Pain Detection From Facial Expressions Based on Transformers and Distillation

Safaa El Morabit, Atika Rivenq

Pain Detection From Facial Expressions Based on Transformers and Distillation

Safaa El Morabit & Atika Rivenq
IEMN DOAE, UMR CNRS 8520
Polytechnic University Hauts-de-France, University of Lille
59300 Valenciennes, France
safaa.elmorabit, Atika.Menhaj @uphf.fr

***Abstract*—Pain assessment is a challenging problem in the field of emotion recognition. Pain represents a complex emotion difficult to detect or to estimate its intensity. This is what makes automatic pain assessment playing an important role in clinical diagnosis. Taking into consideration that pain generally generates spontaneous facial behaviour, these facial expressions could be used to detect the presence of pain. As a matter of fact, previous researches used machine learning and deep learning either to detect pain or to estimate pain level. In this paper, we propose a fine-tuning of pre-trained data-efficient image transformers and distillation (Deit) for pain detection from facial expressions. The effectiveness of the proposed architecture is evaluated on two publicly available databases, namely UNBC McMaster Shoulder Pain and BioVid Heat Pain. The proposed approach achieved promising preliminary results compared to the state of the art.**

***Index Terms*—Pain Detection, Facial Expressions, Vision Transformer model with distillation**

I. Introduction

Facial expressions are important in social interactions. They express spontaneously the emotions of certain persons. Facial expressions therefore provide much information that can be analyzed nowadays not only by humans but also by machines. We can highlight the importance of introducing machines to emotion detection by the fact that in some cases humans are incapable of analyzing facial expressions (for instance, if a person is paralyzed or in case of infants). One of the important applications of computer vision using facial expressions is pain assessment.

Pain presents a complex phenomenon which is not completely understood, starting by its definition as an unpleasant feeling that may be a consequence of numerous causes (for instance, medical causes, emotional or psychological ones [1]). Pain actually generates spontaneous facial expressions. Therefore, in most of the research in pain recognition, the researchers use images of facial expressions [2]. In addition, most of the publicly available databases of pain contain facial images of videos of patients [3] [4] [5].

Regarding the importance of automatic detection of pain from facial expressions, many researchers focus their studies on the detection of pain or no pain task. Others led their researches to the estimation of pain level or chronic versus non chronic pain. Different methodologies have been used. Starting by handcrafted methods, passing by machine learning methods to arrive at the deep learning approaches [2]. In our paper, we introduce a novel method for the automatic detection of pain. We propose a transfer learning using the pre-trained data-efficient image transformers [6] (Deit) for pain detection from facial expressions. This method is based on the transformers [7] that were designed first for Natural Language Processing (NLP). These transformers showed their effectiveness for image recognition with the pioneer work in [8]. We have chosen the Deit [6], as it incorporates distillation that exploits CNNs. To train our proposed architecture, we considered two databases, namely the UNBC McMaster Shoulder Pain [3] and the BioVid Heat Pain [4].

The contribution of this research is to provide an effective pain assessment method based on facial expressions. This report outlines the following contributions:

- Present a fine-tuned data-efficient image transformers (Deit) for pain and no pain detection.
- Highlight the importance of transformers in the images recognition field in general, and in pain tasks more particularly.
- Prove the efficiency of transformers comparing to Convolutional Neural Networks (CNN) while studying the discrimination of pain from no pain task.

This paper is organized as follows. In the Section II, we give a brief overview of the state of the art methods used in pain tasks. Then, Section III outlines the used datasets and explains the method proposed in this paper. The conducted experiments and the results are analyzed in Section IV. Our conclusions are drawn in the final section.

