

Facial Expression Analysis Using Decomposed Multiscale Spatiotemporal Networks

Wheidima Carneiro de Melo^{a,b,*}, Eric Granger^c and Miguel Bordallo Lopez^{a,d}

^aCenter for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland

^bSuperior School of Technology, State University of Amazonas, Brazil

^cLaboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), École de Technologie Supérieure, Canada

^dVTT Technical Research Centre of Finland Ltd, Finland

ARTICLE INFO

Keywords:

Depression Detection
Pain Estimation
Facial Expression Analysis
Deep Learning
Convolutional Neural Networks

ABSTRACT

Video-based analysis of facial expressions has been increasingly applied to infer health states of individuals, such as depression and pain. Among the existing approaches, deep learning models composed of structures for multiscale spatiotemporal processing have shown strong potential for encoding facial dynamics. However, such models have high computational complexity, making for a difficult deployment of these solutions. To address this issue, we introduce a new technique to decompose the extraction of multiscale spatiotemporal features. Particularly, a building block structure called Decomposed Multiscale Spatiotemporal Network (DMSN) is presented along with three variants: DMSN-A, DMSN-B, and DMSN-C blocks. The DMSN-A block generates multiscale representations by analyzing spatiotemporal features at multiple temporal ranges, while the DMSN-B block analyzes spatiotemporal features at multiple spatial sizes, and the DMSN-C block analyzes spatiotemporal features at multiple spatial sizes. Using these variants, we design our DMSN architecture which has the ability to explore a variety of multiscale spatiotemporal features, favoring the adaptation to different facial behaviors. Our extensive experiments on challenging datasets show that the DMSN-C block is effective for depression detection, whereas the DMSN-A block is efficient for pain estimation. Results also indicate that our DMSN architecture achieves competitive performance while requiring 3.51× and 26.55× fewer parameters than the current state-of-the-art models for depression detection and pain estimation, respectively. The code is publicly available at <https://github.com/wheidima/DMSN>.

1. Introduction

Given the population growth, and global shortage of doctors, among others, healthcare applications have been driving the development of automatic systems for medical diagnosis. Such technology can be beneficial to improve the quality of clinical outcomes, and the access to healthcare services. Since face can provide information concerning medical conditions (Thevenot et al. (2018)), there has been a growing interest in developing contact-free, objective, and accurate systems for automatic assistive medical diagnosis from facial videos (Thevenot et al. (2018); Pampouchidou et al. (2019); Werner et al. (2022)). These video-based methods encode the correlations between appearance and dynamics of facial expressions and health states of an individual. For instance, Jaiswal et al. (2017) proposed a method that explores facial expressions, and head pose and movement to predict attention-deficit/hyperactivity disorder.

Two emerging applications for automatic facial expression analysis are depression detection and pain estimation. Depression is defined as a negative state of mind which remains for a long period of time. Such a mental health disorder can affect an individual's emotions, behavior, mind, and physical health (Trivedi (2004)). In severe conditions, depression conducts to substance abuse and suicidal behavior (American Psychiatric Association (2013)). Despite the existence of effective treatment, it is estimated that, in

Europe, about 56% of patients suffering from depression receive no treatment (Purebl and *et al.* (2015)). The reasons for this high number include client fees, and restricted or lack of accessibility to mental healthcare. Studies also show that clinicians have difficulties to diagnose depression (Mitchell et al. (2009); Bostwick and Rackley (2012)). Indeed, the assessment of depression has a subjective nature since it relies on doctor's perception of patient reports. Inaccurate diagnosis of depression has produced an alarming number of false-positives that present grave consequences for the patients (Bostwick and Rackley (2012)).

Pain is an important physical sign associated with the health conditions of an individual. It can be considered as a highly disturbing sensation caused by injury, illness or mental distress, and it is related to depression (Garcia-Cebrian et al. (2006)). The clinical evaluation of pain is mainly determined by patient self-reports (e.g., by using visual analogue scale (Lesage et al. (2012)) or numeric rating scale (Downie et al. (1978))). However, the assessment provided by a patient may not be reliable since patients may have restricted communication potential (e.g., neonates), cognitive impairments or are under the influence of medication. An alternative is the medical staff (e.g., doctors and nurses) perform the assessment. However, observers may overestimate or underestimate pain intensity which impair the treatment (Kappesser and Williams (2010)), and the continuous monitoring is impracticable.

Automatic analysis of facial variations for objective recognition of expressions associated with health states

*Corresponding author

 wheidima.melo@oulu.fi (W.C. de Melo); eric.granger@etsmt1.ca (E. Granger); miguel.bordallo@vtt.fi (M.B. Lopez)

like depression and pain can assist in the reliability and improvement of clinical assessment and monitoring, as well as mitigate issues regarding accessibility and costs. Studies have found facial cues related to depression, such as limited facial expressiveness (Trémeau et al. (2005)), reduced eye contact (Lucas et al. (2015)), smiles with a shorter duration and less intensity (Scherer et al. (2013)), and a small number of mouth movements (Schelde (1998)). In contrast, facial expressions (Werner et al. (2022); Hassan et al. (2021)) involving, e.g., closed eyes, raised cheeks, and a wrinkled nose are relevant indicators of pain. With that, we can claim that a pain event may produce expressions with greater facial variations over time, and a depressive state is linked to expressions with fewer variations over time. Therefore, systems for facial expression analysis based on videos can explore these cues to predict depressive or painful states.

Recently, the emergence of state-of-the-art deep learning (DL) architectures has contributed to significant progress in diverse visual recognition tasks, such as action recognition (Carreira and Zisserman (2017)), image classification (Krizhevsky et al. (2012)), imaging diagnosis (Wu et al. (2023); Li et al. (2022a)), and activity understanding (Kitani et al. (2012)). DL models have also been shown to provide a high level of predictive accuracy for automatic facial expression analysis from videos (Werner et al. (2022); Lopez et al. (2017); de Melo et al. (2022, 2023); Tavakolian and Hadid (2019); Zeng et al. (2018)). Given the availability of pre-trained models for still images, DL models commonly employ 2D Convolutional Neural Networks (CNNs) to leverage spatial correlations, along with an aggregation scheme or a recurrent technique to capture temporal dependencies (de Melo et al. (2019b); He et al. (2021); Jan et al. (2018); Kang et al. (2017); Yu et al. (2019); Rodriguez et al. (2017); Zhou et al. (2016, 2019, 2020); Uddin et al. (2022)). Such an approach has limited capacity in encoding important dynamic information (de Melo et al. (2022, 2023); Tavakolian and Hadid (2019)). Conversely, 3D-CNNs can directly model spatiotemporal variations in facial information from input video clips (de Melo et al. (2019a); Al Jazary and Guo (2021); Gnana Praveen et al. (2020); Zhou et al. (2022); Huang et al. (2022)). However, in addition to high computational complexity, these architectures use basic building blocks that explore a fixed spatiotemporal range, which limits the ability to learn discriminative features since facial expression variations comprise various ranges, and the difference of these variations along distinct levels of a health condition can be small.

In other application domains, efficient architectures have been developed for the modeling of spatiotemporal information (Lin et al. (2019); Qiu et al. (2017); Wang et al. (2019); Xie et al. (2018)). However, these methods also rely on structures with the ability to explore fixed spatiotemporal range. To address this problem, some works (de Melo et al. (2022, 2023); Tavakolian and Hadid (2019)) present effective architectures to model facial expression variations in videos. Such methods explore multiscale spatiotemporal features by using either parallel 3D convolutions with different

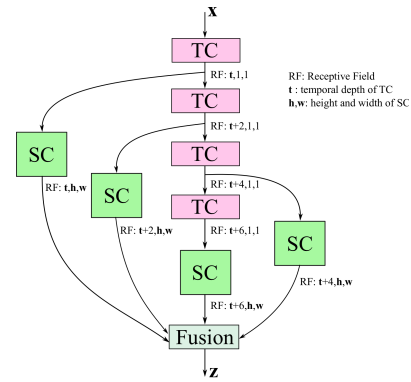


Figure 1: A variant of the proposed DMSN block. Each temporal convolution (pink block) generates features at different scales. Spatial convolutions (green block) complement this operation to explore spatiotemporal features. The fusion stage combines the multiscale spatial and temporal features. A detailed description of this block is presented in Section 3.

kernels (de Melo et al. (2022); Tavakolian and Hadid (2019)) or multiple structures that explore different spatiotemporal ranges (de Melo et al. (2023)). Although these approaches achieve a high level of performance, the use of these multiple elements in the basic building block generates models that have a high number of parameters and computations, even when compared to 3D-CNNs.

