pooling operations are subsequently concatenated and fed into a convolution layer with a sigmoid activation function in order to generate an attention map (see Equation 10). Finally, weighted feature maps are generated by performing an element-wise multiplication of the initial feature maps with the computed attention map (see Equation 11).

More specifically, given a set of feature maps $F \in \mathbb{R}^{1\times W\times C}$ (where $W \in \mathbb{N}_{>0}$ depicts the length of the feature maps and $C \in \mathbb{N}_{>0}$ depicts the number of feature maps) stemming from an intermediate convolution layer, an attention feature map $F_{att} \in$

$\mathbb{R}^{1\times W}$ is generated by using both max-pooling and average-pooling operations applied across the channels of the set of feature maps as follows:

$$F_{att} = \sigma\left(f^{1\times kernel\ size}\left(\left[AvgPool(F), MaxPool(F)\right]\right)\right) \tag{10}$$

where $f^{1\times kernel\ size}$ depicts a 1-dimensional convolution operation with a filter size of $1 \times$ *kernel size* (which is applied on a concatenation of the feature maps resulting from both average- and max-pooling operations) and $\sigma$ depicts the sigmoid function. The weighted feature maps $F_w \in \mathbb{R}^{1\times W\times C}$ are subsequently generated as follows:

$$F_w = F \otimes F_{att} \tag{11}$$

where $\otimes$ depicts an element-wise multiplication. During this operation the values of the attention map are broadcasted along the channels of the feature maps. This specific attention mechanism is integrated into the previously described multi-modal DDCAE architecture (see **Figure 1**) and is applied following each convolution layer in the modality specific encoder architectures.

## 3.3. Self-Supervised Learning of Physiological Signals

In the well-known supervised learning setting, the learner faces a data set $D = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of $N$ pairs of training data $x$ and corresponding labels $y$ and must find a hypothesis that minimizes some loss on the dataset. Collecting labels is expensive and time consuming, especially in the context of pain assessment, as additional experiments have to be conducted. To remedy this, relaxations of supervised learning have been proposed, such as semi-supervised learning (Schwenker and Trentin, 2014), where we only have label information for a subset of data points, and unsupervised learning (Barlow, 1989), where no labels are available. A new approach is Self-Supervised Learning (Jing and Tian, 2020; Jaiswal et al., 2021), which is a form of representation learning (Bengio et al., 2013). Such Self-Supervised Learning algorithms usually consist of a pretext task, that we use to learn a data representation and a downstream task that is the final supervised learning task.

In this study, we follow this line of work and propose an information-theoretic approach to learn such an informative representation. To this end, we propose three additions to the architecture introduced so far. Firstly, we introduce Adaptive Variational Auto-Encoders (AVAE) for self-supervised learning, that *learn* a data dependent prior over their latent representation as opposed to using a fixed prior (Kingma and Welling, 2014). We argue, that this information-processing bottleneck enforces an optimal trade-off between representational capacity and information-processing cost as measured by the Kullback-Leibler divergence ($\mathrm{D_{KL}}$) between the latent posterior $p(h|x)$ and its prior $p(h)$. This idea has been introduced by Hihn et al. (2018) and has shown promising results on low dimensional data. Secondly, we propose to use information processing constraints on the gating layer and the classifier based on a theory of bounded rationality (Ortega et al., 2015). To this end, we follow the work of Hihn and Braun (2020b), where they introduce and motivate such constraints and show their favorable effects on generalization in the meta-learning setting (Hihn and Braun, 2020a). We show that these types of constraints enable efficient representation learning, which is the pretext task. Lastly, we propose to use the learned representation to generate artificial data and use this data to fine-tune the model, which is the downstream task.

### 3.3.1. Adaptive Variational Auto-Encoders

In the following sections we introduce the Adaptive Variational Auto-Encoder and its self-supervised learning application, in order to improve automatic pain assessment while requiring less data points then conventional data augmentation techniques.

Variational Auto-Encoders (VAE) (Kingma and Welling, 2014) are generative models that build on deterministic Auto-Encoder networks. They are best understood as variational Bayesian inference in a latent variable model $p(x|h)$ with a prior distribution $p(h)$, where $x$ represents the observable data, and $h$ the latent variable that explains the data. The goal is to find a set of parameters $\varphi^*$ that maximize the data likelihood $p_\varphi(x) = \int p_\varphi(x|h)p(h)\mathrm{d}h$. We can draw samples from $p_\varphi(x)$ by first sampling $h$ and then draw $x$ from $p_\varphi(x|h)$. As maximum likelihood optimization is intractable due to the integral, we express the likelihood in a different form by defining a variational distribution $q(h|x)$ [also known as Evidence Lower Bound (ELBO)], such that

$$\log p_\varphi(x) \geq \int q(h|x) \log \frac{p_\varphi(x|h)p(h)}{q(h|x)} \mathrm{d}h =: \mathrm{F}(\varphi). \tag{12}$$

Assuming $q(h|x)$ is expressive enough to approximate the true posterior distribution $p_\varphi(h|x)$ well, we can directly maximize the lower bound $\mathrm{F}(\varphi)$ by gradient descent. In VAEs, $q(h|x)$ is the encoder network that generates a latent representation $h$ given and input $x$, and $p(x|h)$ is the decoder that reconstructs $x$ from $h$. We assume all distributions to be isotropic Gaussians.

We extend this approach by removing the fixed prior $p(h)$ and allow for an adaptive prior. To learn a prior that allows for efficient information processing by minimizing the $\mathrm{D_{KL}}$ between $p_\varphi(h|x)$ and $p(h)$, we define $p(h)$ to be the marginal of $p_\varphi(h|x)$ over the inputs $x$:

$$p(h) = \int_{x\in\mathcal{X}} p_\varphi(x|h)\mathrm{d}x. \tag{13}$$

This term is not tractable as the data generating distribution $p(x)$ is unknown. We approximate the true marginal by running an exponential running mean with window length $\tau$ (Hihn and Braun, 2020b; Leibfried and Grau-Moya, 2020):

$$q_{t+1}(h) = (1 - \frac{1}{\tau})q_t(h) + \frac{1}{\tau}p_\varphi(h|x_t), \tag{14}$$

where $q_t(h)$ is the approximated marginal after observing $t$ samples. To find the optimal parameters $\varphi^*$, we optimize the following variational Auto-Encoder objective:

$$\varphi^* = \arg\max_{\varphi} \mathbb{E}_{x\sim p(x),h\sim p_\varphi(h|x)}\left[\log p_\varphi(x|h)\right] - \frac{1}{\beta_1}\,\mathrm{D}_{\mathrm{KL}}\left[p_\varphi(h|x)||q(h)\right], \qquad (15)$$

where $\beta_1$ governs a trade-off between maximizing the log-likelihood and keeping the variational posterior close to the prior $q(h)$. We will refer to $q(h)$ as $p(h)$ to keep notation consistent. There are different interpretations of this approach, e.g., learning with information constraints (Hihn et al., 2018, 2019; Hihn and Braun, 2020b), meta-learning (Hihn and Braun, 2020a), tempered posteriors in variational Bayes applications (Aitchison, 2021), and learning disentangled representations (Higgins et al., 2016). The additional degree of adaptivity introduced allows to learn data dependent priors which can improve the quality of generated samples, as we will show empirically in section 4.3.

#### 3.3.2. Representation Learning

We can interpret the architecture introduced in section 3.1 as an encoder-decoder structure thats maps a high dimensional input signal $x$ to a low dimensional latent representation $h$. Classification is then performed only using a non-linear combination of the low dimensional representation. Training this architecture in an end-to-end fashion will produce values of $h$ that both minimize the reconstruction error (in other words, capture the data well) and the classification error. To further improve this coupling, we propose to impose information-processing constraints on the latent representation, the gating layer, and on the classifier. We argue that such constraints encourage the system to learn representations that discover regularities in the data by discarding all unnecessary information (Hihn and Braun, 2020b). To this end, we formulate an information-theoretic coupling as a two stage hierarchical system (Genewein et al., 2015) with the following objective function:

$$\max_{p(w|h),p(y|w,h)} \mathbb{E}\left[\mathcal{L}(x,y)\right] - \frac{1}{\beta_2}I(W;H) - \frac{1}{\beta_3}I(Y|W;H), \qquad (16)$$

where $\mathcal{L}(x,y)$ is a loss function, $x \in X$ the input, $y \in Y$ the output, $h \in H$ the latent representations produced by the Auto-Encoders, $W$ the weights of the gating layer (see Equation 4), and $I(X;Y)$ is the mutual information between random variables $X$ and $Y$. The hyper-parameters $\beta_2$ and $\beta_3$ are Lagrange multipliers that govern the trade-off between information-processing cost and utility as measured by the loss $\mathcal{L}$. Note that the classifier output $y$ depends on the combined latent variables $h$, as described by Equation (5). We can rewrite Equation (16) into

$$\max_{\theta,\vartheta} \mathbb{E}\left[\mathcal{L}(x,y) - \frac{1}{\beta_2}\log\frac{p_\theta(w|h)}{p(w)} - \frac{1}{\beta_3}\log\frac{p_\vartheta(y|w,h)}{p(y)}\right], \qquad (17)$$

where $h$ is the latent representation, $x$ the input, $y$ the output, $p(w|h)$ is some fusion policy and $p(y|w,h)$ is the output of a decision-maker (e.g., a classifier), and $\theta,\vartheta$ are the parameters. This formulation allows us to perform updates in an on-line manner. As outlined earlier, the optimal priors to find an optimal trade-off are the marginals of the posterior policies $p(w|h)$ and $p(y|w,h)$, which we approximate by a running mean average. Combining VAE and representation learning losses we have the following objective function allowing us to train the system end-to-end (see **Figure 4**):

$$\max_{\varphi,\theta,\vartheta} \mathbb{E}\left[\mathcal{L}(x,y) - \log p_\varphi(x|h) - \frac{1}{\beta_1}\log\frac{p_\varphi(h|x)}{p(h)} - \frac{1}{\beta_2}\log\frac{p_\theta(w|h)}{p(w)} - \frac{1}{\beta_3}\log\frac{p_\vartheta(y|w,h)}{p(y)}\right]. \qquad (18)$$

#### 3.3.3. Self-Supervised Fine Tuning

The method we propose aims to find a representation that allows us to generate informative data samples, without having to go through an expensive data generation and labeling process. The self-supervised fine tuning algorithm we propose consists of a training phase (called the pretext task), followed by a generative and fine tuning phase (also called the downstream task). Firstly, we train the whole system, namely an adaptive variational Auto-Encoder for each modality and the classifier (see **Figure 1**) using the full (but not augmented) dataset (see **Tables 5**, **6** for an account of samples used and generated). The main goal of this phase is for the generative models to learn a latent representation $h$ per modality that is beneficial to the classification task that uses a combination of the latent representations of the Auto-Encoders, i.e., $p(y|w,h)$, where $y$ is the output label, $h$ represents the latent variables, and $w$ represents the weights computed by the gating layer. In this way the learned posterior $p(h|x)$, where $h$ is the latent variable and $x$ the input signal, optimizes both the signal reconstruction architecture and the classification model simultaneously, by optimizing the Auto-Encoder's objective function given by Equation (15). The resulting representation thus captures the structure of the data, as well as the semantic information, making it a suitable candidate for a data generation process. We give an overview of our technique in **Algorithm 1**.