II. Related Work

There is a considerable amount of literature on the automatic pain recognition from facial expressions. First of all, works that focus on the detection of the presence of pain (pain no pain). Other approaches work on the estimation of pain level. These works propose architectures based on handcrafted methods, machine learning or deep learning. Chen et al. [9] proposed a novel architecture for joint pain event detection and locating in video. The authors extracted features applying histogram of oriented gradients (HOG) and used them as an input of Support Vector Machine (SVM). They used the UNBC McMaster Shoulder Pain [3] dataset. Another work by Kaltwang et al [11] aims to estimate pain intensity by using

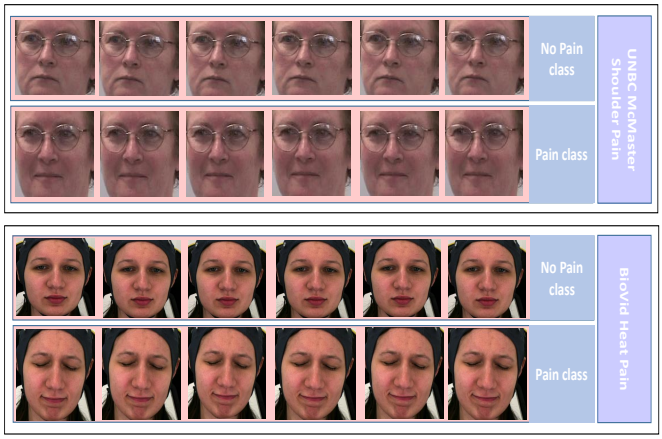

Fig. 1. Examples of some of the sequences from the UNBC-McMaster shoulder pain [3] and from the BioVid Heat Pain Database [4] databases. These sequences show the difference of facial expressions for patients having pain and no pain.

Local Binary Pattern [16] (LBP). The authors of this article divide the facial images into a uniform grid of cells. Then they use the LBP to extract features and use them as facial features to estimate pain. Their model was evaluated on the UNBC McMaster Shoulder Pain [3] dataset.

Bargshady et al [10] proposed a hybrid method by joining a Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). They first used VGGFace to extract deep features from images of the UNBC McMaster Shoulder Pain [3] dataset. In this paper, the authors aim to classify pain into four classes : No pain, weak pain, mild pain and strong pain. One of the state of the art approaches that uses deep learning with deep features is the work of Haque et al [13]. In this study, the authors extracted deep features using CNN and fed forward these features to a Long Short Term Memory network [14] (LSTM). They evaluated early and late fusion strategies for the recognition of pain levels. This study was trained using the Multimodal Intensity Pain [14] (MIntPAIN). This database consists of 20 adults with stimulated electrical pain. In a recent work by Karamitsos et al [15], the authors proposed a novel Convolutional Neural Network (CNN) for automatic pain detection from facial expressions. The proposed CNN consists of a modified version of VGG16 [16] model. They conducted experiments using the UNBC McMaster Shoulder Pain [3] dataset.

It appears that most of the existing methods for the identification of pain from facial expressions, either use machine learning or deep learning methods. To the best of our knowledge, there is not much work yet based on transformers for automatic pain detection. In the section below, we detail our proposed architecture based on transfer learning of the data-efficient image transformer [6] (Deit).

III. Databases and Proposed Method

In this section, we present the two databases used in these experiments. Then, we detail the pre-processing applied to the studied databases. Finally we describe our proposed architecture.

A. Databases and Pre-Processing

The experiments of this study are done on the two Databases: UNBC McMaster Shoulder Pain Database [3] and BioVid Heat Pain Database [4]. Those two Databases are publicly available. In Fig. 1, we present some sequence examples from both databases.

UNBC McMaster Shoulder Pain Database: It consists of 25 adults with shoulder pain. This database includes four parts: first 200 video sequences containing spontaneous facial expressions; Second $48,398$ Facial Action Coding System (FACS) coded frames; Third, associated pain frame-by-frame scores and sequence-level self-report and observer measures; and finally 66-point Active Appearance Model(AAM) landmarks. In our study, we are interested in part two that consists of $48,398$ images. These images are capturing facial expressions, while pain intensity changes. In our case, we are working on a binary representation of pain. Therefore our database is divided into two classes: pain and no pain.

BioVid Heat Pain Database: It is a multimodal database. It contains frontal videos, biomedical signals: Galvanic Skin Response (GSR), Electrocardiography (ECG), and Electromyography (EMG) at trapezius muscle. Pain in this database was stimulated by induced heat pain in four intensities. For each intensity, 20 experiments are done. In our research, we will be interested in frontal videos. In addition, this database is divided into four parts. We will be using part A during our experiments. This part contains 87 subjects with 5 classes (no pain and 4 pain intensities). We convert videos to frames. Thus this database presents a total of 797343 images.

