CS490
Introduction to Research in Computer Science*: Automated Estimation of Shoulder Pain Intensity with Transformer Networks

* Note: This project is carried out within the scope of Bilkent University CS490 course.

Ömer Faruk Akgül
Department of Computer Engineering
Bilkent University
Ankara, Turkey
faruk.akgul@ug.bilkent.edu.tr

Asst. Prof. Hamdi Dibeklioğlu
Department of Computer Engineering
Bilkent University
Ankara, Turkey
dibeklioglu@cs.bilkent.edu.tr

***Abstract*—Measurement of the pain is generally based on the self-report of the patients or assessments of professionals. While the efforts of clinical experts can provide reliable assessments, it requires time-intense labor. On the other hand, not all the patients are able to express their pain clearly. With the recent improvements in Machine Learning techniques, automatic estimation from facial expression videos gained popularity. In this work, we developed an attention-based transformer network to estimate the shoulder pain intensity of patients from their facial expressions. We also proposed a custom loss function, which enforces consistency between the predictions and the data (UNBC-McMaster Pain Archive) for a regression problem. Initial results show improvements compared to the previous works dealing with the same problem. For the next steps of the project, we aim to create a network that is end-to-end differentiable.**

***Index Terms*—Transformer Networks, Pain Estimation, Facial Expression, Attention**

I. Introduction

Pain is defined as an unpleasant sensory and emotional experience associated with actual or potential tissue damage [1]. In order for the treatment applied to the patient to progress in a healthy way, the pain felt by the patient must be accurately determined and adjustments are made in the treatment accordingly. To this end, significant efforts have been put into automated pain estimation from facial expressions [2], [3].

To estimate the pain from facial expressions, there are several widely used metrics. These metrics have been developed to scale the pain level of the patients. Prkachin and Solomon Pain Intensity (PSPI) is a unit of measure that combines the action intensity of the facial regions [1]. However, studies have shown that each patient's pain expressiveness is different [4]. For example, it has been demonstrated that men and women react differently. Several disorders in the nervous system such as dementia reduce the expressiveness of the patient [5]. The facial reactions in children are observed to be very different compared to adults [4]. Therefore, the PSPI score is not an adequate representative metric for the pain intensity.

Metrics such as the Visual Analogue Scale (VAS), Numerical Rating Scale (NRS), Verbal Rating Scale (VRS), on the other hand, contain more subjective results in which patients report their pain level. For instance, the VAS score is obtained by patients marking the pain they feel on a scale [6]. VAS is the most preferred metric in studies because it contains statistically more robust values and provides patient-specific subjectivity [7]. However, it cannot be expected to express self-pain in groups such as children with insufficient motor skills or in unconscious patients. This situation reveals once again the importance of automatic estimation of the pain.

To date, studies on automated estimation of pain intensity from facial expression have mostly been based on the PSPI score [4]. However, as stated above, studies on the VAS score have increased in recent studies, since the main purpose is to estimate self-reported pain [8]–[10]. Similarly, in this work, we focused on estimating the self-reported pain intensity of patients from facial expressions. To this end, we created a framework in which we estimate VAS score by using transformer networks, which is a self-attention based learning method [11].

Our work aims to achieve state-of-the-art performance by deploying the transformer structure into the self-reported pain estimation problem. Recently, the transformer mechanisms are preferred over Recurrent Neural Network models such as LSTMs [12], [13]. The reasons can be listed as

- Transformers introduce a self-attention mechanism that does not suffer from long dependency issues, which means there is no risk of forgetting the past information. Furthermore, the relationship between words (frames in our problem) can be learned thanks to multi-head attention and positional embedding concepts [11].
- As transformers does not require processing in order like LSTM models, it provides a parallelization ability which means processing much more data within the same time interval [11].

As long-term dependency is also a problem in Computer Vision tasks, transformers become a hot-topic in image tasks [14], [15], as well. In addition, the translation invariancy aspect of convolutions causes a loss of information in the global context. Therefore, transformers are preferred to acquire global representation of images [15]. The fact that transformers have not been used in the context of pain estimation before, and the above-mentioned advantages of transformers led us to question whether we can overperform results of previous studies.

