UTRECHT UNIVERSITY
Faculty of Science
Department of Information and Computing Sciences
MSc Artificial Intelligence

VISION TRANSFORMERS FOR PAIN RECOGNITION ON THERMAL IMAGE FRAMES

A THESIS BY

David Pantophlet
7002025

Project supervisor dr. Itır Önal Ertuğrul
Daily supervisor
Second examiner dr. ir. R.W. (Ronald) Poppe

Abstract

Pain remains a phenomenon that is not fully understood scientifically, even though poorly managed pain severely impacts the individuals involved. Therefore, a valid and reliable pain assessment is necessary to manage pain properly. This study investigates the effectiveness of vision transformers in detecting pain from thermal face video frames. In doing so it looks at the effect of incorporating temporal sequences and extracting regions of interest (ROI). Vision transformers (ViT) and video vision transformers (ViViT) models are employed for this analysis. We found that both models can discern pain distinctions, but the models overfit quite easily. However, we did find that the ViViT model trained on sequences of entire thermal images (ViViT whole) shows promise, outperforming other configurations with 60.5% accuracy. ViT ROI was found more effective than ViT whole and ViViT ROI, highlighting the benefit of ROI extraction in the case of single-image pain prediction on thermal images.

Preface

Dear reader, First of all, thank you for taking the time to read this work.

I would like to express my deepest gratitude to Dr. Önal Ertuğrul. Especially for her guidance and patience throughout this research. Life during the writing of this research has not been boring, so to speak. But Professor Önal expertise in the field and understanding of me personally have been instrumental in shaping the direction and quality of this work and gave me the space and motivation to continue.

I would also like to thank my partner for her unwavering support throughout this process while simultaneously carrying our beautiful son into the world.

Disclaimer: This thesis does not contain any studies with human participants or animals performed by the author. Data used in this study were previously collected. The original owner of the data retains ownership of the data during and after the completion of this thesis.

David Pantophlet

Table of Contents

1. Introduction

Even though most people have learned what pain feels like in their earlier parts of life, it is not fully understood scientifically. One accepted definition of pain is “an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage” (John J. Bonica, 1979). However, there is an ongoing discussion about updating the definition (Werner et al., 2022). Also, pain is still often poorly managed (Werner et al., 2022), while pain is the primary reason that makes people seek medical attention (Mäntyselkä et al., 2001), and chronic pain costs the US economy alone more than cancer, heart disease and HIV combined (Lynch, 2011). Untreated pain can lead to chronic pain syndrome, often accompanied by decreased mobility, impaired immunity, decreased concentration, anorexia, and sleep disturbances. Moreover, wrong treatment may lead to problems and risks for the patients. Hence, poorly managed pain severely impacts the individuals involved and our society as a whole.

Therefore, a valid and reliable pain assessment is necessary to manage pain properly. This will help provide comfort and prevent immediate and long-lasting consequences that harm the person’s overall health (Mitchell and Boss, 2002).

Currently, in practice, self-reporting is the standard way in which people get diagnosed with pain. Self-report is done through conscious verbal or physical communication, such as spoken or written language or even pointing. This is considered the most valid way of assessing pain in an individual.

However, for people who cannot speak or adequately express themselves, a doctor or nurse is left to assess the pain in individuals. This is done through guidelines and tools such as the Behavioral Pain Scale (BPS) (Dehghani et al., 2014). Typically, these tools use behavioural signals, such as facial expressions, sounds or body movements, that someone is in pain (Herr et al., 2011).

These tools provide guidelines to try to create objective assessment tools. However, pain assessment through observation by a third person remains incredibly difficult due to the observer’s subjective biases and mistakes. Some studies show that pain is underestimated for certain ethnicities (Mende-Siedlecki et al., 2021). Other studies show that nursing staff systematically underestimates pain in patients (Kappesser and Williams, 2010), where (Achterberg et al., 2013) show that there is a lack of effective assessment and treatment of pain in people with dementia.

Besides these issues, the nursing staff in a hospital cannot monitor every patient continuously. All of these factors have led to research into the development of automated pain detection techniques. Most research has looked into facial expressions for automated pain detection because they are a strong indicator of someone's pain state (Craig, 1992). The Facial Action coding system (FACS), developed by P. Ekman et al. (Ekman and Friesen, 1978), is widely used to annotate pain data. The atomic building blocks of the FACS are action units (AU), which represent specific movements of the face, such as raised eyebrows or lip tightening. Several automated pain detection frameworks have used FACS to train or determine pain, or even Prkachin-Solomon Pain Intensity scores (PSPI) (Prkachin and Solomon, 2008) for individuals (Werner et al., 2022). FACS typically requires a recording of a visible face, which means that these images or videos are typically recorded in the visible light spectrum (e.g. RGB images). However, pain detection using RGB images suffers from some limitations, such as lighting issues, where they do not function when the lighting is too dim or in complete darkness. They do not generalise well between age groups or other groups due to facial variations caused by age, gender, race etc. (Hassan, 2017). Other issues arise when the subject cannot move their face or does not wish to be recorded in general because of privacy concerns.

Therefore research has looked into other physiological responses of the human body to specific emotions, including pain and found that pain and other emotions have a direct effect on the autonomic nervous system (ANS) (Hamunen et al., 2012) (Leone et al., 2006). The ANS can regulate numerous physiological processes, such as heart rate, breathing rate, digestion, blood pressure, body temperature, and metabolism. This means that human behaviours and affects can be reflected from physiological signals (Alqudah, 2021). Some research looked at multi-modal solutions that combine different physiological signals to see if someone is in pain. However, measuring heart rate, blood pressure, and body temperature requires invasive hardware. Thermal data of the face is another technique that got quite some attention, because of its ability to detect pain in humans (Abd Latif et al., 2015). A solution based solely on thermal imaging results is a non-invasive method that deals with many of the limitations that facial images have. For example, it can perform very well under different lighting issues and even in complete darkness.

Some promising work has been done proving that emotion detection solely on thermal image data is possible and, in some cases, even can outperform RGB data.Alqudah (2021) (Ordun et al., 2020). However, a limited amount of work has been done in investigating pain detection, specifically thermal images, especially state-of-the-art techniques like transformer models. The work done using transformer models for emotion prediction shows state-of-the-art performance on this task (Ma et al., 2021) (Arnab et al., 2021). We will investigate the performance of vision transformers on automated pain detection.

1.1 Contributions of the thesis

This work aims to contribute to the development of automated pain detection systems using a minimally invasive technique by answering the question: To what extent can a vision transformer architecture detect pain using the thermal video frames of the face? In answering this question, we will also be looking at the effect of only focusing on the Regions of interest (ROI) as input modality and the effect of incorporating small windows of temporal data. We will use ViT (Dosovitskiy et al., 2020) to process the thermal image frames and ViViT (Arnab et al., 2021) for the small video section. In the end, the main contributions are the use of a transformer model for pain detection on thermal data and testing the effectiveness of inducing inductive bias in a transformer model.

1.2 Research questions

In summary, this project aims to answer the following research questions:

Research question: To what extent can a vision transformer architecture detect pain using thermal video frames of the face?
To answer this question we finetuned ViT and ViViT model on single thermal image frames and sequences of thermal frames of the face respectively and compared their results.

Sub-question 1: How does the use of pain-specific ROI extraction influence the performance of the vision transformers ViT and ViViT? To answer this question we extracted regions of interest for every frame and used that as input for ViT and ViViT, whereafter we compared the results to the same models but with whole thermal images.

Sub-question 2: How does the use of temporal information influence the ability of a transformer model in the pain detection task? To answer this question we compared ViT and ViViT in their ability to predict pain from thermal frames. For ViViT we used two different sequential lengths as input and compared the results.

This work found that, While both models demonstrated the capacity to learn these distinctions, the overall performance of the models left room for improvement. Among the tested configurations, the most promising results were achieved by the ViViT model trained on a sequence of ten frames of whole thermal images (ViViT whole).

In terms of accuracy, the ViViT whole model outperformed the other configurations, achieving an accuracy of 0.605, correctly classifying pain or non-pain instances 60.5% of the time. This model also exhibited superior recall (0.540) in identifying pain cases and precision (0.565) in minimizing false positives. Additionally, the F1 score (0.554) underscored its superior performance, balancing precision and recall effectively.

Comparing models that used regions of interest (ViT ROI and ViViT ROI) versus whole images (ViT whole and ViViT whole), the ViT ROI model outperformed ViT whole, suggesting the benefit of incorporating pain-specific regions of interest in pain detection. However, ViViT whole surpassed ViViT ROI across all metrics, emphasizing the value of incorporating temporal information when more data is available for the task.

In the comparison between models using single images versus temporal sequences, ViT ROI exhibited an accuracy of 0.567, recall of 0.499, precision of 0.533, and an F1 score of 0.547. In contrast, ViViT ROI achieved an accuracy of 0.539, recall of 0.443, precision of 0.453, and an F1 score of 0.499. Notably, ViT ROI slightly outperformed ViViT ROI across all metrics, suggesting that incorporating temporal information did not yield improved results in pain detection when using ROI images as input. Furthermore, the addition of temporal information improved the performance of the ViViT whole model compared to ViT whole highlighting the important role of temporal data in enhancing pain detection from thermal video frames.

The rest of this paper is divided into related work where we deep dive into previous methods and related work to pain detection on facial images. Then we outline the methodology used in this research, including the dataset hardware preprocessing and models and techniques we used to accomplish our result. Thereafter we showcase the results in the experimental results section. Then we will discuss the results, limitations and future work in the discussion section whereafter we end with the conclusions of this paper.

2. Related Work

2.1 Automated pain detection using RGB data

An automated pain detection system usually consists of a pipeline where one has to make several decisions for each pipeline step. First of all, one has to detect the face and normalize the face to the frontal view. From this facial image, a set of features has to be extracted. These features can be handcrafted or learned and result in geometrical features, textural features, or a combination of the two. Textural features are based on pixel values and intensities, whereas geometric features look at the shape of the face. Another decision is whether one uses temporal information across multiple frames or only spatial information from single frames for the detection task. Temporal information can be encoded through textural feature aggregation (Ojala et al., 1994), detecting a sequence of AUs (Valstar et al., 2005) or through the use of several machine learning methods (Werner et al., 2022). Finally, to process these features, a classification algorithm is used to classify signals as pain vs. no-pain.

The main issue in developing automatic pain detection frameworks is the lack of representative data (Werner et al., 2022). One of the first attempts to solve this problem was made by researchers at McMaster University and the University of Northern British Columbia when they developed the UNBC-McMaster Shoulder Pain Expression Archive Database (UNBC-McMaster dataset) (Lucey et al., 2011). This was one of the first publicly available datasets for automated pain recognition. Also, it is still one of the most widely used datasets for automated pain detection, as Werner et al. (2022) shows in their survey. The UNBC-McMaster dataset contains 200 videos showing the facial expressions of 25 participants suffering from shoulder pain. These videos are uni-modal because they only contain RGB images. Other databases such as BioVid (Walter et al., 2013) and BP4D (Zhang et al., 2014) and BP4D+(Zhang et al., 2016) have recorded multiple modalities, which enabled researchers to explore different types of modalities.

2.1.1 Automated pain detection using handcrafted features

Since the UNBC-McMaster dataset is so widely used and was the first publicly available dataset, the first research that used this dataset did so on the RGB images, and the 66-point AAM landmarks (Cootes et al., 1998) provided by the dataset. Therefore, these earlier works use handcrafted features that can be categorised as textural features, geometric features or a combination of these features to use in their classification task. Geometric features are extracted from the landmark position of the face, whereas textural features are extracted from the pixels of the image by different types of algorithms. Commonly used handcrafted features, are Gabor filters (Gabor, 1946), local binary pattern (LBP) (Ojala et al., 1994) and Histogram of oriented gradients (HOG) (Dalal and Triggs, 2005). The resulting feature set of these techniques is often post-processed to increase their discriminative power or to reduce their dimensionality by extracting only the most important information (Werner et al., 2022). Dey Roy et al. (2016) used a Gabor mask with five scales and eight orientations to extract features from the facial region. After a dimension reduction using principal component analysis (PCA), they used a state vector machine (SVM) (Cristianini et al., 2001) to classify between pain and no-pain. They report an accuracy of 94.75% for no-pain images and 96.25% for images depicting pain. One issue in their reporting of the accuracy is that they also balanced out the test set, which should be held out during training. This means that a highly imbalanced dataset distribution is altered, and reported results do not reflect the actual performance on this dataset.