In order to focus on facial expressions, it is in our interest to crop the face of the subjects. First, we use the Multitask Cascaded Convolutional Networks [18] (MTCNN) as a face detector. Second, once the face is detected we align it. Finally we crop it to an image of size 256×256. We divide each database into two classes : one for images that represent no

TABLE I
AMOUNT OF IMAGES IN THE USED DATABASES : UNBC MCMASTER SHOULDER PAIN DATABASE [3] AND BIOVID HEAT PAIN DATABASE [4]. FOR EVERY ONE WE PRECISE THE AMOUNT OF IMAGES FOR EACH CLASS FOR TRAIN VALIDATION AND TEST.

Classes	**UNBC McMaster Shoulder Pain Database**			**BioVid Heat Pain Database**		
	Train	***Validation***	***Test***	***Train***	***Validation***	***Test***
Pain	5 574	1 393	1 402	311 040	77 760	168 480
No Pain	22 344	5 585	12 100	134 688	33 672	71 703
Total	27 918	6 978	13 502	445 728	111 432	240 183

pain and the other one gathers all pain intensities to constitute one class for pain. Table I gives more details about the amount of images in every class and every database.

B. Proposed Method

In our approach, we propose a novel framework based on the transformer. This latter has been introduced in the paper [7]. The transformers are deep learning networks that were conceived first for the Natural Language Processing (NLP) tasks. It represents a sequence-to-sequence architecture. Moreover, the transformers are based on a self-attention mechanism which has the ability to learn the relationship between sequences' components. Self-attention is one of the key ideas of the novelty presented by transformers, in addition to the pre-training on large datasets. Therefore, self-attention is a mechanism that estimates the relevance of an item over another [17]. The attention mechanism can be defined by the equation 1.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

Lets consider an image feature maps $\mathbf{X}$, where $\mathbf{X} \in \mathbf{R}^{n \times d}$. Q is the matrix of the query (vector of one word in NLP tasks and patch in image recognition), K represents the keys (vector of all patches or words in a sequence). V is a vector of values, containing also all the patches or words of a sequence. Therefore, attention mechanisms did bring novelty and efficiency in networks for computer vision in general, and for image recognition in particular. In our case our proposed architecture is based on a transformer that uses distillation knowledge from a Convolutional Neural Network (CNN) as a teacher in addition to attention mechanism.

In this paper we propose a novel approach that consists of transfer learning using the Data-efficient image transformer (Deit) model that was proposed in [6]. It is one of the first papers to show that it is efficient to train transformers for image recognition tasks. The Deit demonstrated interesting results while using mid-size databases [6]. Deit is actually trained using only the ImageNet dataset [19]. The Deit approach is based on distillation [20]. This latter can be defined as a process that transfers the knowledge from a network to another one by some means. We specify a teacher and student models. In particular, in the paper [6] the authors proposed a distillation with state of the art CNN that is pre-trained on ImageNet as a teacher. The student architecture is a modified version of Vision Transformer [8] (ViT). The output of the CNN is also passed as an input to the transformer. The main reason to use the outputs from CNN is to figure out useful representations for input images which will help improve the efficiency of the transformer [21]. While computing the distillation loss, the authors do what is called the hard distillation where the temperature is equal to one. Which means that they literally take the label of the teacher as a true label, then they sum up this distillation loss with a cross entropy of the transformer. Taking a look at Fig. 2, we can notice a distillation token that ensures that the student learns from the teacher through attention. Also, there is the class token that goes through all blocks for the original classification done by the transformer. The patch tokens are obtained from the input image. These three tokens are put through several layers of self-attention and Feed Forward Network (FFN), then obtain the classes on top of the model.

To use the Deit architecture, in our approach we download the pre-trained model available in the original repository of facebook research [22], and adapt this architecture to our task. The last layer was replaced with an output of two classes (pain and no pain) instead of 1000. We also specified ResNet50 [23] as a teacher model. The model is then fine-tuned using Binary Cross Entropy (BCE) on the following databases: UNBC-McMaster shoulder pain [3] and from the BioVid Heat Pain [4]. Fig. 2 presents the used pipeline in our research.