In this paper, we present an efficient alternative for the modeling of facial expression variations captured in videos. The proposed method decomposes the exploration of multiscale spatiotemporal information to improve such modeling and reduce computational costs. Specifically, we introduce a building block called Decomposed Multiscale Spatiotemporal Network (DMSN). The structure consists of a sequence of convolutions to produce multiscale features, where every element operates on a domain, and the branches of this sequence operate on a complementary domain of these elements, allowing to generate multiscale spatiotemporal representations. Inspired by the facial behavior linked to clinical states like depression and pain, we develop three different blocks: DMSN-A, DMSN-B, and DMSN-C. The DMSN-A block learns spatiotemporal features with distinct temporal ranges at a fixed spatial size (see Fig. 1). The DMSN-B block explores diverse spatiotemporal features at distinct ranges. Lastly, the DMSN-C block analyzes spatiotemporal features with different spatial sizes at a fixed temporal range. Our proposed blocks employ residual connections, and are implemented using only 1D and 2D convolutions. Using these three blocks, we design a new deep learning architecture, called DMSN, which has the potential to adapt to different facial behaviors thanks to the different multiscale spatiotemporal representation abilities of the proposed blocks.

The key contributions of this paper are as follows.

- A new building block structure is proposed with three variants – DMSN-A, DMSN-B, and DMSN-C blocks – to decompose the extraction of multiscale spatiotemporal features. Such variants are employed in our

DMSN architecture to provide discriminative representations for different facial behaviors.

- We show empirically that our DMSN-C block is effective for exploring the spatiotemporal dependencies for depression detection, whereas the DMSN-A block is efficient in capturing facial dynamics for pain estimation.
- An extensive set of experiments on the challenging AVEC2013 and AVEC2014 depression datasets, and UNBC-McMaster and BioVid pain datasets, allowing to validate that our DMSN architecture can provide a level of performance that is comparable to state-of-the-art DL models, while significantly reducing the computational costs.
- An analysis of depression and pain features showing that depression features are more useful for pain estimation than pain features are for depression detection.

2. Related Work

The growing interest in analyzing facial expressions captured in videos can be attributed to the psychological studies that indicate the correlation of a health condition and face, and the recent progress in deep learning and computer vision methods. The existing works try to explore non-verbal facial cues in order to infer health conditions. A key challenge is to obtain a robust representation in a scenario of subjective variability of facial expressions across different individuals and capture conditions.

2.1. Models for depression detection

Various authors proposed hand-engineered representations for depression detection. Some examples are the method proposed by Cohn et al. (2009), which employs Active Appearance Model (AAM) features and uses Support Vector Machine (SVM) as classifier, and the one proposed by Gupta et al. (2014), which uses Local Binary Pattern (LBP) features and Support Vector Regressor (SVR). DL models have demonstrated more potential to extract discriminant features from spatiotemporal expressions correlated with depressive states. A common approach is to employ a 2D-CNN and some aggregation technique to explore facial features that are extracted from videos (de Melo et al. (2019b); He et al. (2021); Jan et al. (2018); Kang et al. (2017); Zhou et al. (2019, 2020); Uddin et al. (2022)). For instance, Zhou et al. (2019) used ResNet-50 to explore the appearance information, and attention mechanism to fuse the static facial features. However, such methods have limited ability for the encoding of rich spatiotemporal variations in faces. Two-stream networks (de Melo et al. (2020); Zhu et al. (2018); Chen et al. (2021)) and 3D-CNNs (de Melo et al. (2019a); Al Jazaery and Guo (2021); Zhou et al. (2022)) have also been presented for depression detection. However, these methods are composed of structures that analyze a fixed spatiotemporal range, which reduces the ability to produce discriminative features. Indeed, it has

been shown that a better multiscale capacity is favorable for depression detection which is characterized by small facial expression variations along different levels (de Melo et al. (2022, 2023); Song et al. (2022)). In this context, Song et al. (2022) used spectral representations of behavior signals to analyze multiscale depression patterns. Two recent state-of-the-art methods – Multiscale Spatiotemporal Network (MSN) (de Melo et al. (2022)), and Maximization and Differentiation Network (MDN) (de Melo et al. (2023)) – have shown effectiveness in modeling multiscale spatiotemporal information. The structure of MSN is composed of 3D convolutions with different kernel sizes, whereas the one of MDN is formed using multiple maximization and difference blocks which explore features in diverse ranges. Although these methods achieve a high level of performance, their computational costs are expensive.

2.2. Models for pain estimation

Early methods for pain intensity estimation employed hand-engineered features such as LBP (Kaltwang et al. (2012)), Gabor (Zhao et al. (2016)), Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) (Thiam et al. (2016)), Histograms of Topographical (HoT) (Florea et al. (2014)), Pyramid Histogram of Orientation Gradients (PHOG) (Khan et al. (2013)) and Pyramid Local Binary Pattern (PLBP) (Khan et al. (2013)). In recent years, DL models have been used to encode facial expression variations for pain estimation. Some methods generate deep representations by using frame-wise feature extraction (Rodríguez et al. (2017); Yu et al. (2019); Zhou et al. (2016)). In this case, 2D networks, such as VGG (Simonyan and Zisserman (2014)), and ResNet (He et al. (2016)) or its optimized version (Amelio et al. (2023)), can be used as DL models. For instance, Rodríguez et al. (2017) employed VGG-16 architecture to learn spatial features and Long-Short Term Memory (LSTM) for the capturing of temporal relationships. Other works proposed to model spatiotemporal information within video sequences by employing 3D-CNNs (Gnana Praveen et al. (2020); Wang and Sun (2018); Huang et al. (2022)). Using this approach, Wang and Sun (2018) applied Convolutional 3D (C3D) network, that has as basic structure one $3 \times 3 \times 3$ convolutional layer, to recognize pain expressions. However, these two approaches are frequently ineffective to capture extensive range of facial expression variations. In Tavakolian and Hadid (2019), the authors presented evidence that a multiscale approach is more effective for the modeling of spatial and temporal dependencies related to pain status. They introduced the Spatiotemporal Convolutional Network (SCN) which employs as basic structure parallel 3D convolutions with different temporal depths. SCN obtains high performance for pain estimation, but it requires more than 500M trainable parameters, making its deployment costly. In Table 1, we summarize the deep methods for depression detection and pain estimation.

Table 1

Summary of the existing deep methods for depression detection and pain estimation.

Methods	Application	Advantage	Disadvantage
Deep Trans. Learning (DTL) Kang et al. (2017)	depression	Employ a pre-trained 2D CNN	Average scores of each frame
D. Local Global Attention (DLGA) He et al. (2021)	depression	Use local and global facial regions	Average scores of each frame
DepressNet (four ResNet-50) Zhou et al. (2020)	depression	Find the predominant facial area	Average scores of each frame
Deep Distribution Learning de Melo et al. (2019b)	depression	Learn depression distributions	Average scores of each frame
Volume Local Directional Number (VLDN+CNN+BiLSTM) Uddin et al. (2022)	depression	Explore image, local-patches, and temporal information	Limitation for encoding dynamic information
VGG-16 and Feature Dynamic History Histogram (FDHH) Jan et al. (2018)	depression	Employ different regression techniques	Limitation for encoding dynamic information
Feature Pooling Zhou et al. (2019)	depression	Produce a compact video representation	Limitation for encoding dynamic information
Two C3D de Melo et al. (2019a); Al Jazaery and Guo (2021)	depression	Analyze spatiotemporal variations at two facial regions	Use a structure that explore a fixed spatiotemporal range
C3D with Distribution Learning (DJ-LDML) Zhou et al. (2022)	depression	Learn depression distributions	Use a structure that explore a fixed spatiotemporal range
Two-stream Networks de Melo et al. (2020); Zhu et al. (2018); Chen et al. (2021)	depression	Use a scheme to map temporal variations into an image map	Use a structure that explore a fixed spatiotemporal range
Represen. of Behavior Signals Song et al. (2022)	depression	Explore multiscale features	Use toolbox to detect AUs
MSN de Melo et al. (2022) and MDN de Melo et al. (2023)	depression	Use blocks that explore multiscale spatiotemporal information	Computationally expensive
Recurrent Convolutional Neural Network (RCNN) Zhou et al. (2016)	pain	Insert recurrent operations into convolutions	Limitation for encoding dynamic information
VGG11+LSTM Yu et al. (2019) and VGG16+LSTM Rodriguez et al. (2017)	pain	Investigate RNNs to model static features variations	Limitation for encoding dynamic information
C3D Wang and Sun (2018), I3D Gnana Praveen et al. (2020) and HybNet Huang et al. (2022)	pain	Investigate the use of 3D CNNs	Use a structure that explore a fixed spatiotemporal range
SCN Tavakolian and Hadid (2019) and MDN de Melo et al. (2023)	pain	Use blocks that explore multiscale spatiotemporal information	Computationally expensive

2.3. Spatiotemporal networks

Since 3D-CNNs have an ability to directly model spatial and temporal information, these methods are an intuitive choice for video analysis. Tran et al. (2015) proposed an architecture with 8 convolutional layers, called C3D, to learn spatiotemporal features. Carreira and Zisserman (2017) proposed to inflate all the filters and pooling kernels of 2D Inception model into 3D-CNN generating Inflated 3D-ConvNet (I3D) model. Hara et al. (2018) proposed 3D-CNNs based on residual connections called 3D-ResNet. In Feichtenhofer et al. (2019), the authors introduced the SlowFast network which consists of a slow path to model spatial semantics and a fast path to capture motion at fine temporal resolution. The principal drawbacks of employing 3D-CNNs are the high computational complexity, and the lack of pre-trained backbone models. Recently, diverse architectures have been developed for efficient spatiotemporal modeling. In Wang et al. (2019), Temporal Segment Network (TSN) is introduced to model long-term temporal information employing 2D-CNNs. Qiu et al. (2017) proposed Pseudo-3D residual network (P3D) which factorizes 3D convolution into 2D and 1D convolution. Xie *et al.* studied I3D, I2D, as well as the combination of 2D and 3D methods (Xie et al. (2018)). Lin et al. (2019) proposed the Temporal Shift Module (TSM) to enable 2D-CNNs to explore spatiotemporal dependencies by shifting channels along the temporal dimension. Even though such architectures are efficient for tasks like action recognition, their structure explores a fixed spatiotemporal range, which hinders the capacity of extracting effective representations for facial expression variations. Instead of exploring a fixed range, our proposed blocks explore multiple spatiotemporal ranges favoring the generation of discriminative representations.

