Utrecht University
Faculty of Science

Artificial Intelligence Master Thesis

Interpretable and explainable vision and video vision transformers for pain detection

Project Supervisor:
Itir Önal Ertugrul

Candidate:
Giacomo Fiorentini

Second Examiner:
Albert Ali Salah

Academic Year 2021/2022

Abstract

Automatic detection of facial indicators of pain has many useful applications in the healthcare domain. Vision transformers are a top-performing architecture in computer vision, with little research on their use for pain assessment. In this thesis, we propose the first fully-attentive automated pain assessment pipeline that achieves state-of-the-art performance on direct and indirect pain detection from facial expressions. The models are trained on the UNBC-McMaster dataset, after faces are 3D-registered and rotated to the canonical frontal view. In our direct pain detection experiments we identify important areas of the hyperparameter space and their interaction with vision and video vision transformers, obtaining three noteworthy models. We also test these models on indirect pain detection and direct and indirect pain intensity estimation. Our indirect pain detection models underperform compared to their direct counterparts, but still outperform previous works while providing explanations for their predictions. We analyze the attention maps of one of our direct pain detection models, finding reasonable interpretations for its predictions. We find the models to perform much worse on pain intensity estimation, showing the limits of the simple approach chosen. We also evaluate Mixup, an augmentation technique, and Sharpness-Aware Minimization, an optimizer, with no success. Our presented models for direct pain detection, ViT-1-D (F1 score 0.55 ± 0.15), ViViT-1-D (F1 score 0.55 ± 0.13), and ViViT-2-D (F1 score 0.49 ± 0.04), all outperform earlier works, showing the potential of vision transformers for pain detection.

Contents

1 Introduction

The International Association for the Study of Pain defines pain as "An unpleasant sensory and emotional experience associated with or resembling that associated with, actual or potential tissue damage" [1]. In Europe, one adult in five suffers from moderate to severe chronic pain, with major consequences for their lives and well-being. 20% of them suffer from depression or have lost their job because of pain. 40% are unsatisfied with their treatment, and 30% are not being treated at all. Their ability to sleep, walk, do chores, have sexual relations, live independently, and function normally feels limited or restricted [2]. Pain is a major healthcare problem that medical care needs to overcome.

Pain is a ubiquitous problem for hospital care as well, with a great deal of research dedicated to pain analysis, quantification, and understanding. To quantify pain, visual analogue scales (VAS) [3] and similar metrics are usually employed due to their convenience and simplicity. To measure pain with VAS, the patient has to point at its pain level on a horizontal scale ranging from absence to maximum pain. Unfortunately, this technique has the drawback of being subjective and easily influenced, therefore leaving much to be desired as the gold standard of pain assessment.

Furthermore, under many circumstances patients are unable to report their pain levels, due to their mental or physical condition, making self-reporting techniques unreliable and widely inapplicable [4–6]. On the other hand, in intensive care units, nursing staff can manually check each patient individually, adjusting pain medication on a case-by-case basis, achieving excellent but time-consuming results, although the large amount of staff, time, and money needed to implement this solution render its scope and applicability limited to a small scale [7]. In order to overcome the limitations of VAS and individual checks, automation alongside new metrics have to be employed. In this regard, facial expressions can be an important means of communication for the emotional state of a person, including their pain levels [8].

Facial expressions play an important role in communicating pain. The facial action coding system (FACS) [8] is a framework based on the anatomy of the facial muscles, and divides facial expression into 34 atomic components defined as action units (AU) with scores ranging from A to E depending on their intensity. While by itself this system contains no apparent information on the pain levels of the subject, the Prkachin and Solomon Pain Intensity (PSPI) score [9], identifies six AUs, grouped into four actions, that contain most of the information on pain.

Figure 1: Example frames from the UNBC-McMaster dataset and their PSPI score labels (Fiorentini et al.) [17].

These actions are brow-lowering (AU4), orbital tightening (AU6 and AU7), levator tightening (AU9 and AU10), and eye closure (AU43) [9]. The PSPI score, visible in Figure 1, is computed by taking the highest intensity AU component of each action and summing the numerical equivalent of their intensities (ranging from 0 to 5). As AU43 (eye closure) has only one possible intensity value, PSPI is therefore a 16-point pain scale.

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43 \quad (1)$$

While the PSPI metric, shown in Figure 1, does not rely on self-reporting, eliminating one of the aforementioned limitations, FACS coding requires an average training time of three months, with each trained expert taking on average over two hours to code a single minute of video [10]. In order to overcome this challenging drawback, automation is needed to predict the PSPI scores directly [11]. Desirable properties for such automated pain detection models are spatiotemporal reasoning [12], robustness to occlusion and changes in the environment [13, 14], explainability [15], interpretability [15], and accuracy [16]. Transformer models meet many of these requirements, making them good candidates for pain assessment pipelines.

Transformers are a recently developed deep learning model with roots in natural language processing (NLP) [18], entirely based on attention mechanisms, without convolutions or sequence-aligned recurrent neural networks. These models quickly achieved state-of-the-art performance in the NLP domain and have become its gold standard for many tasks, both thanks to their ability to draw global dependencies within their input and compute most of their operations in parallel, allowing for the training of massive models with over 100 billion parameters [19]. Shortly after its success in the NLP domain, research began to evaluate the use of transformer models for computer vision, obtaining excellent results when trained on large datasets [19].

Transformers have achieved state-of-the-art performances in multiple tasks, but only a few studies have researched their performance for pain detection [20]. The possibility to analyze spatiotemporal relations through video transformers [21], to extract attention maps and generate interpretations intrinsic to the model [22, 23], the ability to fine-tune these models on smaller datasets with good results [24], and their state-of-the-art performance on other computer vision tasks [25], are promising for their application towards pain assessment.

In this thesis, we evaluate the performance of vision transformers (ViT) and video vision transformers (ViViT) for pain detection and pain intensity estimation from facial features using a fully-attentive pipeline. The first part of this research, focusing on direct pain detection and interpretability, was published by Neurips for the VTTA2022 workshop [17]. Training is carried out on the UNBC-McMaster dataset using PSPI and AU labels under a variety of configurations, pinpointing regions of interest in the hyperparameter space during direct pain detection training. Techniques that have been shown to boost transformer performance are evaluated and adapted to the task, attempting to maximize model performance on direct pain detection. The best-performing configurations are then employed for indirect pain detection and direct and indirect pain intensity estimation. We extract attention maps and evaluate them, finding plausible interpretations for the prediction of the tested model. Attention maps and indirect pain assessment make our models achieve interpretability and explainability. We achieve state-of-the-art performance on the task of pain detection for the F1 score metric, even on our explainable models, demonstrating the potential of transformer models for pain assessment, and building foundations for future transformer research on this task.

1.1 Research questions

The main research question is:

- How do vision transformers perform on facial pain assessment?

In order to evaluate the performance of vision transformers on facial pain assessment we carry out four sets of experiments, on a variety of metrics, and compare the results obtained with previous work on the same task. The four tasks considered are direct pain detection, indirect pain detection, direct pain intensity estimation, and indirect pain intensity estimation.

During the overarching process, other important aspects will be evaluated through further analysis and ablation studies. This process will help estimate the impact of various design choices on the model's overall performance and explain its inner workings. Therefore, the sub-research questions are:

- What are the regions of interest in the hyperparameters space for vision transformers?

  We evaluate the performance of our pipeline on the task of direct pain detection over a variety of hyperparameters, finding regions of interest and employing the best-performing models found for successive experiments.

- How does indirect pain assessment, with an interim AU layer prediction, affect the performance of the model?

  We evaluate the performance of our pipeline when predicting AUs and then computing the PSPI formula rather than predicting the PSPI score directly. The former approach makes our models explainable, an important feature in the medical field. We evaluate this approach in two experiments, one on pain detection and the other on pain intensity estimation.

- How do vision transformers perform on facial pain intensity estimation?

  We evaluate the performance of our pipeline when attempting to predict not only the presence of pain but also its intensity. Quantification of pain is an important aspect of pain assessment, allowing a more complete understanding of the subject's pain levels. We test our approach in two experiments, one with direct PSPI score intensity predictions and the other employing an interim AU layer, predicting the intensity of each AU individually, and then summing them according to the PSPI formula.

- How do vision transformers and video vision transformers compare in performance?

  Facial pain assessment literature strongly supports the use of videos over individual frames to capture pain-related facial dynamics and transformer literature offers many approaches to integrating temporal analysis into its architecture. We choose to employ one of ViViT's proposed architectures and test its performance in all our experiments, comparing its results to ViT's.

- Can attention maps be used to generate plausible interpretations of the outputs?

  Interpretability is important to build trust in a model and would allow our model to justify its predictions comparably to a PSPI coding expert. We extract attention maps from one of our direct pain detection models to generate interpretations of its results and qualitatively evaluate their plausibility.

1.2 Outline

In this section, we have introduced the problem of pain, the facial action coding system, the Prkachin and Solomon Pain Intensity score, transformers, and the research questions. Next is Section 2, which thoroughly discusses previous literature on automated pain assessment, transformers, AU detection, facial dynamics, face frontalization, and facial pain dataset balancing. Then, in Section 3, we analyse in-depth the dataset, its label distribution, and outline our pre-processing steps. In Section 4 we discuss the technical aspects of the transformer architecture, our pipeline, and the experiments that will be carried out to answer our research questions. The results of the experiments are reported in Section 5 and compared to previous literature. Finally, in Section 6 we discuss the results and answer the research questions, and summarize our conclusions in Section 7.

2 Related work

Automated pain assessment Recent work has shown that direct pain assessment from facial expressions is a feasible goal, both with shallow and deep learning approaches. Zafar and Khan [26] train k-Nearest Neighbour (kNN) models on 22 facial landmarks, succeeding both in pain intensity estimation and AU detection, proving both to be achievable.

Hammal and Cohn [27] train support vector machines (SVM) on CAPP features extracted from the facial landmarks included in the UNBC dataset for pain intensity estimation. In its conclusion, the paper mentions indirect pain assessment as an untested approach, which we explore in this research. Werner et al. [28] also employ SVM models, focusing however on what they call "activity descriptors", sequence-level feature signals captured through facial landmarks and head pose. Both papers highlight the benefits of temporal integration, which we test with our ViViT model.

Before transformer models, attention could be included in models through long short-term memories (LSTM) and similar techniques. In the paper by Rodriguez et al. [29], the task of pain assessment is approached with a two-model pipeline: the first component is a deep convolutional neural network (DCNN) taking raw images as input, while the second component is an LSTM model, necessary to exploit the spatiotemporal relationship between input images. The authors note that while the DCNN performed well by itself, the introduction of the LSTM model improved the area under the curve (AUC) by almost 4%, reiterating once again the importance of spatiotemporal analysis for performance in facial expression recognition tasks. In this thesis, we compare the performance of ViT and ViViT models to evaluate the benefits of introducing temporal features in transformers.