In order to compare state of the art methods with our proposed model, we implemented GoogleNet [24] for the detection of pain. To do that, we used the pre-trained version of GoogleNet on ImageNet, and adapted the last layer for two classification classes. GoogleNet is a model that gives interesting results while using it in pain detection and estimation [25]. Also it is a model that does not take huge computational resources. Therefore the fine-tuned GoogleNet will be trained on UNBC-McMaster shoulder pain [3] and the BioVid Heat Pain [4] datasets.

IV. EXPERIMENTAL RESULTS

In this section we will carry out implementation details of the proposed methods. Moreover, we aim to present the obtained results of the experiments.

A. Training details

As seen in the subsection III, after the detection, alignment and cropping of the images, we resize them to 256×256. We

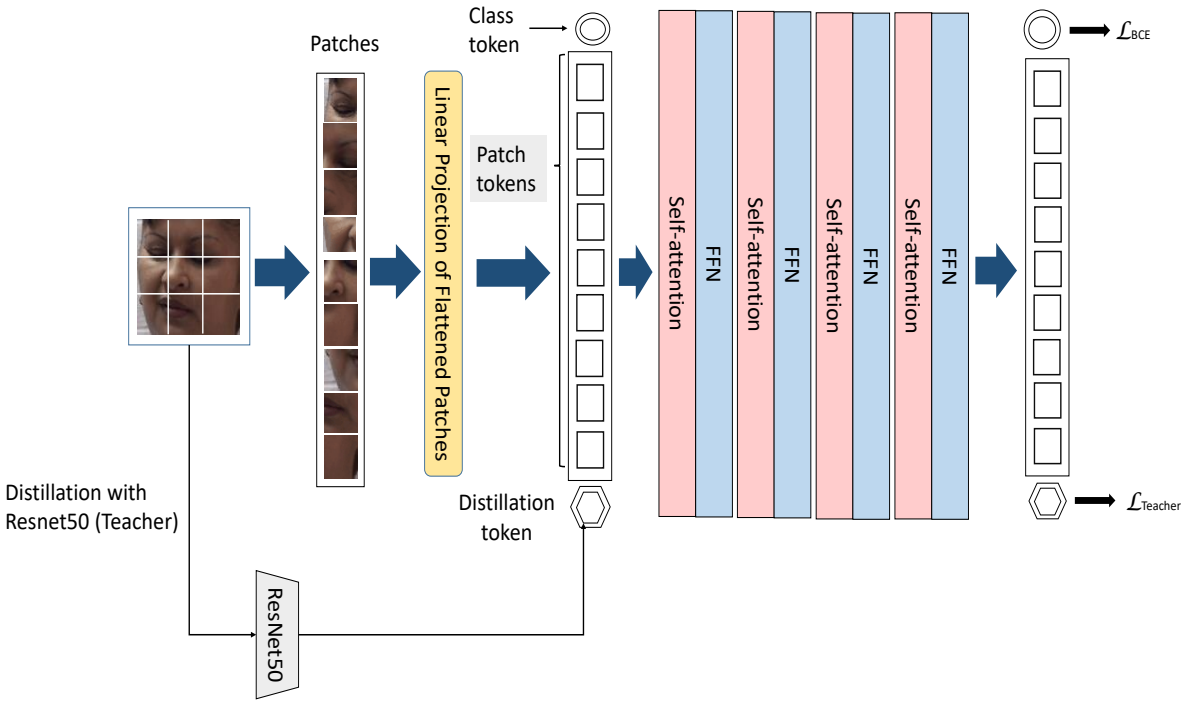

Fig. 2. An overview of the proposed pipeline for Pain Recognition using Data-efficient image transformer (Deit) [6]. (FFN stands for Feed Forward Network. BCE is the Binary Cross Entropy.)

augment our training data using data augmentation. During the training, we fixed the patch size to 32. In addition, the learning rate is 0.00001. The model is trained for 30 epochs with a batch size of 64. And the optimizer we opted for in the standard Adam Optimizer [26]. For classification, we select the best parameters using back-propagation with the binary cross-entropy (BCE) loss. The experiments were done on a machine with two NVIDIA Quadro RTX 5000 GPUs and 32 memory. The training parameters were used to train both Deit [6] and GoogleNet [24] separately on the two datasets: UNBC-McMaster shoulder pain [3] and the BioVid Heat Pain [4].