In contrast to existing methods that use a structure to explore multiscale spatiotemporal information (i.e., MSN, MDN, and SCN), our proposed blocks are designed to efficiently capture such information for the representation of facial videos. To achieve this goal, three distinct variants of the DMSN block are proposed to decompose the extraction of multiscale spatiotemporal features. The use of these three variants allows the design of a cost-effective DMSN architecture. Another unique benefit of our architecture is the ability to adapt to distinct facial behaviors, since it is built employing our three proposed blocks.

3. Decomposed Multiscale Spatiotemporal Network

The dynamics of facial expressions provide rich information for the recognition of facial patterns related to a health condition. Such facial expression variations can be explored, e.g., velocity or intensity, in order to model different levels of a health state. This work aims to develop a deep architecture to capture an extensive range of facial dynamics to produce efficient representations for automatic facial expression analysis. Specifically, we design the Decomposed Multiscale Spatiotemporal Network (DMSN) by introducing three multiscale convolution blocks that employ different strategies to generate multiscale spatiotemporal representations.

The proposed blocks encode both appearance and temporal information in different spatiotemporal ranges. This approach allows the development of deep models that explore facial expression variations linked to different clinical states. Another interesting property of our approach is the capacity of favoring the capturing of subtle facial expression variations. Since there is a high correlation between facial

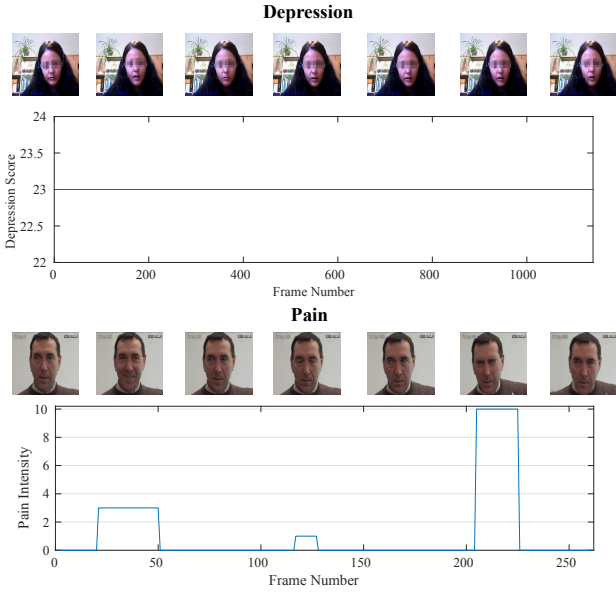


Figure 2: Examples of depression and pain expressions in sequences of consecutive video frames.

expressions displayed in different levels of a condition, e.g., a sad facial expression may be displayed by either a healthy person or depressed person, this ability is essential to generate discriminative representations.

Our DMSN blocks are designed considering that the facial behavior can considerably differ in two distinct health diagnosis applications. As illustrated in Fig. 2, the level of pain can change over time, and the correlated facial expressions can be modified considerably over a short period. On the other hand, the depression level lasts for a longer period and the resulting facial expressions tend to have more gradual variations. A facial behavior that is in the middle term between these two health applications is also considered. Furthermore, an effective architecture for facial expression analysis has the capability of adaptation to distinct facial behaviors. This fact motivates us to build our architecture using blocks with different abilities.

The proposed blocks decompose the exploration of multiscale spatiotemporal information by first employing a sequence of convolutions to increase the range of the region under analysis. This sequence is called the Main Stage sub-block (see Fig. 3). The output of each convolution in the Main Stage is connected as the input to another convolution that operates in a complementary domain to encode spatiotemporal information. The output feature maps of these branches are at different scales, and a $1 \times 1 \times 1$ convolution fuses these features to generate multiscale spatiotemporal representations.

The architectural design of our DMSN block allows the investigation of different strategies to extract multiscale spatiotemporal features. Since the Main Stage sub-block is responsible for the multiscale ability, it is able to employ convolutions on either the same or different domains, which can be beneficial in the elaboration of more efficient multiscale representations. In this context, we derive three

variants of our proposed block (see Fig. 3). In the sequence, we present a detailed description of these variants.

3.1. DMSN-A block

Considering that the pain level can vary more rapidly over time, its level can last for different periods, and it can produce sudden facial expression variations, we define the Main Stage of the DMSN-A block as a sequence of $3 \times 1 \times 1$ temporal convolutions. This sub-block is formed using four 1D temporal convolutions in order to explore the short, medium, and long temporal ranges. The output \mathbf{T}_i of each 1D temporal convolution (\mathbf{M}_i^t) is given by:

$$\mathbf{T}_i = \begin{cases} \mathbf{M}_i^t(\mathbf{x}) & i = 1 \\ \mathbf{M}_i^t(\mathbf{T}_{i-1}) & 2 \leq i \leq 4 \end{cases} \quad (1)$$

Each 1D convolution increases the temporal range explored by this sub-block. Branches of the Main Stage employ $1 \times 3 \times 3$ spatial convolutions which generate spatiotemporal features at multiple temporal ranges. The output \mathbf{ST}_j of each 2D spatial convolution (\mathbf{M}_j^s) is defined by:

$$\mathbf{ST}_j = \mathbf{M}_j^s(\mathbf{T}_j) \quad 1 \leq j \leq 4 \quad (2)$$

3.2. DMSN-B block

This block employs the Main Stage sub-block to increase the explored regions in both domains by using $1 \times 3 \times 3$ spatial convolutions and $3 \times 1 \times 1$ temporal convolutions. With the purpose of maintaining a similar computational complexity in comparison with DMSN-A block, the DMSN-B block employs four convolutions in the Main Stage. The output \mathbf{Y}_i of each element in this sub-block is calculated by:

$$\mathbf{Y}_i = \begin{cases} \mathbf{M}_i^s(\mathbf{x}) & i = 1 \\ \mathbf{M}_{i-1}^s(\mathbf{Y}_{i-1}) & i = 3 \\ \mathbf{M}_{i/2}^t(\mathbf{Y}_{i-1}) & i = 2, 4 \end{cases} \quad (3)$$

Each element of this sub-block increases the spatiotemporal receptive field size in analysis. Branches of the Main Stage use complementary convolution (in relation to domain) to generate spatiotemporal features at multiple ranges. Specifically, the output \mathbf{ST}_j of each branch is given by:

$$\mathbf{ST}_j = \begin{cases} \mathbf{M}_{j+3-(j+1)/2}^t(\mathbf{Y}_j) & j = 1, 3 \\ \mathbf{M}_{j+2-j/2}^s(\mathbf{Y}_j) & j = 2, 4 \end{cases} \quad (4)$$

3.3. DMSN-C block

Given that depressive states can present less facial expression variations over time, and the depression level of a subject in a video tends to be constant, the DMSN-C block employs the Main Stage to produce multiscale spatial features. The sub-block is constituted by a sequence of $1 \times 3 \times 3$ spatial convolutions where each element increases