Previous works have also successfully achieved indirect pain assessment [11, 30] by training models on AU labels rather than PSPI scores. Kaltwang et al. [11] employ three sets of features for the purpose of direct and indirect pain intensity estimation. The first set contains facial landmark points included in the dataset and extracted with an active appearance model (AAM), the second set contains features extracted from aligned facial images by applying the Discrete Cosine Transform, and the third set contains Local Binary Pattern features also extracted from the aligned images, with the best result obtained through late fusion of all these features. The paper claims that error propagation can be a key issue in indirect pain detection, as errors in AU detection compound into poor PSPI score predictions, a possible problem for our research as well.

However, AU predictions can also be employed as middle steps of longer pain assessment pipelines, a major departure from the approach of human experts. In the work of Xu et al. [30], statistical features are extracted from the interim AU predictions and then further processed before finally outputting a VAS score prediction, achieving great performance with indirect AU detection.

In our work, the model is trained on end-to-end pain-related AU detection and indicates both the AUs detected and the result of the PSPI formula applied to them, producing explainable results, in line with facial pain assessment literature, similarly the paper by Kaltwang et al. [11], and comparably to a human expert. In our experiments, we compare the results of indirect pain assessment with direct pain assessment to determine how our approach is affected by this variable.

Vision transformers After their success in neural machine translation [18], transformers have been used as standard in several NLP tasks. Yet, their application to vision-related tasks is relatively new. Dostovitskiy et al. have proposed vision transformers (ViT) and have shown that ViT outperforms CNN once it is trained on very large databases [19]. In comparison, earlier approaches to purely-attentive pipelines by Prajit et al. and Wang et al. failed to outperform CNNs, proving the effectiveness of this new architecture.

Specifically, Prajit et al. substitute all convolutions within a ResNet model with self-attention modules, outperforming the original ResNet implementation while remaining more computationally efficient, proving that self-attention can be an effective primary primitive for neural networks even in the complete absence of convolutions.

Wang et al. instead achieve competitive performance while factorizing 2D attention into vertical and horizontal 1D attentions, reducing computational complexity and allowing the model to compute attention globally rather than locally. Both approaches however are surpassed by ViT, which applies minimal modifications to the original NLP implementation, and remains one of the most generic and simple approaches to transformers for computer vision, making it our model of choice. Furthermore, video sequences can easily be processed by ViT models by making some simple adjustments to the tokenization process [21].

Recently, video vision transformers (ViViT) have been proposed to model spatiotemporal information and have been shown to achieve state-of-the-art performance on activity recognition in several settings. ViViT has outperformed earlier approaches that model spatiotemporal information [31, 32] and other temporal extensions of ViT [21]. ViViT is therefore not only the natural extension of ViT, requiring minimal changes to the positional embeddings to capture the temporal dimension, but also outperforms earlier works on multiple tasks, including the TimeSformer transformer model, which employs a similar architecture with variations in attention computation [33]. However, in the vision transformer domain, the architectures chosen are not state-of-the-art. Soon after their release, new models introduced convolutions [34], different types of attention [35], and alternative spatiotemporal embeddings [36], often surpassing the performance of the originals.

Wu et al. [34] introduce convolutions to transformer models for token embedding and before multi-headed attention on a reshaped 2d token map, outperforming ViT and achieving state-of-the-art performance on a variety of datasets for image classification. Ali et al. [35] replace self-attention with transposed attention, which operates across the feature dimension, achieving complexity linear to the number of tokens rather than quadratic and obtaining competitive performance on multiple tasks while remaining computationally efficient. Wang et al. [36] employ overlapping patch embedding alongside convolutions to improve the modelling of local context, surpassing their previous model with these improvements [37] and performing competitively compared to other works. ViT and ViViT however, are purely-attentive and represent the most general implementations of vision and video vision transformers, suitable to demonstrate the potential of purely-attentive architectures on end-to-end pain assessment under a variety of setups.

Transformers for pain detection Several works have shown the success of using vision transformers for facial expression recognition [38], and facial action unit detection [39]. However, their application in automated pain assessment is very scarce. To the best of our knowledge, the only existing work based on transformer technology for pain intensity estimation is by Xu and Liu [20].

The pipeline presented in this work focuses on end-to-end pain intensity estimation and includes both a CNN and a transformer. Pain-related features are first identified and extracted from the input images by a ResNet architecture with bottleneck attention modules, then processed by a transformer model that predicts pain intensity. The successful performance of our model on a similar task, while only fine-tuning a pre-trained transformer, contradicts their finding that a transformer alone does not work for pain assessment.

Transformers for AU detection As previously mentioned, recent works have begun exploring the potential of vision transformers for AU detection. Wang and Wang [39] propose a pipeline composed of a ResNet-based convolutional neural network and a two-branched transformer. A CNN is employed to extract feature maps of both fine and coarse resolution, each to be fed to a different branch of the transformer component. The results of the two branches are then recombined in the multi-layer perceptron head for AU prediction. Performance is tracked through the F1 score, and the BP4D and DISFA datasets are used for training. In comparison, our research on AU detection will be the first to show the potential of a purely-attentive pipeline for AU detection.

Facial pain dynamics Modeling temporal information has been shown to be crucial as a static approach based on Relevance Vector Regression [11] could not distinguish between eye blinks and eye closures, which are pivotal for pain intensity estimation. Recurrent convolutional neural networks (RCNN) [40] and a combination of CNNs with a long short-term memory (LSTM) networks [29] have been used to model spatiotemporal relationship among successive frames. They have shown superior performance compared to the static approaches. Feature engineering can also be used to introduce the temporal dimension directly to the dataset, by transforming the features into signals as Werner et al. have done [28]. Inspired by these findings, we compare the performance of ViT and ViViT on automated pain detection.

Face frontalization Face frontalization consists in tracking facial landmarks across a variety of subjects, poses, and environments, and aligning them to a frontal mask. Many earlier works employ no frontalization [29, 41] or 2D frontalization [42], which treats faces as 2D objects for the purposes of registration and alignment. The performance of these models however rapidly degrades with variations in pose due to self-occlusion [43]. In order to avoid this drawback, in this paper we employ 3D frontalization, a technique which treats faces as 3D objects to be fit into a 3D mask.

A variety of approaches were evaluated before choosing PRnet [44]. For 2D frontalization, dougsouza's face-frontalization software [45] was tested, but it rapidly proved to perform poorly on self-occluded images, in line with face frontalization literature. Then, OSTeC [46] was employed as the first attempt at 3D frontalization, but the generative process was both very computationally expensive and failed to preserve subject identity, a very important feature in facial pain assessment. Cleardusk's 3DDFA_V2 performed the most similarly to PRnet, however, we perceived the images to be less robust to self-occlusion, with large texture artifacts appearing on the 3D model compared to the results of the latter model.

PRnet [44] in comparison with the aforementioned methods is fast, robust to self-occlusion, and preserves subject identity making it our 3D registration software of choice for this research. Therefore, we employ PRnet to align facial components, and establish semantic correspondence between visual words across frames and subjects for our vision transformer model. Our approach improves on previous works by requiring no subject-specific training [47], achieving good results on unseen data.

Pain dataset imbalance Due to the delicate nature of the task and the high costs associated with PSPI coding, images containing painful facial expressions labelled with PSPI scores are very scarce. In order to overcome this obstacle, previous works successfully employed undersampling of the majority class [29, 48]. In our work, we instead begin by employing oversampling of the minority class for direct pain detection and, given the rapid convergence of our model, proceed to undersample the majority class for the remaining tasks.

Issues with dataset imbalance are under-reported and tackled without consistent standards across works. Due to the lower number of pain samples, especially for the most extreme pain score, earlier works employed 4 levels [27, 48–50], 6 levels [20, 29], 7 levels of pain [51], and 16 levels of pain [11, 30, 40], rarely justifying the rationale behind the choice. While it may seem that 4 pain levels are more common overall, the pain scores within those 4 levels are not standardized, even for works from the same authors [48, 49]. This problem further extends to binary classification, with works not using only PSPI scores of 0 as no pain [52], a departure from PSPI literature.

Our models employ PSPI scores of 0 as no pain labels, following many of the earlier works on literature, and employ 16 pain levels due to the lack of concrete grouping standards, enabling comparisons with earlier works that employ the same approach. Nevertheless, the lack of standards in pain assessment research remains an unsolved issue.

To the best of our knowledge, no work has reported on the effect of long sequence lengths on the dataset distribution and the metrics used, despite the impact that dataset imbalance can have on some metrics, such as the F1 score [53]. Previous works by Rezaei et al. [13] and Xu et al. [20] reported their best results on sequences of 40 frames or more, however, due to pain samples most often appearing in the middle of sequences in the UNBC-McMaster dataset, long sub-sequence lengths naturally remove from the dataset many no pain samples, altering the distribution of the test set significantly. In our work we employ single-frame and four-frame sequences, affecting the distribution of the dataset minimally and reducing the bias of sub-sequence length on our metrics.

Metrics In order to determine the overall performance of video transformers for pain assessment a variety of metrics will be employed, including F1 score, area under curve (AUC), mean square error (MSE), and mean absolute error (MAE). The choice of using a variety of metrics was dictated by the lack of a distinctly superior metric in current literature [53] and due to the large variety of metrics used by previous studies in the same domain, therefore maximizing the possible comparisons with previous research.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{2}$$

For direct and indirect pain detection we employ the F1 score and AUC, both on the overall pain predictions and the individual AU detection experiments. We compute the F1 score metric as twice the number of true positive samples, divided by twice the number of positive samples plus the number of false positive and false negative samples, where the positive class is the pain class. We employ the F1 score to demonstrate the ability of the model to predict pain presence, despite the imbalanced number of no pain frames in the dataset, to test our model on the harder task of pain detection. We employ a threshold of 0 on the output of the transformer head to determine the predicted label for AU detection, a balanced value which might have a significant effect on this metric. The AUC metric measures the ability of a model to correctly rank two randomly sampled frames, one of each class, and is not affected by dataset imbalance [53], making it a valuable metric on this dataset.

For direct and indirect pain intensity estimation we employ MSE and MAE. The former is computed as the summation of the square of the difference between the predicted intensity value and the real label. This metric is highly sensitive to outliers, unlike MAE, which is computed as the summation of the absolute difference between the predicted intensity value and the real label. Both provide important information on the performance of the model, either penalizing more the mistakes made on outliers with MSE, or focusing more on average samples with MAE.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \overline{y})^2 \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|y_i - \overline{y}| \tag{4}$$

3 Dataset

The models are trained on the UNBC-McMaster dataset [7], one of the most commonly used datasets for facial pain assessment. It consists of 200 video sequences and 48398 video frames, from 25 patients suffering from shoulder-related pain, captured as the patients performed active and passive range-of-motion tests with each of their limbs. The sizes of each frame vary, with 15838 having a resolution of 320x240 and 32560 having 352x240 instead. The dataset contains 12 male and 13 female subjects, resulting balanced in this regard.

Table 1: Grouped PSPI score frequency

PSPI score	0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16
Frequency	40029	5260	2214	512	132	99	124	23	5

Label distribution Reported in Table 1 are the statistics of the PSPI labels in the dataset; the distribution of the labels is clearly heavily imbalanced towards lower scores, with the no pain label taking up 83% of the total samples. In Table 2 are instead reported the frequencies of the relevant AU labels in the dataset; their distribution is heavily imbalanced with AU9 and AU10 appearing in less than 2% of the total samples present in the database, while AU6 is comparably far more common, appearing in 11% of the frames. Clearly, to avoid the model overfitting, biased sampling of the classes is necessary, as previous works have done [29, 48].