Specifically, we fed the features extracted using the pre-trained convolutional neural network into our transformer network, dictating more attention to significant parts in the video sequences. For our work, which we define as the regression problem, we have implemented a new loss function by modifying the chi-square loss [16]. With this custom loss, we aimed to maximize the consistency of the histograms we created for predictions and labels while minimizing the Mean Absolute Error (MAE) in estimations. While our work resulted in better performance compared to previous studies on shoulder pain intensity estimation, it provided efficient training time thanks to the parallelization ability provided by the transformer mechanism.

II. Related Work

With the recent advancements in Machine Learning and Computer Vision techniques, and release of the pain datasets such as UNBC-McMaster Shoulder Pain Archive (Pain Archive) [17], BioVid Heat Pain [18] and EmoPain [19], research focusing on automated estimation of pain from facial images gained popularity. Among the mentioned datasets, Pain Archive provides PSPI scores for each frame in a video [1], so a wide range of works focuses on estimation of the PSPI scores from pain images in Pain Archive [20], [21] instead of self-reported pain VAS. Also, while initial works aimed at detecting whether the patient feels pain or not (binary classification problem) [1], more recent studies aim at assessing the level of pain intensity [22]. Here is the outline of the leading studies.

[23], [24] used Artificial Neural Networks to discriminate pain and no-pain images. [25] used Support Vector Machines (SVM) using Histogram of Oriented Gradients (HOG) as frame features, again for detection of pain occurrence. Later, SVMs are used to leverage the level of pain from images with a one-versus-all approach [22]. For detection of pain level from videos, [26] have proposed a semi-supervised model to detect the frame with the highest pain level to estimate the entire sequence. Similarly, [27] have attempted to find the most demonstrative (i.e the highest pain) frame with a multiple-instance learning method.

However, none of the listed papers above has dealt with the video-based estimation of VAS scores, which is considered the golden standard for self-reported pain intensity. To the best of our knowledge, there have been three recent studies in this direction. In [9], Martinez and his colleagues propose a model consisting of two main steps for personalized estimation of VAS scores. First, they employ a bi-directional Long Short-Term Memory to estimate PSPI scores from image sequences. To transform PSPI based scores into a personalized manner, Individual Facial Expressiveness Score is used to augment features. Then, they fed these features into Hidden Conditional Random Fields for self-reported VAS estimations. As this method does not generalize to unseen patients, [8] proposes another method for VAS estimation, the so-called DeepFaceLIFT. In this method, personalization is enabled via the computation of the set of predefined features such as gender, age and skin tone. These features are used together with image features and the neural network is trained. Then, the Gaussian Regression model is fed with a combination of the output of neural networks and a set of sequence-level statistics. Similar to the previous method, DeepFaceLIFT is a two-stage learning approach. Erekat and her colleagues [10] propose a spatiotemporal end-to-end CNN-GRU model for VAS estimation. They also introduce a custom loss function that enforces consistency between different pain estimation scales (VAS, the Sensory Scale (SEN), and the Affective Motivational Scale (AFF), and the Observer Pain Intensity (OPI)).

On the other hand, transformers are first introduced by Vaswani et al. for language machine translation tasks. After demonstrating exemplary performance on a broad range of language tasks, researchers started using transformer models in other domains. Vision Transformer (ViTs) [15] structure is first introduced to show that image classification tasks can be performed with transformers without relying on CNNs. Later, transformers are used for object detection [28] and image segmentation tasks [29]. Sun and colleagues [30] proposed a model called VideoBERT for activity recognition and video captioning tasks. Similarly, [31] aims at classifying actions in spatiotemporal data by using a transformer network. In this work, we also dealt with spatiotemporal data for the estimation of VAS scores with transformers.

III. METHOD

We proposed a two-stage learning framework for our VAS estimator. At the first step, frames are passed through a pre-trained convolutional neural network to extract features. Then, the features are pre-processed into a compatible form to feed the transformer network. The model is illustrated in Figure 1.