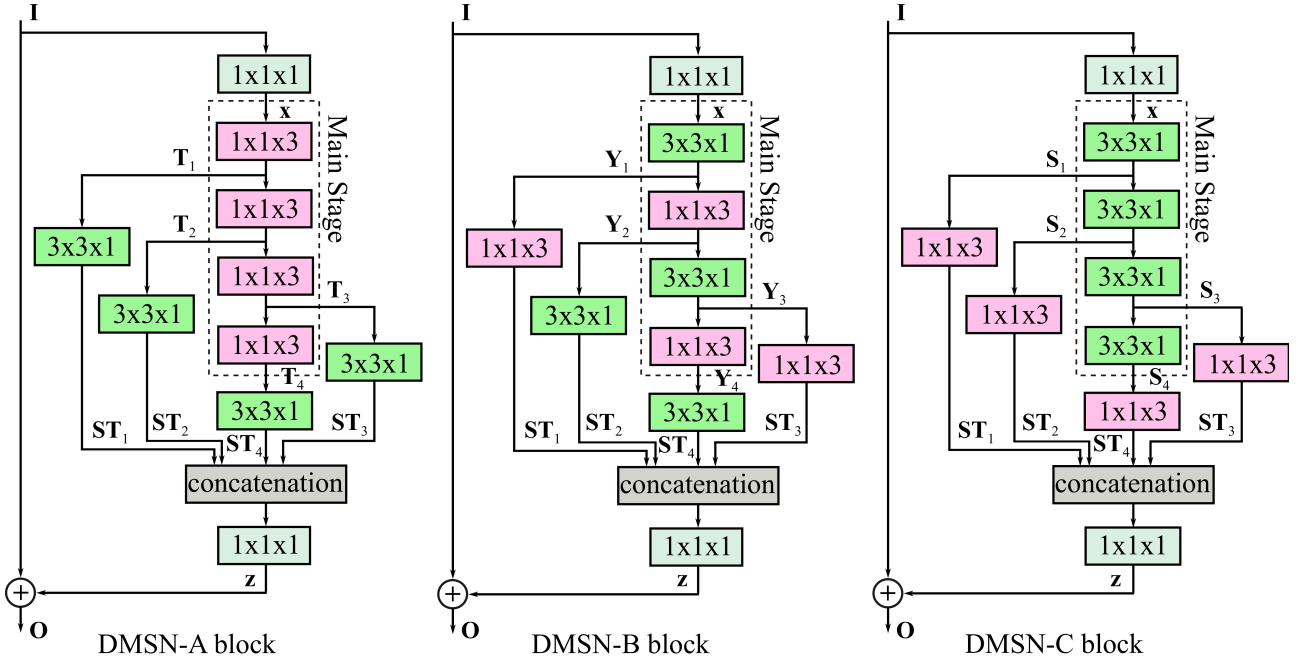


Figure 3: The proposed DMSN building blocks. Pink blocks represent temporal convolutions whereas green blocks represent spatial convolutions.

the spatial receptive field size. The output S_i of each 2D spatial convolution (M_i^s) is defined by:

$$S_i = \begin{cases} M_i^s(\mathbf{x}) & i = 1 \\ M_i^s(S_{i-1}) & 2 \leq i \leq 4 \end{cases} \quad (5)$$

Branches of the Main Stage use $3 \times 1 \times 1$ temporal convolution to produce spatiotemporal features at multiple spatial sizes. The output ST_j of each 1D temporal convolution (M_j^t) can be given by

$$ST_j = M_j^t(S_j) \quad 1 \leq j \leq 4 \quad (6)$$

Furthermore, for DMSN-A, DMSN-B, and DMSN-C blocks, the first element of the Main Stage reduces the number of channels by half in comparison with the number of output channels of the first $1 \times 1 \times 1$ convolution whereas the convolutions in the branches reduce this number by one quarter (i.e., 1 divided by the number of branches).

3.4. DMSN architecture

We construct the DSMN architecture using our three blocks. In this way, our model has structures with diverse capacities favoring the creation of a model that can perform well in different applications. In Table 2, we provide the details of our proposed model. The output feature map is defined as a tensor $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C}$, where T , H , W , and C are the temporal depth, height, width, and number of channels, respectively. The model size and the number of blocks in each layer are defined similarly to ResNet-50. The DMSN blocks are employed in the residual layers (res) and the regression layer outputs a value related to pain or depression score. Moreover, we develop three models which

are named according to the DMSN block they employ, e.g., DMSN-A model uses only DMSN-A blocks. With that, we can understand the contributions of each DMSN block for a given application.

4. Experimental Methodology

4.1. Depression datasets

We conduct experiments on two public benchmarking datasets for depression, called the Audio-Visual Emotion Challenge 2013 and 2014 – AVEC2013 (Valstar et al. (2013)) and AVEC2014 (Valstar et al. (2014)) depression sub-challenge datasets. This AVEC sub-challenge consisted of predicting the depression severity of subjects on Beck Depression Inventory (BDI-II). The severity of depression can be determined in accordance to BDI-II score as follows: minimal (0–13), mild (14–19), moderate (20–28), and severe (29–63). Although there exist other depression datasets such as AVEC2016 (Valstar et al. (2016)), to the best of our knowledge, AVEC2013 and AVEC2014 are the only datasets that provide raw video data.

The AVEC2013 dataset is composed of 150 videos from a group of individuals which the average age is 31.5 years. The individuals were recorded during an interaction with a computer carrying out 14 tasks, including counting from 1 to 10. The dataset is divided into three partitions: training, development, and test subsets. Every subset is comprised of 50 videos, where each video has a BDI-II score as a discrete-value label which indicates the level of depression of an individual. The maximum duration of the videos is 50 minutes, the minimum is 20 minutes, and the average length is 25 minutes.

Table 2

Properties of the proposed DMSN architecture. DMSN-A, DMSN-B, and DMSN-C models are built using only an instance of the DMSN block.

Layer	Output Size	Number of Channels	Structure	Number of Layers
input	16×112×112	3		×1
conv1	16×56×56	64	7×7×7	×1
MaxPool	8×28×28	64	3×3×3	×1
res2	8×28×28	128	DMSN-A DMSN-B DMSN-C	×1
res3	8×14×14	256	DMSN-A DMSN-B DMSN-C DMSN-A	×1
res4	8×7×7	512	DMSN-A DMSN-B DMSN-C	×2
res5	8×4×4	1024	DMSN-A DMSN-B DMSN-C DMSN-A	×1
regression	1×1	spatial AvgPool, FC, AvgPool		



Figure 4: Examples of facial frames from depression and pain datasets.

The AVEC2014 dataset contains videos of individuals performing two tasks: Freeform and Northwind. In the first one, the individuals answer questions like discussing a sad childhood memory. In the second one, individuals read audibly an excerpt from a fable. In total, there are 150 videos of each task with a ground truth label (BDI-II score) for each video. For both tasks, the videos are distributed in three partitions: training, development, and test subsets. The videos have length between 6 and 248 seconds. Samples from both datasets are exhibited in Fig. 4. Due to privacy concerns, all samples of depression shown in this work are blurred.

4.2. Pain datasets

To evaluate the performance of our proposed approach on pain estimation, we conduct experiments on two publicly available datasets: UNBC-McMaster Shoulder Pain Expression Archive (Lucey et al. (2011)), and BioVid Heat Pain (Walter et al. (2013)).

The UNBC-McMaster dataset has been largely employed for pain estimation from facial information. It consists of 200 face videos of 25 individuals with a total number of 48,398 frames. Fig. 4 presents some facial frames from this dataset. Each video is labeled using Prkachin and Solomon Pain Intensity (PSPI) scores in a frame-level fashion on a range of 16 discrete levels ranging from 0 (no pain) to 15 (maximum pain). Since the input of our proposed model is a clip, we follow the works in de Melo et al. (2023); Tavakolian and Hadid (2019); Gnana Praveen et al. (2020); Yu et al. (2019); Rajasekhar et al. (2021); Ruiz et al. (2018); Tavakolian et al. (2020), which define a label for each clip. Specifically, we use the average of the pain intensity of each frame inside the clip as a label. Moreover, since the dataset is highly imbalanced (82.7% of frames have pain score of 0), we adopted the common quantization strategy, which maps the pain levels to 6 ordinal levels as: 0:0, 1:1, 2:2, 3:3, 4-5:4, 6-15:5.

The BioVid Heat Pain dataset contains videos and bio-signals that were acquired during acute heat-induced pain experiments in healthy adults. Pain was induced in four distinct intensities in the right arm of each individual. Although the dataset includes bio-signals such as Skin Conductance Level (SCL), electrocardiogram (ECG), electromyography (EMG), and electroencephalogram (EEG), our experiments only consider Biovid part A which has 8,700 videos of 87 individuals. Each video is labeled with a pain stimulus level

Table 3

Performance of the proposed methods against spatiotemporal models for estimating of depression scores on AVEC2013 and AVEC2014 datasets.

Architecture	AVEC2013		AVEC2014		#Param ↓	FLOPs ↓
	RMSE	MAE	RMSE	MAE		
3D-ResNet Hara et al. (2018)	8.81	6.92	8.40	6.79	63.0M	12.22G
TSN Wang et al. (2019)	8.89	6.21	8.72	6.45	23.5M	16.45G
TSM Lin et al. (2019)	8.89	6.41	8.53	6.29	23.5M	16.45G
P3D Qiu et al. (2017)	8.50	6.24	8.63	6.80	24.9M	8.56G
DMSN-A (Ours)	7.98	6.32	8.13	6.48	19.0M	10.26G
DMSN-B (Ours)	7.92	6.59	7.86	6.24	23.6M	10.83G
DMSN-C (Ours)	7.77	6.14	7.66	6.10	25.9M	11.53G
DMSN (Ours)	7.66	6.14	7.50	5.69	22.1M	11.29G

which ranges from 0 (no pain) to 4 (severe pain). A sample from this dataset is shown in Fig. 4.