B. Performance analysis

We conducted experiments on both UNBC-McMaster shoulder pain [3] and the BioVid Heat Pain [4] datasets. To evaluate the proposed fine-tuned Deit model for the recognition of pain, we used the accuracy as a metric. First, we train the proposed method separately on the two datasets. As shown in Table II, we obtained an accuracy of 84.15% while the Deit-PNP is trained on the UNBC-McMaster shoulder pain [3]. Surprisingly, the accuracy achieved when we used the BioVid Heat Pain [4] dataset is 72.11%. Although the BioVid Heat Pain [4] dataset contains more data, the UNBC-McMaster shoulder pain [3] achieved better results. Our findings appear to confirm that when using Trasnformers the size of dataset improves results if it's about huge difference of the size.

We kept the same parameters used to evaluate our proposed method for the experiments concerning the state of the art model: pre-trained GoogleNet for detection of pain from no pain that we are denoted as GoogleNet-PNP. We have conducted two experiments, first on the UNBC-McMaster shoulder pain [3] dataset, second on BioVid Heat Pain [4] dataset. As reported in Table II, the GoogeNet-PNP achieved 80.01% accuracy while trained on UNBC-McMaster shoulder pain [3] dataset. Therefore, these findings confirm that our method achieves better results in the differentiating between pain and no pain task. We have also performed experiments using the GoogleNet-PNP on the BioVid Heat Pain [4] dataset. In this case the accuracy achieved 65.75%. This result reinforces what we claimed above. The solution we proposed in this paper, that is based on transformers, proved the importance of transformers in the detection of pain and no pain from patients' facial expressions.

It is worthwhile noting that the two databases are not balanced. The first one: UNBC-McMaster shoulder pain [3], the amount of images belonging to no pain class is much bigger that the one belonging to pain class. This can be noticed in Table I. Concerning the second database: BioVid Heat Pain [4] is also unbalanced. Contrary to the first one, this database contains images of pain more than the ones with no pain. The fact that the databases are not balanced is a potential cause of the difference obtained in accuracy. Despite this problem, we can still state that our proposed architecture

TABLE II

RESULTS OF THE DIFFERENT EXPERIMENTS OF THE PROPOSED METHOD TO DETECT PAIN FROM NO PAIN, COMPARED TO THE STATE OF THE ART MODEL: GOOGLENET. THE EXPERIMENTS ARE DONE USING THE PUBLICLY AVAILABLE DATASETS: UNBC-MCMASTER SHOULDER PAIN [3] AND THE BIOVID HEAT PAIN [4]. WE NOTE THE PROPOSED ARCHITECTURE DEIT-PNP TO DESIGN THE FINE-TUNED DEIT FOR DETECTION OF PAIN FROM NO PAIN. THE GOOGLENET FOR PAIN AND NO PAIN IS ALSO NOTED AS GOOGLENET-PNP.

Method	**Input size**	**Epochs**	**Accuracy %**	
			UNBC McMaster Shoulder Pain Database [3]	***BioVid Heat Pain Database [4]***
GoogleNet-PNP	256×256	30	80.01	65.75
Deit-PNP *Proposed method*	256×256	30	**84.15**	**72.11**

of a fine-tuned Deit for pain recognition exceeds the state of the art methods in terms of accuracy.

V. CONCLUSION

In this article, we have presented a novel architecture for the binary recognition of pain from facial expressions. This architecture is based on the data-efficient image transformers [6] (Deit). We used fine-tunig of the pre-trained Deit model on the ImageNet [19] dataset. The Deit model achieved interesting results compared to the state of the art. In our paper, we trained the proposed method using UNBC-McMaster shoulder pain [3] and BioVid Heat Pain [4] datasets. These two datasets contain facial images of subjects expressing pain, and no pain. Moreover, to compare our proposed architecture with the state of the art, we implemented a pre-trained model to discriminate pain from no pain facial expressions. We chose the GoogleNet [24] model. At the end of the experiments, our proposed method showed promising results compared to the state of the art method. These results demonstrate the importance of the transformers [7] in image recognition. Finally, the present findings might help to have more insights into the importance of the transformers in the pain field.

REFERENCES

[1] WILLIAMS, Amanda C. de C. et CRAIG, Kenneth D. Updating the definition of pain. Pain, 2016, vol. 157, no 11, p. 2420-2423.