Lu et al. (2008) used a similar pipeline as Dey Roy et al. (2016) but used a different method for dimension reduction, namely AdaBoosted Gabor filters (Shen and Bai, 2006) to reduce the dimension of the feature vectors. They used cross-validation on a more balanced dataset and ended up with a recognition rate of 85.29% on the pain detection task.

Yang et al. (2016b) divided the facial images into several regions and calculated the LBP features for each patch. Then they concatenated these images to encapsulate spatial information. Furthermore, they use the same logic on three orthogonal planes (between subsequent frames) to encapsulate temporal information. They report an accuracy of 59.08 on the frame level and 63.72 on the sequence level. So in this paper, they used several other techniques to encapsulate temporal information to compare it to single-frame classification and showed that temporal information holds extra valuable information for pain detection.

Nanni et al. (2010) did a comparative study on variants of LBP for classifying pain states. They looked at Local Ternary Patterns (LTP), a generalisation of LBP that represents the differences between pixels in a ternary value instead of a binary value and generally obtains a more robust descriptor than LBP. They also introduced a novel descriptor called Elongated Ternary Patterns (ELTP) and showed that it outperformed regular LTP descriptors.

Instead of using only single features Khan et al. (2013) combined a pyramid LBP (PLBP) with pyramid HOG (PHOG), which is a spatial shape descriptor. This technique extracts shape features with the PHOG and appearance features with PLBP from RGB images. They achieve respectable results when combining the extracted features with a two nearest neighbours classification algorithm based on Euclidean distance.

Rathee and Ganotra (2016) went a step further and proposed a face descriptor that uses a fusion of LBP, HOG, and Gabor features. The idea is that the combination of features gives even more types of information about the face that is described. An SVM was trained as a classifier and a fusion of the different modalities was performed before the learning step. The results were comparable with other literature they reported.

Other research by Neshov and Manolova (2015) used the Supervised Descent Method (SDM) algorithm (Xiong and De la Torre, 2013) to locate facial landmarks and localise the face. Then, they used SIFT features to form local gradient histograms around each landmark, with the idea that the deformation of the face when someone experiences pain causes variations in the direction of the gradients of image intensities. From these SIFT features, they build a feature vector that is reduced in dimension using principled component analysis, a general algorithm for dimension reduction of data. Finally, linear regression or SVM is used to either map the reduced feature vector to a continuous value representing pain intensity or to a discrete value for pain detection. This paper reported high accuracy on the binary classification task. However, their data set was highly imbalanced. So accuracy is not a right measure[1].

Singh and Singh (2015) did a comparative study between SIFT and sped-up robust features (SURF) (Bay et al., 2006) for the pain detection task. SURF is a robust and fast local feature detector inspired by SIFT (Singh and Singh, 2015). Singh and Singh (2015) demonstrate that SURF, along with an SVM classifier, can improve the pain recognition task's performance and is faster than SIFT.

Besides textural feature extraction methods, other works focused on geometric feature extraction methods for pain detection. These features describe changes in the shape and geometrical properties of different facial features, such as lowering eyebrows or stretching the lips (Werner et al., 2022). They typically use facial landmarks that describe the shape of the face in terms of point-based shape description schemes. More specifically, they point to the placement and orientation of facial indicators such as eyebrows, eyes, nose etc. Approaches differ in the use of these landmark positions. They might use facial landmark distances, angles, or a combination of the two (Werner et al., 2022). The resulting feature vectors might be used as inputs to various machine learning methods (Rupenga and Vadapalli, 2016) to recognise the corresponding facial expression or facial AU.

For example, Rupenga and Vadapalli (2016) tried estimating pain intensities from RGB images. They used plain 66-point active appearance models for feature extraction. These

[1] accuracy is not the right measure for imbalanced data sets because if a dataset contains 97% of one class, an accuracy of 97% sounds good; however, the algorithm could have just classified everything to the majority class, which means the algorithm is not working well

landmarks are represented as coordinates and directly fed to an SVM and a multi-layer perceptron (MLP). Their research aimed to compare the performance of an SVM and a neural network with one hidden layer in the pain intensity estimation task. They found that the neural network outperformed the SVM in both frame-level and sequence-level predictions. However, the performances of both models deteriorated in their prediction on the sequence level. This can be explained by the fact that an MLP and SVM are both architectures that do not inherently model temporal information. In other words, a previous prediction does not influence the next prediction. The temporal data is entirely encapsulated in the feature extraction method. Besides using plain landmark coordinates as their feature set, another thing is important to note here, namely that the performance of the models deteriorated when using sequential input. However, as mentioned earlier by Yang et al. (2016b), they mentioned an increase in performance when encapsulating temporal information in their feature set. It makes sense to model a pain expression as a temporal event because of the temporal character of pain expressions, which can be useful to reduce the false positive rate (Werner et al., 2022). Hence the use of sequential temporal information is an important factor in the process of automated pain detection frameworks.

2.1.2 Automated pain detection using temporal information

Several studies have encoded temporal information from videos for automated pain detection. Temporal information is thought of as the aggregation of frame-level features. For example, (Yang et al., 2016b) used a 3D neighbourhood (two dimensions for spatial neighbourhood and one dimension for temporal neighbourhood) while extracting textural features. In their work, Schmid et al. (2012) learned AU sequences that typically depict pain. They aimed to create a collection of grammars that typically depict pain. They used ABL (Zaanen, 2000), an unsupervised learning approach designed to extract grammars from text, to extract a collection of AU sequences that describe pain. This work intended to explore using AUs as a context-free grammar, for classification. The results were promising for the time and mainly showed that including temporal data can be useful in pain and pain vs no-pain classification task.

Furthermore, Chen et al. (2015), combined spatial, textural and spatiotemporal features. They use HOG to extract spatial features from video frames. They extracted appearance features from neighbourhoods around the facial fiducial points (e.g. areas around the nose, eyes, mouth etc.) by applying HOG from these neighbourhoods. Also, they use HOG from three Orthogonal planes (HOG-TOP) to represent dynamic textures of frame sequences. So they calculate the pixel gradient orientation over a 3D plane instead of 2D. They fused the two types of features before training an SVM for the pain detection task. They achieved promising results at the time.

Yang et al. (2016a) compared the performance of several other spatiotemporal textural features in pain detection of facial images, such as LBP-TOP, LPQ-TOP11, BSIF-TOP12, and their combinations. They found that the use of spatiotemporal information consistently performs better in the pain detection task than pain detection with single frames. Also, they found that the combination of spatial features improves results but does increase the feature dimensions.

2.1.3 Automated pain detection using deep learning

An issue with previously mentioned "conventional" feature extraction methods is that they are inherently limited in their ability to model a dynamic interaction over longer periods. As Zhou et al. (2016) point out, traditional features like LBP, extracted from separate frames, tend to lack the ability to pick up on relevant dynamic information. For example, people tend to close their eyes when in pain. However, these conventional features and features extracted from single frames cannot differentiate between a blink and an eye closure caused by pain. This results in very unstable pain estimations over periods of time. Also, Zhang et al. (1998) found geometric features to perform poorly compared to Gabor filters for the recognition in the task of expression recognition. Even though geometric features perform similarly in recognition of facial action units Valstar et al. (2005), at a lower resolution, geometric features are difficult to extract from the image, as Li (2004) shows in their work. Texture features perform better than geometric features at lower resolutions if and only if the face is properly aligned (Li, 2004).

Recent advances in machine learning, particularly the emergence of deep learning, have enabled the automatic detection of important features from data without prior handcrafting. Since then, deep learning methods have been used for automated pain detection, including Convolutional neural networks (CNN), of which the architecture was first proposed by Lecun et al. (1998).

Xiang and Wang (2017) used a pretrained CNN (Wen et al., 2016) to predict pain intensities from single images of the UNBC dataset. They removed the last layer and swapped it with an MLP with a softmax output layer. They used a small learning rate so that the convolutional layers did not change too much in the finetuning process. They showed that deep features outperformed handcrafted ones. However, they also note that the baseline of always predicting "no-pain" performs better than some state-of-the-art algorithms. This might mean that the data is highly imbalanced. However, they do not report the ratio between pain and no-pain frames, showing that one should choose proper evaluation measures that account for imbalanced labels to evaluate models' performance.

More promising are the results shown by research that combine CNNs with Recurrent

Neural Networks (RNN) architectures (Rumelhart et al., 1985), which enables the use of temporal information to make predictions. For example Zhou et al. (2016) built on the work of Xiang and Wang (2017) and used a Recurrent CNN (RCNN) for continuous pain estimation prediction. Their results were on par with the state-of-the-art methods in mean squared error (MSE) and had a higher Pearson coefficient correlation (PCC) (Benesty et al., 2009) with the ground truth and had a notably lower runtime. An RNN has hidden layers that process the current time step together with several previous time steps. In other words, an RNN makes a prediction based on previous predictions and the current input. The power of an RNN is that it preserves the temporal information in sequences, which makes it well-suited for sequential tasks. Zhou et al. (2016) combine this architecture with the CNN architecture, which results in an architecture that can process temporal visual information. The power of such a network is the recurrent convolutional layers. These layers propagate their previous estimation to themselves, meaning that each layer unfolds in several layers taking their previous output into account together with the new input. Another special thing (Zhou et al., 2016) did was swapping the last softmax layer with a simple linear transformation between the weights and the output of the second to last MLP layer. This approach performed well in predicting pain intensity scores over time, as Zhou et al. (2016) reported that they have a much smaller mean squared error than the previous state-of-the-art model, which used earlier "traditional" methods. They also reported a Pearson correlation coefficient similar to the previous state-of-the-art.

2.1.4 Limitations of using RGB data for pain detection

All research mentioned above suffers from similar issues because they focused on facial expression in the visible light spectrum (e.g., red-green-blue referred to as RGB domain). This spectrum has certain limitations when extracting information from these images. It is very sensitive to variations in illumination conditions. For example, using dim light limits the ability of an RGB image to capture the intricacies of the human face. Or, more extreme, a system designed to extract facial features based on RGB images cannot detect informative information in complete darkness.

Another drawback of using RGB data for pain prediction is that systems trained on a particular age group do not generalise well to other age groups (Hassan, 2017). Similar results were found in human observers when Hess et al. (2012) found that humans have a harder time recognising emotions such as disgust or fear in elderly people. Hess et al. (2012) found that this is mainly because wrinkles and folds in the face tend to reduce the clarity of the emotion displayed and affect the behavioural intentions communicated through facial expressions. In automated pain detection frameworks that use RGB data, this results in great confusion, for example, where the neutral faces of elderly people are confused with other emotions (Hassan, 2017). Disgust, one of the emotions that is hard to predict for systems trained on younger faces, is very similar in expression to pain (Kunz et al., 2013). Therefore, when training a system that classifies faces on such a sensitive topic, this generalisation between age groups must be flawless.

Besides age, other facial characteristics such as race, ethnicity, gender and personal history can cause great variation in visible facial features such as the shape of the head, scarring on the face, skin colour etc. (Hassan, 2017). You could imagine that a system that tries to extract textural features from a face with a very dark skin tone with bad lighting cannot detect any detailed textures from the face. Hence it would be challenging to extract any meaningful features.

Also, people who lost the ability to express emotions cannot be analysed within this light spectrum because there will be no changes in texture or geometry in the face after a pain sensation. Yet, people who have a form of facial paralysis can still experience pain (Webber, 1876). This means that there is no way for a system that uses the RGB spectrum as its only source for feature extraction can extract any relevant features regarding the face.