4.3. Training of the model

The model analyzes faces that are detected and extracted from video frames of datasets employing MTCNN (Zhang et al. (2016)). Each facial image is resized to form a bounding-box sample with the size of $112 \times 112 \times 3$ that is fed to the model. Usually, datasets for facial expression analysis have a limited amount of training data, which can hinder the generalization ability of a deep architecture. To avoid this problem, deep models are normally pre-trained on large datasets and then fine-tuned on the target dataset. Following the works in de Melo et al. (2022, 2023), our proposed model is pre-trained on the VGGFace2 dataset (Cao et al. (2018)) that contains 3.31 million images of more than 9,000 subjects. In this process, the model is optimized using Stochastic Gradient Descent (SGD) with a momentum of 0.9, weight decay 0.0001, and an initial learning rate of 0.01. The learning rate is divided by 10 after every 10 epochs. The RGB input images are normalized by using the mean channel subtraction. In the fine-tuning process, the ADAM optimization algorithm is adopted. For depression detection task, the initial learning rate is defined as 0.005, then, in the second epoch, this rate is modified to 0.0005. The training is stopped after 3 epochs. For pain estimation task, we define the learning rate equal to 0.001 under two epochs training. In the data augmentation process, we follow the same strategy as in de Melo et al. (2023, 2022).

4.4. Performance measures

For depression detection, an input video from the test subset is segmented into non-overlapped clips of 16 frames. The model generates a depression score for each clip and the median of these values defines the final predicted score for the input video. In order to provide a fair comparison with state-of-the-art methods, we report the performance of the proposed architecture in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are commonly used for depression detection (Zhou et al. (2020); de Melo et al. (2019b); He et al. (2021); de Melo et al. (2023); Al Jazaery and Guo (2021); de Melo

et al. (2019a)). For pain estimation, we perform leave-one-subject-out cross-validation to evaluate the performance of our proposed model. For fair comparison with state-of-the-art methods, the performance of our architecture is measured in terms of Mean Square Error (MSE), and MAE, which are widely used for pain estimation (Tavakolian and Hadid (2019); de Melo et al. (2023); Rodriguez et al. (2017); Yu et al. (2019)). The computational complexity of models is assessed in terms of the number of parameters (memory complexity), and the number of Floating Point Operations (FLOPs) for the processing of a clip (time complexity).

5. Results and Discussion

5.1. Analysis of the DMSN blocks

In order to demonstrate the potential of the proposed DMSN blocks, we generate results using the three models that are named according to the DMSN block they employ, e.g., DMSN-A uses DMSN-A blocks. We also compare these models with our proposed DMSN architecture to show the benefits of using all DMSN blocks. Finally, we compare our architecture in terms of performance and computational complexity with 3D-ResNet (Hara et al. (2018)) and three other efficient spatiotemporal models: TSN (Wang et al. (2019)), TSM (Lin et al. (2019)), and P3D (Qiu et al. (2017)). For fair comparison, all these models follow the same training process that our proposed architecture, i.e., first pre-train on VGGFace2 dataset, then fine-tune on depression or pain datasets.

5.1.1. Depression detection

Table 3 reports the results for our three models on AVEC2013 and AVEC2014 datasets. When compared with DMSN-A, DMSN-B achieves better performance, except for AVEC2013 in terms of MAE. As can be seen, the best performance is obtained by DMSN-C. Regarding the computational complexity, it is possible to observe that DMSN-A employs fewer parameters and requires fewer FLOPs, whereas DMSN-C is more computationally expensive in comparison with DMSN-A, and DMSN-B. Among our three models, DMSN-C provides the best trade-off between performance and computational complexity since this model improves the results with slightly more resources. From these results, we

Table 4

Performance of our proposed approach against spatiotemporal models on UNBC-McMaster and BioVid datasets.

Architecture	UNBC-McMaster		BioVid	
	MSE	MAE	MSE	MAE
3D-ResNet Hara et al. (2018)	0.75	0.56	2.28	1.30
TSN Wang et al. (2019)	0.58	0.53	2.07	1.21
TSM Lin et al. (2019)	0.46	0.49	1.94	1.20
P3D Qiu et al. (2017)	0.67	0.50	2.04	1.23
DMSN-A (Ours)	0.43	0.39	1.68	1.08
DMSN-B (Ours)	0.41	0.37	1.70	1.09
DMSN-C (Ours)	0.44	0.38	1.71	1.09
DMSN (Ours)	0.38	0.35	1.54	1.04

can claim that the DMSN-C block is effective to explore facial expression variations for depression detection.

Table 3 also shows the performance of our proposed DMSN architecture which employs DMSN-A, DMSN-B, and DMSN-C blocks. The use of our three blocks in our architecture provides an improvement of results over DMSN-A, DMSN-B, and DMSN-C models (except for AVEC2013 in terms of MAE where DMSN-C achieves the same result). Observe that DMSN architecture has lower computational costs than the DMSN-C model. Although our architecture has higher FLOPs than DMSN-B and is more expensive than DMSN-A, DMSN significantly improves the performance on depression detection when compared with these two models. These results demonstrate that the diversity of multiscale spatiotemporal features explored by our DMSN architecture enhances the representation for recognition of depressive states.

We also compare our DMSN architecture with the 3D-ResNet, TSN, TSM, and P3D models in Table 3. DMSN improves the results by more than 1.0 in terms of RMSE on AVEC2013 and in terms of MAE on AVEC2014 when compared with 3D-ResNet. DMSN also outperforms TSN, TSM, and P3D where the difference in results on AVEC2014 is significant. DMSN employs fewer parameters than these models and has fewer FLOPs, except for P3D. As indicated by the results, DMSN has the potential to generate efficient spatiotemporal representations for depression detection.

5.1.2. Pain Estimation

Table 4 presents the performance of our DMSN-A, DMSN-B, and DMSN-C models on UNBC-McMaster and BioVid datasets. As can be seen, the three DMSN models achieve comparable results on UNBC-McMaster pain dataset. On the other hand, DMSN-A exhibits a better performance on BioVid pain dataset when compared to DMSN-B and DMSN-C. Given that the DMSN-A model requires fewer parameters and FLOPs, the DMSN-A block, which has the capacity to explore diverse spatiotemporal features at different temporal ranges, can be considered an efficient strategy to capture spatiotemporal variations for pain estimation.

In Table 4, we also show the results of our DMSN architecture for pain estimation. The employment of the three DMSN blocks in our architecture produces better results

Table 5

Performance analysis of our proposed DMSN architecture for different input depths on AVEC2014 and UNBC-McMaster datasets.

Depth	AVEC2014		UNBC-McMaster		FLOPs↓
	RMSE	MAE	MSE	MAE	
8	8.84	6.71	0.49	0.40	5.64G
16	7.50	5.69	0.38	0.35	11.29G
24	7.50	5.80	0.43	0.36	16.93G
32	7.72	5.96	0.38	0.40	22.57G

in comparison with DMSN-A, DMSN-B, and DMSN-C models. Consequently, the construction of our architecture using different strategies to learn multiscale spatiotemporal features favors a performance improvement for an application with greater facial expression variations as in pain estimation, and one with fewer facial variations as in depression detection.

A comparison between our DMSN architecture and 3D-ResNet, TSN, TSM, and P3D models is also presented in Table 4. Compared with P3D, DMSN improves the results by 0.5 and 0.29 in terms of MSE on BioVid and UNBC-McMaster datasets, respectively. In summary, DMSN outperforms these methods and the difference in results is higher on BioVid dataset, indicating that DMSN has good ability to encode facial dynamics for pain estimation.

5.2. Analysis of temporal depth of input

The proposed DMSN architecture is designed to explore a wide range of facial expression variations. Consequently, the temporal depth of input is an important factor in the performance of the model. In Table 5, we provide evaluations considering inputs with 8, 16, 24, 32 frames. Since the pain and depression datasets are composed of similar face videos, and the evaluations involve a long training process, we carry out this analysis on AVEC2014 and UNBC-McMaster datasets. For depression detection, using sequences with 8 frames significantly degrades the performance of our model. In fact, very short sequences increase the level of ambiguity along the depression levels, making harder to generate effective representations. The model sustains the highest levels of performance for a clip size of 16 and 24 frames, and worsens for 32 frames. For pain estimation, the model maintains a comparable level of performance for all sequences employed, but the worst results are obtained using clips with 8 frames. Furthermore, as the clip size increases, the model requires more FLOPs to generate an output.

5.3. Analysis of multiscale spatiotemporal ability

Given that facial dynamics comprise different spatiotemporal variations, it is essential that our architecture has a multiscale spatiotemporal representation ability to encode such variations. To demonstrate this capacity, we evaluate the performance of our proposed architecture by changing the number of branches in the Main Stage sub-block. To maintain a comparable computational complexity, when the number of branches is reduced, we increase the number of

Table 6

Evaluation of our DMSN architecture considering different number of branches in the Main Stage sub-block on AVEC2014 and UNBC-McMaster datasets.

Number of branches	AVEC2014		UNBC-McMaster		#Param↓	FLOPs↓
	RMSE	MAE	MSE	MAE		
2	8.45	6.64	0.63	0.52	18.0M	9.64G
3	7.71	6.08	0.45	0.42	20.1M	10.48G
4	7.50	5.69	0.38	0.35	22.1M	11.29G

Table 7

Performance of proposed and state-of-the-art methods for estimating of depression scores on AVEC datasets.