Table 2: Frequency of AUs relevant for the PSPI score

AUs	AU4	AU6	AU7	AU9	AU10	AU43
Frequency	2.21%	11.48%	6.95%	0.87%	1.08%	5.02%

Label preprocessing For the purposes of binary classification, we divide the dataset into two categories, 0 (no pain) and 1 (pain), the latter category including images with a PSPI score above 0. During training, the pain class is over-sampled to prevent overfitting on the majority class. For the purposes of AU detection, pain intensity estimation, and AU intensity estimation, we instead employ under-sampling of the no pain class.

Earlier works have at times grouped certain pain levels together to reduce the extreme unbalance between classes, however, due to the lack of standards in literature, groupings are inconsistent and varied between studies, and we have therefore decided not to combine any pain levels. During training, labels for regressions tasks are normalized between 0 and 1, while their metrics are computed on the non-normalized scores, allowing us to compare our results with previous works.

Video preprocessing The use of multiple subsequent frames aims to capture the dynamics of the facial expressions, enabling the model to distinguish between the subject shutting their eyes due to pain (AU43) and blinking and other critical dynamics of facial pain. Delving deeper into the analysis of the dataset, we computed the minimum, maximum and median length of consecutive identical AUs in order to estimate how long a sub-sequence would have to be to effectively capture AUs. Reported in Table 3 are the values recorded. As can be seen in the first column of Table 3, the minimum length of the AUs is between one and four frames, sub-sequence lengths of 4 and above should therefore be able to capture the entirety of the dynamics of the shorter AU sequences.

Table 3: Analysis of continuous AU frames length

	Length		
AUs	Minimum	Maximum	Median
AU4	4	191	41
AU6	2	278	51
AU7	1	290	47
AU9	1	27	7
AU10	1	14	5
AU43	1	135	27

As can be seen in column 2 and 3 of Table 3, the median and maximum lengths are often rather large and would have a significant impact on the number of samples and the distribution of the dataset if employed as sub-sequence lengths e.g. in order to generate sub-sequences with a length of 41, the median value for AU4, 40 samples from the beginning of each sequence would be lost. Due to the dataset having 200 sequences, in total, this would be a loss of 8000 samples or around 17% of the dataset. Furthermore, due to pain samples being mostly distributed towards the middle of the sequences, the distribution of the labels is also significantly affected, which might in turn bias the results on certain metrics [53]. We therefore process the dataset for the purposes of video vision transformer training - grouping subsequent images in 2×2 grids and labelling according to the label of the last image of the 4-frame sequence.

Data augmentation Another important aspect of the dataset is the use of data augmentation. Previous work by Steiner et al. [54] on the topic of data augmentation for transformers tested a variety of setups combining Mixup [55] and RandAugment [56], achieving results akin to increasing the dataset size tenfold. However, thanks to our pre-processing steps, faces in the samples are frontalized and positioned centrally within our images across all frames, making RandAugment unfit for our dataset.

Figure 2: An example of mixup, the frames are mixed with a blending value of 0.5 and their resulting label is [0.5, 0.5].

On the other hand, Mixup is an augmentation technique that takes a percentage of the dataset and generates hybrid images by blending frames with distinct labels. e.g. A painful frame labelled [1,0] is combined with a painless frame labelled [0,1], with a blending value of 0.2. The resulting frame is labelled [0.8, 0.2], and consists of the sum of the pixel intensity values from the painful and painless picture, the former at 80% opacity, and the latter at 20%. The hybridity of the images is controlled by the parameter α, with higher values providing images that borrow information equally from the two samples, and lower values causing most images to have mostly data from a single frame. An example of Mixup augmentation can be seen in Figure 2. Ultimately, we find Mixup to offer data augmentation suitable for our task, and employ it on 20% of the dataset in some of our experiments.

4 Method

4.1 Technical description

Before describing how our pipeline processes its inputs, we first explain the inner workings of ViT and NLP transformers. The approach used for NLP transformers can be implemented with little changes for vision transformers as well, granting the ViT architecture the same scalability and efficiency as the original one.

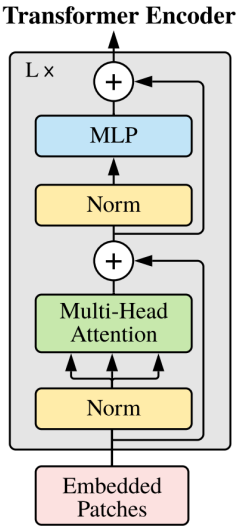

Figure 3: The architecture of a transformer encoder within the transformer model (Dosovitskiy et al.) [19].

Tokenization In order to process an image, it is first split into patches of fixed size by the feature extractor, then, the patches are flattened into a 1-dimensional sequence of tokens, the "visual word" equivalent of the vision transformer. Due to self-attention being computed globally inside transformers and the lack of other strong local biases within the architecture, 1D positional embeddings are used to capture local context. The embeddings are trained as part of the original ViT model and at initialization contain no information on their position on the 2D grid. Local context is therefore learnt as part of transformer pre-training, and the positional embedding weights are frozen and unchanged for the entirety of the experiments. Finally, transformer models can add a classifier token to its inputs or compute the mean of all token outputs as input for the transformer head, with our paper employing the former approach.

A series of Transformer encoders, visible in Figure 3, are then successively applied to the token input, each consisting of layer normalization, multi-headed attention, residual connections, layer normalization, a multi-layer perceptron, and another layer of residual connections. We will now introduce each of these components and their role in encoding.

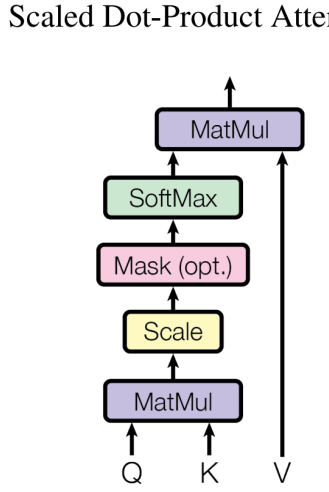

Figure 4: Self-attention computation (Vaswani et al.) [18].

Self-attention While not an explicit component of transformer models, understanding self-attention is necessary to understand multi-headed attention, so it will be introduced here. The type of self-attention employed by transformers is known as "scaled dot-product attention" and employs an input vector and three trained matrices to generate the Query, Key, and Value matrices, which are then used to compute the output of the self-attention layer.

The three matrices are computed by multiplying the input vector and the three trained matrices, while the output is computed by applying a softmax function to the product of the Query and Key matrices, scaled by the root of their length d_k, and then by multiplying by the Value matrix. The process is visible in Figure 4.

$$\begin{aligned} Q &= Input * Q_{trained} \\ K &= Input * K_{trained} \\ V &= Input * V_{trained} \\ Attention(Q, K, V) &= softmax(\frac{Q * K}{\sqrt{d_k}}) * V \end{aligned} \tag{5}$$

Multi-headed attention Multi-headed attention is the core of transformer models and where self-attention takes place. Multi-headed attention functions by applying a number of parallel self-attention layers, each with its own trained matrices, concatenating the results and multiplying the resulting matrix to an additional trained matrix, obtaining the final output. The main advantage of the multi-headed approach is having access to multiple representation sub-spaces as shown in Figure 5.

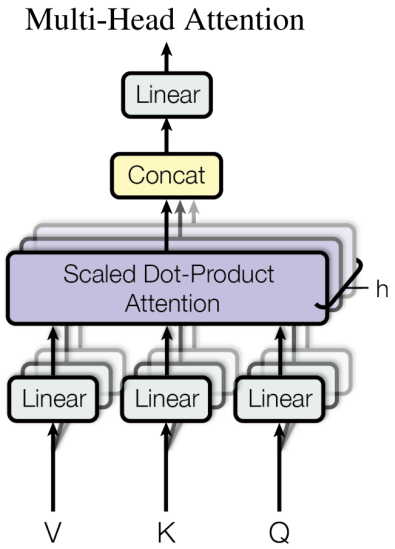

Figure 5: Multi-head attention and its multiple representation sub-spaces (Vaswani et al.) [18].

Layer normalization Layer normalization is a process introduced by Wang et al. [57] for efficient learning of deep encoding transformers, it functions by normalizing the input sample, thus avoiding normalizing the entire batch.

Residual connections Residual connections are the sum of the output of a layer with its input, with the overall structure of residual connections and layer normalization being credited to the work by Baevski and Auli [58].

Multi-layer perceptron The multi-layer perceptron (MLP) inside the transformer is composed of two layers with a Gaussian Error Linear Units (GELU) non-linearity, while last the multi-layer perceptron, also called the head of the transformer model, has a hidden layer when training the model, and a single linear layer when fine-tuning it during pre-training. NLP transformers have two branches of transformers instead, with the first encoding the input and the second decoding the output. The latter branch similarly employs an MLP with a linear output head.

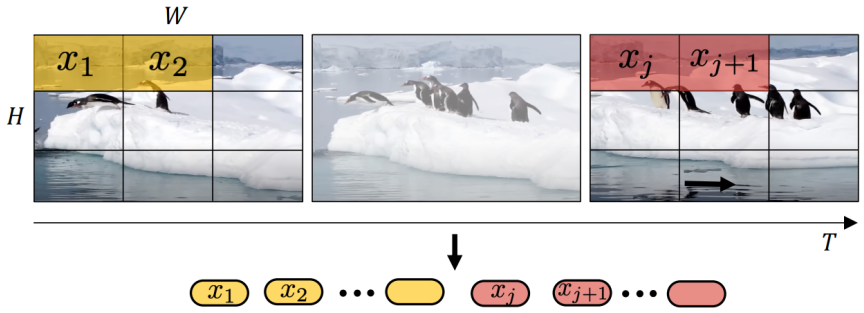

Figure 6: Combined spatiotemporal tokenization (Arnab et al.) [21].

Spatiotemporal attention ViViT transformers work identically to ViT transformers, except patches are extracted from each individual frame and then concatenated, enabling the computation of temporal attention. Variants of the model can compute spatial and temporal attention either together (Figure 6 and Figure 7a), spatially per frame then combined temporally (Figure 7b), sequentially (Figure 7c), or in parallel (Figure 7d). The former approach has been initially explored in the ViViT paper [21] and suffers from quadratic complexity. Despite this drawback, it is by far the most straightforward and general video vision transformer approach and has the best results among the models presented, making it our approach of choice.