## 3.4. Data Sets

The BioVid Heat Pain Database (Part A) (Walter et al., 2013) is a multi-modal data set consisting of 87 healthy participants subjected to four levels of gradually increasing and individually calibrated thermal pain elicitation $(T_1, T_2, T_3, T_4)$. Several modalities were recorded during the experiments including video streams, EDA, ECG, and EMG signals. Each single level of pain elicitation was randomly elicited a total of 20 times, with each elicitation lasting 4 s (sec), followed by a recovery phase of randomized duration (lasting between 8 and 12 s). During this recovery phase, a baseline temperature $T_0$ of 32°C was applied (see **Figure 5**). The data set specific to each participant consists of a total of $20 \times 5 = 100$ samples, summing up to a database of $87 \times 100 = 8,700$ samples. Each sample is labeled with its corresponding level of thermal pain elicitation $(T_0, T_1, T_2, T_3, T_4)$. The proposed approaches are evaluated uniquely on the physiological signals EMG, ECG, and EDA.

Analogously to the BioVid Heat Pain Database, the SenseEmotion Database (Velana et al., 2017) consists of 45 healthy individuals subjected to 3 levels of individually calibrated

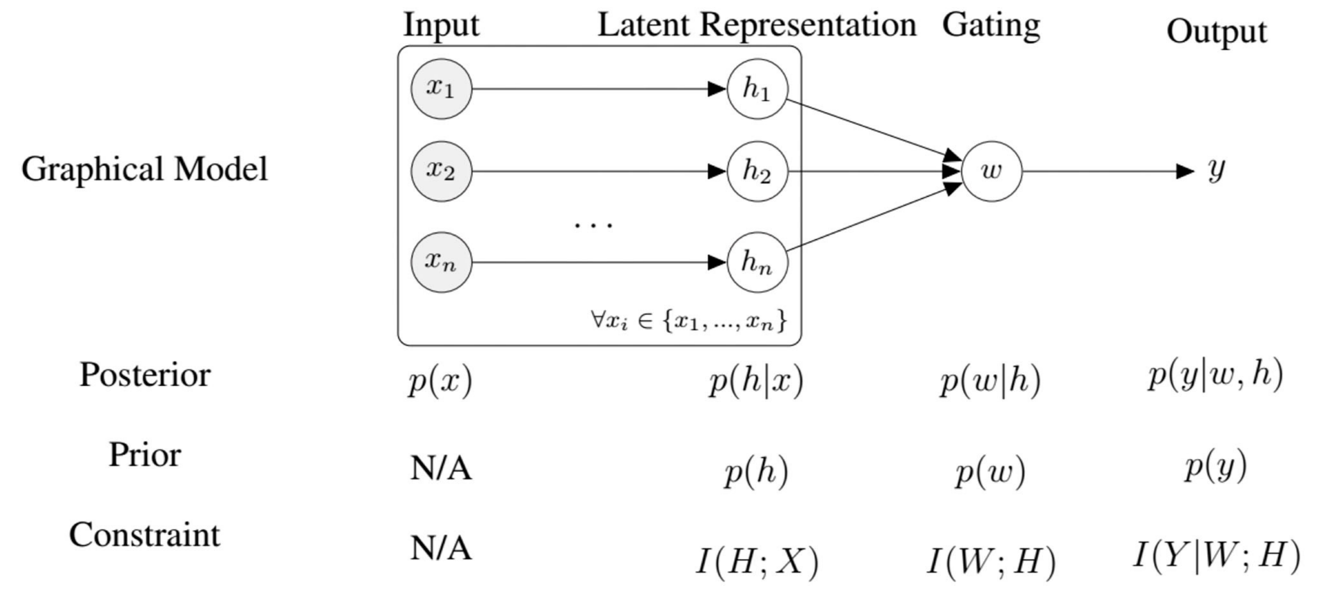

**FIGURE 4 |** An overview of our model for self-supervised fine-tuning and the corresponding information-processing constraints, where $n$ is the number of modalities.

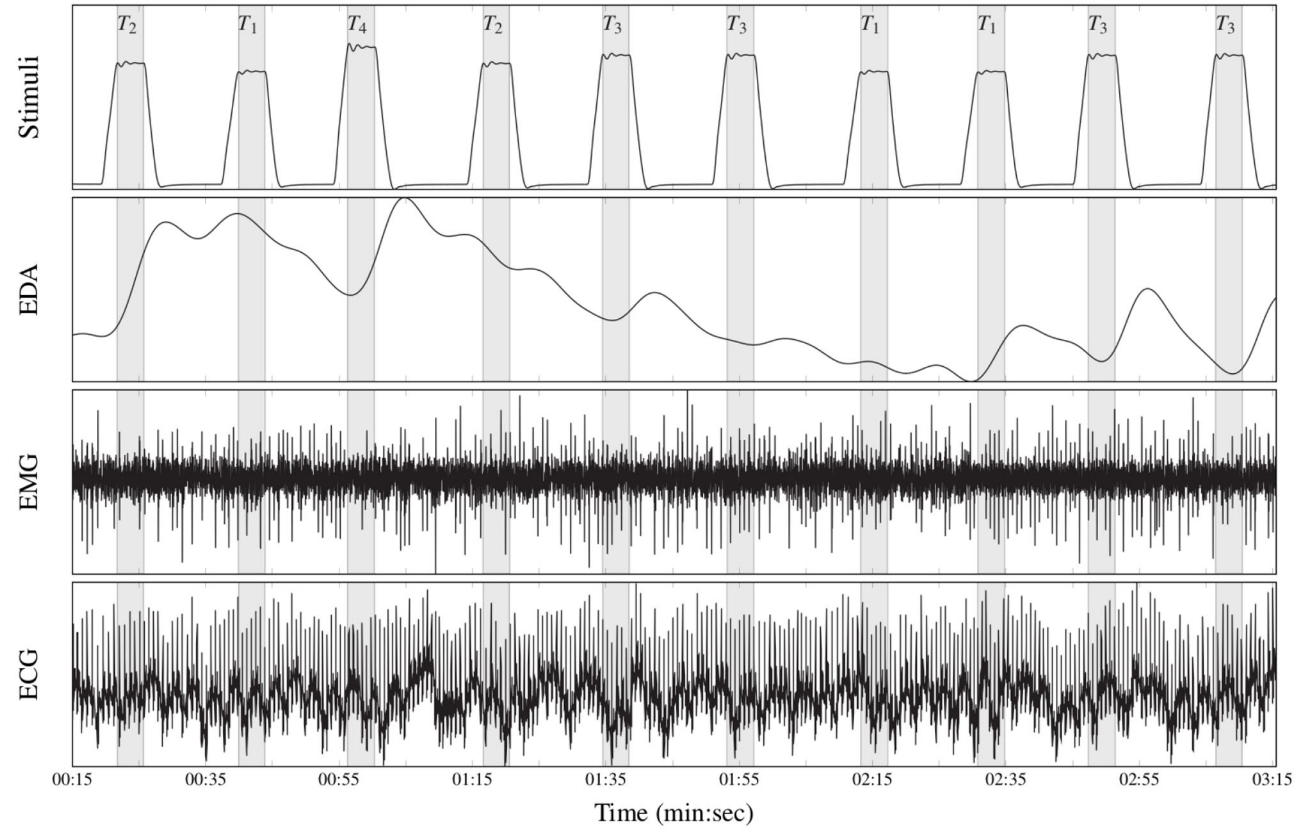

**FIGURE 5 |** Recorded physiological data (BioVid Heat Pain Database, thermal pain elicitation). From top to bottom: stimuli ($T_1$: pain threshold temperature, $T_2$: first intermediate elicitation temperature, $T_3$: second intermediate elicitation temperature, $T_4$: pain tolerance temperature); EDA ($\mu$S); EMG ($\mu$V); ECG ($\mu$V).

**Algorithm 1** We split Self-Supervised Learning into three phases: (i) training, (ii) self-supervised fine-tuning, and (ii) evaluation. $\mathcal{D}_{-j}$ denotes the dataset without data from subject $j$.

```
Input: Dataset $\mathcal{D}$ with data from $N_s$ subjects with $L$ modalities
Hyper-parameters: prior penalty parameters $\beta_1$, $\beta_2$, $\beta_3$, number of samples $K$ to generate, training episodes $N$, fine-tuning training episodes $M$
for $j$ = 1, 2,..., $N_s$ do
    Initialize Auto-Encoders and classifier parameters
    Train Auto-Encoders and classifier for $N$ episodes using subset $\mathcal{D}_{-j}$ with parameters $\beta_1, \beta_2, \beta_3$
    $\mathcal{D}_{\text{self-super}} = \emptyset$
    for $i$ = 1, 2,..., $K$ do
        for $l = 1, 2, ..., L$ do
            Generate latent representation $h_l$ using prior $p_l(h)$
            Reconstruct samples $\tilde{X}_l$ by using the decoders $g_{\phi_l}(h)$
        end for
        Combine all $\tilde{X}_l$ into $\tilde{X}_i$
        Classify sample using $f_\varphi(\mathbf{h})$, where $\mathbf{h}$ is the output of the gating layer, to obtain label $\tilde{y}_i$
        $\mathcal{D}_{\text{self-super}} = \mathcal{D}_{\text{self-super}} \cup (\tilde{X}_i, \tilde{y}_i)$
    end for
    Fine-tune system using only $\mathcal{D}_{\text{self-super}}$ for $M$ episodes with parameters $\beta_1, \beta_2, \beta_3$
    Evaluate data from subject $j$ and collect metrics
end for
return evaluation metrics
```

and gradually increasing thermal pain elicitation ($T_1$, $T_2$, $T_3$) and a baseline level $T_0$ set identically for all participants to 32°C (corresponding to no pain elicitation). The modalities recorded during the performed experiments consist of audio signals, 3 video streams, the trapezius EMG signal, RSP, ECG, and EDA signals. The performed experiments consist of two 40 min sessions, during which the piece of hardware used to perform the thermal pain elicitations was attached to a specific forearm (once on the right forearm and once on the left forearm). The calibration of the temperatures of elicitation as well as the thermal elicitation procedure were carried out identically to the BioVid Heat Pain Database, with the only difference being the total number of stimuli per pain level. Each pain level was randomly elicited a total of 30 times with a pause of about 8–12 s between the elicitations. Due to technical issues during the experiments 5 participants were excluded from the data set because of missing or erroneous data. We therefore evaluate the proposed approaches on a reduced subset consisting of 40 participants and a data set consisting of a total of $40 \times 30 \times 4 \times 2 \approx 9,600$ samples. The assessment of the proposed approaches is performed uniquely on the physiological signals EMG, ECG, EDA, and RSP.