2.2 Pain detection using thermal data

The limitations of using RGB data for automated pain detection have pushed research to look into different modalities for recognising human emotion in automated pain detection systems, one of which is thermal data. Abd Latif et al. (2015) found that facial cutaneous temperature, picked up by thermal cameras and its topographic distribution, displayed specific features correlated to individual emotional states. They conclude that thermal imaging can function as a contactless and non-invasive method for assessing an individual's emotional arousal in psychophysiology. Also, Erel and Özkan (2017) found that thermal cameras can pick up a significant difference in facial temperature between an individual experiencing pain and no-pain. The use of thermal data for automated pain detection has advantages compared to RGB data because it can overcome earlier limitations. First, thermal data is light variation insensitive (Kumar and Pai N, 2017). Thus, even in complete darkness, features can be extracted for automated pain detection (Nguyen et al., 2018).

Work in automated pain detection using thermal data has been very scarce, especially in works that used thermal data specifically for facial emotion recognition. In their review, Ordun et al. (2020) note that they only found 11 works that use thermal data in combination with machine learning approaches, whereas Li and Deng (2018) found 74 papers using RGB data in combination with several machine learning methods. Ordun et al. (2020) associate the lack of use of thermal data in the research on FER to the lack of readily available public data. According to Haque et al. (2018), the only noticeable publicly available dataset that contained multiple modalities was BioVid (Walter et al., 2013), which did not contain any thermal data modality. Therefore, Haque et al. (2018) created the Multimodal Intensity Pain (MIntPAIN) database, which contained RGBDT, Where besides RGB data, thermal and depth data were included. This dataset used 20 participants, and five pain levels were annotated for each frame. For their baseline performance, they used a standard 2-step approach, where they first applied a 2D-CNN for frame-wise feature extraction and pain recognition and, secondly, implemented an RNN architecture called to estimate the temporal to perform sequence level pain recognition. For the 2D-CNN, they finetuned the VGG-FACE model (Bellantonio et al., 2017), and for the RNN architecture, they implemented an LSTM architecture (Bellantonio et al., 2017). They found that only using thermal data with this architecture yields an accuracy of 18.33. However, using all three modalities with early fusion yields the best results with an accuracy of 36.55%. The authors suggest that future work should focus on the use of different modalities as well as analysing the performance of different models and fusion strategies.

The BP4D+ dataset provided by Zhang et al. (2016) contains 140 participants from different ethnicities. Participants were recorded in different modalities, under which RGB, 3D, thermal and physiological data, while they performed a task that provoked certain emotional reactions including pain. To validate the thermal data, they calculated the error of the tracked landmarks on the thermal data and the underlying ground truth for only 60 of the 140 participants, with which they achieved a 91% accuracy. Limited work has been reported on the subset of thermal images of this dataset. Liu and Yin (2017) presented a new thermal video representation called trajectory-pooled fisher vector descriptor (TFD). To get the local energy and temperature changes of the thermal images. They report a 78% accuracy which shows promise for using thermal data in the FER task, including pain.Prawiro et al. (2020) did an empirical study where they compared a 2D CNN, 3D CNN and 2D CNN + Temporal Shift Module architecture. Their 3D CNN performed best with an average accuracy of 72.83%. Whereas their design used on RGB data performed better with 79.5% accuracy. In their study, the RGB model performed slightly better than thermal data. To fill the research gap on this modality in the specific task of pain detection Olczyk and Önal (2021) set out to create a benchmark for pain recognition on thermal images. They used the CNN+RNN architecture, comparing CNNs trained exclusively on thermal data and finetuned CNNs that were pretrained on RGB images. For the RNN architectures, they compared the LSTM and GRU architectures. They found that the best performing model was the combination of the finetuned CNN combined with the LSTM with a weighted accuracy of 84.37%.

2.3 The use of ROIs in thermal pain detection

Human emotions influence the facial temperature by working on the autonomic nervous system (ANS) responses. Another factor that influences the temperature of the face is facial contractions (Alqudah, 2021). Different emotions have different effects on facial temperature. Pain usually causes a decrease in temperature in the forehead and a decrease in the maxillary (Alqudah, 2021). This means that in using thermal images for pain prediction, extracting only information about the ROIs could be useful. The use of thermal data also has some drawbacks because some areas of the face do not respond to emotional changes or infrared rays in the area around the eyes when wearing glasses. As a result, the eye area is picked up as being colder than in reality (Lopez et al., 2017), which is another reason that using ROIs can be useful in pain detection. Unfortunately, little research is done on using dynamically determined ROIs on thermal images for pain detection. However, research has been done on using ROIs for thermal images in a general emotion detection task.

For example, Delisle Rodriguez et al. (2019) located the regions of interest based on RGB images and transferred them over to thermal images. They used manually marked reference frames, which domain experts mark, to correct the placement of the ROI on the thermal image. After a feature reduction using PCA and Latent Dirichlet Allocation (LDA) to infer emotions, the results show that the ROI tracking works and that using ROIs is relevant for recognising emotions.

Also, (Nguyen et al., 2014) used ROIs on thermal images to make their emotion classification algorithm. Different from (Delisle Rodriguez et al., 2019), they dynamically detect the regions of interest in the thermal images. They fused visible images with thermal images to strengthen their predictive power and used PCA and Eigen-space Method based on class features (EMC) to classify their feature set into seven different emotions. They proved that their fusion method has an average improvement of 2.74% in performance over only using RGB data. They also found that only using PCA on thermal images increases performance than using RGB images.

Moreover, Nguyen et al. (Nguyen et al., 2018) improved over the previous technique (Nguyen et al., 2014) even though they only used thermal images. They included the change in pixel intensity in the grey scale of thermal images. However, they did not use any features from RGB images. They found an increase in accuracy over the previously mentioned technique, even though their classification techniques were similar in that they also used PCA, EMC, and a combination of the two for the emotion prediction task and demonstrated the effectiveness and superior performance of their proposed method.

Above mentioned research shows that the use of thermal data in combination with the use of ROIs can improve the performance of the model. However, they used conventional feature extraction methods, such as LDA or PCA. Whereas, as mentioned earlier, deep learning techniques, are known for learning relevant features in the learning process. Lopez et al. (2017) show that using only predefined regions of interest for fatigue prediction on thermal images in combination with a CNN-based feature extraction method yields comparable and, in some cases, even superior performance compared to using the whole face.

2.4 Transformers and feature learning

A crucial part of any affective state prediction, particularly pain prediction, is a robust classification method that can maintain high accuracy. In the realm of pain detection, CNN with RNNs has been successful. However, an underlying problem remains due to the recurrence in any RNN architecture, such as LSTMs or GRUs. Passing information through an extended time series using recurrent connections means losing long-distance information (Jurafsky and Martin, 2020). Also, the RNNs have an inherent sequential nature which inhibits the use of parallel computing effectively (Jurafsky and Martin, 2020). These considerations led to the development of Transformers, first introduced by Vaswani et al. (2017). Transformer models were first developed for the task of natural language generation. The main benefit of the transformer model is that it does not use any recurrences or convolutions, making the model superior in quality while being more parallelisable and requiring significantly less time to train (Vaswani et al., 2017).

The key innovation of transformer models is the idea of self-attention. Self-attention allows the model to extract information from an arbitrarily large context without using recurrent units. In the basic models, when an item of the input is processed, it has access to all previous contexts, where all items are processed independently. Simply put, for every input item, it generates three vectors. The first vector describes the item as being the input that other items use as context (key). The second vector describes the item as being the currently processed item (query). Finally, the last vector describes the item by itself (value). To calculate the output vector $\mathbf{Y_i}$ for a given input item $\mathbf{X_i}$, a dot product is taken between the query vector of input item $\mathbf{X_i}$ and the key vector of the preceding items. The resulting vector of these dot products describes the relevance of the preceding items to the item $\mathbf{X_i}$. Then simply put a weighted sum between the dot product of these relevance vectors and the value vectors of each item is taken to calculate the outcome $\mathbf{Y_3}$. All output vectors can be calculated simultaneously using parallel processes because the calculation of the output vectors of each input item depends on the input vectors of all preceding items and not on their output vectors.

A transformer model consists of transformer blocks. Its simplest form consists of a self-attention layer and a fully connected feedforward layer, where both outputs of each layer are normalised. An illustration of a transformer model can be seen in Figure 1. A fully functional transformer block usually uses multiple self-attention layers called ahead. The benefit is that each head can learn distinct relations for each word relating to other words in the sentence, instead of one head trying to learn all different relations for all words in a sentence (Jurafsky and Martin, 2020).

The use of transformer models has transcended beyond natural language generation and has also been used for image classification tasks and emotion detection tasks. Dosovitskiy et al. (2020) introduced the use of the transformer model for image classification tasks. The general power but also a weakness of transformer models is that it has no inductive biases like previous models such as CNNs and LSTMs. This means a transformer model can learn any task and outperform and generalise better than models with a high inductive bias, such as CNNs and LSTMs if fed sufficient data. However, if data scarcity is a concern, models that already have extracted relevant features for a task will outperform a transformer model (Dosovitskiy et al., 2020). There are other benefits of using a transformer model for image classification tasks other than increased classification performance. (Dosovitskiy et al., 2020) also point out that the cost of pretraining their transformer-based model has lower computational cost than models using recurrence and convolutions while attaining state of the art on most recognition benchmarks (Dosovitskiy et al., 2020).

Ma et al. (2021) show superior performance in the affective state recognition task with RGB images over other methods, with the use of transformer models. They used a pretrained ResNet18 to extract feature maps. They fuse these extracted features with their attentional selective fusion to get representative visual words. The input visual words are obtained by simply flattening the spatial dimensions of the feature maps and projecting them to the specific dimension. They feed these to their transformer model.

Where Ma et al. (2021) only use RGB images as their input modality, Zhang et al. (2022) also uses audio recordings, text and static frames as input modalities for their transformer-based fusion model. They outperform other methods in the AU detection and FER task that use transformer models. They also show that using fewer modalities than all four negatively impacts the architecture's performance. To deal with the temporal information of the non-static features, they use GRUs to extract en encode this temporal information.

In contrast, (Arnab et al., 2021) showed that one could use a pure transformer model to encapsulate temporal information when they developed ViViT.They tested three architectures that factorize different components of the transformer encoder over the spatial- and temporal dimensions. Because the datasets for videos are not as extensive as images, they pretrained their models on images, whereafter they finetuned it on videos. They achieved state-of-the-art performance across five popular datasets. To our knowledge, the ViViT model has not yet been used for emotion prediction.

We contribute to the research of pain detection by exploring the use of a pure transformer model on thermal image frames for pain detection and exploring the use of ROIs together with a transformer model in pain detection.