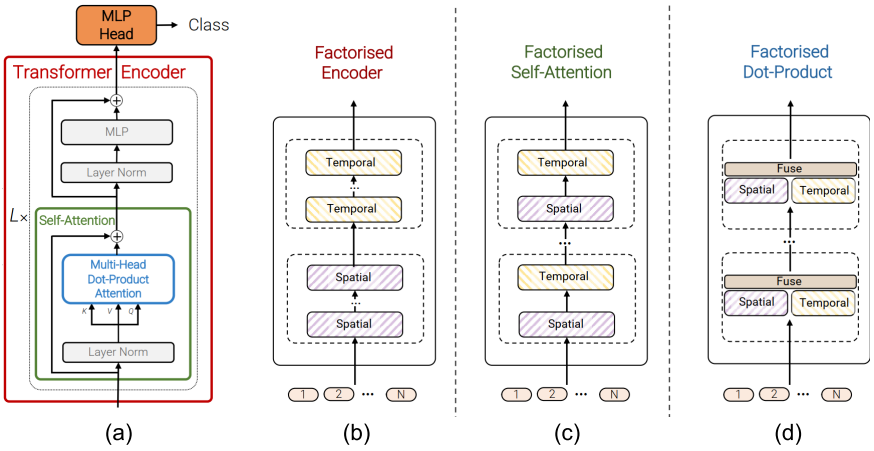

Figure 7: Attention types described in the ViViT paper (Arnab et al.) [21].

4.2 Pipeline

Pre-trained model The base transformer model is pre-trained on ImageNet-21k [59] and fine-tuned on ImageNet [60] at a resolution of 224x224 by HuggingFace [61]. It consists of 12 attention and 12 fully-connected layers, and employs 12 attention heads, for a total of 86 million parameters. At runtime all fully connected layers are frozen, massively reducing the training time required. Next, we describe the steps taken by our pipelines to make a prediction on an unprocessed image from the dataset.

Figure 8: An example of frame frontalization.

Face frontalization The first step in our pipeline is 3D frontalization, visible in Figure 8. We perform 3D registration using PRNet [44], which gets a 2D face image as input, performs 3D registration without requiring person-specific training, and outputs a dense 3D mesh of the face. The result is achieved by regressing the UV position map, a structure that records 3D coordinates of a complete facial point cloud, from the input image. We then use the face3d tool [62] to rasterize 2D image from frontalized 3D facial structure generated by PRNet as shown in Figure 10a. After this step, semantic correspondence is established across frames and subjects. Consequently, visual words used in vision transformers are aligned as given in Figure 10b. This step produces a dataset of frontalized facial images of size 256x256 for each frame.

Figure 9: The same sample in the single-frame and four-frame datasets, they share the same label.

Sub-sequence generation Images are further pre-processed for the purpose of training and testing video vision transformers on the task of pain prediction. Images are concatenated in 2×2 grids with their immediately preceding frames, creating a video sub-sequence frame with a resolution of 512x512 as visible in Figure 9 . This step generates a second distinct dataset, employed exclusively for video vision transformers.

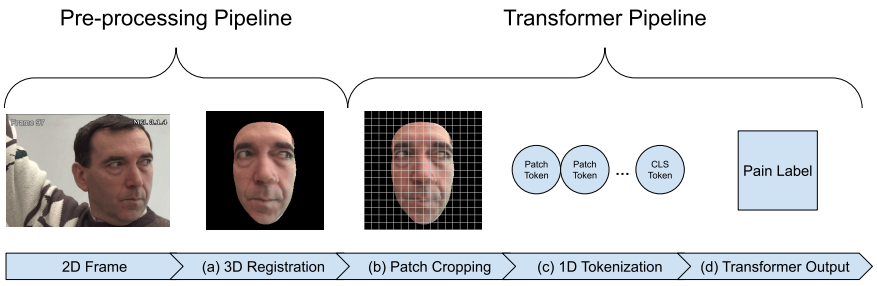

Figure 10: Data transformations through the pre-processing and transformer pipeline (Fiorentini et al.) [17].

Tokenization At runtime, the frontalized images are resized to 224x224, or 448x448 in the case of ViViT, the original size the transformer was trained on and then split into 14x14 patches, or 28x28 in the case of ViViT, of 16x16 pixels each as visible in Figure 10b. When the patches are combined with their corresponding pre-trained positional embedding, they become transformer-ready tokens. Finally, an additional classifier token to be interpreted by the MLP head of the transformer is added to the input tokens as visible in Figure 10c. When Mixup augmentation is active, the images and their respective output labels are mixed before tokenization at training time.

In order to process the 2×2 image grid for video vision transformers, the original pre-trained embeddings are joined with themselves into 2×2 grids. This approach processes each patch according to its original position within the sample and not the resulting one in the grid, leaving temporal reasoning solely to the attention mechanism. By the end of this step, the input is ready to be processed by the transformer model.

Encoding The transformer model processes the token input through 12 encoder transformers, each having 12 attention heads. The output of each layer has exactly the same size as the original input due to the linear ViT architecture. At training time, the weights of the MLPs within the transformer encoders are frozen, leaving the multi-headed attention layers to be fine-tuned and a 10% dropout is applied before the decoding step. The output of this step is 14x14 tokens, or 28x28 in the case of ViViT, and one classifier token to be interpreted.

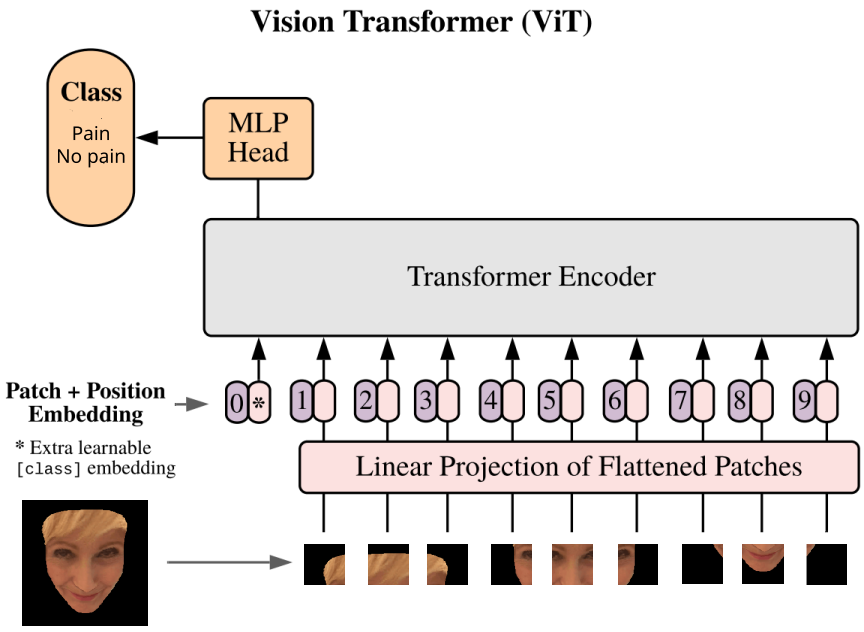

Figure 11: The architecture of our direct pain detection ViT model. (Dosovitskiy et al.) [19].

Decoding As our model does not employ mean pooling of the output tokens and instead uses a classifier token, all tokens except the latter are discarded. The MLP head consists of a single linear layer and interprets the classifier token to generate the predicted output.

In the case of direct pain detection, the output is a one-hot encoding of the pain/no pain prediction and the label with the largest value is considered the predicted output. The probability of the pain prediction is computed with the softmax function for AUC computation purposes. A summary of our direct pain detection model can be found in Figure 11.

In the case of direct pain intensity estimation, the output is a single value. The mean squared error and mean absolute error are computed on the difference between the predicted value and the PSPI label.

In the case of indirect pain detection, the output of the six models, each predicting one AU, is a single value. If any of the model outputs is above 0, then it is considered a pain label prediction, and no pain otherwise. The probability of the AU prediction of each individual model is computed with the sigmoid function for AUC computation purposes.

In the case of indirect pain intensity estimation, the output of the six models predicting each individual AU is a single value. The mean squared error and mean absolute error are computed on the difference between the result of the PSPI formula applied to the predicted values and the true PSPI label.

4.3 Experiments

To determine the performance of vision and video vision transformers in automated pain assessment, we conduct four sets of experiments.

Direct pain detection The first set of experiments evaluates the performance of ViT and ViViT on the task of direct pain detection. It consists in tuning a single hyperparameter and saving its best-performing value to be used while tuning the next parameter. The hyperparameters tuned are the number of attention layers trained, the learning rate, the use of Mixup, and the use of SAM. The pipeline employed is visible in Figure 11.

Due to the extreme imbalance of the labels, we evaluate the performance of the model using the F1 score on the minority class while oversampling it on the training set. This way, we ensure that the model prioritizes performance on the more difficult task of pain detection and is trained on a balanced number of samples per class. Furthermore, earlier studies carried out on this task used the F1 score metric, making it possible to compare results.

We have tested 14 possible configurations, the first six seek the optimal number of unfrozen attention layers for the transformer model, then four to determine the optimal learning rate of the Adam optimizer, one to quantify the effects of the Sharpness-Aware Minimization in combination with Adam, and three for the impact of the Mixup augmentation [63] on the performance of the transformer. All 14 configurations have been tested separately for the single-image and the 2×2 grid datasets, with the rationale that the use of vision or video vision transformers is unlikely to be independent of each individual hyperparameter.

First, all the fully-connected layers are frozen, leaving 12 attention layers to be fine-tuned. However, while too few attention layers cannot be effectively fine-tuned on a specific task, a higher number does not necessarily lead to a better performance [64], necessitating the model to be evaluated with varying amounts of unfrozen layers. Next, the learning rate of the Adam optimizer is tuned, ahead of the introduction of the Sharpness-Aware Minimization (SAM) optimizer.

Transformer models work best with large amounts of data, nevertheless, this weakness might be mitigated with techniques such as SAM [65] and Mixup [63]. The SAM optimizer works in conjunction with the original optimizer, in our case Adam, to prevent the model from converging to sharp local minima. While it could potentially reduce overfitting on the small UNBC-McMaster dataset, it also requires a second forward-backwards pass, almost doubling the training time required.

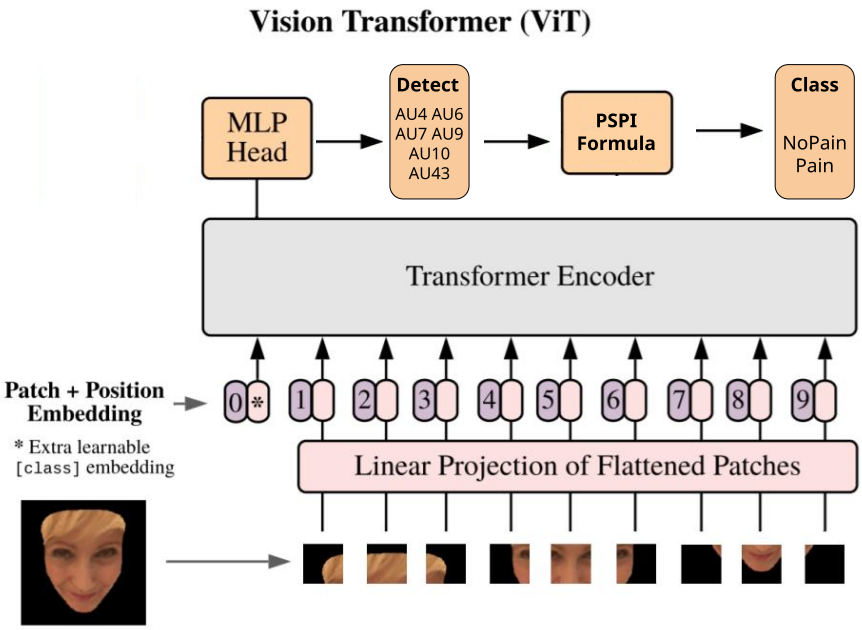

Figure 12: The architecture of our indirect pain detection ViT model. The six action unit detection models and their MLP heads are condensed into a single unit for ease of presentation. (Dosovitskiy et al.) [19].