[2] WERNER, Philipp, LOPEZ-MARTINEZ, Daniel, WALTER, Steffen, et al. Automatic recognition methods supporting pain assessment: A survey. IEEE Transactions on Affective Computing, 2019.

[3] LUCEY, Patrick, COHN, Jeffrey F., PRKACHIN, Kenneth M., et al. Painful data: The UNBC-McMaster shoulder pain expression archive database. In : 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). IEEE, 2011. p. 57-64.

[4] WERNER, Philipp, AL-HAMADI, Ayoub, NIESE, Robert, et al. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In : Proceedings of the British Machine Vision Conference. 2013. p. 1-13.

[5] AUNG, Min SH, KALTWANG, Sebastian, ROMERA-PAREDES, Bernardino, et al. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset. IEEE transactions on affective computing, 2015, vol. 7, no 4, p. 435-451.

[6] TOUVRON, Hugo, CORD, Matthieu, DOUZE, Matthijs, et al. Training data-efficient image transformers & distillation through attention. In : International Conference on Machine Learning. PMLR, 2021. p. 10347-10357.

[7] VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, et al. Attention is all you need. In : Advances in neural information processing systems. 2017. p. 5998-6008.

[8] DOSOVITSKIY, Alexey, BEYER, Lucas, KOLESNIKOV, Alexander, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[9] CHEN, Junkai, CHI, Zheru, et FU, Hong. A new framework with multiple tasks for detecting and locating pain events in video. Computer Vision and Image Understanding, 2017, vol. 155, p. 113-123.

[10] BARGSHADY, Ghazal, SOAR, Jeffrey, ZHOU, Xujuan, et al. 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. p. 52-56.

[11] KALTWANG, Sebastian, TODOROVIC, Sinisa, et PANTIC, Maja. Doubly sparse relevance vector machine for continuous facial behavior estimation. IEEE transactions on pattern analysis and machine intelligence, 2015, vol. 38, no 9, p. 1748-1761.

[12] AHONEN, Timo, HADID, Abdenour, et PIETIKÄINEN, Matti. Face recognition with local binary patterns. In : European conference on computer vision. Springer, Berlin, Heidelberg, 2004. p. 469-481.

[13] HAQUE, Mohammad A., BAUTISTA, Ruben B., NOROOZI, Fatemeh, et al. Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities. In : 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018. p. 250-257.

[14] HOCHREITER, Sepp et SCHMIDHUBER, Jürgen. Long short-term memory. Neural computation, 1997, vol. 9, no 8, p. 1735-1780.

[15] KARAMITSOS, Ioannis, SELADJI, Ilham, et MODAK, Sanjay. A Modified CNN Network for Automatic Pain Identification Using Facial Expressions. Journal of Software Engineering and Applications, 2021, vol. 14, no 8, p. 400-417.

[16] SIMONYAN, Karen et ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[17] CHAUDHARI, Sneha, MITHAL, Varun, POLATKAN, Gungor, et al. An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, vol. 12, no 5, p. 1-32.

[18] XIANG, Jia et ZHU, Gengming. Joint face detection and facial expression recognition with MTCNN. In : 2017 4th international conference on information science and control engineering (ICISCE). IEEE, 2017. p. 424-427.

[19] DENG, Jia, DONG, Wei, SOCHER, Richard, et al. Imagenet: A large-scale hierarchical image database. In : 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009. p. 248-255

[20] HINTON, Geoffrey, VINYALS, Oriol, et DEAN, Jeff. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[21] KHAN, Salman, NASEER, Muzammal, HAYAT, Munawar, et al. Transformers in vision: A survey. arXiv preprint arXiv:2101.01169, 2021.

[22] https://github.com/facebookresearch/deit

[23] HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, et al. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[24] SZEGEDY, Christian, LIU, Wei, JIA, Yangqing, et al. Going deeper with convolutions. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1-9.

[25] EL MORABIT, Safaa, RIVENQ, Atika, ZIGHEM, Mohammed-Ennadhir, et al. Automatic pain estimation from facial expressions: a comparative analysis using off-the-shelf CNN architectures. Electronics, 2021, vol. 10, no 16, p. 1926.

[26] KINGMA, Diederik P. et BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.