## 3.5. Data Preprocessing

Similar preprocessing operations were applied on the recorded physiological signals of both datasets. First of all, the sampling rate of the recorded signals was reduced to 256 Hz in order to significantly reduce the amount of computational requirements. Next, the amount of noise and artifacts within each signal was reduced by applying specific signal processing techniques. For both datasets, a low-pass Butterworth filter of order 3 with a cut-off frequency of 0.2 Hz was applied on the EDA signals. Concerning the BioVid Heat Pain Database, EMG signals were filtered using a fourth order bandpass Butterworth filter with a frequency range of $[20, 250]$ Hz, while ECG signals were filtered with a third order bandpass Butterworth filter with a frequency range of $[0.1, 250]$ Hz. Subsequently, piecewise detrending of the filtered ECG signals was performed, by subtracting a fifth degree polynomial least-squares fit from the filtered signals (as proposed in Thiam et al., 2019a). Concerning the SenseEmotion Database, the RSP signals were smoothed using a third order low-pass Butterworth filter with a cut-off frequency of 0.8 Hz. Both EMG and ECG signals were preprocessed by applying a third order bandpass Butterworth filter with respective frequency ranges of $[0.05, 25]$ and $[0.1, 25]$ Hz, followed by a similar piecewise detrending as in the case of the BioVid Heat Pain Database. The resulting filtered signals were subsequently segmented, and each segment in combination with its corresponding level of pain elicitation was used to perform the assessment of the proposed approaches.

In the case of the BioVid Heat Pain Database, the assessment is performed on windows of length 4.5 s with a shift of 4 s from the elicitations' onset (see **Figure 6**). Analogously, in the case of the SenseEmotion Database, the assessment of the proposed approaches is performed on widows of length 6.5 s with the same shift of 4 s from the elicitations' onset. Each signal within these specific windows consists of a one-dimensional array of size $m = 4.5 \times 256 = 1,152$ for the BioVid Heat Pain Database, and $m = 6.5 \times 256 = 1,664$ for the SenseEmotion Database. Moreover, since a huge amount of parameters specific to the multi-modal DDCAE architectures has to be optimized, data augmentation was performed by shifting the 4.5 s (6.5 s, respectively) window of segmentation backward and forward in time with small shifts of 250 ms and a maximum total window shift of 1 s in each direction. These shifts were performed, starting from the initial position of the windows (as depicted in **Figure 6**). This procedure was performed uniquely during the training phase of the proposed architectures, resulting in generating nine times the total amount of training samples specific to the initial windows of segmentation. Following the optimization of the multi-modal DDCAE architectures, the evaluation of the trained architectures was performed on the initial windows of 4.5 s (6.5 s, respectively) with a shift of 4 s from the elicitations' onset.

# 4. RESULTS

In the following section, a description of the results relative to the proposed multi-modal DDCAE architecture is provided. An assessment of the performance of the architecture is performed

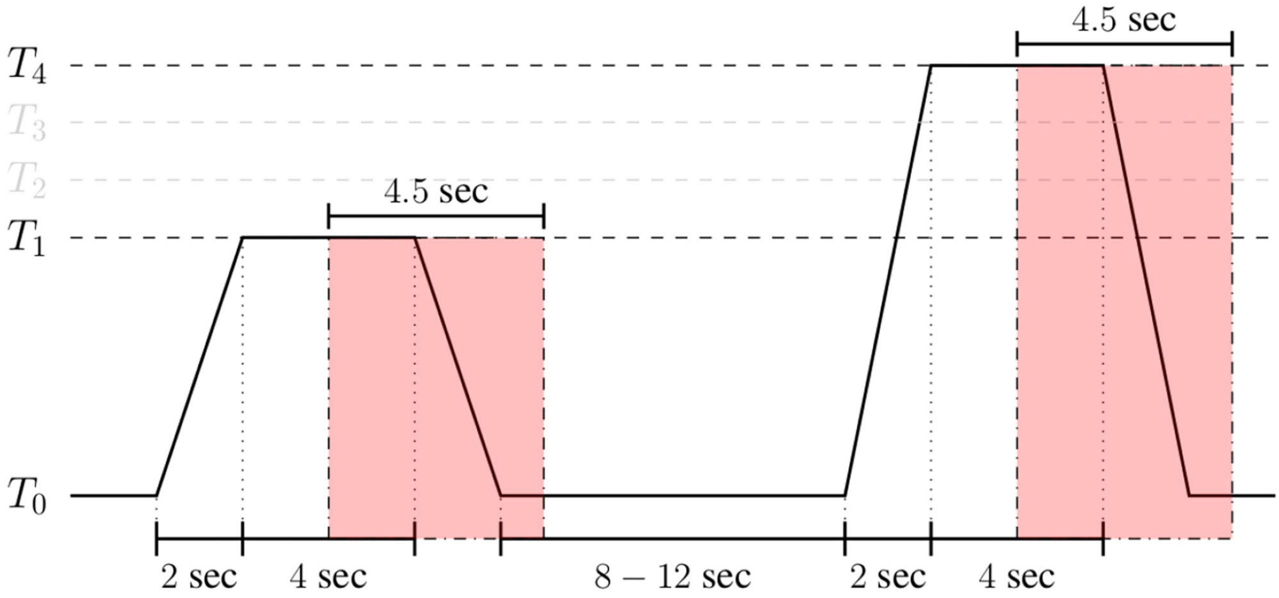

**FIGURE 6 |** Signal segmentation (BioVid Heat Pain Database). Experiments are carried out on windows of length 4.5 s with a temporal shift of 4 s from the elicitations' onset.

and a comparison of the results achieved with and also without the proposed attention mechanism is conducted. Finally, the results specific to the proposed self-learning algorithm as well as a comparison of the achieved results between self-learning approaches and fully supervised learning approaches are described and discussed.

## 4.1. Experimental Settings

In the current work, the multi-model DDCAE architecture consists of one-dimensional convolutional operations. The Exponential Linear Unit (ELU) (Clevert et al., 2016) function defined in Equation (19) (with $\alpha = 1$) is used in both convolutional and fully connected layers as activation function, except for the output layer of both classification and regression models. In the case of a classification model, a softmax activation function is used, while a linear activation function is used in the case of a regression model.

$$elu_{\alpha}(x) = \begin{cases} \alpha\left(exp(x) - 1\right), & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases} \tag{19}$$

Similar encoders' and decoders' architectures are used for each physiological signal, with the only difference being the size of the convolutional kernel for each modality specific DDCAE: in the case of EDA, a fixed convolutional kernel with a size of 3 and a stride of 1 is set empirically. In the case of EMG, ECG and RSP, the size of the convolutional kernel is set to 11 with a stride set also to 1. The dimensionality of the resulting latent representation specific to each modality specific DDCAE is set empirically to $\eta = 256$. The designed architectures are summarized in **Table 1**. The convolutional kernel of the attention mechanism (see Equation 10) is set empirically to 3 with a stride of 1 (*kernel size* = 3).

**TABLE 1 |** DDCAE architecture: the MaxPooling and UpSampling operations are performed with an identical pooling size set to 2 and a stride set to 2.

| **Encoder** | |
|---|---|
| **Layer** | **No. kernels/Units** |
| 2 × Conv1D and MaxPooling | 8 |
| 2 × Conv1D and MaxPooling | 16 |
| 2 × Conv1D and MaxPooling | 32 |
| Flatten | – |
| Fully connected | 256 |
| **Decoder** | |
| **Layer** | **No. kernels/Units** |
| Fully connected | 576 |
| Reshape | – |
| 2 ×Conv1D and UpSampling | 32 |
| 2 ×Conv1D and UpSampling | 16 |
| 2 ×Conv1D and UpSampling | 8 |
| 1 ×Conv1D | 1 |
| **Inference (classification or regression)** | |
| **Layer** | **No. kernels/Units** |
| Fully connected | 128 |
| Dropout | – |
| Fully connected | *c* |

*The Dropout rate is set empirically to 0.25. Concerning the underlying inference task, in case of a classification task c depicts the number of classes, while in the case of a regression task, c = 1.*

All architectures are trained using the Adaptive Moment Estimation (Adam) (Kingma and Ba, 2015) optimization algorithm with a fixed learning rate set empirically to $10^{-5}$. The training process is performed through a total of 100 epoches. The batch size is set to 40 in the case of a 2 *Classes* classification task and 100 in the case of a 5 *Classes* classification task, for the BioVid Heat Pain Database. Concerning the SenseEmotion Database, the batch size is set to 120 for 2 *Classes* classification tasks, and 480 for 4 *Classes* classification tasks. The same batch sizes are used for regression tasks. The regularization parameter in Equation (6) is set as follows: $\lambda = 10^{-3}$. The regularization weights of the objective function defined in Equation (9) are set as follows: $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$ and $\alpha_\Psi = 0.4$, for the BioVid Heat Pain Database; $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.15$ and $\alpha_\Psi = 0.4$, for the SenseEmotion Database. The regularization weight specific to the inference model is set higher than the others in order to focus more on the inference performance of the whole architecture. Moreover, noisy signals are generated by adding some Gaussian noise to the unaltered original signals. The parameters of the distribution specific to the Gaussian noise consist of a standard deviation set to 0.1 and a mean set to 0. The implementation and the evaluation of the proposed approaches were performed with the libraries Tensorflow (Abadi et al., 2016), Keras (Chollet et al., 2015), and Scikit-learn (Pedregosa et al., 2011). The evaluation of the approaches was performed by applying a *Leave One Subject Out* (LOSO) cross-validation evaluation, which means that the data specific to each single participant is used once to evaluate the performance of the trained model and is never seen during the training process, while the data specific to the remaining participants is used to optimize or train the model. This results in a total of 87 experiments in the case of the BioVid Heat Pain Database, and 40 experiments in the case of the SenseEmotion Database. The results specific to each inference task that are depicted in the following sections are therefore averaged across the totality of the performed experiments.