Architecture	AVEC2013		AVEC2014		#Param ↓
	RMSE	MAE	RMSE	MAE	
Baseline-AVEC2013 Valstar et al. (2013)	13.61	10.88	-	-	-
Baseline-AVEC2014 Valstar et al. (2014)	-	-	10.86	8.86	-
MHH+LBP Meng et al. (2013)	11.19	9.14	-	-	-
LPQ+Geo+CCA Kaya and Salah (2014)	9.72	7.86	-	-	-
Two-stream GoogLeNet Zhu et al. (2018)	9.82	7.58	9.55	7.47	-
Two C3D Al Jazaery and Guo (2021)	9.28	7.37	9.20	7.22	≈64.2M
VLDN+CNN+Bi-LSTM Uddin et al. (2022)	8.93	7.04	8.78	6.86	-
Two C3D de Melo et al. (2019a)	8.26	6.40	8.31	6.59	≈64.2M
VGG-16+FDHH Jan et al. (2018)	-	-	8.04	6.68	≈138.0M
DTL Kang et al. (2017)	-	-	9.43	7.74	-
ResNet-50+pooling Zhou et al. (2019)	-	-	8.43	6.37	≈23.5M
Four ResNet-50 Zhou et al. (2020)	8.28	6.20	8.39	6.21	≈94.0M
ResNet-50 de Melo et al. (2019b)	8.25	6.30	8.23	6.15	≈23.5M
Behavior signals Song et al. (2022)	8.10	6.16	8.30*	6.78*	-
DLGA-CNN He et al. (2021)	8.39	6.59	8.30	6.51	-
DJ-LDML Zhou et al. (2022)	8.37	6.63	8.30	6.59	≈298M
DeepFusion Chen et al. (2021)	-	-	8.13	6.16	≈47M
Two-stream ResNet-50 de Melo et al. (2020)	7.97	5.96	7.94	6.20	≈47M
MSN de Melo et al. (2022)	7.90	5.98	7.61	5.82	≈77.7M
MDN de Melo et al. (2023)	7.55	6.24	7.65	6.06	≈52M
DMSN (Ours)	7.66	6.14	7.50	5.69	22.1M

* Results of the method for Freeform task.

channels in the branches of the Main Stage sub-block. As seen in Table 6, for depression detection and pain estimation, when more spatiotemporal ranges are explored (i.e., increasing the number of branches), the performance of DMSN improves, indicating a boost in the ability of encoding facial variations. It is worth noting that by increasing the number of branches in the Main Stage sub-block, the architecture not only enhances the capacity of exploring spatiotemporal features in different ranges, but it also increases the diversity of this exploration, since DMSN employs DMSN-A, DMSN-B, and DMSN-C blocks, which use different strategies to learn multiscale spatiotemporal features.

5.4. Comparison with state-of-the-art

5.4.1. Depression detection

Table 7 compares the performance of our proposed architecture with state-of-the-art methods on AVEC2013 and AVEC2014 depression datasets. DMSN outperforms the method based on LPQ features (Kaya and Salah (2014)) and other related descriptors (Valstar et al. (2013, 2014); Meng et al. (2013)). The methods in de Melo et al. (2019b); He et al. (2021); Jan et al. (2018); Kang et al. (2017); Zhou

et al. (2019, 2020); Uddin et al. (2022) are based on 2D-CNNs followed by an aggregation technique. The methods in de Melo et al. (2019a); Al Jazaery and Guo (2021); Zhou et al. (2022) employ 3D-CNNs to explore spatiotemporal information. The authors in de Melo et al. (2020); Zhu et al. (2018); Chen et al. (2021) infer depressive states by using two-stream networks. Our DMSN outperforms these methods (except for the method in de Melo et al. (2020) in terms of MAE on AVEC2013, but this approach employs 24.9M more parameters). These results confirm findings in de Melo et al. (2022, 2023); Song et al. (2022), which underscore the importance of a multiscale approach for facial depression recognition. We also observe that DMSN achieves better results than the method in Song et al. (2022), which explores behavioral primitives (facial action units, head pose, and gaze directions). When compared with MSN (de Melo et al. (2022)) and MDN (de Melo et al. (2023)), DMSN outperforms both models on AVEC2014, and achieves competitive results on AVEC2013, while requiring 3.51× and 2.35× fewer parameters than MSN and MDN, respectively. This comparison with MSN is important because it demonstrates

Table 8

Performance of our proposed DMSN architecture against state-of-the-art methods on UNBC-McMaster dataset.

Architecture	MSE	MAE	PCC	#Param ↓
RVR+DCT Kaltwang et al. (2012)	1.39	-	0.59	-
HoT Florea et al. (2014)	1.21	-	0.53	-
OSVR Zhao et al. (2016)	-	0.81	0.60	-
RCNN Zhou et al. (2016)	1.54	-	0.64	-
VGG11+LSTM Yu et al. (2019)	1.22	0.58	0.40	≈133M
VGG16+LSTM Rodriguez et al. (2017)	0.74	0.5	0.78	≈138M
C3D Tavakolian and Hadid (2019)	0.71	-	0.81	≈32M
I3D Gnana Praveen et al. (2020)	-	0.80	0.44	≈13M
HybNet Huang et al. (2022)	0.76	0.40	0.82	≈35.1M
MDN de Melo et al. (2023)	0.68	0.42	-	≈52M
SCN Tavakolian and Hadid (2019)	0.32	-	0.92	≈586.8M
DMSN (Ours)	0.38	0.35	0.83	22.1M

Table 9

Performance of our proposed DMSN architecture against state-of-the-art methods on BioVid dataset.

Method	Modality	MSE	MAE
I3D Rajasekhar et al. (2021)	Video	-	1.42
Fusion Kächele et al. (2017)	Multimodal	1.16 (RMSE)	0.99
DMSN (Ours)	Video	1.54	1.04

that our approach is more efficient than the use of parallel 3D convolutions. In other words, the decomposition of the exploration of diverse multiscale spatiotemporal features facilitates the learning of depression patterns and minimizes the problems with overfitting. Overall, these results show that our DMSN architecture can provide a cost-effective solution for depression detection.

5.4.2. Pain estimation

Table 8 compares the performance of our proposed architecture with state-of-the-art methods on UNBC-McMaster dataset. Since this dataset is highly imbalanced, we additionally report Pearson Correlation Coefficient (PCC). As can be seen, our DMSN outperforms different schemes for pain expression recognition. For instance, DMSN achieves better results than the method in Rodriguez et al. (2017), which uses VGG-16 architecture and LSTM, while requiring around 6.2 times fewer parameters. The comparison with the SCN method (Tavakolian and Hadid (2019)) is interesting because the basic block of this architecture is composed of parallel 3D convolutions with diverse temporal depths to explore multiscale spatiotemporal information. DMSN presents similar performance as this method with significant reduction of parameters (DMSN has around 26.55× fewer parameters). These results indicate that our DMSN architecture is also an efficient option for pain estimation.

Table 9 compares our DMSN architecture with state-of-the-art methods on BioVid dataset. DMSN outperforms the method in Rajasekhar et al. (2021) which also explores facial expressions variations from videos. In Kächele et al. (2017), the authors explore diverse features from ECG, EMG, and SCL as well as face videos. As we can see, DMSN obtains comparable results, demonstrating that facial expression analysis can provide essential information for the estimation of pain intensities.

Table 10

Performance of the proposed method in cross-dataset setting.

Training set	Test set	RMSE	MAE	MSE
AVEC2013	AVEC2014	7.78	6.18	-
AVEC2014	AVEC2013	8.36	6.62	-
UNBC	BioVid	-	1.19	1.92
BioVid	UNBC	-	0.63	0.91
AVEC2013	UNBC	-	0.62	0.92
AVEC2014	UNBC	-	0.61	0.90
AVEC2013	BioVid	-	1.19	1.95
AVEC2014	BioVid	-	1.21	1.99
UNBC	AVEC2013	11.13	9.41	-
UNBC	AVEC2014	11.24	9.40	-
BioVid	AVEC2013	11.10	9.27	-
BioVid	AVEC2014	10.93	9.13	-

5.5. Cross-corpus database analysis

Depression and Pain states are closely related to each other (Garcia-Cebrian et al. (2006); Von Korff and Simon (1996)). For example, a person suffering from depression may experience headache, backache, and stomach ache (Borgman et al. (2020); Stahl (2002)). This fact motivates the study of the applicability of depression/pain features, which are extracted by our DMSN architecture, to the pain/depression recognition task. To carry out this study, we define that the source and target databases belong to different tasks (e.g., AVEC2013 is the source database, and UNBC-McMaster is the target database). In this case, since the labels of pain and depression datasets are different, we replace the regression layer of DMSN to properly evaluate the representations generated by the model. Moreover, we assess the generalization capabilities of our DMSN by performing cross-database experiments.

In Table 10, we present the results of these experiments. When the evaluations are performed on the same task (e.g., depression detection), the model achieves reasonable results, indicating a robust representation for facial videos. In the results between tasks, we can observe that the representations learned on depression datasets allow DMSN to achieve good results on pain datasets. On the other hand, when DMSN is trained on pain datasets and then evaluated on depression detection task, there is a higher degradation in performance. One reason for this result is the high level of ambiguity in depressive states which makes it difficult to directly apply the features of other applications.