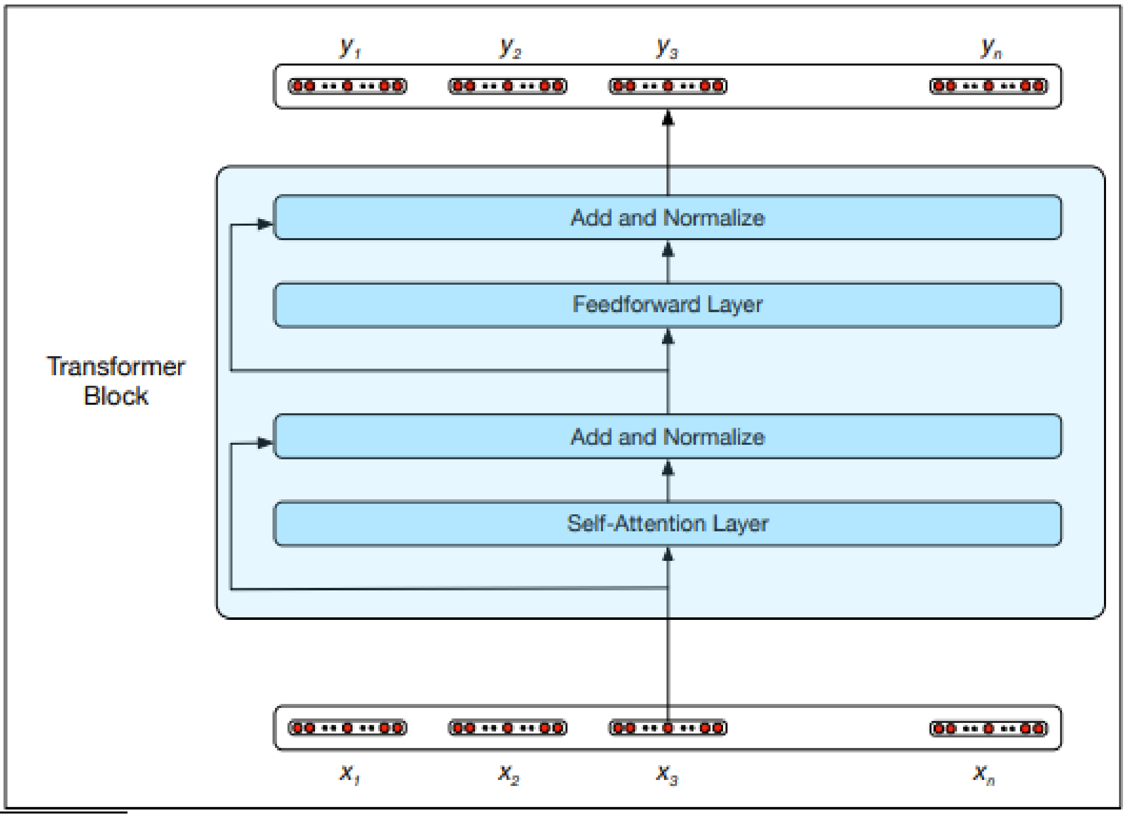


Figure 1. A simple transformer block depiction for sequence to sequence prediction, where X_i represents a part of the input sequence as an embedded feature vector and Y_1 represents a part of the output sequence, from (Jurafsky and Martin, 2020)

3. Methodology

3.1 Dataset

We used a subset of data provided by BP4D+ (Zhang et al., 2016) in our experiments. This subset included thermal videos of spontaneous reactions from 140 participants, with 58 being male and 82 being female. The participants had diverse racial/ethnic backgrounds, including black, white, Asian, Hispanic/Latino, and others (e.g., Native American).

Each participant had been subjected to 10 different tasks corresponding to 10 different emotions. We used only the thermal videos that corresponded to the pain task, where participants were asked to place their hands in ice water to provoke a pain response. The original resolution of the image frames was 726 x 480. However, we resized the images to 128 for memory efficiency reasons.

The subset of thermal pain videos contained frames in which the subjects expressed a pain reaction (pain frames), frames where participants showed no-pain (non-pain frames), and frames that were unusable because participants were partially out of the screen or untracked. We employed the formula from the Prkachin-Solomon Pain Intensity (PSPI) scale (Prkachin and Solomon, 2008), often used to quantify pain intensity through facial expressions. The formula for pain intensity was:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43.$$

We can read the specific action of the different AUs in Tables 1 and 2, where Table 1 describes the fundamental action units that express pain and Table 2 describe the AUs that are associated with pain expression (Rojo et al., 2015a).

AU	Action unit description
AU4	Lower eyebrows and /or lower forehead
AU6	Cheek raising and eyelid compression
AU7	Eyelid tightening
AU9	Nose wrinkling
AU10	Lifting upper lip

Table 1. Basic facial action units in the expression of pain (Rojo et al., 2015a)

AU	Action description
AU1	Raising the medial eyebrow area
AU2	Raising the distal eyebrow area
AU12	Lifting the lip corners
AU14	Appearance of dimples
AU17	Chin lift
AU25	Lips slightly parted
AU26	Relaxed jaw with mouth open
AU27	Mouth open with effort
AU43	Eye closure
AU45	Blinking
AU46	Wink

Table 2. Facial action units associated with the expression of pain (Rojo et al., 2015a)

A frame was labelled as "pain" if PSPI > 0 and "no-pain" if PSPI = 0. As the dataset did not come with the pain intensity scores the presence of one of the basic AUs or AU43 would make the PSPI > 0, hence a frame was labeled as pain if one of the basic AUs or AU43 was present. Our analysis revealed that the labels were fairly distributed, Out of the 41925 frames 18530 of them 44.2% were labelled as pain whereas the remaining frames (55.8%) were labelled as no-pain.

3.2 Hardware and Software

For all preprocessing steps and training, we relied on the Google Colab notebooks Bisong (2019), which is currently running Python version 3.10.12. For the training of ViT, we used a Nvidia Tesla T4 GPU and for the training of ViViT, we used a Nvidia A100 GPU, due to the memory constraints of the other available machines. We used the HuggingFace API for the ViT and ViViT models (Huggingface, 2022b) (Huggingface, 2022a), which are written in PyTorch (pyt, 2019).

3.3 Preprocessing

Following Olczyk and Önal (2021), each video was divided into separate frames using the OpenCV library by Bradski and Kaehler (2000). For each frame, we aligned the face to a frontal position by using the retina face package by Serengil and Ozpinar (2021) Serengil and Ozpinar (2020) Deng et al. (2019). This algorithm used deep learning techniques to find the landmarks of the face and align the entire face such that the eyes are horizontally in the image. For the sequences, we took 10 frames of each video and used the last frame for classification, practically dividing each video into 10-frame videos. As the literature suggests longer temporal sequences up to 128 frames tend to perform best for the ViViT transformer according to Arnab et al. (2021). However, when Arnab et al. (2021) used the base model for testing they ran into memory constraints after a depth of 48 frames. For Rojo et al. (2015b) a frame length of 32 frames worked best, but they used a convolutional neural network architecture. However, Rojo et al. (2015b) also show that for depths of 8 to 16 accuracy was around 88% and 96% respectively, where their model that uses 35 frames scores 98.53%. Showing that the biggest performance increase was between 8 and 16 frames. For this reason and to minimize the risk of memory constraints we choose to use shorter frame sequences namely 10 frames. Furthermore, every image was normalized by dividing every image channel by 255. We divided the data into a subject-dependent train validation and test set, such that none of the sets shares data from the same subject. The test set comprised 18,3% of the entire dataset (22 subjects), while the remaining data was divided into an 85-15% split between the training and validation sets.

3.4 Extraction of ROI

The regions of interest were extracted following Nguyen et al. (2018). We chose this technique because it successfully showed that the results of their ROI extraction method resulted in a subset of pixels that successfully determined the underlying emotion, while only using PCA to make the final inference.

This technique locates patches of different sizes and different areas of the face. It aims to find those regions that contain the most information about the underlying emotion of the subject depicted in the frame. Nguyen et al. (2018) describe their approach as follows.

They use three parameters found through the following functions:

$$\Delta T_F = T_{max} - T_{min}$$
$$\delta T_f = \Delta T_F / max(g(i,j))$$

where T_{max} is the maximum temperature recorded and T_{min} denotes the minimum temperature recorded in the frame. $max(g(i,j))$ is the maximum pixel intensity of the image in greyscale, and $g_{i,j}$ is the value of the pixel (i,j) in the greyscale image. A mask denoting the ROIs is extracted as follows. Firstly, two terms are defined namely: $T_{min} + \delta T_f \cdot g_{i,j}$ and $Tmax + \delta T_f \cdot g_{i,j}$. Let them be named term A and term B respectively. The group of pixels making up the final image mask is defined by all the pixels where the temperature value of the pixel, denoted as $h(i,j)$, is equal or larger than term A and equal or smaller than term B. This is also explained by the following function:

ROI per frame = $\{(i,j) \in F_{thermal} | T_{min} + \delta T_f \cdot g_{i,j} \leq h(i,j) \leq T_{max} + \delta T_f \cdot g_{i,j} \}$

Where F_{temp} are the pixels in the thermal image frame. This technique was used to extract ROIs per frame, which were then used as a mask and overlayed on the original images. This approach resulted in images where only these regions were visible for the transformer. These new images could be used as single input frames or as sequences. An example of such an image can be seen in Figure 2

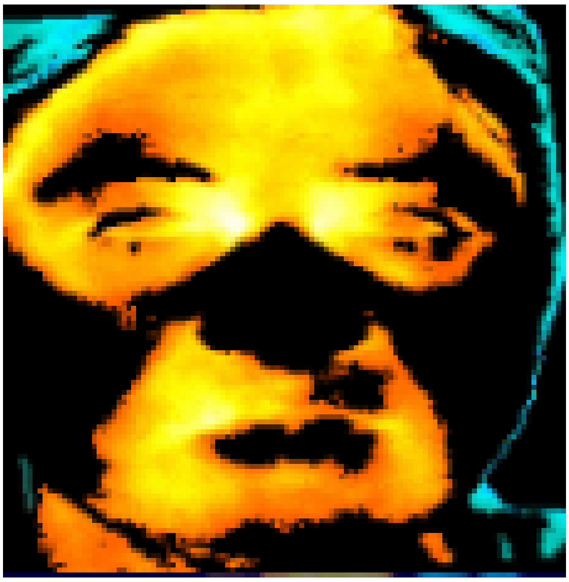


Figure 2. A thermal image of a face after ROI extraction

3.5 Models

3.5.1 ViT

As described earlier the vision transformer model leverages a pure transformer block to extract relevant features from images. To make it capable of processing image data it splits an image into fixed-size patches of 16*16*3 and linearly embeds each of them. Afterwards, a position embedding is added to each patch such that the model can understand the spatial orientation of each patch. This embedding is fed into a standard transformer encoder to extract features from the image. Finally, in the case of image classification, the output of this encoder is pushed through a standard classification layer to make the final prediction, see Figure 3 for a visualization.

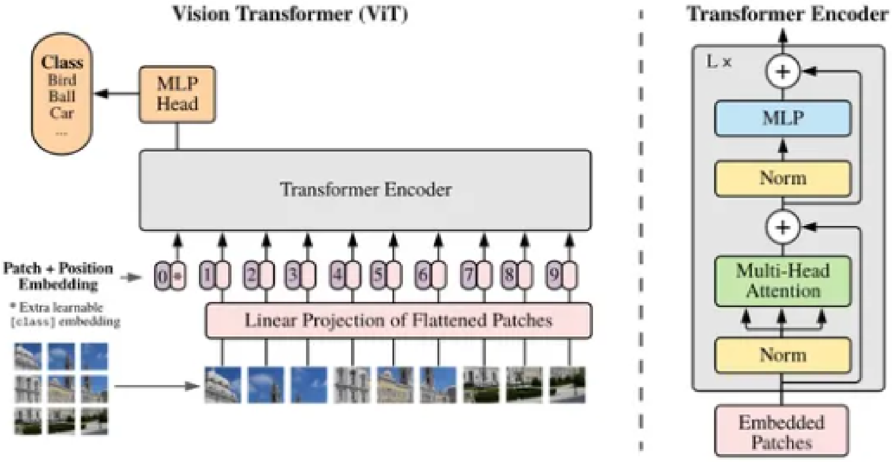


Figure 3. A visual representation of the vision transformer architecture for image classification Dosovitskiy et al. (2020).

3.5.2 ViViT

ViViT (Arnab et al., 2021) is a derivative of the ViT model by Dosovitskiy et al. (2020), with the main adaptation being how the input data is embedded, which can be seen in Figure 4. Since the input data is a sequence rather than a single image Dosovitskiy et al. (2020) used two main ways of embedding the input data, namely "Uniform frame sampling" and "tubelet embedding". In this paper, we use the tubelet embedding technique, because Arnab et al. (2021) show in their ablation study that tubelet encoding outperforms uniform frame sampling in their classification task. This technique extracts "tubes" from the input video of a defined dimension width * height * time, which can be visualized in Figure 5. By doing so, it inherently embeds spatio-temporal information during tokenizations (Arnab et al., 2021), meaning that information of a given patch of the image is embedded over time. After embedding the video using tubelet embedding, similar to the ViT architecture, a positional token is added and fed through a pure transformer block which then will pass the output to the classifier. It is important to note that we used a pure transformer block as depicted in Figure 4. The paper by Arnab et al. (2021) develops 3 other model configurations that tend to perform better than the base configuration, however, we chose to use the basic model to make a better comparison between ViT and ViViT and the influence of temporal data on pain detection on thermal face images.