## 4.2. Multi-Modal Deep Denoising Convolutional Auto-Encoder: Results

In the current section, an assessment of the classification performance of the proposed multi-modal DDCAE approach is performed and described. The assessment consists of performing several binary and multi-class classification tasks based on both BioVid Heat Pain Database and SenseEmotion Database. A comparison of the achieved performances of the multi-modal DDCAE approach, respectively without (w/o) and with the proposed attention mechanism, is also provided. The performance measures used to conduct the assessment are described in **Table 2**.

In the case of a binary classification task (e.g., $T_0 vs. T_4$), *true positives* (*tp*) correspond to the number of correctly classified samples of the positive class (e.g., $T_4$), while true negatives corresponds to the number of correctly classified samples of the negative class (e.g., $T_0$). Analogously, *false positives* (*fp*) correspond to the number of incorrectly classified samples of the negative class, while *false negatives* (*fn*) correspond to the number of incorrectly classified samples of the positive class. These values stem from the confusion matrix of an evaluated classification model and are used to define and compute the performance measures.

**TABLE 2 |** Classification performance measures.

| Measure | Binary classification | Multi-class classification |
|---|---|---|
| Accuracy | $\frac{tp+tn}{tp+tn+fp+fn}$ | $\frac{1}{c}\sum_{i=1}^{c}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}$ |
| Precision | $\frac{tp}{tp+fp}$ | $\frac{1}{c}\sum_{i=1}^{c}\frac{tp_i}{tp_i+fp_i}$ |
| Recall | $\frac{tp}{tp+fn}$ | $\frac{1}{c}\sum_{i=1}^{c}\frac{tp_i}{tp_i+fn_i}$ |
| F1-Score | $\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$ | |

*In the case of multi-class classification experiments: $tp_i$ corresponds to true positives, $tn_i$ corresponds to true negatives, $fp_i$ corresponds to false positives and $fn_i$ corresponds to false negatives in the confusion matrix associated with the ith class. Since the data sets used to perform the evaluation of the proposed approaches are balanced, the macro-averaged F1-score is used in the case of multi-class classification.*

First of all, a summary of the results specific to the signal reconstruction performance of the proposed approach in terms of Mean Squared Error (MSE) averaged across the performed LOSO cross-validation evaluation ($\forall i, \ \text{MSE}_i = \frac{1}{\mathcal{T}}\sum_{j=1}^{\mathcal{T}}\left\|X_{i,j} - \widetilde{X}'_{i,j}\right\|_2^2$, with $\mathcal{T} \in \mathbb{N}_{>0}$ being the size of the testing set specific to the $i^{th}$ modality) is provided in **Table 3**. Concerning the BioVid Heat Pain Database, the attention mechanism improves the performance of the DDCAE architectures and in most cases significantly, with regards to both EDA and ECG signals. Concerning the EMG signal, both DDCAE architectures (without and with attention mechanism) perform similarly, with the approach without attention mechanism slightly outperforming the one with the attention mechanism, however not significantly. Concerning the SenseEmotion Database, the DDCAE architectures with the attention mechanism outperform those without attention mechanism in most cases, with regards to the EDA, ECG, and EMG signals. The reconstruction error of the RSP signal is significantly higher than those of the other signals, regardless of the applied approach. A similar reconstruction error performance between both approaches with and without attention mechanism can also be seen across all classification tasks. Overall, the proposed attention mechanism has a positive effect on the reconstruction performance of the multi-modal DDCAE architecture, and helps further reducing the MSE between the output of the model and the original unaltered input signals.

Furthermore, the performance of the jointly trained classification model for each classification task is depicted in **Figure 7** for the BioVid Heat Pain Database and in **Figure 8** for the SenseEmotion Database. Moreover, a summary of the classification results is provided in **Table 4**. Concerning the BioVid Heat Pain Database, the proposed attention mechanism improves the performance of the multi-modal DDCAE across all classification tasks. The performance improvement is also significant in most cases in terms of F1-score. Concerning the SenseEmotion Database, the attention mechanism significantly improves the performance of the proposed approach regarding the binary classification task $T_0 vs. T_3$.

In the case of the binary classification task $T_1 vs. T_3$, the attention mechanism also improves the overall performance

**TABLE 3 |** Signal reconstruction performance [Mean Squared Error (MSE)] comparison in a *Leave One Subject Out* (LOSO) cross-validation evaluation setting: Average MSE in % (standard deviation in %).

**BioVid Heat Pain Database (Part A)**

| Task | $T_0 vs. T_4$ | | $T_1 vs. T_4$ | | $T_0 vs. T_1 vs. T_2 vs. T_3 vs. T_4$ (5 Classes) | |
|---|---|---|---|---|---|---|
| Model | DDCAE | | DDCAE | | DDCAE | |
| | W/o attention | With attention | W/o attention | With attention | W/o attention | With attention |
| EDA | 04.82 (05.26) | **03.61 (03.90)*** | 05.13 (05.47) | **03.79 (04.46)*** | 03.18 (04.14) | **03.14 (04.14)** |
| ECG | 09.23 (07.12) | **08.57 (07.32)*** | 09.31 (07.54) | **07.75 (06.92)*** | 06.25 (05.81) | **05.92 (05.72)*** |
| EMG | **17.38 (32.39)** | 17.40 (32.65) | 17.37 (31.80) | **17.13 (31.63)** | **16.28 (31.34)** | 16.89 (31.50) |

**SenseEmotion Database**

| Task | $T_0 vs. T_3$ | | $T_1 vs. T_3$ | | $T_0 vs. T_1 vs. T_2 vs. T_3$ (4 Classes) | |
|---|---|---|---|---|---|---|
| Model | DDCAE | | DDCAE | | DDCAE | |
| | W/o attention | With attention | W/o attention | With attention | W/o attention | With attention |
| EDA | 03.75 (04.32) | **03.46 (03.88)** | 04.03 (04.73) | **03.86 (04.50)** | 03.81 (04.37) | **03.77 (03.85)** |
| ECG | 05.83 (02.91) | **05.52 (03.13)*** | 05.74 (03.09) | **05.43 (03.32)*** | **05.61 (03.23)** | 05.54 (03.23) |
| EMG | **07.42 (09.05)** | 07.62 (09.41) | 07.32 (08.32) | **07.04 (08.72)*** | 07.53 (09.31) | **07.77 (08.40)** |
| RSP | 34.71 (89.97) | **33.63 (87.89)** | **34.37 (82.40)*** | 35.11 (86.38) | **35.39 (94.68)** | 35.97 (97.17) |

*The best achieved performance is depicted in bold. An asterisk (∗) depicts a significant performance improvement. The significance test is performed using a two-sided Wilcoxon-Signed-Rank test with a significance level of* 5%.

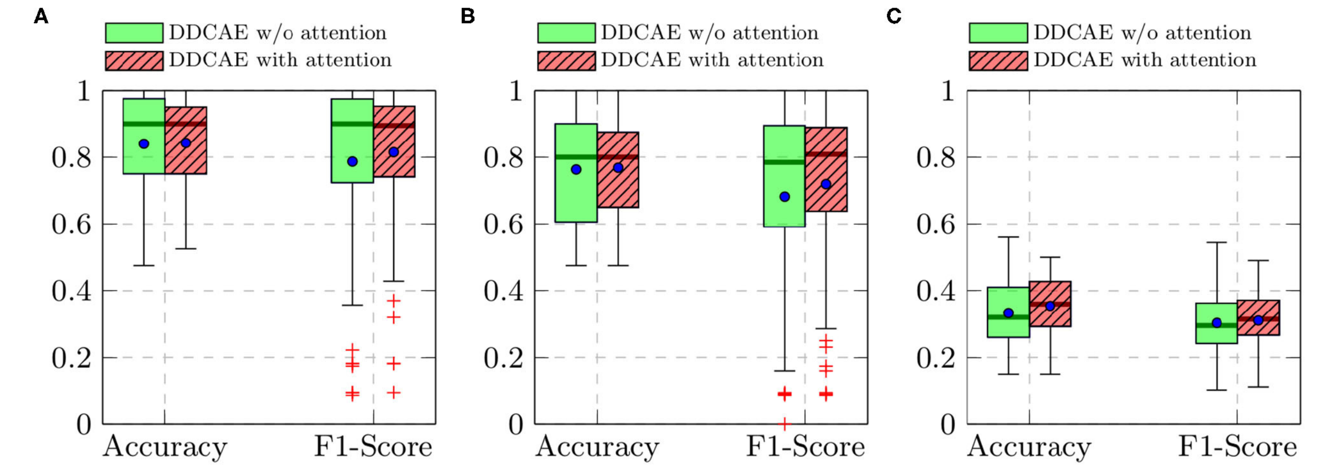

**FIGURE 7 |** BioVid Heat Pain Database (Part A): classification performance comparison in a *Leave One Subject Out* (LOSO) cross-validation evaluation setting. Within each box plot, the mean and the median classification performance are depicted respectively with a dot and a horizontal line. **(A)** $T_0 vs. T_4$, **(B)** $T_1 vs. T_4$, **(C)** 5 classes.

of the proposed architecture. In the case of the multi-class classification task ($T_0 vs. T_1 vs. T_2 vs. T_3$), the improvement of the performance can only be seen in terms of accuracy. Overall, the proposed attention mechanism improves the performance of the designed multi-modal DDCAE across all classification tasks. Moreover, the depicted results show that applying a gated approach for the generation of a single weighted latent representation is not only beneficial for the reduction of the dimensionality of the final representation, but also, due to the optimized weighting parameters, the generated representation can significantly improve the performance of the classification system. In summary, the proposed gating layer is able to successfully perform an aggregation of the latent representations specific to each of the modalities and the resulting aggregated representation can be jointly used to optimize an effective inference model. Moreover, the proposed attention mechanism is able to improve the overall performance of the proposed multi-modal DDCAE model in terms of classification accuracy as well as reconstruction MSE. In the following sections, if not mentioned otherwise, the experiments are carried out with a version of the multi-modal DDCAE extended with the proposed attention mechanism.