As previously mentioned, we employ Mixup as an augmentation technique for our dataset. To better integrate Mixup with the pre-processed UNBC-McMaster dataset, one further restriction is applied, allowing only images from the same patient to be combined for the experiment. The intensity of Mixup can be adjusted through its α parameter, causing images to be increasingly hybrid, and has been configured according to previous research on Mixup and transformers [66]. Although these samples could allow for a more nuanced and linear function of pain for the model to learn from, they might also result too noisy and unnatural compared to other samples, further degrading the already limited data available.

Indirect pain detection The second set of experiments exploits the knowledge of the hyperparameter space obtained during the first experiment by testing its best-performing models on indirect pain detection. Indirect pain detection is achieved by training six models, one per AU relevant to the PSPI score, and testing their ability to detect a specific AU. The results of the detection are then aggregated and used to generate an F1 score prediction on the minority class. The pipeline is visible in Figure 12.

There are works in literature [11] that claim that an interim layer reduces the performance of pain detection models, following the general trend that explainability and performance are at odds. There are also others that employ interim AU layers, for example as part of a larger pipeline, achieving competitive results instead [30]. Clearly, there is no consensus in facial pain assessment literature on the superiority of direct or indirect pain detection, however, indirect pain measurements enable the model to explain its pain prediction through the AUs detected, similarly to how a human expert would justify the PSPI score label they choose for a given frame.

This type of model therefore can offer human-like explanations and focuses on providing the exact data that a human FACS-coder could use in order to justify its predictions. Work by Amann et al. [67] identifies textual and visual explanations as useful tools to assist medical staff in evaluating the predictions of a model and earning their trust, making it an important aspect of this research. Another work on the topic underlines [68] the importance of "why" explanations for medical predictions and describes them as a balancing act between promoting trust in the model while avoiding over-reliance on it. Our goal in this regard is for the model to produce results comparable to those of a human coder and promote a similar level of trust. We therefore employ indirect pain detection due to the importance of explainability in the context of medical care.

Direct pain intensity estimation The third set of experiments also exploits the knowledge of the hyperparameter space obtained during the first experiment by testing its best-performing models on direct pain intensity estimation. Direct pain intensity estimation is achieved by training the model on the original 16 levels of the PSPI score and outputting a single prediction to be compared with the original label using the MSE and MAE metrics.

The task is far more complex than the one described in the first two sets of experiments, as not only the positive samples are a lower percentage of the total dataset, but they are now further split into 15 smaller pain levels and distributed unevenly across them, while pain levels must be understood and predicted more accurately than the broader pain vs no pain labelling.

The main advantage of this approach is the quantification of pain itself. In a hospital context, the mere presence of pain might not be sufficient to alarm medical staff, who might be more interested in spikes in pain levels or continuous medium levels of pain. It is therefore important to test the capabilities of the model on the task of pain intensity estimation rather than solely on pain detection.

Indirect pain intensity estimation The fourth and final set of experiments also exploits the knowledge of the hyperparameter space obtained during the first experiment by testing its best-performing models on indirect pain intensity estimation. Our approach starts by training six models, one per AU relevant to the PSPI score, and testing their ability to assess the intensity of a specific AU. The results of the detection are then summed according to the PSPI formula and used to compute the MAE and MSE metrics on the test set.

This task is also far more difficult than the one described in the first two sets, however, due to the intensity level of AUs being distributed across 6 levels rather than 16, it could result simpler to train on in comparison with predicting the PSPI score directly. Furthermore, as was mentioned earlier, AU prediction enables the model to justify its prediction akin to a human expert, making this model the most complete of all the previous ones, able to quantify pain and justify its predictions. However, the distribution of the AUs, especially in their intensity, is still quite skewed, making it a difficult challenge for the transformer model. Furthermore, previous works in literature warn of AU intensity assessment mistakes compounding into even larger errors when computing PSPI intensity [11], a problem which our approach might encounter as well.

Cross-validation During training, for all but the preliminary experiments, the frames are divided into five folds, each containing samples from exactly five patients. The splits are generated with the aim of maintaining a similar number of pain samples across folds, achieved by pairing patients with the fewest and most pain samples and shuffling four patients between folds to further balance them, ensuring that each has a reasonable number of samples for the minority class. Five-fold cross-validation guarantees that the models learn to generalize painful features rather than overfitting on specific patients.

Loss functions Cross-entropy loss is employed in the direct and indirect pain detection experiments, while Huber loss [69] with δ equivalent to the difference between normalized categories was used for direct and indirect pain intensity estimation. Huber loss behaves similarly to mean squared error for prediction errors greater than δ and similarly to mean absolute error for smaller ones, and was therefore chosen to mitigate the impact of samples with outlier values on the overall training of the model.

$$Huber\ loss = \begin{cases} 0.5 * (x_n - y_n)^2 & if\ |x_n - y_n| < \delta \\ \delta\ * (|x_n - y_n| - 0.5 * \delta) & otherwise \end{cases} \tag{6}$$

5 Results

5.1 Direct pain detection

Preliminary experiments have shown reasonable values for various parameters such as the learning rate (2E-04), the batch size (16), and the number of epochs (1). Other important parameters for the initial training of the model are the drop-out rate before the classification head (0.10), the β values (0.9, 0.999) and ϵ (1e-08) of the Adam optimizer, weight decay (0), and the ρ (0.05) of the SAM optimizer. Finally, on this task, we employ oversampling and Cross Entropy loss.

Number of unfrozen attention layers The first step consists in identifying the optimal number of unfrozen layers. The results can be seen in Figure 13. In total, 12 models are trained for the vision and video vision transformer with multiples of two as the number of layers, from 2 to 12. For ViT, fine-tuning 12 layers performs best (F1 score 0.47) while fine-tuning 6 layers achieves the second-best performance (F1 score 0.45). For ViViT, fine-tuning 6 layers performs best (F1 score 0.55), while fine-tuning 12 layers achieves the second-best performance (F1 score 0.53).

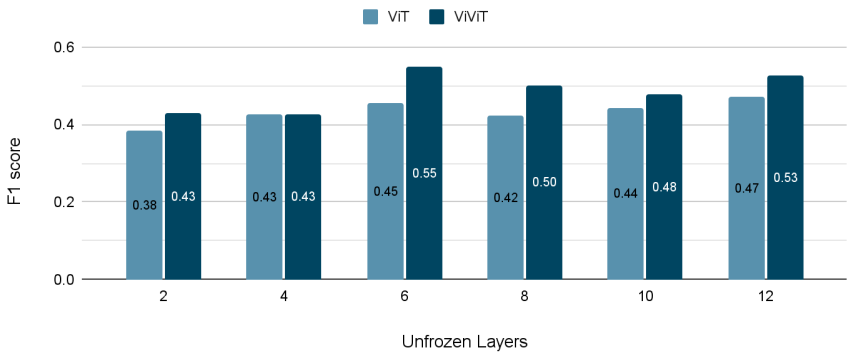

Figure 13: Performance (F1 score) of ViT and ViViT models with different numbers of unfrozen (fine-tuned) attention layers (Fiorentini et al.) [17].

Learning Rate For the second step, a large range of learning rates is tested to identify regions of interest in the hyperparameter space. The initial learning rate of 0.0002 is both increased and decreased tenfold and hundredfold. Performances of the resulting models can be seen in Figure 14.

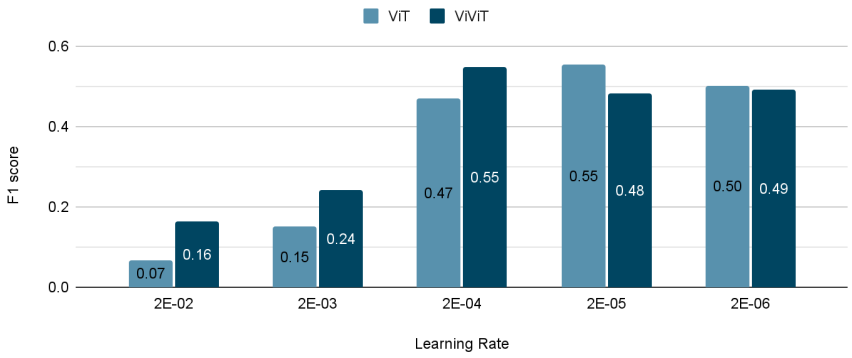

Figure 14: Performance (F1 score) of ViT and ViViT models with different learning rates (Fiorentini et al.) [17].

ViT performance peaks with a learning rate of 2E-05 (F1 score 0.55, model ViT-1-D), followed by 2E-06 (F1 score 0.50). ViViT performs the best with a learning rate of 2E-04 (F1 score 0.55, model ViViT-1-D), and obtains its second-best performance with a learning rate of 2E-06 (F1 score 0.49, model ViViT-2-D). However, a peculiar trait emerges from the latter model, an extremely low standard deviation across folds of the F1 score as visible in Table 4.

The models ViT-1-D (attention layers (al) = 12, learning rate (lr) = 2E-05) and ViViT-1-D (al = 6, lr = 2E-04) are the best-performing models of their type across all experiments, while ViViT-2-D (al = 6, lr = 2E-06) is the second best-performing video vision transformer and has a uniquely low standard deviation. Visible in Figure 15 is a comparison of the best two ViViT models, showcasing the good performance across all folds for ViViT-2-D compared to ViViT-1-D.

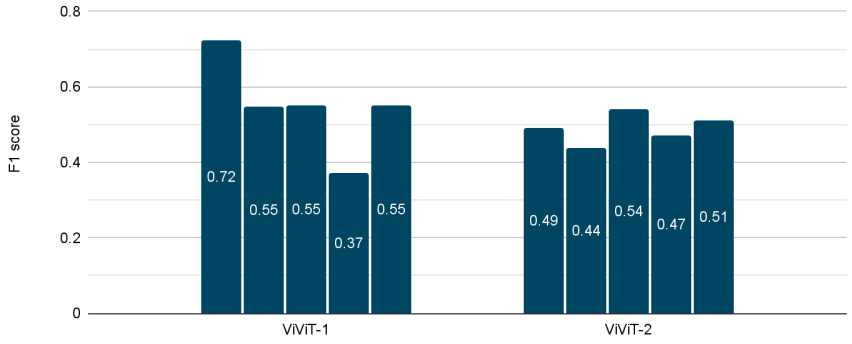

Figure 15: Performance (F1 score) per fold of the two best performing ViViT models (Fiorentini et al.) [17].

Sharpness-Aware Minimization The third step of experimentation introduces SAM to the model's training, however, this addition not only almost doubles the training time necessary but also worsens the performance of ViViT (F1 score 0.40) and ViT (F1 score 0.50), as shown in Figure 16.