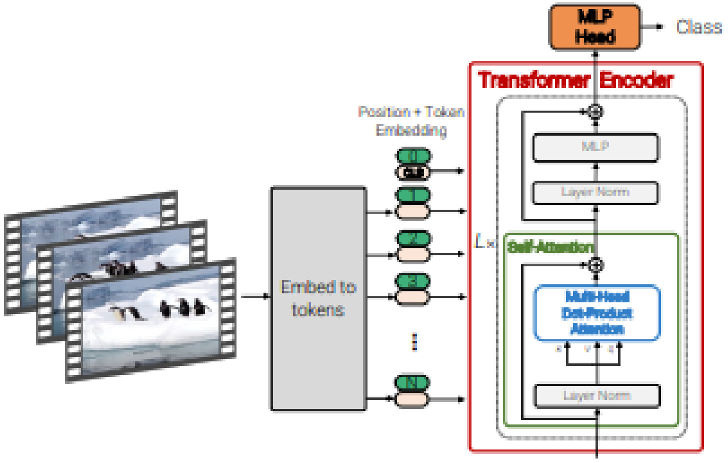


Figure 4. A visualization of the ViViT encoder block using the base encoder and the embedding of video frames
Arnab et al. (2021).

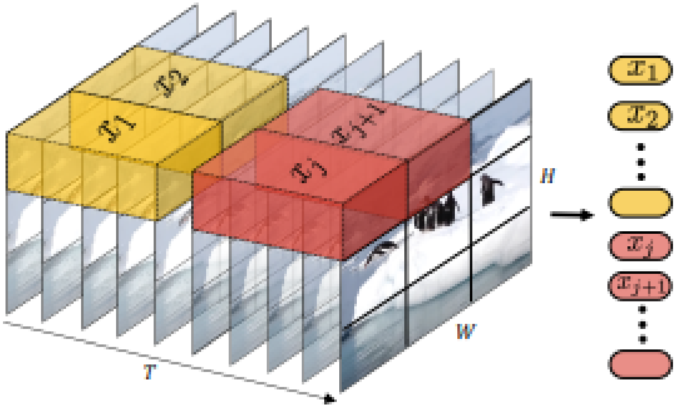


Figure 5. A visualization of the tubelet embedding of video frames Arnab et al. (2021).

3.6 Model Finetuning

To fine-tune the two models ViT and ViViT we used the pretrained models provided by the HuggingFace API at the following sources (Huggingface, 2022b) (Huggingface, 2022a). ViT was pretrained on ImageNet (Deng et al., 2009) and ViViT was pretrained on Kinetics400 (Kay et al., 2017). We modified the ViT architecture (Dosovitskiy et al., 2020) for frame-level binary classification by replacing the last softmax layer with a binary sigmoid layer representing the pain and no-pain classes. The same modification was made for the ViViT (Arnab et al., 2021) architecture. For both models none of the layers were frozen as suggested by the guide provided by (Huggingface, 2023), meaning the whole model including the body was retrained.

3.6.1 Hyper-parameter Optimization with Ray Tune and Ax Search Algorithm

To search for the best set of hyper-parameters for the four models, we chose to leverage the Ray Tune framework (Liaw et al., 2018), which is a platform that provides several implementations of different hyper-parameter search algorithms.

From Ray Tune, we used the Ax search algorithm, which is a Bayesian optimization technique (Balandat et al., 2019) (Facebook opensource, 2023) that is particularly well-suited for optimizing complex functions with noisy evaluations, such as those often encountered in deep learning task. Apart from that, we chose this algorithm because of its concept of adaptive experimentation. This means that it adapts and tunes hyper-parameters every iteration to identify the most promising configurations, leading to a more efficient exploration of our parameter space.

The parameter space we explored for our hyper-parameter search was defined as in Table 3.

Table 3. Hyperparameter Configuration

Hyperparameter	**Values**
Learning Rate (lr)	loguniform(1×10^{-7}, 1×10^{-4})
Batch Size	randint(10, 40)
Attention Dropout	choice(0.1, 0.2, 0.3)
Hidden Dropout Probability	choice(0.1, 0.15, 0.2, 0.25)
Hidden Layers	choice(1, 2, 3, 4, 6)

For every model, a search of ten samples was conducted with Ax search. Every sample was a model trained with a maximum of 50 epochs and a patience of 10 epochs. We considered a range of learning rates, batch sizes, and dropout probabilities for both the attention mechanism and hidden layers. Additionally, the number of hidden layers in the model varied among a set of discrete choices, as shown in Table 3.

3.7 Evaluation

To validate the performance of the models, we reported the mean accuracy, recall, precision, and F1 score. Accuracy was chosen because the pain/no-pain classes were fairly balanced, with 41.03% belonging to the pain class. Recall and precision were reported to assess the types of mistakes (false positives/false negatives) that were most prominent in every structure and to gain further insights into the models' behaviour. Despite the fairly balanced class distribution, to mitigate any influence of the slight skew, F1 was included, which is a harmonic mean between precision and recall and is not influenced by class imbalances.

For visual evaluation, we used gradient attention maps by Gildenblat (2022) to gain insight into the models' internal decision-making. This technique calculates the gradients of the model's output with respect to the intermediate feature maps in the Vision Transformer. These gradients are used to compute weighted heatmaps, which highlight image regions that the model focuses on for specific class predictions. darker red areas in the heatmap correspond to regions that strongly influence the model's decision while blue regions have lower influence.

4. Experimental Results

We conducted four experiments using two variations of the ViT model and its derivative ViViT model. Both models were trained and tested with two types of input data: conventional thermal images of the face and images in which only ROIs were visible. The final results of these experiments are displayed in Table 5. This table highlights the best test results achieved by the four model configurations: ViT whole, ViViT whole, ViT ROI, and ViViT ROI. To find the optimal hyperparameter configuration for each of the four models we did a hyperparameter search and found the hyperparameters for each model shown in Table 4.

Table 4. The best hyperparameter configuration found per model by the ax hyperparameter search.

	learning rate	batch size	attention dropout	hidden dropout	hidden layers
ViT whole	1.79919e-05	26	0.2	0.1	1
ViT ROI	1.0094e-05	34	0.15	0.1	2
ViViT whole	3.23522e-06	34	0.2	0.2	2
ViViT ROI	4.37541e-07	18	0.3	0.1	3

Table 5. All test results after hyperparameter tuning and finetuning of the models

	ViT whole	ViViT whole	ViT ROI	ViViT ROI	Basline
Accuracy	0.519	**0.605**	0.567	0.539	0.640
recall	0.449	**0.540**	0.499	0.443	0.541
precision	0.479	**0.565**	0.533	0.453	1.0
F1	0.491	**0.554**	0.547	0.499	0.702

The results in Table 5, show the performance of the four model configurations: ViT whole, ViViT whole, ViT ROI, and ViViT ROI, in the task of pain detection from thermal video frames of human faces. The evaluation metrics employed include accuracy, recall, precision, and F1 score.

First of all we can see that all models underperform in comparison to the baseline. However, when we compare the models with each other, we find that ViViT whole stands out as the top performer among the four models, with an accuracy of 0.605 on the test set, meaning that it correctly classified pain or non-pain instances in thermal video frames 60.5% of the time. In comparison, it achieves an 8.6%, 3.8%, and 6.6% increase in accuracy over ViT whole, ViT ROI, and ViViT ROI respectively.

We also measured a 9.1%, 4.1%, and a 9.7% increase in the recall score over ViT whole ViT ROI and ViViT ROI respectively. Recall highlights the model's ability to minimize false negatives, which in the context of pain detection is crucial.

Also, the ability of ViViT whole to minimize false positives shows a similar trend. It outperforms the other models with a precision score of 0.565, which means that it pulled off an increase in precision of 8.6%, 3.2%, and 11.2% over ViT whole ViT ROI and ViViT ROI respectively.

Finally, the F1 score, which balances the trade-off between precision and recall, reiterates the superior performance of ViViT whole over the other three models, it achieves a score of 0.554. This model strikes a better balance between correctly classifying pain instances while minimizing both false positives and false negatives than the other models. The other models, ViT whole, ViT ROI, and ViViT ROI, obtain F1 scores of 0.491, 0.547, and 0.499, respectively. However the baseline f1 score is 0.702 (if we always predict the majority class), meaning that it is 14.8 % higher than that of ViViT whole.

4.1 Overfitting

In Table 6, we can see the results of the training metrics of the best-tuned models. Here we see that all models overfitted on the training data, as we observe a discrepancy of around 30% to 40% between the training and validation set accuracy. This suggests that the models learned features that are only relevant to the training data specifically and do not generalize well over unseen data.

Table 6. Train and test accuracies of the different ViT and ViViT configurations presented in the study

	training accuracy	test accuracy
ViT whole	0.891	0.519
ViT ROI	0.954	0.567
ViViT whole	0.867	0.605
ViViT ROI	0.869	0.539

4.2 ROI vs Whole image

To gain insight into the effect of ROI extraction on the pain detection task from thermal face images we first look at ViT models and their performance on whole images versus ROI images. There was an overall increase in performance when ROI frames were used. More specifically, the results show a 4.2% increase in accuracy and a 5% increase in recall meaning that ViT ROI is better at finding the positive cases of pain than ViT whole. Also, we found a 5.4% and a 5.6% increase in precision and F1 respectively, showing that it also becomes better at predicting pain classes better and the trade-off between precision and recall became better.

However, when we introduced temporal information we can see that ViViT whole outperforms ViViT ROI across all evaluated metrics by quite a margin. specifically, we can see a 6.6% increase in accuracy. Also, we observe a 9.7%, 11.2%, and 10.5% increase in the recall, precision, and F1 score, respectively of ViViT whole compared to ViViT ROI. This shows us the opposite trend in performance between ROI and the whole image compared to ViT.

4.3 Comparing the use of single images and a sequence of frames

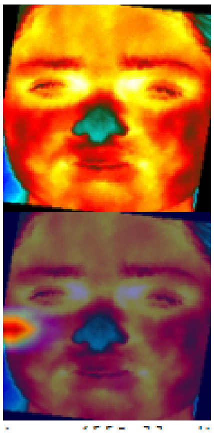
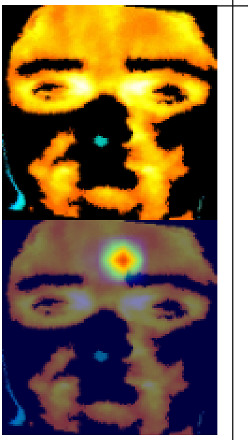
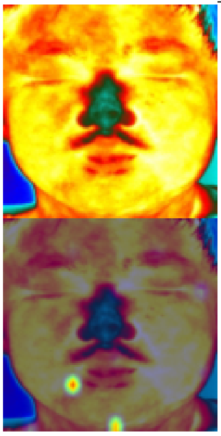
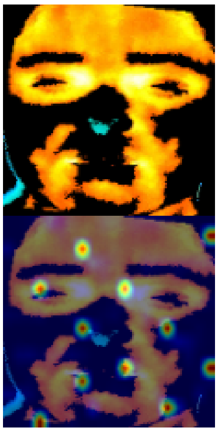
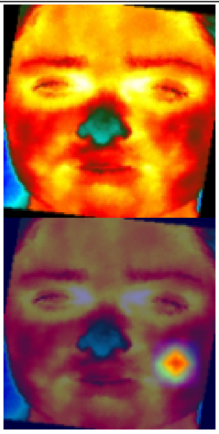
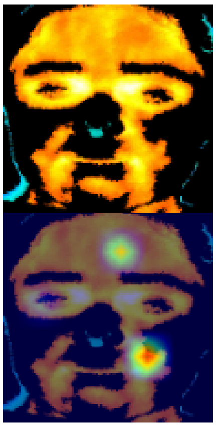
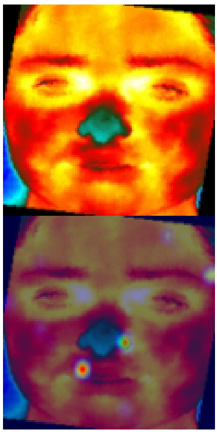
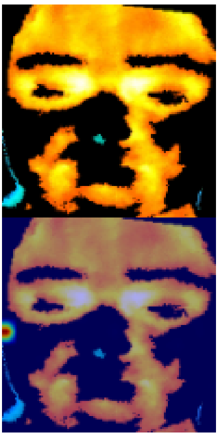
When we compare the performance of ViT ROI and ViViT ROI, we see a decrease in all evaluated metrics when adding temporal sequences to the ROI algorithm. ViT outperforms ViViT by a slight 2.8% in accuracy but a larger 5.6% in recall, 8% in precision and 4.8% in F1 score. Thus in the context of pain detection from thermal video frames, the incorporation of temporal information in the ViViT ROI model did not yield improved results compared to the use of regions of interest in ViT ROI.