5.6. Qualitative results

To interpret the performance differences for depression detection between our DMSN architecture and DMSN-A, DMSN-B, DMSN-C models as well as P3D, we present the class activation maps (CAMs) employing the Grad-CAM method (Selvaraju et al. (2017)). In the visualizations of Fig. 5, lighter colors represent those regions that are most relevant for a model's predictions. Considering the most activated regions, the models appear to explore the eyes and mouth regions. In fact, these regions convey important information about depressive states. As we can see, our

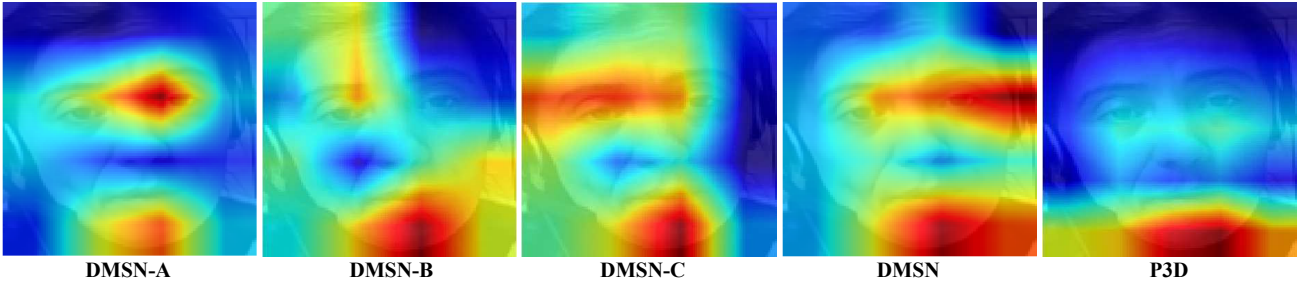


Figure 5: Example of the CAMs showing the facial regions activated by our proposed DMSN models and P3D on a facial image from the AVEC2014 dataset.

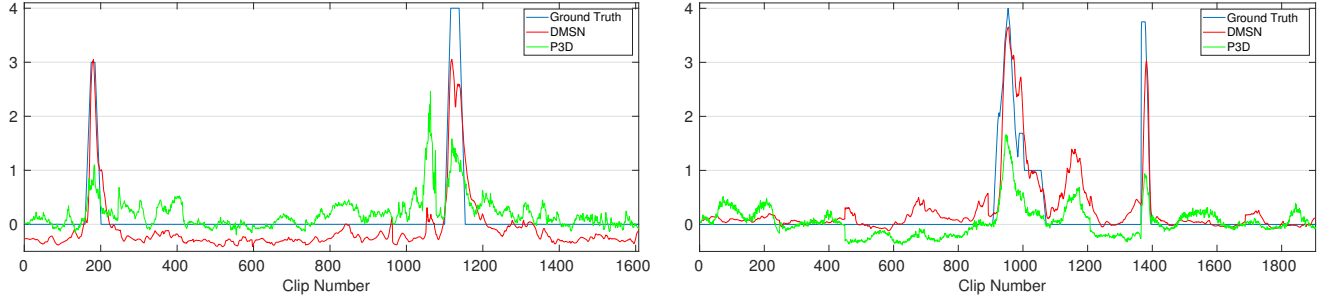


Figure 6: Performance analysis of our proposed DMSN architecture and P3D on samples of two individuals from UNBC-McMaster dataset.

approach is more effective in exploring such areas than P3D. In comparison with DMSN-A, DMSN-B, and DMSN-C, DMSN seems to be more successful in capturing face expression variations from these areas. We understand that this capacity of DMSN is a decisive factor for the good performance in depression detection.

Fig. 6 shows the effectiveness of our approach for pain estimation by comparing the predictions of our architecture with that of P3D and ground truth. It can be observed that our architecture can satisfactorily identify the occurrence of pain, which is an important characteristic for clinical application, whereas P3D has lower accuracy. DMSN presents a better performance than P3D in recognizing changes of pain levels, which is due to a better multiscale spatiotemporal ability of DMSN. In general, our architecture has a good ability to follow the variations of pain levels, meaning that DMSN is effectively modeling transitions in facial pain expressions.

5.7. Limitation of the DMSN Architecture

The previous experiments showed that our DMSN architecture is a cost-effective solution for facial analysis applications, which are characterized for a range from fewer facial variations over time, as in depression detection, to greater variations, as in pain estimation. However, the deployment of our model on some resource-constrained device for real-time applications can be challenging. To exemplify this point, let us consider the deployment of our DMSN on iPhone 6. On this mobile device, studies show that a model that requires 25.6M parameters and 4.01G FLOPs has an inference latency of 2386 ms (Almadan and Rattani (2021)) (for comparison, the same model has a latency of 3 ms on

iPhone 12 (Li et al. (2022b))). One reason for this difference is that the GPU performance of iPhone 12 is significantly better than the one on iPhone 6). As our DMSN has 22.1M parameters and 11.29G FLOPs, its deployment on iPhone 6 would also produce a high inference latency, like that model, which would be a problem for a real-time task.

6. Conclusion

In this paper, we propose a structure called Decomposed Multiscale Spatiotemporal Network (DMSN) to decompose the exploration of multiscale spatiotemporal features from facial expressions in videos. Three variants of the DMSN block are introduced, which employ different strategies to effectively and efficiently capture facial dynamics. We design our DMSN architecture using these blocks to explore a variety of multiscale spatiotemporal features, which favors the adaptation to different facial behaviors. In our extensive experiments on AVEC2013 and AVEC2014 depression datasets, and UNBC-McMaster and BioVid pain datasets, we show that exploring the spatiotemporal information at multiple spatial sizes (DMSN-C block) is effective for depression detection, whereas capturing spatiotemporal features at multiple temporal ranges (DMSN-A block) is efficient for pain estimation. We also show that our architecture achieves competitive performance against state-of-the-art approaches for depression and pain expression detection, yet requires significantly fewer model parameters. Moreover, we demonstrate that depression features are more useful for pain estimation than pain features are for depression detection. In future work, we plan to analyze the performance of our DMSN architecture for stress detection, which is another

important facial analysis task, as well as to investigate the applicability of stress features for pain estimation and depression detection.