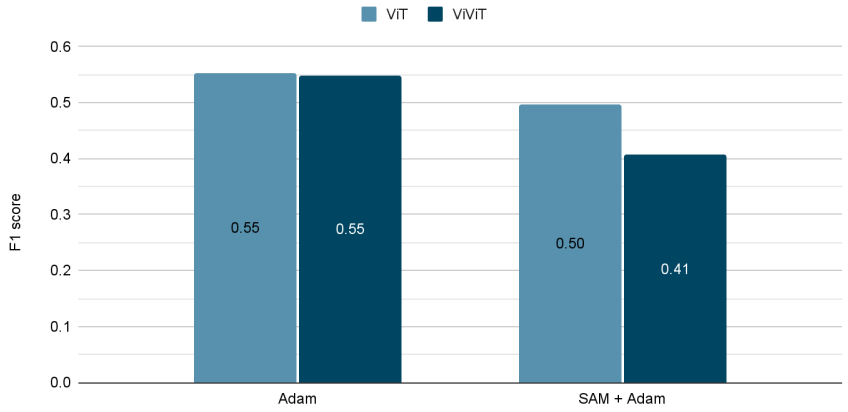

Figure 16: Performance (F1 score) of ViT and ViViT models optimized with and without the Sharpness-Aware Minimization (SAM) (Fiorentini et al.) [17].

Mixup The fourth experimental step augments 20% of the dataset with the Mixup technique, with three different α configurations. Mixup, even with the additional restriction of combining images belonging to the same patient, fails to contribute to the model's performance even with its best parameter ($\alpha = 0.8$) for the ViT model (F1 score 0.52) and ViViT model (F1 score 0.52), as shown in Figure 17.

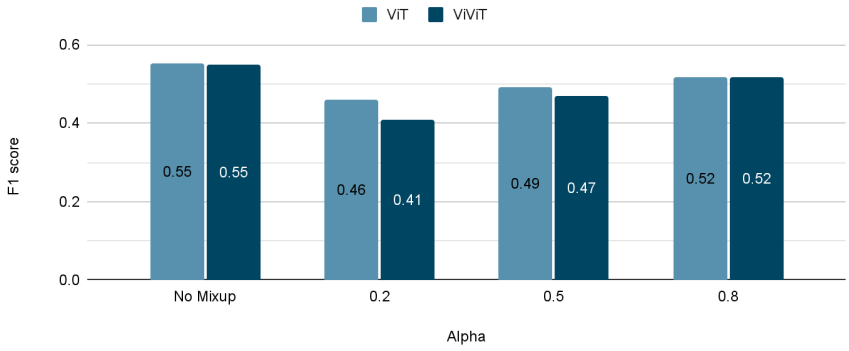

Figure 17: Performance (F1 score) of ViT and ViViT models fine-tuned with Mixup augmentation (Fiorentini et al.) [17].

Comparison with previous works To our knowledge, recent work by Rudovic et al. [70] is the state-of-the-art for pain detection on the F1 score metric. Their experimental setup uses a CNN baseline (CDL) but focuses on federalized learning (PFDL), achieving its best-performing model with this technique. As can be seen in Table 4 our method not only achieves better performance on the F1 score metric with our best-performing ViT and ViViT models, but their method is also outperformed by our ViViT-2-D model, which trades off performance for more consistent results across folds.

Table 4: Model results on direct and indirect pain detection

Model Name	F1 score	AUC
CDL [70]	0.46 ± 0.18	-
PFDL [70]	0.47 ± 0.20	-
SPTS+CAPP [7]	-	0.84
SPTS+SAPP+CAPP [71]	-	0.85
ViT-1-D	**0.55** ± 0.15	**0.88**
ViT-1-I	0.50 ± 0.11	-
ViViT-1-D	**0.55** ± 0.13	0.86
ViViT-1-I	0.50 ± 0.07	-
ViViT-2-D	0.49 ± **0.04**	0.76
ViViT-2-I	0.34 ± 0.03	-

The F1 score is affected by the skew in the labels but AUC is not [53]. Given that our labels are highly imbalanced, we also report AUC values and compare our results with the works that also report AUC. We compare our top-performing models ViT-1-D (AUC 0.88) and ViViT-1-D (AUC 0.86) against SPTS + CAPP (AUC 0.84) [7] and SPTS + SAPP + CAPP (AUC 0.85) [71], and find them to outperform previous works despite not being optimized for this metric.

5.2 Indirect pain detection

For all remaining experiments, we train 3 sets of models using the parameters of the ViT-1-D, ViViT-1-D, and ViViT-2-D models. We take advantage of the exploration of the hyperparameter space of the first experiment by employing parameters of models that exhibited important qualities in the previous task. Moreover, we employ Cross Entropy loss again while undersampling the majority class for two epochs. The latter change is motivated by the rapid convergence of the model noticed during the first experiment.

In Table 4 are visible the results of the three models on indirect pain detection. Performance is consistently lower across all models compared to the direct approach counterparts, in particular, the ViViT-2-I (F1 score 0.34) model clearly underperforms on this task. However, the ViT-1-I (F1 score 0.50) and ViViT-1-I (F1 score 0.50) models outperform earlier works with a comparatively low standard deviation.

Table 5: Model results on individual AU detection.

	AUC					
Model Name	AU4	AU6	AU7	AU9	AU10	AU43
SPTS+CAPP [7]	0.57	0.85	0.80	0.85	**0.89**	0.88
SPTS+SAPP+CAPP [71]	0.54	0.86	0.70	0.80	0.75	0.91
ViT-1-I	0.78	**0.89**	0.79	**0.86**	0.78	0.95
ViViT-1-I	**0.87**	0.88	**0.86**	**0.86**	0.84	**0.96**
ViViT-2-I	0.69	0.79	0.73	0.72	0.77	0.86

In Table 5 are collected the results of the models trained on individual AU detection. ViViT-2-I slightly underperforms across the board, partially reflecting the poor results achieved on the F1 score metric. ViT-1-I and ViViT-1-I instead outperform or perform comparably to previous models on all AUs except AU10. Comparing ViT-1-I and ViViT-1-I, the latter performs better, with superior performance across half of the action units and comparable in the others.

5.3 Direct pain intensity estimation

We will now discuss the performance of the ViT and ViViT models on the task of direct pain intensity estimation. On this task, we employ Huber loss (δ 0.0625) and undersample the majority class across 10 epochs. The increase in epochs is motivated by the complexity of regression compared to binary label classification. Finally, we undersample according to the pain and no pain classes rather than per pain level due to the extreme distribution of labels.

The models clearly underperform on this task compared to earlier works, both on the MSE and the MAE metric as shown in Table 6. However, the direct models still outperform some approaches and the baseline, a model which solely predicts the majority class 0, on the MSE metric, proving that the task is not completely out of reach for purely-attentive models.

Table 6: Model results on direct and indirect pain intensity estimation

Model Name	MSE	MAE
0-Baseline	2.06	0.44
PTS-D [11]	2.59	-
PTS-I [11]	2.53	-
PTS+DCT+LBP-D [11]	1.37	-
PTS+DCT+LBP-I [11]	1.48	-
VGG-CNN-SVR-D [40]	1.70	-
RCNN-D [40]	1.54	-
VGG16-D [30]	-	0.84
ViT-1-D	1.75	0.96
ViT-1-I	28.33	5.08
ViViT-1-D	1.72	0.88
ViViT-1-I	28.67	5.08
ViViT-2-D	1.99	1.11
ViViT-2-I	28.76	5.11

5.4 Indirect pain intensity estimation

Finally, we report the results of the models on indirect pain intensity estimation. On this task, we employ Huber loss (δ 0.16) and undersample the majority class across only 2 epochs, to reduce the massive training time required to train one model per AU. None of the models obtain a sufficient performance on this task, with massive error values for each of the parameters chosen.

5.5 Model interpretability

Previous works have shown that attention maps can be used to generate visual interpretations for the predictions of vision transformers [23, 39]. To demonstrate this feature we will perform a qualitative analysis of the attention maps of ViViT-1-D given the sample visible in Figure 18, whose pain label is correctly predicted by the model.

Figure 18: The pain sample used to qualitatively analyze the interpretability of ViViT-1-D. The first frame in the top-left has a PSPI score of 0, while the remaining three have a PSPI score of 10 (Fiorentini et al.) [17].

As shown in Figure 19, the last attention layer of ViViT-1-D has disentangled representations across its attention heads. These representations partially overlap with AU43 (eyes closed, head 0), AU4 (brow-lowering, head 1) AU6 (orbital tightening, head 2), while others capture a large area of the face (head 3).

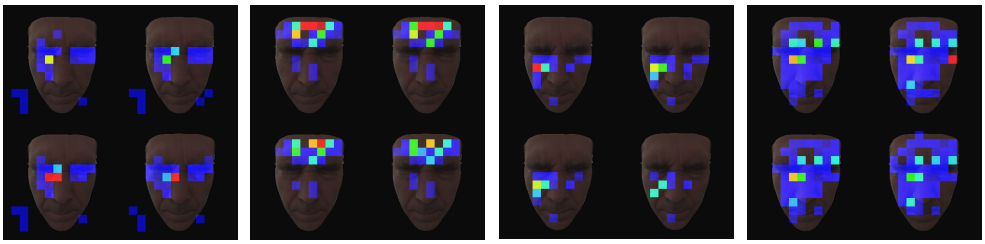

Figure 19: Attention maps of individual heads of the final attention layer. The activations are thresholded between 0.7 (blue) and 1 (red) (Fiorentini et al.) [17].

In Figure 20a, the maximum value of the attention patches for the ViViT-1-D model is shown, obtained with attention rollout [72]. Attention rollout is a transformer technique that combines information from every attention layer, capturing its flow through the model. In Figure 20b, we show instead the combined maximum values of the heads of the last attention layer for ViViT-1-D. While the strongest activations are found in the forehead and cheek area for the final layer (b), the flow of information instead clearly originates from the inner brow, lip corner and cheek area (a), all of which are areas of significance according to facial pain assessment literature [8].

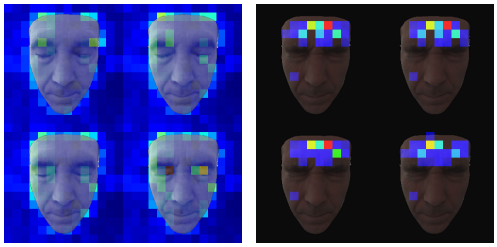

Figure 20: Attention maps obtained with max rollout (a) and maximum values of the last layer (b). Frame (b) is thresholded between 0.7 (blue) and 1 (red) (Fiorentini et al.) [17].

The model is clearly capable of generating intrinsic and plausible interpretations for its predictions. Facial regions having higher attention weights in the attention maps are the ones that show a change in appearance and shape during several actions that are observed during a painful expression. It shows that the model effectively detects pain from actions of relevant facial regions.

5.6 Model explainability

According to facial pain literature, PSPI scores are not meant to be predicted directly but instead are the result of the PSPI formula applied to the individual AU detected within an image. An expert, therefore, can pinpoint areas of the face in which pain-related AUs were detected to justify its overall pain prediction, but must also declare which AUs were detected and their intensity.

As we have shown previously, quantifying pain and indicating regions of interest within an image are feasible goals. Now, we show that generating explanations to justify the pain predictions of our indirect pain detection model is possible with our pipeline. Shown in Figure 21 is the sample we use to demonstrate the explainability of our ViVIT-1-I model, which correctly predicts the presence of AU6 and AU7 in the image and the absence of all the other pain-related AUs.