When comparing ViT whole and ViViT whole, we see the largest increase between the two models in the table as ViViT whole is the best-performing model whereas ViT whole is the worst.

4.4 Attention maps

In the image grid shown in Table 7, we present the attention maps of the four models generated by gradient attention rollout (Gildenblat, 2022). These maps are created by single image predictions and might not generalise overall observations.

Table 7. All attention maps created with gradient attention rollout for sample images with and without pain

	ViT whole	ViT ROI	ViViT whole	ViViT ROI
Pain				
no-pain				

In Table 7 we can observe the following in the attention maps. ViT Whole focused on the upper cheek area when detecting pain in this frame which which may be related to AU6 (cheek raising).

In the case of ViT ROI, we observed that it used the cheek, the inner eyebrows, and the forehead, which are areas associated with AU4 (brow lowerer), and AU12 (Lip corner puller). ViViT Whole paid attention to the mouth corners, chin, and nose area. ViViT ROI examined a lot of areas when it made this pain detection. These regions correspond to AUs that signal pain, such as AU4 (brow lowerer), AU6 (cheek raiser), and AU7 (eyelid tightener). Notably, ViViT ROI paid minimal attention to these areas when it did not predict pain.

In Table 8 and Table 9 we present confusion matrices generated by 100 random sample images, overlaid on a sample face image, to create more generalisable attention maps. When looking at these confusion matrices we also observe some interesting patterns. In the case of ViT whole, it directs its attention to the left upper cheek, and right cheek area when correctly identifying instances of pain, which are associated with AU6 (Cheek raiser) and AU14 (dimpeler) respectively. Interestingly enough the same areas are used when it identifies pain wrong however the focus on the lips becomes greater which is associated with AU10 (upper lip raiser) and AU25 (lips slightly parted). Also, when it identifies no-pain correctly it uses the forehead, left upper cheek, chin, mouth, and right outer eye corner. These areas are also associated with AU4 (brow lowerer), AU6 (cheek raiser) AU10 (upper lip raiser). We can see a roughly similar pattern when it incorrectly identifies no-pain, however, it does not use the outer right eye corner, chin, and mouth, but the right lower cheek and right mouth corner/cheek area instead.

On the other hand, ViT ROI generally looks at the nose and inner eyebrows to make its predictions which are associated with AU4 (brow lowerer) and AU9 (nose wrinkler) respectively. However, when it correctly detects pain it only uses the nose and part of the inner eyebrows, whereas, in the other attention maps, we can see that it also considers other parts of the face including the corners of the image. Notably, When it incorrectly identifies pain its attention is concentrated on the inner eyebrow as well as the nose. It also uses the cheeks and chin when it incorrectly detects pain. Between the true prediction of no-pain and the false prediction of this class, there are no notable differences in the attention patterns.

In terms of ViViT whole, we can observe that it looked at smaller but more areas of the image than ViT when it made these predictions. The attention is spread out across the face in particular the areas such as the forehead, cheeks and nasolabial furrows which become visible during AU10 (upper lip raiser), which again are relevant areas to pain detection from thermal images of the face (Alqudah, 2021). In the case of true positives, we see that it focuses on the forehead and eyebrow areas as well as the outer eye, cheek, and mouth areas, which are associated with AU17 (chin raiser), AU1 (inner brow raiser), and AU6 (cheek raiser). However, in the case of false positives, it focuses on a diagonal of the face and several other spots such as the outer chin, forehead, eyes, and eyebrows, showing that it uses several different areas often when it makes mistakes. However, the attention maps of the true negatives and false negatives are very similar.

Observing the attention maps of ViViT ROI we see very little attention being paid in general. When it correctly identifies pain it looks at the left cheek, nose, and forehead, which are associated with AU6 (cheek raiser) and AU4 (brow lowerer). The nose area that is detected is not related to any pain-relevant AU though. On the other hand, it does not look at the forehead but at the left cheek, nose, and right eyebrow when correctly identifying no-pain, albeit the attention on the left cheek is not as predominant as the other used areas. Also, when it makes incorrect predictions it looks at the nose and eyebrow, but we see less attention being paid to the eyebrow and there is no use of the cheek area in these instances.

Table 8. Confusion matrix of attention maps created by 100 samples, for the ViT models

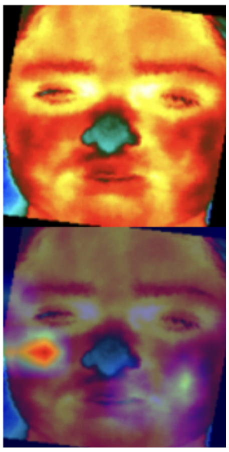
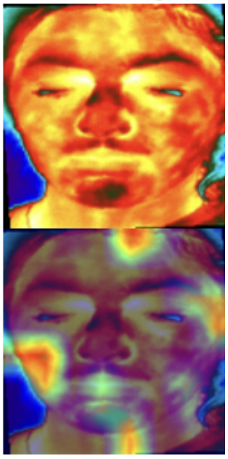
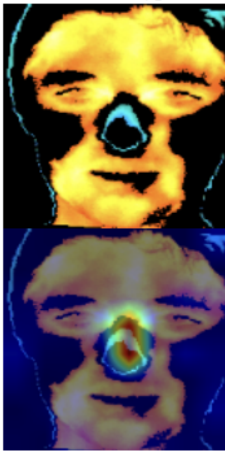
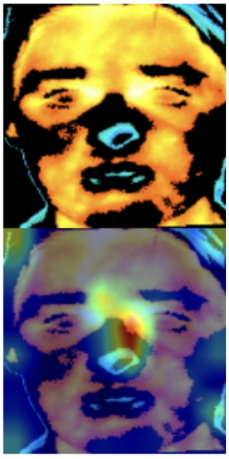
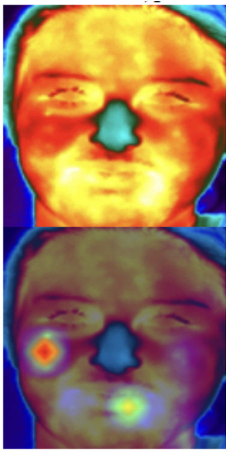
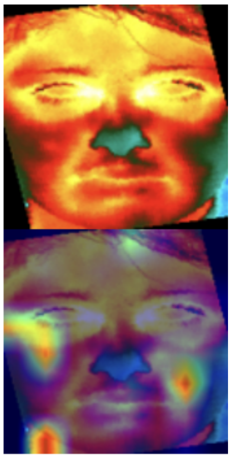
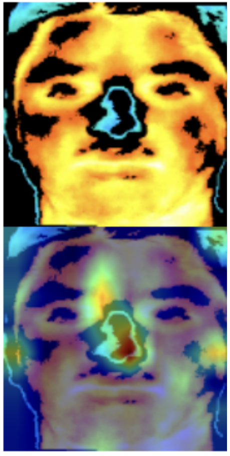
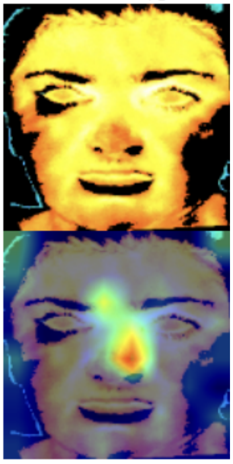
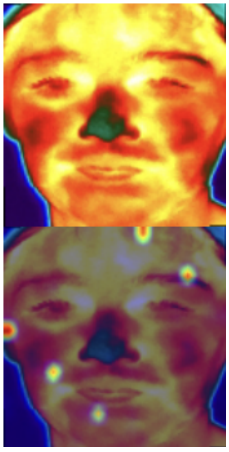
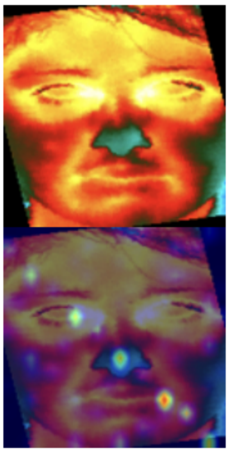
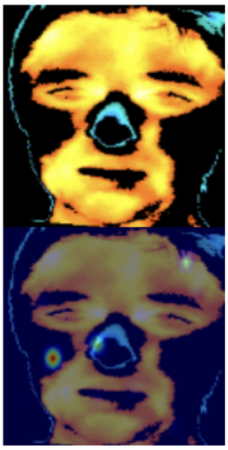
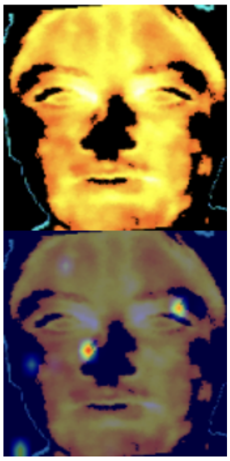
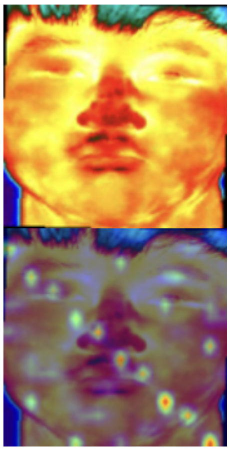
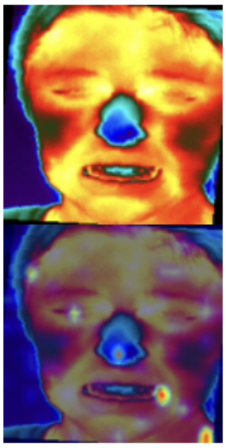
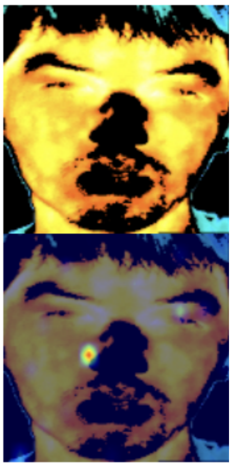
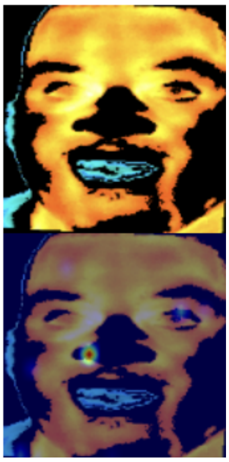
	ViT whole		**ViT ROI**	
	Positive	**Negative**	**Positive**	**Negative**
True				
False				

Table 9. Confusion matrix of attention maps created by 100 samples, for the ViViT models

	ViViT whole		**ViViT ROI**	
	Positive	**Negative**	**Positive**	**Negative**
True				
False				

5. Discussion

In this study, we aimed to answer the question "To what extent can a vision transformer discern between pain and no-pain from thermal images" and its subquestions "How does the use of pain-specific ROI extraction influence the performance of the vision transformers ViT and ViViT" and "How does the use of temporal information influence the ability of a transformer model in the pain detection task?". To answer these questions we conducted four experiments where we compared ViT and ViViT models to compare single frame predictions to image sequence predictions and we compared the same models using the images where we extracted the regions of interest and used these as input.

The results show us several insights that we want to discuss below. However, it is important to note that these insights emerge in a context where the overall performance of the models is underwhelming, compared to previous research. Nevertheless, because all models were trained in equal conditions, we believe that the results can tell us something about the impact of ROIs and temporal sequences when we compare the models directly with each other.