## 4.3. Self-Supervised Approach: Results

To evaluate our self-supervised learning (SSL) algorithm we perform experiments on two datasets in the self-supervised setting: the BioVid Heat Pain Database (Part A) (Walter et al., 2013) (see **Table 5**) and the SenseEmotion Database (Velana

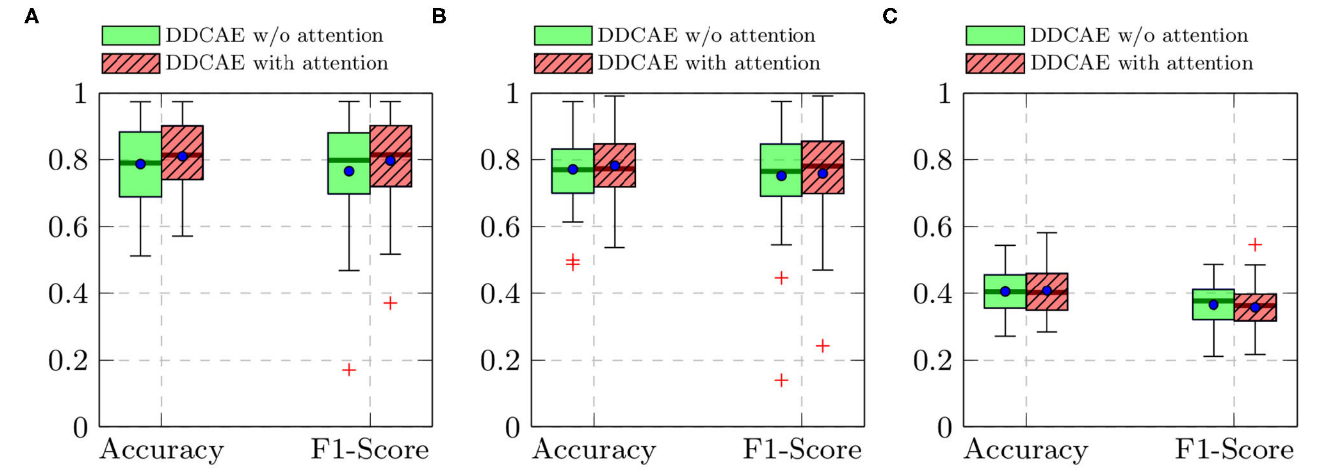

**FIGURE 8 |** SenseEmotion Database: classification performance comparison in a *Leave One Subject Out* (LOSO) cross-validation evaluation setting. Within each box plot, the mean and the median classification performance are depicted, respectively with a dot and a horizontal line. **(A)** $T_0 vs. T_3$, **(B)** $T_1 vs. T_3$, **(C)** 4 classes.

**TABLE 4 |** Classification performance comparison in a *Leave One Subject Out* (LOSO) cross-validation evaluation setting: Average Performance in % (Standard Deviation in %).

| **BioVid Heat Pain Database (Part A)** | | | | | | |
|---|---|---|---|---|---|---|
| Task | $T_0 vs. T_4$ | | $T_1$ vs. $T_4$ | | $T_0$ vs. $T_1$ vs. $T_2$ vs. $T_3$ vs. $T_4$ (5 Classes) | |
| Model | DDCAE | | DDCAE | | DDCAE | |
| | W/o attention | With attention | W/o attention | With attention | W/o attention | With attention |
| Accuracy | 83.99 (15.58) | **84.25 (13.82)** | 76.29 (16.09) | **76.81 (15.08)** | 33.31 (09.35) | **35.44 (08.66)**$^{*}$ |
| F1-Score | 78.66 (25.43) | **81.48 (20.55)**$^{*}$ | 68.18 (29.08) | **71.92 (24.81)**$^{*}$ | 30.31 (09.30) | **31.04 (07.91)** |
| **SenseEmotion Database** | | | | | | |
| Task | $T_0$ vs. $T_3$ | | $T_1$ vs. $T_3$ | | $T_0$ vs. $T_1$ vs. $T_2$ vs. $T_3$ (4 Classes) | |
| Model | DDCAE | | DDCAE | | DDCAE | |
| | W/o attention | With attention | W/o attention | With attention | W/o attention | With attention |
| Accuracy | 78.76 (11.37) | **81.05 (10.73)**$^{*}$ | 77.19 (11.70) | **78.26 (10.21)** | 40.48 (06.88) | **40.77 (07.00)** |
| F1-Score | 76.64 (15.99) | **79.78 (13.30)**$^{*}$ | 75.26 (15.70) | **75.95 (14.71)** | **36.53 (06.61)** | 35.75 (06.69) |

*The best achieved performance is depicted in bold. An asterisk (*$*$*) depicts a significant performance improvement. The significance test is performed using a Two-Sided Wilcoxon-Signed-Rank test with a significance level of 5%.*

et al., 2017; Thiam et al., 2019b) (see **Table 6**). In all our SSL experiments we kept the architecture as described in **Table 1**, with the exception of the output layer of the encoder network, where we have 512 units [256 for the mean and 256 for the log-variance of $p(h|x)$].

We designed the experiments to investigate the influence of our self-supervised tuning method in combination with adaptive variational Auto-Encoders on the number of training samples required. To this end, we evaluate our approach with an adaptive prior as described in section 3 and with a fixed prior, i.e., a standard normal distribution as $p(h)$. We evaluate the adaptive prior approach on the original data, on the augmented data (see section 3.5 for details) and in a self-supervised learning scenario, while we show results for the fixed prior on the augmented dataset. In all settings we trained the model for 100 epochs on the fully augmented data and evaluated on test data [*Leave One Subject Out* (LOSO) cross-validation evaluation]. For the self-trained setting, we trained the model on non-augmented data for 100 epochs, generated 25% additional data using the learned variational priors and classified them using model predictions, and fine tuned the previously trained model for 100 additional epochs using *only* the generated data. In the case of regression experiments, the performance measures are both Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), defined as follows:

$$\text{MAE} = \frac{1}{\mathcal{T}} \sum_{j=1}^{\mathcal{T}} |y_j - \widehat{y}_j| \tag{20}$$

$$\text{RMSE} = \sqrt{\frac{1}{\mathcal{T}} \sum_{j=1}^{\mathcal{T}} \left(y_j - \widehat{y}_j\right)^2} \tag{21}$$

where $y_j$ is the ground-truth class value of the $j^{th}$ sample, $\widehat{y}_j$ is the corresponding regression output value, and $\mathcal{T}$ is the number of

**TABLE 5 |** Results for the BioVid dataset in the binary, five classes, and regression setting in a *Leave One Subject Out* (LOSO) cross-validation evaluation setting: average performance in % (standard deviation in %).

| | | | | | | |
|---|---|---|---|---|---|---|
| **BioVid Heat Pain Database (Part A)** | | | | | | |
| **Task** | $T_0 vs. T_4$ | | | | | |
| **Model** | **DDCVAE** | | | | **DDCAE** | |
| **VAE prior** | **Adaptive** | | | **Fixed** | **N/A** | **N/A** |
| **method** | **Orig. data** | **Aug. data** | **Self-trained** | **self-trained** | **Aug. data** | **Orig. data** |
| **Samples** | 3,440 | 30,960 | 3,440 + 860 | 3,440 + 860 | 30,960 | 3,440 |
| **Accuracy** | 69.3 (14.7) | 84.0 (15.2) | 83.0 (15.9)† | 76.1 (16.8) | **84.2 (13.7)**† | 69.0 (15.0) |
| **F1-Score** | 63.8 (22.4) | 80.0 (23.0) | 78.0 (24.5) | 65.8 (28.9) | **81.5 (20.0)*** | 63.6 (22.9) |
| **Task** | $T_0 vs. T_1 vs. T_2 vs. T_3 vs. T_4$ | | | | | |
| **Model** | **DDCVAE** | | | | **DDCAE** | |
| **VAE prior** | **Adaptive** | | | **Fixed** | **N/A** | **N/A** |
| **method** | **Orig. data** | **Aug. data** | **Self-trained** | **self-trained** | **Aug. data** | **Orig. data** |
| **Samples** | 8,600 | 77,400 | 8,600 + 2,150 | 8,600 + 2,150 | 77,400 | 8,600 |
| **Accuracy** | 25.8 (5.0) | **35.5 (7.9)** | 32.9 (6.8) | 29.0 (6.9) | 35.4 (8.6)* | 25.1 (5.5) |
| **F1-Score** | 15.8 (5.0) | **31.6 (7.5)** | 20.0 (6.0) | 16.9 (5.7) | 31.0 (7.9)* | 17.5 (5.6) |
| **Task** | **Regression** | | | | | |
| **Model** | **DDCVAE** | | | | **DDCAE** | |
| **VAE prior** | **Adaptive** | | | **Fixed** | **N/A** | **N/A** |
| **Method** | **Orig. data** | **Aug. data** | **Self-trained** | **self-trained** | **Aug. data** | **Orig. data** |
| **Samples** | 8,600 | 77,400 | 8,600 + 860 | 8,600 + 860 | 77,400 | 8,600 |
| **MAE** | 1.21 (0.08) | 0.97 (0.18) | 1.00 (0.18) | 1.03 (0.18) | **0.97 (0.19)*** | 0.99 (0.21) |
| **RMSE** | 1.41 (0.09) | 1.16 (0.20) | 1.18 (0.20)† | 1.20 (0.16) | **1.16 (0.21)**† | 1.17 (0.18) |

† *p-Value of two-sided Wilcoxon-Signed-Rank Test is not signifying a statistically significant difference.*
$*$*p-Value of one-sided W-Test is signifying a statistically significant difference. For classification the alternative.*
*is "greater" and for regression "less".*
*Bold font signifies the best result. The Wilcoxon-Signed-Rank tests have been carried out between the results of the self-supervised fine-tuning approach using only the original dataset ("DDCVAE Adaptive Prior Self-Trained") and the DDCAE approach using the augmented dataset ("DDCAE Aug. Data").*

samples. In all self-supervised fine-tuning experiments we used $\beta_1 = 0.001, \beta_2 = 0.001, \beta_3 = 0.001, \lambda = 0.9995$. All Gaussian distributions were assumed to be isotropic, i.e., independent dimensions, which allows us to learn only the diagonal of the covariance matrices instead of the full matrix. We compute the variance of the Gaussian priors via the soft-plus activation of the corresponding output of encoder network, defined by $softplus(x) = \log[1 + \exp(x)]$.