Figure 21: The pain sample used to qualitatively analyze the explainability of ViViT-1-I. The sample is labelled with AU6 and AU7, with an intensity of 4 and 3 respectively.

AU6 and AU7 are part of the orbital tightening action, meaning that when both are present, only the strongest of the two signals is considered for pain intensity estimation. Specifically, AU6 is the "cheek raise" action unit and AU7 is the "lids tightener" action unit. Both these AUs are strictly related to the eye area, but neither cause the eye to close completely, meaning there is a similarity and possible ambiguity between this action unit group and AU43, the "eye closure" action unit, another important signal for facial pain assessment.

In total, AU6 and AU7 last for 37 consecutive frames with the same intensity before suddenly ceasing, and of those 37 frames, our model correctly predicts 32. The mistaken predictions of the model are all concentrated in the first 6 frames of the sequence, during the onset of the action unit, which appears to prove particularly difficult for the model to detect, perhaps due to the fusion of spatial and temporal attention, which prevent it from correctly identifying emerging AUs.

Nevertheless, our model also correctly predicts the immediate stop of all action units immediately after the last frame, successfully identifying a complete lack of pain-related AUs. In summary, the model has an overall strong understanding of the temporal mechanics of the 4 frames involved despite the initial mistakes and is capable of generating correct explanations to support its predictions.

6 Discussion

6.1 Direct pain detection

Number of unfrozen layers The pre-trained transformer model is successful across a large variety of parameters for the task of direct pain detection, contrary to earlier findings on the topic [20]. While the models perform consistently no matter the number of layers trained, the region around 6 and 12 layers stands out as the better choice both for ViT and ViViT, warranting a deeper investigation of similar parameters and setting a precedent for future work. Despite sharing the top two configurations, the vision and video vision transformers perform best with different numbers of layers, distinguishing their configurations for the following steps.

Learning rate Comparably, learning rate proves to be a far more delicate parameter, with two of the configurations achieving the worst performance overall across all experiments for ViT and ViViT. Learning rates lower than 2E-03 are generally high-performing, with ViT peaking around 2E-05, achieving the best model performance across all configurations, and ViViT performing best with 2E-04, a much larger learning rate.

Furthermore, ViViT scores its second-best performance at the learning rate of 2E-06 with an extremely low standard deviation across folds, a sign of good generalization. Consistency is a desirable trait for all models, and even more so for delicate tasks in the medical field. Therefore, while achieving only the second-best performance according to the metric chosen for this experiment, it possesses a desirable trait and indicates a second region of interest for the tuning of this parameter.

Sharpness-Aware Minimization and Mixup The SAM optimizer and Mixup augmentation fail to improve the model's performance across a variety of configurations for both ViT and ViViT. While SAM decreases the standard deviation across folds for the ViViT model, it does so by affecting all folds negatively rather than pushing their performance towards an average. Mixup appears ineffective in generating meaningful samples for the model to learn from despite the constraints applied, perhaps due to the delicate nature of FACS and PSPI encoding.

ViT and ViViT ViViT fails to outperform ViT on direct pain detection despite the strong case for facial dynamics in pain detection literature. Overall, it achieves comparable performance on its best model to ViT and produces a model with slightly lower performance but incredibly low standard deviation, which is an important trait.

While the configurations we use may be limiting the performance of the architecture, we believe that ViViT models which compute spatio-temporal attention separately, such as the other ViViT implementations described in the original paper [21], might ultimately suit this task better. Separating the temporal and spatial attention would benefit the model by allowing larger sequence lengths while avoiding quadratically increased computational time.

Furthermore, temporally-local attention might track better the delicate facial dynamics necessary for pain detection. Overall, both ViT and ViViT models achieve state-of-the-art performance on the F1 score metric for pain detection, proving the effectiveness of transformers on this task under a variety of hyperparameters and the potential for even better results with more powerful transformer architectures. How to best employ video vision transformers on this task and the best approaches for augmentation and optimization remain open questions for future research.

6.2 Indirect pain detection

The indirect pain detection models fail to outperform their direct counterparts on the F1 score metric, showing that explainability comes at the cost of performance with this approach. Furthermore, the ViViT-2-I model underperforms on this task both on the F1 score and the AUC metric, underlining a degree of difference between the direct and indirect hyperparameter spaces.

On the other hand, the ViT-1-I and ViViT-1-I performances still significantly outperform earlier works, with a lower standard deviation across folds, while maintaining explainability. Moreover, on the task of individual AU detection, the models perform comparably or better on almost all AUs on the AUC metric.

On this task, ViViT partially outperforms ViT. On individual AU detection it performs better than ViT on half the AUs, while on the others they have comparable performance. Their ability to confidently and correctly detect the presence of AUs is likely an important factor in their overall superior performance compared to previous approaches. In fact, the great performance on the AUC metric does not appear to be always proportionate to the F1 score metric, raising concerns regarding the thresholding value of 0 chosen for prediction labelling.

Future work could explore different thresholds for label prediction in order to maximize the F1 score metric. Overall, indirect pain detection remains a competitive approach, which inherits many of the characteristics of the direct pain detection hyperparameter space, while transformers show potential for explainable pain assessment.

6.3 Direct pain intensity estimation

The models fail to outperform the state-of-the-art on the task of direct pain intensity estimation, likely due to the simplicity of the architecture chosen. However, the model successfully estimates pain and outperforms some previous works and the 0-baseline, contrary to earlier findings on transformers for pain intensity estimation [20]. Nevertheless, we believe the visual words of the ViT architecture to be too coarse and imprecise to effectively capture the delicate dynamics necessary to distinguish the varying levels of pain in a regression task, which would justify most of the gap in performance between pain detection and pain intensity estimation.

However, there might also be other minor components such as the hyperparameters chosen, the use of 5-fold validation rather than leave-one-subject-out validation, and other approaches that could have boosted the model's performance further. We are therefore not excluding the possibility of making this model architecture competitive on this task, however, we believe that more powerful and precise transformer architectures would easily outperform our models under similar circumstances.

6.4 Indirect pain intensity estimation

All three models perform terribly on this task, with massive errors on both metrics employed. While the reduction in epochs and hyperparameter chosen might one hand influence the results, it is also clear that the models come nowhere near a working solution to the problem of indirect intensity estimation, therefore, a change in approach is necessary. Furthermore, bad predictions on individual AUs often compound into greater mistakes on PSPI scores, as Kaltwang [11] had encountered in earlier works.

The delicate features necessary to detect AUs and estimate their intensity are likely beyond the capabilities of the simple and coarse architecture employed. Due to the extremely limited amount of samples, this task would also benefit greatly from leave-one-subject-out validation, increasing the meagre number of samples by 20%. Furthermore, the increase in complexity and specialization due to the introduction of regression and indirect pain assessment might benefit from model-specific parameter-tuning on a separate validation set, a change that would work best alongside the aforementioned cross-validation technique. Overall, we don't believe that the current approach and architecture can work on this task without significant changes in the approach.

6.5 Research questions

How do vision transformers perform on facial pain assessment? Transformers perform competitively for facial pain assessment, under a variety of hyperparameters, achieving state-of-the-art performance on the F1 score metric on direct and indirect pain detection. We report good results on the AUC metric for pain detection and individual AU detection. We demonstrate interpretability on our best-performing models and successfully achieve explainability while maintaining competitive results, with a slight reduction in performance. Regression results range from underperformance to insufficiency, and we have identified and reported their likely causes and possible solutions.

What are the regions of interest in the hyperparameters space for vision transformers? We identify 3 regions of interest in the hyperparameter space, in the proximity of the three models ViT-1, ViViT-1, and ViViT-2. Specifically, ViT models perform best with 12 fine-tuned attention layers and 2E-05 as learning rate, while ViViT models perform best when fine-tuning 6 attention layers with learning rates of 2E-04 and 2E-06.

We note that learning rates around 2E-06 appear to impact significantly the variance in performance across folds for the ViViT model, enabling it to perform much more consistently overall, an especially desirable trait in the medical field. More broadly, we identify as high-performing all models with a number of fine-tuned attention layers of 6 and greater, and with learning rates in the proximity of 2E-04 and lower.

How does indirect pain assessment, with an interim AU layer prediction, affect the performance of the model? The results of our models trained on interim AU layers are consistently inferior to the ones predicting PSPI score directly, showing that the interim AU layer overall reduces the model's performance with this approach. We prove that our model can generate meaningful explanation for individual predictions, and show its consistency throughout a 37-frame sub-section containing continuous AU6 and AU7 labels, with exact AU predictions for 32 out of 37 frames, furthermore predicting correctly the end of the AUs shown. Indirect pain intensity estimation fails completely with the current approach, whereas we find that while indirect pain detection models underperform slightly compared to direct pain detection models, the approach is competitive and effectively provides explainability.

How do vision transformers perform on facial pain intensity estimation? Vision transformers consistently fail to achieve state-of-the-art performance on the task of direct pain assessment, however, it appears to be a feasible task for our purely-attentive pipeline. contradicting earlier findings on the topic [20]. While the model achieves sub-standard results overall, it still manages to outperform some earlier approaches and the baseline proposed, proving that the task is not beyond the employed architecture. On the other hand, the model is completely unsuccessful on indirect pain assessment, obtaining absolutely insufficient results.

Even though minor adjustments such as the use of leave-one-subject-out validation and finer hyperparameter tuning could meaningfully boost the results of the direct pain intensity estimation models given their current performance, we believe that employing more powerful transformer architectures, capable of extracting finer details from the input images, to be a better path going forward. We would recommend a similar approach for indirect pain intensity estimation given the inability of the current architecture, with our approach, to indirectly predict pain intensity.

How do vision transformers and video vision transformers compare in performance? Despite the importance of facial dynamics for pain assessment, we fail to find large performance differences between vision transformers and video vision transformers, finding both approaches similarly competitive, sufficient, or insufficient, depending on the task. ViViT does nonetheless show some interesting features, such as very low standard deviation in one of its models for direct pain detection, and does partially outperform ViT on the AUC metric for individual AU detection. Future work will likely benefit from testing longer video sub-sequences and different approaches to computing spatiotemporal attention, which might better capture the delicate temporal dynamics of facial expressions.

Can attention maps be used to generate plausible interpretations of the outputs? We extract attention maps from one of our best-performing models and use them to provide plausible interpretations of our model's predictions according to our qualitative evaluation. The attention map activations depicted are computed from high-intensity attention activations on the last attention layer and through attention rollout, a technique which captures the flow information across the whole transformer architecture. We find that the activations correspond to areas of high significance according to facial pain assessment literature, both in the flow of information and their maximal intensity in the last attention layer, and therefore succeed in providing plausible interpretability to our models.

7 Conclusion

In this thesis, we have used vision and video vision transformers trained on the UNBC-McMaster dataset for direct and indirect pain detection and intensity estimation.

We have achieved state-of-the-art performance using the F1 score, identified regions of interest in the transformer hyperparameter space, compared the performance of vision and video transformers on this task, made competitive and explainable facial pain detection models, evaluated the feasibility of pain intensity estimation, and obtained plausible intrinsic interpretations for the performance of the model.