First of all, our experiments demonstrate that the inclusion of temporal data has a positive effect on the quality of models, particularly in the context of whole images. ViViT whole, showed superior performance over all other model configurations across all evaluation metrics. The higher accuracy, recall, precision, and F1 score achieved by ViViT whole shows us that the inclusion of temporal context enhances the model's ability to detect pain from thermal video frames. This observation is consistent with recent literature that underscores the significance of temporal data in image analysis tasks (Arnab et al., 2021).

More surprisingly, the ViT ROI also performed quite well on the pain detection task compared to ViT whole and ViViT ROI, outperforming both configurations in all four metrics. This result shows the impact of region-specific information extraction in enhancing model efficacy for single-image pain prediction.

In contrast, highlighting the performance difference between ViViT whole and ViViT ROI shows that only for single image prediction ROI extraction proved effective in the pain detection task. This might be explained by the fact that ViViT uses image features over time to make its prediction, because, as we can also observe in the attention maps of the ViViT models, this results in different patches interacting with each other to make a final prediction. By extracting ROIs you may disrupt this interaction of relevant parts, that the

ROI algorithm sees as irrelevant, losing these important interactions, and resulting in a loss of predictive power.

On the other hand, for single image prediction, the image is the entire context. This means that all interactions happen in that one image. When we apply ROI extraction to these images we do not have the risk of destroying useful information for the next image in a sequence, which means that in contrast to ViViT, ROI extraction can help ViT to focus the attention on the most relevant parts of the image.

In addition, it is important to note, that we used a dynamic ROI extraction method, meaning that the patches that we extracted changed per subject and might even change per frame making it more likely that relevant information over time might change or become unavailable over time. Potentially, if we were to extract a standard handcrafted ROI pattern, which would result in the same areas of the face being extracted from the images, the results might be better for ViViT ROI.

Furthermore, when analyzing the attention maps, it can be observed that ViT whole focuses on the mouth and cheeks, whereas ViT ROI uses the nose and the inner eyebrows for its predictions. They both seem to use areas associated with relevant AUs to pain. However, when the models start making mistakes they tend to include areas that are not as relevant for pain prediction, such as the left chin corner for ViT whole and the image corners and outer cheeks for ViT ROI. On the one hand, this shows the models do learn useful patterns, possibly due to the relevant AUs in these areas. On the other hand, however, the reliance on information that does not convey any AU signals is still present and seems to contribute to the models' mistakes.

Interestingly enough ViT whole tends to use different areas between its pains versus no-pain predictions whereas ViT ROI mainly focuses on the nose and inner brow area. This might also explain the edge that ViT ROI has over ViT whole because it mainly focuses on two areas relevant to pain detection in the face. It focuses on regions relevant to AU4 (lowering eyebrows) and AU9 (nose wrinkling) which might enable it to find criteria that discern pain versus no-pain better. This is underscored by the observation that it tends to make mistakes when it uses information from other less relevant areas of the face.

However, we can observe that ViViT whole looks at multiple relevant AUs when it makes correct predictions, which might make it more robust to noise in the thermal space. In the case of the false positives by ViViT whole, we see that it looks at all kinds of areas and not the set pattern we see in the true positive predictions. Notably, it heavily depends on a part of the neck when it makes mistakes, showing that the model starts to use patches that do not convey any AU information when it makes mistakes.

Furthermore, we can also observe that for both ViT and ViViT, it seems that they find similar patterns between their respective true and false predictions. This means that the models find a pattern when they believe it is pain or no-pain, that remains the same between true and false predictions, even though the ground truth might be different. This might be because the variability of the "texture" of the face is very high, meaning that some faces are very bright yellow while others might be very dark in comparison. This results in the models finding useful patterns looking at a specific area such as the cheek or nose area, but because of the variability in appearance, it might find that it looks the same as another pain case, however, if a larger context was used it might have found that it is not a pain frame.

Also, the fact that the models seem to rely on AUs to make their predictions might impact performance negatively, because it looks at an image where the wrinkles or deformations, that would telegraph AUs quite clearly in the RGB space, are not as present in the thermal space. This means that the probability of making a mistake by honing in on a cheek area, for example, might be more prone to mistake, especially when not including a larger context. Moreover, the presence of other patterns such as thermal differences in the neck or other parts of the face that might not convey any AU information might become noise instead of useful information, making the model focus on parts of the image that do not convey trustworthy information, causing the mistakes.

Based on previous observations, it is noteworthy that we discussed one of the benefits of pain prediction on thermal face images as that it can be used for people who are unable to move their face partially or completely. However, both ViT and ViViT seem to use mostly AU information to make their prediction as we can see that they tend to focus on relevant AU areas when making predictions. Even though these areas often correspond to relevant thermal areas for pain namely the maxillary and the forehead, due to the overlap we cannot determine if it uses the thermal information as is, or that it uses the thermal textures that are created by facial expressions. This means that one of the benefits of using thermal face images for pain prediction could be compromised if it solely uses AUs to make predictions.

Finally, it is also worth pointing out that our results provide a useful comparison to the work by Olczyk and Önal (2021), who used whole videos as sequences for pain detection instead of single frames and short sequences. The model architecture used in this paper was also different, namely, they used a pretrained CNN combined with an LSTM and achieved an accuracy of 84.37%. This not only underscores the advantages of having longer temporal sequences for the task of pain detection on thermal data but might also indicate that for pain detection in the thermal domain sequences, ViViT might not be the preferred model architecture. Furthermore, it is important to point out that the task at hand was slightly different. Where Olczyk and Önal (2021) investigated the question "What have you seen in the past frames?", whereas this paper investigated the question "based on what you have seen in the past frames, what are you seeing now?". It might be the case that the latter question is harder, for a model, to answer than the one asked by Olczyk and Önal (2021) because in the latter question, the model might find the same pattern in the first nine frames but the label is different because the last frame differs slightly. Whereas in the question answered by Olczyk and Önal (2021) a similar pattern results in the same label, because if the pain has been depicted somewhere in the video the label is pain.

5.1 Limitations

In our study, it is important to acknowledge certain limitations, particularly in the context of hyperparameter tuning.

One important limitation of our study was the observed tendency of our models to overfit on the training data. While we identified that employing smaller models, lower learning rates, and dropout rates was effective in reducing overfitting, our exploration in this area was limited. There may exist other approaches or combinations of hyperparameters that could offer improved results. Making our overall results not conclusive on our first research question. To address this limitation, further research could involve a wider range of hyperparameter settings and model variations to gain a more comprehensive understanding of the mitigation of overfitting with ViT and ViViT models.

Additionally, it's worth noting that we did not employ cross-validation in our analysis. This means that the results presented in this study might be skewed to the upside or the downside due to the train-validation-test data split, that we used. We acknowledge that cross-validation could provide a more robust assessment of the model's generalization capabilities.

Furthermore, we constrained our hyperparameter search space to a relatively small range, primarily due to resource limitations. A broader exploration, starting with an initial search space of 20-50 samples and gradually narrowing down promising regions, has the potential to yield additional insights. For example, our findings suggested that smaller model architectures tended to reduce overfitting, a more comprehensive search with smaller models would be promising. In a more resource-rich setting, we could have extended our investigation to include a broader array of hyperparameters and model configurations, thus enabling a more comprehensive analysis and the possible discovery of additional optimization opportunities.

It's important to note that the primary research question in our study revolved around the extent to which a vision transformer can discern between pain and non-pain thermal images of human faces. However, the limited search space for hyperparameter tuning means that the results presented in this study may not provide a conclusive answer to this question. Further investigation with an expanded hyperparameter search and larger computational resources could help address this central question more comprehensively.

Moreover, including a comparison between fine-tuned models, models trained from scratch, and models where the part of the body was frozen could have enriched our findings and would have given for a more robust exploration of the ViT and ViViT capability to predict pain from thermal images.

Finally, it is essential to recognize that the subset of the dataset that was used in this study consisted solely of pain and neutral (no-pain) frames. We chose this subset to maintain a balanced dataset, ensuring equal representation of pain and no-pain cases. However, this results in that the algorithms developed and tested in this research might face challenges when confronted with the broader spectrum of human emotions. In real-world scenarios, human facial expressions encompass a wide range of emotional nuances, from joy to sadness, anger, and more. The models' generalization capabilities in such a landscape will probably be quite poor. Future investigations might benefit from expanding the dataset to encompass a broader array of emotions, thereby allowing for a more comprehensive assessment of the model's robustness and applicability in a wider range of emotional contexts.

5.2 Future work

The shortcomings mentioned above also provide room for some future improvement and work.

First of all, to improve our model's overall performance, longer training times and a more extensive hyperparameter search quite possibly will improve the performance of the models. Also, using cross-validation will provide a more robust evaluation of the model's ability to generalize over new data. In addition, further research into smaller model architectures, as well as the strategic freezing of layers, can be explored to mitigate overfitting.

When we compare the results of the models in this paper with the results of previous work on the same data by Olczyk and Önal (2021), two other opportunities for future research are worth exploring. The first one is to explore the impact of different lengths of temporal sequences on model performances. This work primarily focuses on short temporal sequences of thermal images, namely ten frames. However, further investigations could involve incorporating longer temporal sequences, to assess how extended temporal context affects ViViT's ability to predict pain from thermal facial frames. This research may offer insights into the trade-offs between computational complexity and improved recognition accuracy, providing valuable information for the development of more robust and versatile models in thermal image analysis, that can be deployed in the real world.

The second one is the investigation of different model architectures on the ROI image data, because our study demonstrated the efficacy of ROI extraction, it is worthwhile to investigate whether other model architectures may interact more harmoniously with the ROI data. In a sense extracting ROI patches already is a type of attention mechanism it only presents the algorithm with "relevant data" patches. Another model architecture such as a CNN in combination with a LSTM or GRU for temporal data and just with a linear layer for single image prediction might prove to work better on this type of data because it uses the whole image to make its predictions, rather than use a subset of these features.

Furthermore, our current approach employs dynamic regions of interest, but future work could assess the performance of ROIs when working with default, static regions. It could prove beneficial to use predefined ROIs, especially with the ViViT architectures because the information that is present at the beginning of the sequence would be still available at the end.

ROI-based methods may find more significant results in tasks involving emotion distinction rather than binary pain classifications. Emotion recognition tasks consist of a wide variety of emotions all working on the autonomous nervous system in their own way resulting in higher variability in facial expressions and temperature patterns. This might make the colour temperature differences in the face more pronounced. Future studies could investigate the potential advantages of ROI-based data in combination with ViT and ViViT models in addressing the broader spectrum of human emotions.

6. Conclusions

This research tried to find the answer to the questions: "To what extent can a vision transformer differentiate between pain and no-pain in thermal images?" and its subquestions, 2) "How does the use of pain-specific ROI extraction impact the performance of the vision transformers ViT and ViViT?" and 3) "How does the inclusion of temporal information affect the transformer model's effectiveness in pain detection?"

To answer these questions, four models were fine-tuned: ViT whole, ViT ROI, ViViT whole, ViViT ROI. Notably, we observed that all models tended to overfit. However, smaller models with a depth of 1 to 3 layers, lower learning rates, and the incorporation of small dropout rates for both hidden and attention layers, turned out to be a step in the right direction in terms of model performance.

When we compared the different model's performances, we found that including temporal sequences improved the model's performance. But more surprisingly we found that extracting relevant ROIs from images improved model performance on single image predictions. However, the combination of these two techniques showed less pronounced results. We discussed various avenues for further research, with one of the first improvements in this current research being a more comprehensive hyperparameter search for all four models.

In conclusion, when using transformer-based methods, this research showed that it proves quite difficult to prevent over-fitting when using ViT or ViViT in the pain detection task using thermal data. However, we were able to showcase the effectiveness of including temporal sequences to enhance the model's ability to predict pain from thermal frames. Also, we showed that in the context of single-image prediction, the extraction of relevant ROIs before inputting data into the ViT model improves its ability to detect pain from thermal images of the face.

Bibliography

(2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A.-F. E. R., editor, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Abd Latif, M. H., Md. Yusof, H., Sidek, S., and Rusli, N. (2015). Thermal imaging based affective state recognition. In *2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*, pages 214–219.