In all three settings (binary classification, all-vs.-all, and regression) we were able to achieve results that are only marginally lower compared to the fully augmented dataset, while only using a fraction of the samples (see column labeled "Self-Trained" of **Tables 5**, **6**), thus showing the effectiveness of our method. All classification results were in the margin of 1–3% while only requiring an additional 25% of data points. The performance in terms of accuracy concerning the binary classification task of the BioVid Heat Pain Database was not significantly worse compared to the fully augmented case. Importantly, this does not hold for the fixed prior: the performances drop significantly compared to adaptive priors. We argue that this shows that our self-supervised approach enables the system to learn a informative representation for each of the modalities. Additionally, through the coupling introduced by the information-processing constraints, these representations are tuned in such a way that they improve the overall classification and regression performance, by discovering regularities in the data that can be exploited efficiently.

## 5. DISCUSSION

We introduced and evaluated a novel approach to deep multi-modal pain intensity assessment. We evaluated our approach on two complex pain intensity assessment datasets and were able to achieve results comparable to current state-of-the-art methods (see **Table 7**). The results specific to the proposed multi-modal DDCAE architecture show that the joint optimization of a single latent representation for each specific input channel and a gating layer (with trainable parameters) to generate a weighted latent representation (that is subsequently fed into a jointly trained model to perform an inference task), can improve the overall performance of an entire architecture by multiple percent. Additionally the reconstruction of the input signals is also performed at a satisfactory extent. Furthermore, when combined with an appropriate attention mechanism, the performance of the entire architecture can be further significantly improved. Therefore, feature learning can be considered as

**TABLE 6 |** Results for the SenseEmotion dataset in the binary, four classes, and regression setting in a *Leave One Subject Out* (LOSO) cross-validation evaluation setting: average performance in % (standard deviation in %).

| **SenseEmotion Database** | | | | | | |
|---|---|---|---|---|---|---|
| **Task** | $T_0 vs. T_3$ | | | | | |
| **Model** | **DDCVAE** | | | | **DDCAE** | |
| **VAE prior** | **Adaptive** | | | **Fixed** | **N/A** | **N/A** |
| **method** | **Orig. data** | **Aug. data** | **Self-trained** | **self-trained** | **Aug. data** | **Orig. data** |
| **Samples** | 4,680 | 42,120 | 4,680 + 1,170 | 4,680 + 1,170 | 42,120 | 4,680 |
| **Accuracy** | 77.7 (11.3) | 79.0 (11.0) | 78.5 (11.4) | 68.1 (11.8) | **81.0 (10.7)*** | 52.3 (4.9) |
| **F1-Score** | 74.2 (16.4) | 75.5 (15.9) | 73.8 (17.9) | 51.8 (22.6) | **79.8 (13.3)*** | 65.4 (7.9) |
| **Task** | $T_0 vs. T_1 vs. T_2 vs. T_3$ | | | | | |
| **Model** | **DDCVAE** | | | | **DDCAE** | |
| **VAE prior** | **Adaptive** | | | **Fixed** | **N/A** | **N/A** |
| **method** | **Orig. data** | **Aug. data** | **Self-trained** | **self-trained** | **Aug. data** | **Orig. data** |
| **Samples** | 9,314 | 83,835 | 9,314 + 2,328 | 9,314 + 2,328 | 83,835 | 9,314 |
| **Accuracy** | 38.4 (6.6) | 40.1 (6.5) | 39.1 (5.9) | 34.8 (6.3) | **40.8 (7.0)*** | 31.5 (6.8) |
| **F1-Score** | 30.0 (5.4) | 36.6 (5.6) | 27.4 (4.5) | 23.3 (6.1) | **35.8 (6.7)*** | 21.3 (10.8) |
| **Task** | **Regression** | | | | | |
| **Model** | **DDCVAE** | | | | **DDCAE** | |
| **Prior** | **Adaptive** | | | **Fixed** | **N/A** | **N/A** |
| **method** | **Orig. data** | **Aug. data** | **Self-trained** | **self-trained** | **Aug. data** | **Orig. data** |
| Samples | 9,314 | 83,835 | 9,314 + 2,328 | 9,314 + 2,328 | 83,835 | 9,314 |
| MAE | 0.82 (0.10) | 0.80 (0.10) | 0.81 (0.11) | 0.85 (0.09) | **0.80 (0.10)*** | 1.04 (0.07) |
| RMSE | 0.97 (0.11) | 0.96 (0.11) | 0.96 (0.12) | 1.01 (0.11) | **0.96 (0.11)*** | 1.18 (0.09) |

*†p-value of two-sided Wilcoxon-Signed-Rank Test is not signifying a statistically significant difference.*

**p-value of one-sided W-Test is signifying a statistically significant difference. For classification the alternative.*

*hypothesis is "greater" and for regression "less".*

*The Wilcoxon-Signed-Rank tests have been carried out between the results of the self-supervised fine-tuning approach using only the original dataset ("DDCVAE adaptive prior self-trained") and the DDCAE approach using the augmented dataset ("DDCAE Aug. Data").*

a sound alternative to manual feature engineering, since the designed architecture is able to autonomously generate a set of relevant parameters without the need of expert knowledge in this particular area of application. As potential future work, an investigation of the temporal aspect of the physiological signals through the introduction of recurrent neural networks such as LSTMs should be undertaken. Moreover, since we set most of the hyper-parameters involved in the performed assessment of the proposed approaches empirically, methods designed to perform the fine-tuning of such hyper-parameters (Feurer and Hutter, 2019) may automate this step. The introduction and investigation of generative models [such as Generative Adversarial Networks (GANs)] for data augmentation in the case of bio-physiological data should also be undertaken and a comparison of the performances achieved by such approaches, with those achieved through SSL approaches could provide further insights into the dynamics involved in multi-modal inference tasks.

In the Self-Supervised Learning setting we showed that our approach is able to drastically reduce the required number of training samples compared to classic data augmentation techniques. We achieved this by training a Deep Denoising Convolutional Adaptive Variational Auto-Encoder on each modality during a primary training period. We then use the learned latent prior for each modality to artificially generate new data samples, classify them with hard labels and perform a second fine-tuning training phase. Our method requires only 25% additional data with only a small performance loss. In an ablation study we were able to show that our adaptive VAE outperforms a classic VAE with a fixed prior, indicating that the additional flexibility allows to learn disentangled representations [encoded by their prior $p(h)$] for each modality. Given that we are working with temporal data it could be a promising research direction to investigate recurrent neural networks in the representation learning part, as these are known to be able to extract temporal dependencies in data. Furthermore, our self-supervised fine-tuning approach is independent of the underlying problem structure and we can therefore apply it to a variety of learning tasks. As potential future work one could investigate our method in the reinforcement learning setting to improve sample efficiency. A drawback of our method is that it requires careful tuning of the hyper-parameters $\beta_1, \beta_2$, and $\beta_3$, as they have a drastic impact on the results. Chosen too small, the posterior can never diverge from the prior and thus no learning is possible, while a large value leads to large divergence and thus rendering the prior useless, as it does not capture the posterior. Meta-learning techniques may prove useful to automatically tune these hyper-parameters.

**TABLE 7 |** Classification performance comparison with previous works (BioVid Heat Pain Database: $T_0$ *vs*. $T_4$).

| **BioVid heat pain database (Part A)** | |
|---|---|
| **Approach** | **Accuracy (%)** |
| Early fusion with random forests (Werner et al., 2014) | 74.10 |
| Multi-modal DDCAE with a shared latent representation (Thiam et al., 2020a) | 76.90 |
| Multi-modal DDCAE with a concatenated latent representation (Thiam et al., 2020a) | 77.24 |
| Early fusion with random forests (Kächele et al., 2016, 2017) | 82.73 |
| Deep neural network ensemble with a weighted aggregation layer (Thiam et al., 2019a) | 84.40 |
| Multi-modal DDCAE with a gated latent representation (w/o attention) | 83.99 |
| Multi-modal DDCAE with a gated latent representation (with attention) | 84.25 |
| Multi-modal DDCVAE with a gated latent representation & SSL | 83.00 |

*The performances of the proposed multi-modal DDCAE approaches are compared to other information fusion architectures involving the exact same modalities with an evaluation performed in a Leave One Subject Out (LOSO) cross-validation evaluation setting.*

## 6. CONCLUSION

Even though the results depicted in the current work are very promising, pain recognition remains a very complex inference task. Several parameters have to be taken in consideration in order to ensure the effectiveness of the developed approaches. In the current work, the assessment of the proposed approaches is performed on data sets characterized by thermal pain elicitations. However, the outcome of the performed experiments can be biased by both the nature of the stimuli applied and the types of sensors used to perform the recordings. An assessment of the proposed approaches in diverse settings, using different types of painful stimuli such as pressure, cold or electrical stimuli, should therefore be conducted. Moreover, both data sets were recorded in controlled environments. Hence, the implementation and evaluation of the proposed approaches in real world settings would provide valuable insights and bring the whole research community one step further toward the goal of autonomously and effectively performing the assessment of different levels of pain. Such a technology would substantially improve the effectiveness of pain management in a clinical setting.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the University of Ulm (Helmholtzstraße 20, 89081 Ulm, Germany) (ethical committee approval was granted: 196/10-UBB/bal). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

PT, FS, and HK conceived and designed the multi-modal DDCAE architecture. PT implemented and evaluated the architecture and collected and preprocessed the data used for the evaluation of the designed architectures. HH and DB designed the self-supervised architecture. HH implemented the self-supervised experiments and collected, interpreted, and evaluated the data. PT and HH interpreted and evaluated the results and wrote the first version of the manuscript. Funding Acquisition: FS, DB, and HK. All authors wrote, read, and approved the manuscript.

## FUNDING

The research leading to these results has received funding from the Federal Ministry of Education and Research (BMBF, SenseEmotion) to FS (BMBF, e:Med, CONFIRM, ID 01ZX1708C) to HK, and the Ministry of Science and Education Baden-Württemberg (Project ZIV) to HK. FS is supported by the project Multi-modal recognition of affect over the course of a tutorial learning experiment (SCHW623/7-1) funded by the German Research Foundation (DFG). The work of HH and DB is supported by the European Research Council, grant number ERC-StG-2015-ERC, Project ID: 678082, BRISC: Bounded Rationality in Sensorimotor Coordination.

## ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

## REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA: USENIX Association), 265–283.

Abe, S. (2010). *Support Vector Machines for Pattern Classification. Advances in Pattern Recognition, 2nd Edn*. London: Springer-Verlag. doi: 10.1007/978-1-84996-098-4

Aitchison, L. (2021). "A statistical theory of cold posteriors in deep neural networks," in *Proceedings of the International Conference on Learning Representations* (Vienna).

Argüello Prada, E. J. (2020). The Internet of Things (IoT) in pain assessment and management: an overview. *Inform. Med. Unlock*. 18:100298. doi: 10.1016/j.imu.2020.100298

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems, Vol. 33*, eds H. Larochelle, M.

Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Vancouver, BC: Curran Associates, Inc.), 12449–12460

Barlow, H. B. (1989). Unsupervised learning. *Neural Comput*. 1, 295–311. doi: 10.1162/neco.1989.1.3.295

Bellmann, P., Thiam, P., and Schwenker, F. (2018). "Multi-classifier-systems: architectures, algorithms and applications," in *Computational Intelligence for Pattern Recognition, Vol. 777*, eds W. Pedrycz and S. M. Chen (Cham: Springer International Publishing), 83–113. doi: 10.1007/978-3-319-89629-8_4

Bellmann, P., Thiam, P., and Schwenker, F. (2020). "Dominant channel fusion architectures - an intelligent late fusion approach," in *2020 International Joint Conference on Neural Networks (IJCNN)* (Glasgow), 1–8. doi: 10.1109/IJCNN48605.2020.9206814

Bellmann, P., Thiam, P., and Schwenker, F. (2021). "Using meta labels for the training of weighting models in a sample-specific late fusion classification architecture," in *2020 25th International Conference on Pattern Recognition (ICPR)* (Milan), 2604–2611. doi: 10.1109/ICPR48806.2021.9412509

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Campbell, E., Phinyomark, A., and Scheme, E. (2019). Feature extraction and selection for pain recognition using peripheral physiological signals. *Front. Neurosci*. 13:437. doi: 10.3389/fnins.2019.00437

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., et al. (2017). "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA: IEEE Computer Society), 6298–6306. doi: 10.1109/CVPR.2017.667

Chollet, F., et al. (2015). *Keras*. Available online at: https://keras.io (accessed June 6, 2021).

Chu, Y., Zhao, X., Han, J., and Su, Y. (2017). Physiological signal-based method for measurement of pain intensity. *Front. Neurosci*. 11:279. doi: 10.3389/fnins.2017.00279

Chu, Y., Zhao, X., Yao, J., Zhao, Y., and Wu, Z. (2014). Physiological signals based quantitative evaluation method of the pain. *IFAC Proc*. 47, 2981–2986. doi: 10.3182/20140824-6-ZA-1003.01420

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2016). "Fast and accurate deep neural network learning by exponential linear units (ELUs)," in *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016*, eds Y. Bengio and Y. LeCun (San Juan).

Das, A., Mock, J., Huang, Y., Golob, E., and Najafirad, P. (2021). "Interpretable self-supervised facial micro-expression learning to predict cognitive state and neurological disorders," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35* (Palo Alto, CA), 818–826.

Dickey, J. P., Pierrynowski, M. R., Bednar, D. A., and Yang, X. Y. (2002). Relationship between pain and vertebral motion in chronic low-back pain subjects. *Clin. Biomech*. 17, 345–352. doi: 10.1016/S0268-0033(02)00032-3

Donahue, J., Krähenbühl, P., and Darrell, T. (2017). "Adversarial feature learning," in *5th International Conference on Learning Representations, ICLR 2017* (Toulon). Available online at: https://openreview.net/forum?id=BJtNZAFgg

Eccleston, C., Blyth, F. M., Dear, B. F., Fisher, E. A., Keefe, F. J., Lynch, M. E., et al. (2020). Managing patients with chronic pain during the COVID-19 outbreak: considerations for the rapid introduction of remotely supported (eHealth) pain management services. *Pain* 161, 889–893. doi: 10.1097/j.pain.0000000000001885

Eckard, C., Asbury, C., Bolduc, B., Camerlengo, C., Gotthardt, J., Healy, L., et al. (2016). The integration of technology into treatment programs to aid in the reduction of chronic pain. *J. Pain Manage. Med*. 2:118. doi: 10.35248/2684-1320.16.2.118

Feurer, M., and Hutter, F. (2019). "Hyperparameter optimization," in *Automated Machine Learning: Methods, Systems, Challenges*, Springer Series on Challenges in Machine Learning (Cham: Springer International Publishing), 3–33. doi: 10.1007/978-3-030-05318-5_1

Fisher, R. A. (1936). The use of multiple measurement in taxonomic problems. *Ann. Eugen*. 7, 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Galashov, A., Jayakumar, S. M., Hasenclever, L., Tirumala, D., Schwarz, J., Desjardins, G., et al. (2019). "Information asymmetry in KL-regularized RL," in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019* (New Orleans, LA).

Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: an information-theoretic optimality principle. *Front. Robot. AI* 2:27. doi: 10.3389/frobt.2015.00027

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems, Vol. 27*, eds Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Montreal, QC: Curran Associates, Inc.), 2672–2680.

Grau-Moya, J., Leibfried, F., and Vrancx, P. (2019). "Soft Q-learning with mutual-information regularization," in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019* (New Orleans, LA).

Hawker, G. A., Mian, S., Kendzerska, T., and French, M. (2011). Measures of adult pain: visual analog scale for pain (VAS pain), numeric rating scale for pain (NRS pain), McGill pain questionnaire (MPQ), short-form McGill pain questionnaire (SF-MPQ), chronic pain grade scale (CPGS), short form-36 bodily pain scale (SF-36 BPS), and measure of intermittent and constant osteoarthritis pain (ICOAP). *Arthr. Care Res*. 63, S240–S252. doi: 10.1002/acr.20543

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2016). "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017* (Toulon).

Hihn, H., and Braun, D. A. (2020a). "Hierarchical expert networks for meta-learning," in *4th ICML Workshop on Life Long Machine Learning* (Vienna).

Hihn, H., and Braun, D. A. (2020b). Specialization in hierarchical learning systems. *Neural Process. Lett*. 52, 2319–2352. doi: 10.1007/s11063-020-10351-3

Hihn, H., Gottwald, S., and Braun, D. A. (2018). "Bounded rational decision-making with adaptive neural network priors," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Vol. 11081*, eds L. Pancioni, F. Schwenker, and E. Trentin (Cham: Springer International Publishing), 213–225. doi: 10.1007/978-3-319-99978-4_17

Hihn, H., Gottwald, S., and Braun, D. A. (2019). "An information-theoretic on-line learning principle for specialization in hierarchical decision-making systems," in *2019 IEEE 58th Conference on Decision and Control (CDC)* (Nice), 3677–3684. doi: 10.1109/CDC40024.2019.9029255

Hinton, G. E., and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hinton, G. E., and Zemel, R. S. (1994). "Autoencoders, minimum description length and helmholtz free energy," in *Advances in Neural Information Processing Systems, Vol. 6*, eds J. Cowan, G. Tesauro, and J. Alspector (Denver, CO: Morgan-Kaufmann), 3–10.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. (2016). "VIME: variational information maximizing exploration," in *Advances in Neural Information Processing Systems, Vol. 29*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Barcelona: Curran Associates, Inc.), 1117–1125.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2021). A survey on contrastive self-supervised learning. *Technologies* 9:2. doi: 10.3390/technologies9010002

Jing, L., and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* doi: 10.1109/TPAMI.2020.2992393. [Epub ahead of print].

Kächele, M., Amirian, M., Thiam, P., Werner, P., Walter, S., Palm, G., et al. (2017). Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evol. Syst*. 8, 71–83. doi: 10.1007/s12530-016-9158-4

Kächele, M., Thiam, P., Amirian, M., Schwenker, F., and Palm, G. (2016). Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE J. Select. Top. Signal Process*. 10, 854–864. doi: 10.1109/JSTSP.2016.2535962

Kächele, M., Werner, P., Walter, S., Al-Hamadi, A., and Schwenker, F. (2015). "Bio-visual fusion for person-independent recognition of pain intensity," in *Multiple Classifier Systems (MCS)*, *Vol. 9132 of Lecture Notes in Computer Science*, eds

F. Schwenker, F. Roli, and J. Kittler (Cham: Springer International Publishing), 220–230. doi: 10.1007/978-3-319-20248-8_19

Kessler, V., Thiam, P., Amirian, M., and Schwenker, F. (2017). "Multimodal fusion including camera photoplethysmography for pain recognition," in *2017 International Conference on Companion Technology (ICCT)* (Ulm), 1–4. doi: 10.1109/COMPANION.2017.8287083

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, eds Y. Bengio and Y. LeCun (San Diego, CA).

Kingma, D. P., and Welling, M. (2014). "Auto-encoding variational Bayes," in *Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014*, eds Y. Bengio and Y. LeCun (Banff, AB).

Kittler, J., and Roli, F., (eds.). (2000). *Multiple Classifier Systems, Vol. 1857 of Lecture Notes in Computer Science*. Berlin; Heidelberg: Springer-Verla. doi: 10.1007/3-540-48219-9

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, Inc. doi: 10.1002/0471660264

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). "ALBERT: a lite BERT for self-supervised learning of language representations," in *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020* (Addis Ababa).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Leibfried, F., and Grau-Moya, J. (2020). "Mutual-information regularization in Markov decision processes and actor-critic learning," in *Proceedings of the Conference on Robot Learning, Vol. 100 of Proceedings of Machine Learning Research*, eds L. P. Kaelbling, D. Kragic, and K. Sugiura (Osaka), 360–373.

Leibfried, F., Pascual-Diaz, S., and Grau-Moya, J. (2019). "A unified bellman optimality principle combining reward maximization and empowerment," in *Advances in Neural Information Processing Systems, Vol. 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC: Curran Associates, Inc.).