Results show that pre-trained transformers can be applied toward pain detection with good results, after one or two epochs of training and on a small unbalanced dataset, while pain intensity estimation still requires further research. Future work could include different augmentation techniques, leave-one-patient-out validation, runtime hyperparameter tuning on a validation set, longer video sub-sequences, and more efficient transformer architectures.

References

[1] SN Raja, DB Carr, M Cohen, NB Finnerup, H Flor, S Gibson, FJ Keefe, JS Mogil, M Ringkamp, KA Sluka, et al. "The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. Pain". In: *press. doi* 10 (2020).

[2] Harald Breivik, Beverly Collett, Vittorio Ventafridda, Rob Cohen, and Derek Gallacher. "Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment". In: *European journal of pain* 10.4 (2006), pp. 287–333.

[3] MH Hayes. "Experimental developement of the graphics rating method". In: *Physiol Bull* 18 (1921), pp. 98–99.

[4] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. "The painful face–pain expression recognition using active appearance models". In: *Image and vision computing* 27.12 (2009), pp. 1788–1796.

[5] Zakia Hammal and Jeffrey F Cohn. "Automatic, objective, and efficient measurement of pain using automated face analysis". In: *Social and interpersonal dynamics in pain*. Springer, 2018, pp. 121–146.

[6] Amanda C de C Williams, Huw Talfryn Oakley Davies, and Yasmin Chadury. "Simple pain rating scales hide complex idiosyncratic meanings". In: *Pain* 85.3 (2000), pp. 457–463.

[7] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. "Painful data: The UNBC-McMaster shoulder pain expression archive database". In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE. 2011, pp. 57–64.

[8] Paul Ekman and Wallace V Friesen. "Facial action coding system". In: *Environmental Psychology & Nonverbal Behavior* (1978).

[9] Kenneth M Prkachin and Patricia E Solomon. "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain". In: *Pain* 139.2 (2008), pp. 267–274.

[10] Elizabeth A Clark, J'Nai Kessinger, Susan E Duncan, Martha Ann Bell, Jacob Lahne, Daniel L Gallagher, and Sean F O'Keefe. "The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review". In: *Frontiers in psychology* 11 (2020), p. 920.

[11] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. "Continuous pain intensity estimation from facial expressions". In: *International Symposium on Visual Computing*. Springer. 2012, pp. 368–377.

[12] Zara Ambadar, Jonathan W Schooler, and Jeffrey F Cohn. "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions". In: *Psychological science* 16.5 (2005), pp. 403–410.

[13] Siavash Rezaei, Abhishek Moturu, Shun Zhao, Kenneth M Prkachin, Thomas Hadjistavropoulos, and Babak Taati. "Unobtrusive pain monitoring in older adults with dementia using pairwise and contrastive training". In: *IEEE Journal of Biomedical and Health Informatics* 25.5 (2020), pp. 1450–1462.

[14] Yue Sun, Caifeng Shan, Tao Tan, Xi Long, Arash Pourtaherian, Svitlana Zinger, et al. "Video-based discomfort detection for infants". In: *Machine Vision and Applications* 30.5 (2019), pp. 933–944.

[15] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. "What clinicians want: contextualizing explainable machine learning for clinical end use". In: *Machine learning for healthcare conference*. PMLR. 2019, pp. 359–380.

[16] Steffen Walter, Sascha Gruss, Stephan Frisch, Joseph Liter, Lucia Jerg-Bretzke, Benedikt Zujalovic, and Eberhard Barth. ""What About Automated Pain Recognition for Routine Clinical Use?" A Survey of Physicians and Nursing Staff on Expectations, Requirements, and Acceptance". In: *Frontiers in medicine* (2020), p. 990.

[17] Giacomo Fiorentini, Itir Onal Ertugrul, and Albert Ali Salah. "Fully-attentive and interpretable: vision and video vision transformers for pain detection". In: *arXiv preprint arXiv:2210.15769* (2022).

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[20] Haochen Xu and Manhua Liu. "A Deep Attention Transformer Network for Pain Estimation with Facial Expression Video". In: *Chinese Conference on Biometric Recognition*. Springer. 2021, pp. 112–119.

[21] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "Vivit: A video vision transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6836–6846.

[22] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. "When vision transformers outperform ResNets without pre-training or strong data augmentations". In: *arXiv preprint arXiv:2106.01548* (2021).

[23] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. "Transformer in transformer". In: *Advances in Neural Information Processing Systems* 34 (2021).

[24] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. "Segmenter: Transformer for semantic segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 7262–7272.

[25] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. "Pre-trained image processing transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12299–12310.

[26] Zuhair Zafar and Nadeem Ahmad Khan. "Pain intensity evaluation through facial action units". In: *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 4696–4701.

[27] Zakia Hammal and Jeffrey F Cohn. "Automatic detection of pain intensity". In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. 2012, pp. 47–52.

[28] Philipp Werner, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C Traue. "Automatic pain assessment with facial activity descriptors". In: *IEEE Transactions on Affective Computing* 8.3 (2017), pp. 286–299.

[29] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M Gonfaus, Kamal Nasrollahi, Thomas B Moeslund, and F Xavier Roca. "Deep pain: Exploiting long short-term memory networks for facial expression classification". In: *IEEE transactions on cybernetics* (2017).

[30] Xiaojing Xu, Jeannie S Huang, and Virginia R De Sa. "Pain evaluation in video using extended multitask learning from multidimensional measurements". In: *Machine Learning for Health Workshop*. PMLR. 2020, pp. 141–154.

[31] Quanfu Fan, Chun-Fu Richard Chen, Hilde Kuehne, Marco Pistoia, and David Cox. "More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation". In: *Advances in Neural Information Processing Systems* 32 (2019).

[32] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. "Stm: Spatiotemporal and motion encoding for action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2000–2009.

[33] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is space-time attention all you need for video understanding". In: *arXiv preprint arXiv:2102.05095* 2.3 (2021), p. 4.

[34] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. "Cvt: Introducing convolutions to vision transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 22–31.

[35] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. "Xcit: Cross-covariance image transformers". In: *Advances in neural information processing systems* 34 (2021).

[36] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "PVT v2: Improved baselines with Pyramid Vision Transformer". In: *Computational Visual Media* (2022), pp. 1–10.

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 568–578.

[38] Fuyan Ma, Bin Sun, and Shutao Li. "Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion". In: *IEEE Transactions on Affective Computing* (2021).

[39] Chongwen Wang and Zicheng Wang. "Progressive Multi-Scale Vision Transformer for Facial Action Unit Detection". In: *Frontiers in Neurorobotics* 15 (2021).

[40] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. "Recurrent convolutional neural network regression for continuous pain intensity estimation in video". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 84–92.

[41] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, and Saida Bouakaz. "Pain detection through shape and appearance features". In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2013, pp. 1–6.

[42] Paola Casti, Arianna Mencattini, Joanna Filippi, Michele D'Orazio, Maria Colomba Comes, Davide Di Giuseppe, and Eugenio Martinelli. "Metrological characterization of a pain detection system based on transfer entropy of facial landmarks". In: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–8.

[43] László A Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F Cohn. "The first 3d face alignment in the wild (3dfaw) challenge". In: *European Conference on Computer Vision*. Springer. 2016, pp. 511–520.

[44] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. "Joint 3d face reconstruction and dense alignment with position map regression network". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 534–551.

[45] dougsouza. *face-frontalization*. `https://github.com/dougsouza/face-frontalization`. 2018.

[46] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. "Ostec: one-shot texture completion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7628–7638.

[47] Yunfeng Zhu, Fernando De la Torre, Jeffrey F Cohn, and Yu-Jin Zhang. "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior". In: *IEEE transactions on affective computing* 2.2 (2011), pp. 79–91.

[48] Ghazal Bargshady, Xujuan Zhou, Ravinesh C Deo, Jeffrey Soar, Frank Whittaker, and Hua Wang. "Enhanced deep learning algorithm development to detect pain intensity from facial expression images". In: *Expert Systems with Applications* 149 (2020), p. 113305.

[49] Ghazal Bargshady, Jeffrey Soar, Xujuan Zhou, Ravinesh C Deo, Frank Whittaker, and Hua Wang. "A joint deep neural network model for pain recognition from face". In: *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. IEEE. 2019, pp. 52–56.

[50] Neeru Rathee and Dinesh Ganotra. "Multiview distance metric learning on facial feature descriptors for automatic pain intensity detection". In: *Computer Vision and Image Understanding* 147 (2016), pp. 77–86.

[51] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. "Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields". In: *International Symposium on Visual Computing*. Springer. 2013, pp. 234–243.

[52] Junkai Chen, Zheru Chi, and Hong Fu. "A new framework with multiple tasks for detecting and locating pain events in video". In: *Computer Vision and Image Understanding* 155 (2017), pp. 113–123.

[53] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. "Facing imbalanced data–recommendations for the use of performance metrics". In: *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE. 2013, pp. 245–251.

[54] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. "How to train your vit? data, augmentation, and regularization in vision transformers". In: *arXiv preprint arXiv:2106.10270* (2021).

[55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[56] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. "Randaugment: Practical automated data augmentation with a reduced search space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 702–703.

[57] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. "Learning deep transformer models for machine translation". In: *arXiv preprint arXiv:1906.01787* (2019).

[58] Alexei Baevski and Michael Auli. "Adaptive input representations for neural language modeling". In: *arXiv preprint arXiv:1809.10853* (2018).

[59] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. "Imagenet-21k pretraining for the masses". In: *arXiv preprint arXiv:2104.10972* (2021).

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[61] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[62] YadiraF. *face3d*. https://github.com/YadiraF/face3d. 2018.

[63] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[64] Jaejun Lee, Raphael Tang, and Jimmy Lin. "What would elsa do? freezing layers during transformer fine-tuning". In: *arXiv preprint arXiv:1911.03090* (2019).

[65] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. "Sharpness-aware minimization for efficiently improving generalization". In: *arXiv preprint arXiv:2010.01412* (2020).

[66] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. "When vision transformers outperform ResNets without pre-training or strong data augmentations". In: *arXiv preprint arXiv:2106.01548* (2021).

[67] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC Medical Informatics and Decision Making* 20.1 (2020), pp. 1–9.

[68] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. "The role of explanations on trust and reliance in clinical decision support systems". In: *2015 international conference on healthcare informatics*. IEEE. 2015, pp. 160–169.

[69] Peter J Huber. "Robust estimation of a location parameter". In: *Breakthroughs in statistics*. Springer, 1992, pp. 492–518.

[70] Ognjen Rudovic, Nicolas Tobis, Sebastian Kaltwang, Björn Schuller, Daniel Rueckert, Jeffrey F Cohn, and Rosalind W Picard. "Personalized federated deep learning for pain estimation from face images". In: *arXiv preprint arXiv:2101.04800* (2021).

[71] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. "Automatically detecting pain in video through facial action units". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.3 (2010), pp. 664–674.

[72] Samira Abnar and Willem Zuidema. "Quantifying attention flow in transformers". In: *arXiv preprint arXiv:2005.00928* (2020).