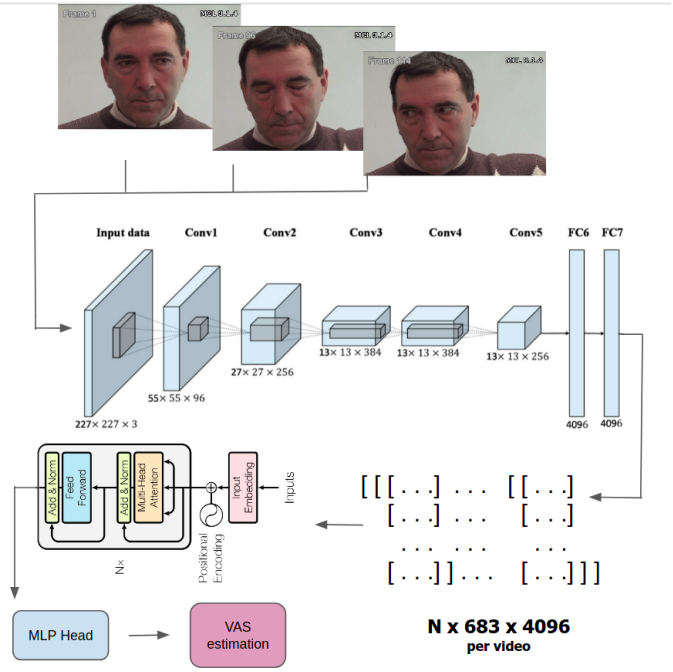

Fig. 1. Visualization of two-stage learning framework. The first block shows the feature extraction with AlexNet. The second one displays the transformer network given the zero-padded outputs of CNN architecture. The output is the VAS score estimation. N is the total number of frames per video.

A. Extraction of Frame Features from Normalized Face Appearances

In our work, we used the normalized face images created by Erakat et al. [10]. They use 66 facial points distributed with the UNBC-McMaster Pain Archive to extract the normalized face appearance of each video frame.

We used weights of AlexNet pre-trained CNN (composed of 5 convolutional layers, a max-pooling layer, and 2 fully-connected layers) [32] to learn spatial features of the video frames. At the current stage, we did not want to spend our resources on extracting features of images with a custom model to avoid computational costs. Thus, we used AlexNet weights after removing the fully connected layer. Choosing AlexNet in our model was significant as it provides a comparison of transformer network with the CNN-GRU model at [10]. The normalized images of size 128x117x3 are preprocessed and converted into 224x224x3 as AlexNet architecture accepts inputs of that size. At the end of the first stage, we had 4096 features per frame.

B. Transformer Network

The extracted frame features are then stacked and zero-padded to the maximum sequence length (i.e 683) in the database. Then, the transformer network with the dense dimension of 8 and 2 heads is fed with frame features. The number of dense dimensions and heads are selected based on our experimental results. Note that during the back-propagation, only the weights of the transformer is updated as pre-trained CNN is used only for extracting the spatial features.

Although there are 10 levels of pain in the VAS scale and it is possible to design the model as a classification problem, we set the output dimension of the model as 1. Our reasoning is that the pain is continuous and not independent from each other. As a loss function, we selected Mean Absolute Error (MAE) for the initial experiments to avoid focusing on the outlier data. We also scaled target VAS scores to the 0-1 range to make the learning process more stable.

C. Custom Loss Function

In our model, we aimed to maximize the consistency between the distributions of labels and predictions while minimizing MAE. To this end, we proposed the following objective function.

$$L = \alpha \cdot L_{MAE} + (1 - \alpha) \cdot L_c \tag{1}$$

where α is a learnable parameter given to the model. L_{MAE} is the Mean Absolute Error and can be formulated as given below.

$$L_{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

where n is the total number of videos, y_i is the label of video and $\hat{y}_i$ is the prediction of network for i^{th} sample.

L_C represents the consistency term inspired by the Chi-Square loss proposed in [16]. As we designed our problem as a regression task, we cannot directly apply chi-square loss. Instead, we use a simple trick modifying predictions of our network by rounding off to the nearest integer value. In other words, continuous-valued predictions are transformed into a histogram of 11 classes (from 0 to 10, just like VAS scores). Then the average distance between predicted and ground-truth histogram is calculated and added to the loss function. This new loss function punishes the inconsistency of these two histograms. Also, the relative importance of the added term is determined during the training process with the coefficient of $1 - \alpha$.

$$L_c = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{10} \frac{\hat{y_j}^2}{y_j} - 1 \tag{3}$$

where j represents the the class (i.e the column of the histogram) with VAS score j. Proof of the unbiasedness of chi-square loss can be found in [16].