References

- Al Jazaery, M., Guo, G., 2021. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing* 12, 262–268.
- Almadan, A., Rattani, A., 2021. Towards on-device face recognition in body-worn cameras, in: *IEEE International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6.
- Amelio, A., Bonifazi, G., Cauteruccio, F., Corradini, E., Marchetti, M., Ursino, D., Virgili, L., 2023. Representation and compression of residual neural networks through a multilayer network based approach. *Expert Systems with Applications* 215, 119391.
- American Psychiatric Association, 2013. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Publishing.
- Borgman, S., Ericsson, I., Clausson, E.K., Garmy, P., 2020. The relationship between reported pain and depressive symptoms among adolescents. *The Journal of School Nursing* 36, 87–93.
- Bostwick, J.M., Rackley, S., 2012. Recognizing mimics of depression: the '8 ds'. *Current Psychiatry* 11, 31–36.
- Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. Vggface2: A dataset for recognising faces across pose and age, in: *FG*, pp. 67–74.
- Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: *CVPR*, pp. 4724–4733.
- Chen, Q., Chaturvedi, I., Ji, S., Cambria, E., 2021. Sequential fusion of facial appearance and dynamics for depression recognition. *Pattern Recognition Letters* 150, 115–121.
- Cohn, J.F., Kruez, T.S., Matthews, I., Yang, Y., Nguyen, M.H., Padilla, M.T., Zhou, F., De la Torre, F., 2009. Detecting depression from facial actions and vocal prosody, in: *ACII*, pp. 1–7.
- Downie, W., Leatham, P., Rhind, V., Wright, V., Branco, J., Anderson, J., 1978. Studies with pain rating scales. *Annals of the Rheumatic Diseases* 37, 378–381.
- Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition, in: *ICCV*, pp. 6202–6211.
- Florea, C., Florea, L., Vertan, C., 2014. Learning pain from emotion: transferred hot data representation for pain intensity estimation, in: *ECCVW*, pp. 778–790.
- Garcia-Cebrian, A., Gandhi, P., Demyttenaere, K., Peveler, R., 2006. The association of depression and painful physical symptoms—a review of the european literature. *European Psychiatry* 21, 379–388.
- Gnana Praveen, R., Granger, E., Cardinal, P., 2020. Deep weakly supervised domain adaptation for pain localization in videos, in: *FG*, pp. 473–480.
- Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., Narayanan, S., 2014. Multimodal prediction of affective dimensions and depression in human-computer interactions, in: *AVEC'14*, pp. 33–40.
- Hara, K., Kataoka, H., Satoh, Y., 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: *CVPR*, pp. 6546–6555.
- Hassan, T., Seuß, D., Wollenberg, J., Weitz, K., Kunz, M., Lautenbacher, S., Garbas, J.U., Schmid, U., 2021. Automatic detection of pain from facial expressions: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1815–1831.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
- He, L., Chan, J.C.W., Wang, Z., 2021. Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing* 422, 165–175.
- Huang, Y., Qing, L., Xu, S., Wang, L., Peng, Y., 2022. Hybnet: a hybrid network structure for pain intensity estimation. *The Visual Computer* 38, 871–882.
- Jaiswal, S., Valstar, M.F., Gillott, A., Daley, D., 2017. Automatic detection of adhd and asd from expressive behaviour in rgbd data, in: *FG*, pp. 762–769.
- Jan, A., Meng, H., Gaus, Y.F.B.A., Zhang, F., 2018. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems* 10, 668–680.
- Kächele, M., Amirian, M., Thiam, P., Werner, P., Walter, S., Palm, G., Schwenker, F., 2017. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems* 8, 71–83.
- Kaltwang, S., Rudovic, O., Pantic, M., 2012. Continuous pain intensity estimation from facial expressions, in: *ISVC*, pp. 368–377.
- Kang, Y., Jiang, X., Yin, Y., Shang, Y., Zhou, X., 2017. Deep transformation learning for depression diagnosis from facial images, in: *CCBR*, pp. 13–22.
- Kappesser, J., Williams, A.C.d.C., 2010. Pain estimation: Asking the right questions. *Pain* 148, 184–187.
- Kaya, H., Salah, A.A., 2014. Eyes whisper depression: A cca based multimodal approach, in: *MM'14*, p. 961–964.
- Khan, R.A., Meyer, A., Konik, H., Bouakaz, S., 2013. Pain detection through shape and appearance features, in: *ICME*, pp. 1–6.
- Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M., 2012. Activity forecasting, in: *ECCV*, pp. 201–214.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: *NeurIPS*.
- Lesage, F.X., Berjot, S., Deschamps, F., 2012. Clinical stress assessment using a visual analogue scale. *Occupational Medicine* 62, 600–605.
- Li, H., Zeng, N., Wu, P., Clawson, K., 2022a. Cov-net: A computer-aided diagnosis method for recognizing covid-19 from chest x-ray images via machine vision. *Expert Systems with Applications* 207, 118029.
- Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J., 2022b. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems* 35, 12934–12949.
- Lin, J., Gan, C., Han, S., 2019. Tsm: Temporal shift module for efficient video understanding, in: *ICCV*, pp. 7082–7092.
- Lopez, M.B., del Blanco, C.R., Garcia, N., 2017. Detecting exercise-induced fatigue using thermal imaging and deep learning, in: *Proc. International Conference on Image Processing Theory, Tools and Applications*, pp. 1–6.
- Lucas, G.M., Gratch, J., Scherer, S., Boberg, J., Stratou, G., 2015. Towards an affective interface for assessment of psychological distress, in: *ACII*, pp. 539–545.
- Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I., 2011. Painful data: The unbc-mcmaster shoulder pain expression archive database, in: *FG*, pp. 57–64.
- de Melo, W.C., Granger, E., Hadid, A., 2019a. Combining global and local convolutional 3d networks for detecting depression from facial expressions, in: *FG*, pp. 1–8.
- de Melo, W.C., Granger, E., Hadid, A., 2019b. Depression detection based on deep distribution learning, in: *ICIP*, pp. 4544–4548.
- de Melo, W.C., Granger, E., Hadid, A., 2022. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE Transactions on Affective Computing* 13, 1581–1592.
- de Melo, W.C., Granger, E., Lopez, M.B., 2020. Encoding temporal information for automatic depression recognition from facial analysis, in: *ICASSP*, pp. 1080–1084.
- de Melo, W.C., Granger, E., López, M.B., 2023. Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Transactions on Affective Computing* 14, 578–590.
- Meng, H., Huang, D., Wang, H., Yang, H., AI-Shuraifi, M., Wang, Y., 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression, in: *AVEC'13*, p. 21–30.
- Mitchell, A.J., Vaze, A., Rao, S., 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 374, 609–619.
- Pampouchidou, A., Simos, P.G., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M., Tsiknakis, M., 2019. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing* 10, 445–470.
- Purebl, G., et al., 2015. Depression, suicide prevention and e-health: situation analysis and recommendations for action. *The Joint Action on Mental Health and Well-being*.

- Qiu, Z., Yao, T., Mei, T., 2017. Learning spatio-temporal representation with pseudo-3d residual networks, in: ICCV, pp. 5534–5542.
- Rajasekhar, G.P., Granger, E., Cardinal, P., 2021. Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos. *Image and Vision Computing* 110, 1–10.
- Rodriguez, P., Cucurull, G., González, J., Gonfau, J.M., Nasrollahi, K., Moeslund, T.B., Roca, F.X., 2017. Deep pain: exploiting long short-term memory networks for facial expression classification. *IEEE Transactions on Cybernetics*, 1–12.
- Ruiz, A., Rudovic, O., Binefa, X., Pantic, M., 2018. Multi-instance dynamic ordinal random fields for weakly supervised facial behavior analysis. *IEEE Transactions on Image Processing* 27, 3969–3982.
- Schelde, J.T.M., 1998. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease* 186, 133–140.
- Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., Morency, L.P., 2013. Automatic behavior descriptors for psychological disorder analysis, in: FG, pp. 1–8.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: ICCV, pp. 618–626.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Song, S., Jaiswal, S., Shen, L., Valstar, M., 2022. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing* 13, 829–844.
- Stahl, S.M., 2002. Does depression hurt? *Journal of Clinical Psychiatry* 63, 273–274.
- Tavakolian, M., Bordallo Lopez, M., Liu, L., 2020. Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recognition Letters* 140, 26–33.
- Tavakolian, M., Hadid, A., 2019. A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics. *International Journal of Computer Vision* 127, 1413–1425.
- Thevenot, J., López, M.B., Hadid, A., 2018. A survey on computer vision for assistive medical diagnosis from faces. *IEEE Journal of Biomedical and Health Informatics* 22, 1497–1511.
- Thiam, P., Kessler, V., Walter, S., Palm, G., Schwenker, F., 2016. Audio-visual recognition of pain intensity, in: MPRSS, pp. 110–126.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: ICCV, pp. 4489–4497.
- Trémeau, F., Malaspina, D., Duval, F., Corrêa, H., Hager-Budny, M., Coin-Bariou, L., Macher, J.P., Gorman, J.M., 2005. Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects. *American Journal of Psychiatry* 162, 92–101.
- Trivedi, M.H., 2004. The link between depression and physical symptoms. *Primary care Companion to the Journal of Clinical Psychiatry* 6, 12–16.
- Uddin, M.A., Joolee, J.B., Lee, Y.K., 2022. Depression level prediction using deep spatiotemporal features and multilayer bi-lstm. *IEEE Transactions on Affective Computing* 13, 864–870.
- Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M., 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge, in: AVEC'16, p. 3–10.
- Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M., 2014. AVEC 2014: 3d dimensional affect and depression recognition challenge, in: AVEC'14, p. 3–10.
- Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M., 2013. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge, in: AVEC'13, p. 3–10.
- Von Korff, M., Simon, G., 1996. The relationship between pain and depression. *The British Journal of Psychiatry* 168, 101–108.
- Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H.C., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A.O., Moreira da Silva, G., 2013. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system, in: CYBCO, pp. 128–131.
- Wang, J., Sun, H., 2018. Pain intensity estimation using deep spatiotemporal and handcrafted features. *IEICE Transactions on Information and Systems* 101, 1572–1580.
- Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2019. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2740–2755.
- Werner, P., Lopez-Martinez, D., Walter, S., Al-Hamadi, A., Gruss, S., Picard, R.W., 2022. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing* 13, 530–552.
- Wu, P., Wang, Z., Zheng, B., Li, H., Alsaadi, F.E., Zeng, N., 2023. Aggn: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion. *Computers in Biology and Medicine* 152, 106457.
- Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K., 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: ECCV, pp. 305–321.
- Yu, J., Kurihara, T., Zhan, S., 2019. Frame by frame pain estimation using locally spatial attention learning, in: IbPRIA, pp. 229–238.
- Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M., 2018. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* 273, 643–649.
- Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 1499–1503.
- Zhao, R., Gan, Q., Wang, S., Ji, Q., 2016. Facial expression intensity estimation using ordinal information, in: CVPR, pp. 3466–3474.
- Zhou, J., Hong, X., Su, F., Zhao, G., 2016. Recurrent convolutional neural network regression for continuous pain intensity estimation in video, in: CVPRW, pp. 1535–1543.
- Zhou, X., Huang, P., Liu, H., Niu, S., 2019. Learning content-adaptive feature pooling for facial depression recognition in videos. *Electronics Letters* 55, 648–650.
- Zhou, X., Jin, K., Shang, Y., Guo, G., 2020. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing* 11, 542–552.
- Zhou, X., Wei, Z., Xu, M., Qu, S., Guo, G., 2022. Facial depression recognition by deep joint label distribution and metric learning. *IEEE Transactions on Affective Computing* 13, 1605–1618.
- Zhu, Y., Shang, Y., Shao, Z., Guo, G., 2018. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing* 9, 578–584.