Lim, H., Kim, B., Noh, G.-J., and Yoo, S. K. (2019). A deep neural network-based pain classifier using a photoplethysmography signal. *Sensors* 19:384. doi: 10.3390/s19020384

Lopes, N., and Ribeiro, B. (2015). "Deep belief networks (DBNs)," in *Machine Learning for Adaptive Many-Core Machines - A Practical Approach, Vol. 7 of Studies in Big Data* (Cham: Springer International Publishing), 155–186. doi: 10.1007/978-3-319-06938-8_8

Luck, S. J., and Ford, M. A. (1998). On the role of selective attention in visual perception. *Proc. Natl. Acad. Sci. U.S.A.* 95, 825–830. doi: 10.1073/pnas.95.3.825

Olugbade, T. A., Singh, A., Bianchi-Berthouze, N., Marquardt, N., Aung, M. S. H., and Williams, A. C. d. C. (2019). How can affect be detected and represented in technological support for physical rehabilitation? *ACM Trans. Comput. Hum. Interact.* 26, 1–29. doi: 10.1145/3299095

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., and Tishby, N. (2015). Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*. Available online at: https://arxiv.org/abs/1512.06789

Palm, G., and Schwenker, F. (2009). "Sensor-fusion in neural networks," in *Harbour Protection Through Data Fusion Technologies*, eds E. Shahbazian, G. Rogova, and M. J. DeWeert (Dordrecht: Springer Netherlands), 299–306. doi: 10.1007/978-1-4020-8883-4_35

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). "Context encoders: feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2536–2544. doi: 10.1109/CVPR.2016.278

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peng, Z., Genewein, T., Leibfried, F., and Braun, D. A. (2017). "An information-theoretic on-line update principle for perception-action coupling," in *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC), 789–796. doi: 10.1109/IROS.2017.8202240

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., et al. (2020). "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6989–6993. doi: 10.1109/ICASSP40776.2020.9053569

Ricken, T., Steinert, A., Bellmann, P., Walter, S., and Schwenker, F. (2020). "Feature extraction: a time window analysis based on the X-ITE pain database," in *Artificial Neural Networks in Pattern Recognition, Vol. 12294 of Lecture Notes on Computer Science*, eds F. P. Schilling and T. Stadelmann (Cham: Springer International Publishing), 138–148. doi: 10.1007/978-3-030-58309-5_11

Rodriguez, P., Cucurull, G., González, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B., et al. (2017). Deep pain: exploiting long short-term memory networks for facial expression classification. *IEEE Trans. Cybern*. doi: 10.1109/TCYB.2017.2662199. [Epub ahead of print].

Roli, F. (2009). "Multiple classifier systems," in *Encyclopedia of Biometrics*, eds S. Z. Li and A. Jain (Boston, MA: Springer US), 981–986. doi: 10.1007/978-0-387-73003-5_148

Salakhutdinov, R., and Hinton, G. (2009). "Deep boltzmann machines," in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, Vol. 5 of Proceedings of Machine Learning Research*, eds D. van Dyk and M. Welling (Clearwater Beach, FL), 448–455.

Schobel, J., Volz, M., Hörner, K., Kuhn, P., Jobst, F., Schwab, J. D., et al. (2021). Supporting medical staff from psycho-oncology with smart mobile devices: insights into the development process and first results. *Int. J. Environ. Res. Publ. Health* 18:5092. doi: 10.3390/ijerph18105092

Schwenker, F., Dietrich, C. R., Thiel, C., and Palm, G. (2006). Learning of decision fusion mappings for pattern recognition. *Int. J. Artif. Intell. Mach. Learn.* 6, 17–21. Available online at: https://scholar.google.com/scholar_lookup?title=Learning%20of%20decision%20fusion%20mappings%20for%20pattern%20recognition&publication_year=2006&author=F.%20Schwenker&author=C.R.%20Dietrich&author=C.%20Thiel&author=G.%20Palm

Schwenker, F., and Trentin, E. (2014). Pattern classification and clustering: a review of partially supervised learning approaches. *Pattern Recogn. Lett.* 37, 4–14. doi: 10.1016/j.patrec.2013.10.017

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. (2020). "Planning to explore via self-supervised world models," in *Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research*, ed A. Singh (Vienna), 8583–8592.

Sharma, M., Tan, R.-S., and Acharya, U. R. (2019). Automated heartbeat classification and detection of arrhythmia using optimal orthogonal wavelet filters. *Inform. Med. Unlock*. 16:100221. doi: 10.1016/j.imu.2019.100221

Sharma, M., Tan, R.-S., and Acharya, U. R. (2020). Detection of shockable ventricular arrhythmia using optimal orthogonal wavelet filters. *Neural Comput. Appl.* 32, 15869–15884. doi: 10.1007/s00521-019-04061-8

Subramaniam, S. D., and Dass, B. (2021). Automated nociceptive pain assessment using physiological signals and a hybrid deep learning network. *IEEE Sensors J.* 21, 3335–3343. doi: 10.1109/JSEN.2020.3023656

Tavakolian, M., and Hadid, A. (2019). A spatiotemporal convolutional neural network for automatic pain estimation from facial dynamics. *Int. J. Comput. Vis.* 127, 1413–1425. doi: 10.1007/s11263-019-01191-3

Tavakolian, M., Lopez, M. B., and Liu, L. (2020). Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recogn. Lett.* 140, 26–33. doi: 10.1016/j.patrec.2020.09.012

Thiam, P., Bellmann, P., Kestler, H. A., and Schwenker, F. (2019a). Exploring deep physiological models for nociceptive pain recognition. *Sensors* 19:4503. doi: 10.3390/s19204503

Thiam, P., Kessler, V., Amirian, M., Bellmann, P., Layher, G., Zhang, Y., et al. (2019b). Multi-modal pain intensity recognition based on the sense emotion database. *IEEE Trans. Affect. Comput.* 2019:2892090. doi: 10.1109/TAFFC.2019.2892090

Thiam, P., Kessler, V., Walter, S., Palm, G., and Scwenker, F. (2017). "Audio-visual recognition of pain intensity," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Vol. 10183 of Lecture Notes in Computer Science*, eds F. Schwenker and S. Scherer (Cham: Springer International Publishing), 110–126. doi: 10.1007/978-3-319-59259-6_10

Thiam, P., Kestler, H. A., and Schwenker, F. (2020a). "Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals," in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM), Vol. 1* (Valletta: SciTePress), 289–296. doi: 10.5220/0008896102890296

Thiam, P., Kestler, H. A., and Schwenker, F. (2020b). Two-stream attention network for pain recognition from video sequences. *Sensors* 20:839. doi: 10.3390/s20030839

Thiam, P., Meudt, S., Palm, G., and Schwenker, F. (2018). A temporal dependency based multi-modal active learning approach for audiovisual event detection. *Neural Process. Lett*. 48, 709–732. doi: 10.1007/s11063-017-9719-y

Thiam, P., and Schwenker, F. (2017). "Multi-modal data fusion for pain intensity assessement and classification," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)* (Montreal, QC), 1–6. doi: 10.1109/IPTA.2017.8310115

Thiam, P., and Schwenker, F. (2019). "Combining deep and hand-crafted features for audio-based pain intensity classification," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Vol. 11377 of Lecture Notes in Computer Science*, eds F. Schwenker and S. Scherer (Cham: Springer International Publishing), 49–58. doi: 10.1007/978-3-030-20984-1_5

Tsai, F.-S., Hsu, Y.-L., Chen, W.-C., Weng, Y.-M., Ng, C.-J., and Lee, C.-C. (2016). Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions. *Interspeech* 2016, 92–96. doi: 10.21437/Interspeech.2016-408

Tsai, F.-S., Weng, Y.-M., Ng, C.-J., and Lee, C.-C. (2017). "Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (San Francisco, CA), 313–318. doi: 10.1109/ACII.2017.8273618

Tung, H.-Y. F., Tung, H.-W., Yumer, E., and Fragkiadaki, K. (2017). "Self-supervised learning of motion capture," in *Advances in Neural Information Processing Systems, Vol. 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Curran Associates, Inc.), 5242–5252.

Uddin, M. T., and Canavan, S. (2020). "Multimodal multilevel fusion for sequential protective behavior detection and pain estimation," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (Buenos Aires), 844–848. doi: 10.1109/FG47880.2020.00073

Velana, M., Gruss, S., Layher, G., Thiam, P., Zhang, Y., Schork, D., et al. (2017). "The SenseEmotion database: a multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Vol. 10183 of Lecture Notes in Computer Science*, eds F. Schwenker and S. Scherer (Cham: Springer International Publishing), 127–139. doi: 10.1007/978-3-319-59259-6_11

Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H. C., Crawcour, S., et al. (2013). "The BioVid heat pain database: data for the advancement and systematic validation of an automated pain recognition system," in *2013 IEEE International Conference on Cybernetics (CYBCO)* (Lausanne), 128–131. doi: 10.1109/CYBConf.2013.6617456

Walter, S., Gruss, S., Frisch, S., Liter, J., Jerg-Bretzke, L., Zujalovic, B., et al. (2020). "What about automated pain recognition for routine clinical use?" A survey of physicians and nursing staff on expectations, requirements, and acceptance. *Front. Med*. 7:990. doi: 10.3389/fmed.2020.566278

Walter, S., Gruss, S., Limbrecht-Ecklundt, K., Traue, H. C., Werner, P., Al-Hamadi, A., et al. (2014). Automatic pain quantification using autonomic parameters. *Psychol. Neurosci*. 7, 363–380. doi: 10.3922/j.psns.2014.041

Werner, P., Al-Hamadi, A., Gruss, S., and Walter, S. (2019). "Twofold-multimodal pain recognition with the X-ITE pain database," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (Cambridge), 290–296. doi: 10.1109/ACIIW.2019.8925061

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., and Traue, H. C. (2017). Automatic pain assessment with facial activity descriptors. *IEEE Trans. Affect. Comput*. 8, 286–299. doi: 10.1109/TAFFC.2016.2537327

Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., and Traue, H. C. (2014). "Automatic pain recognition from video and biomedical signals," in *2014 22nd International Conference on Pattern Recognition* (Stockholm), 4582–4587. doi: 10.1109/ICPR.2014.784

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich). doi: 10.1007/978-3-030-01234-2_1

Zhang, R., Isola, P., and Efros, A. A. (2017). "Split-brain autoencoders: unsupervised learning by cross-channel prediction," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI), 645–654. doi: 10.1109/CVPR.2017.76

Zhao, P., Zhang, J., Fang, W., and Deng, S. (2020). SCAU-Net: spatial-channel attention u-net for gland segmentation. *Front. Bioeng. Biotechnol*. 8:670. doi: 10.3389/fbioe.2020.00670

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.