Achterberg, W., Pieper, M., Dalen-Kok, A., Waal, M., Husebø, B., Lautenbacher, S., Kunz, M., Scherder, E., and Corbett, A. (2013). Pain management in patiens with dementia. *Clinical interventions in aging*, 8:1471–1482.

Alqudah, M. (2021). Affective state recognition using thermal-based imaging: A survey. *Computer Systems Science and Engineering*, 37:47–62.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., and Schmid, C. (2021). Vivit: A video vision transformer. *CoRR*, abs/2103.15691.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2019). BoTorch: Programmable Bayesian Optimization in PyTorch. *arxiv e-prints*.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. volume 3951, pages 404–417.

Bellantonio, M., Haque, M., Rodriguez, P., Nasrollahi, K., Telve, T., Escalera, S., Gonzàlez, J., Moeslund, T., Rasti, P., and Anbarjafari, G. (2017). Spatio-temporal pain recognition in cnn-based super-resolved facial images. pages 151–162.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bisong, E. (2019). *Google Colaboratory*, pages 59–64.

Bradski, G. and Kaehler, A. (2000). The opencv library. *Dr. Dobb's Journal of Software Tools*, pages 120; 122–125.

Chen, J., Chi, Z., and Fu, H. (2015). A new approach for pain event detection in video. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 250–254. IEEE.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In *European conference on computer vision*, pages 484–498. Springer.

Craig, K. D. (1992). The facial expression of pain better than a thousand words? *APS Journal*, 1(3):153–162.

Cristianini, Nello, and Shawe-Taylor, J. (2001). An introduction to support vector machines and other kernel-based learning methods. repr. *Introduction to Support Vector Machines and other Kernel-Based Learning Methods*, 22.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

Dehghani, H., Tavangar, H., and Ghandehari, A. (2014). Validity and reliability of behavioral pain scale in patients with low level of consciousness due to head trauma hospitalized in intensive care unit. *Archives of trauma research*, 3(1).

Delisle Rodriguez, D., Goulart, C., Valadão, C., Funayama, D., Favarato, A., Baldo, G., Binotte, V., Caldeira, E., and Bastos, T. (2019). Visual and thermal image processing for facial specific landmark detection to infer emotions in a child-robot interaction. *Sensors*, 19:2844.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. (2019). Retinaface: Single-stage dense face localisation in the wild.

Dey Roy, S., BHOWMIK, M., Saha, P., and Ghosh, A. (2016). An approach for automatic pain detection through facial expression. *Procedia Computer Science*, 84:99–106.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Erel, V. K. and Özkan, H. S. (2017). Thermal camera as a pain monitor. *Journal of pain research*, 10:2827.

Facebook opensource (2023). Why ax? `https://ax.dev/docs/why-ax.html`.

Gabor, D. (1946). Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93:429–441.

Gildenblat, J. (2022). Exploring explainability for vision transformers. `https://jacobgil.github.io/deeplearning/vision-transformer-explainability`. Accessed on sep 01, 202.

Hamunen, K., Kontinen, V., Hakala, E., Talke, P., Paloheimo, M., and Kalso, E. (2012). Effect of pain on autonomic nervous system indices derived from photoplethysmography in healthy volunteers. *BJA: British Journal of Anaesthesia*, 108(5):838–844.

Haque, M. A., Bautista, R. B., Noroozi, F., Kulkarni, K., Laursen, C. B., Irani, R., Bellantonio, M., Escalera, S., Anbarjafari, G., Nasrollahi, K., Andersen, O. K., Spaich, E. G., and Moeslund, T. B. (2018). Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 250–257.

Hassan, T. C. (2017). Recognizing emotions conveyed through facial expressions. Technical report, Fachbereich Informatik.

Herr, K., Coyne, P. J., McCaffery, M., Manworren, R., and Merkel, S. (2011). Pain assessment in the patient unable to self-report: position statement with clinical practice recommendations. *Pain management nursing*, 12(4):230–250.

Hess, U., Adams, R. B., Simard, A., Stevenson, M. T., and Kleck, R. E. (2012). Smiling and sad wrinkles: Age-related changes in the face and the perception of emotions and intentions. *Journal of experimental social psychology*, 48 6:1377–1380.

Huggingface (2022a). Video vision transformer (vivit). `https://huggingface.co/docs/transformers/main/en/model_doc/vivit`. Accessed on dec 01, 2022.

Huggingface (2022b). Vision transformer (vit). `https://huggingface.co/docs/transformers/model_doc/vit`. Accessed on dec 01, 2022.

Huggingface (2023). Fine-tuning a pretrained model. `https://huggingface.co/docs/transformers/v4.15.0/training`. accesed on 2023-10-19.

John J. Bonica, M. (1979). Editorial the need of a taxonomy. *PAIN*, 6(3):247–252.

Jurafsky, D. and Martin, J. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 3.

Kappesser, J. and Williams, A. C. d. C. (2010). Pain estimation: asking the right questions. *Pain*, 148(2):184–187.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The kinetics human action video dataset.

Khan, R. A., Meyer, A., Konik, H., and Bouakaz, S. (2013). Pain detection through shape and appearance features. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Kumar, C. and Pai N, S. (2017). A survey on human emotion analysis using thermal imaging and physiological variables. *International Journal of Scientific Research Growth*.

Kunz, M., Peter, J., Huster, S., and Lautenbacher, S. (2013). Pain and disgust: The facial signaling of two aversive bodily experiences. *PloS one*, 8:e83277.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Leone, M., Cecchini, A., Mea, E., Tullo, V., Curone, M., and Bussone, G. (2006). Neuroimaging and pain: A window on the autonomic nervous system. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 27 Suppl 2:S134–7.

Li, S. and Deng, W. (2018). Deep facial expression recognition: A survey. *CoRR*, abs/1804.08348.

Li, Y. (2004). Evaluation of face resolution for expression analysis. pages 82– 82.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Liu, P. and Yin, L. (2017). Spontaneous thermal facial expression analysis based on trajectory-pooled fisher vector descriptor. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 835–840, Los Alamitos, CA, USA. IEEE Computer Society.

Lopez, M. B., del Blanco, C. R., and Garcia, N. (2017). Detecting exercise-induced fatigue using thermal imaging and deep learning. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6.

Lu, G., Li, X., and Li, H. (2008). Facial expression recognition for neonatal pain assessment. *2008 International Conference on Neural Networks and Signal Processing*, pages 456–460.

Lucey, P., Cohn, J., Prkachin, K., Solomon, P., and Matthews, I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. pages 57–64.

Lynch, M. E. (2011). The need for a canadian pain strategy.

Ma, F., Sun, B., and Li, S. (2021). Robust facial expression recognition with convolutional visual transformers. *CoRR*, abs/2103.16854.

Mende-Siedlecki, P., Lin, J., Ferron, S., Gibbons, C., Drain, A., and Goharzad, A. (2021). Seeing no pain: Assessing the generalizability of racial bias in pain perception. *Emotion*.

Mitchell, A. and Boss, B. J. (2002). Adverse effects of pain on the nervous system of newborns and young children: a review of the literature. *Journal of Neuroscience Nursing*, 34(5):228.

Mäntyselkä, P., Kumpusalo, E., Ahonen, R., Kumpusalo, A., Kauhanen, J., Viinamäki, H., Halonen, P., and Takala, J. (2001). Pain as a reason to visit the doctor: A study in finnish primary health care. *Pain*, 89:175–80.

Nanni, L., Brahnam, S., and Lumini, A. (2010). A local approach based on a local binary patterns variant texture descriptor for classifying pain states. *Expert Systems with Applications*, 37(12):7888–7894.

Neshov, N. N. and Manolova, A. H. (2015). Pain detection from facial characteristics using supervised descent method. *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 1:251–256.

Nguyen, H., Chen, F., Kotani, K., and Le, B. (2014). Fusion of visible images and thermal image sequences for automated facial emotion estimation. *J. Mobile Multimedia*, 10:294–308.

Nguyen, T., Tran, K., and Nguyen, H. (2018). Towards thermal region of interest for human emotion estimation. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 152–157.

Ojala, T., Pietikainen, M., and Harwood, D. (1994). Performance evaluation of texture measures with classification based on kullback discrimination of distributions. *Proceedings of 12th International Conference on Pattern Recognition*, 1:582–585 vol.1.

Olczyk, A. and Önal, I. (2021). Pain recognition from thermal videos using deep neural networks. 33rd Benelux Conference on Artificial Intelligence and the 30th Belgian Dutch Conference on Machine Learning , BNAIC/BENELEARN 2021 ; Conference date: 10-11-2021 Through 12-11-2021.

Ordun, C., Raff, E., and Purushotham, S. (2020). The use of AI for thermal emotion recognition: A review of problems and limitations in standard design and data. *CoRR*, abs/2009.10589.

Prawiro, H., Pan, T.-Y., and Hu, M.-C. (2020). An empirical study of emotion recognition from thermal video based on deep neural networks. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 407–410.

Prkachin, K. M. and Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274.

Rathee, N. and Ganotra, D. (2016). Multiview distance metric learning on facial feature descriptors for automatic pain intensity detection. *Computer Vision and Image Understanding*, 147:77–86.

Rojo, R., Prados-Frutos, J. C., and López-Valverde, A. (2015a). Pain assessment using the facial action coding system. a systematic review. *Medicina Clínica (English Edition)*, 145(8):350–355.

Rojo, R., Prados-Frutos, J. C., and López-Valverde, A. (2015b). Pain assessment using the facial action coding system. a systematic review. *Medicina Clínica (English Edition)*, 145(8):350–355.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Rupenga, M. and Vadapalli, H. (2016). Automatic spontaneous pain recognition using supervised classification learning algorithms. *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6.

Schmid, U., Siebers, M., Seuß, D., Kunz, M., and Lautenbacher, S. (2012). Applying grammar inference to identify generalized patterns of facial expressions of pain.

Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE.

Serengil, S. I. and Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.

Shen, L. and Bai, L. (2006). Mutualboost learning for selecting gabor features for face recognition. *Pattern Recognition Letters*, 27(15):1758–1767. Vision for Crime Detection and Prevention.

Singh, S. K. and Singh, S. (2015). Sift and surf performance evaluation for pain assessment using facial expressions. *J Biol Eng Res Rev*, 2(1):6–14.

Valstar, M., Patras, I., and Pantic, M. (2005). Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. volume 3, pages 76–76.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H. C., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A. O., and Moreira da Silva, G. (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)*, pages 128–131.

Webber, S. G. (1876). Pain as a symptom in facial paralysis, and its cause. *The Boston Medical and Surgical Journal*, 95(26):750–755.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. volume 9911, pages 499–515.

Werner, P., Lopez-Martinez, D., Walter, S., Al-Hamadi, A., Gruss, S., and Picard, R. W. (2022). Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 13(1):530–552.

Xiang, X. and Wang, F. (2017). Transferring a face verification network for expression intensity regression.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. pages 532–539.

Yang, R., Tong, S., Bordallo, M., Boutellaa, E., Peng, J., Feng, X., and Hadid, A. (2016a). On pain assessment from facial videos using spatio-temporal local descriptors. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.

Yang, R., Tong, S., Bordallo Lopez, M., Boutellaa, E., Peng, J., Feng, X., and Hadid, A. (2016b). On pain assessment from facial videos using spatio-temporal local descriptors.

Zaanen, M. (2000). Abl: Alignment-based learning. pages 961–967.

Zhang, W., Zhang, Z., Qiu, F., Wang, S., Ma, B., Zeng, H., An, R., and Ding, Y. (2022). Transformer-based multimodal information fusion for facial expression analysis.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706. Best of Automatic Face and Gesture Recognition 2013.

Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J. F., Ji, Q., and Yin, L. (2016). Multimodal spontaneous emotion corpus for human behavior analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446.

Zhang, Z., Lyons, M., Schuster, M., and Akamatsu, S. (1998). Comparison between geometry-based and gabor wavelets-based facial expression recognition using multi-layer perceptron. pages 454 – 459.

Zhou, J., Hong, X., Su, F., and Zhao, G. (2016). Recurrent convolutional neural network regression for continuous pain intensity estimation in video.