When (1) is used, we needed to apply a label smoothing to avoid division by zero in (3). Label smoothing also provides avoiding overconfidence and improves generalization ability [16]. Here is the formula for label smoothing.

$$y_i^{LS} = y_i(1-\beta) + \frac{\beta}{k} \tag{4}$$

k represents the number of classes and equals to 11 in our setting. β is a hyper-parameter and we used $\beta = 0.1$ in our model.

First, we trained our proposed architecture with MAE to get the basement performance. Weights of the transformer network are randomly initialized and updated to minimize MAE. Hyper-parameters are determined based on the minimum loss value and preferred values are displayed in Table 1. Then, the proposed loss function is applied and the model is expected to tune the α parameter (see Equation (1)). The table also includes the selected epoch number, which is determined by observing the printed loss values.

TABLE I
LIST OF THE TUNED HYPER-PARAMETERS FOR THE OPTIMIZATION OF TRANSFORMER MODEL

Hyper-parameter	Values
Learning rate	0.001
Dropout rate	0.3
Number of epochs	250
Optimizer	Nadam[a]

[a] Adam with Nesterov momentum.

IV. EXPERIMENTS AND RESULTS

A. The UNBC-McMaster Pain Archive

The UNBC-McMaster Pain Archive [17] is used to address the need for a well-annotated facial expression database for pain assessment from video. The UNBC-McMaster Pain Archive is collected from patients with shoulder pain [19]. Participants were video recorded during the abduction and flexion movements of their affected and unaffected shoulders. The publicly available portion of the database consists of 200 videos from 25 different participants (with 48398 frames [19]) After each record, patients marked the highest pain level they felt on the VAS scale, which is ranging from 0 to 10.

B. Experimental Setup

We use two-level 5-fold cross-validation to provide each sample has a contribution to the training and evaluation performance of our model. The data is divided into five independent folds and one of them is used as a test set while one of the remaining four sets is used as a validation set. If a participant is used as part of the training fold, he/she is not included in the test fold to ensure subject independence. We dealt with the problem as a regression problem rather than a multi-class classification problem as pain scores are continuous and not independent. The regression model is optimized (i.e., searching for the best hyper-parameters of the transformer model) on the validation set by minimizing the validation loss (i.e. mean absolute error, MAE) in the first experiment. Similarly, the custom loss function is minimized on the validation set during the second experiment. In addition, labels of the videos are scaled to 0-1 range to ensure a more stable learning. When calculating the MAE and plotting the results, labels and predictions are re-scaled back to 0-10 range (While plotting predictions and labels in terms of VAS, numbers are rounded to the closest integer values).

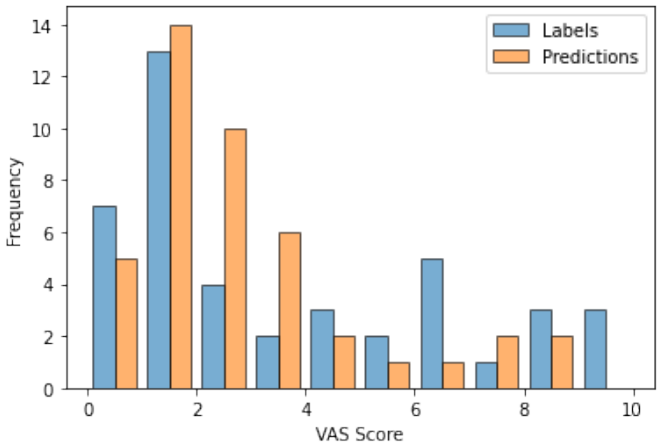

Fig. 2. Histogram of Transformer Model (with objective function MAE) Predictions and Labels.

C. Experiment with MAE Loss and Comparison with State of the Art

We used Mean Absolute Value loss as an objective function in our first experiment to provide comparison of transformer model with CNN-GRU model at [10]. The histogram of predictions and labels of transformer model is given in Figure 2. [10] shares their results of CNN-GRU model with pure MAE loss (i.e without consistency term of their custom loss function) obtained when difference of predictions and labels for each pain scale is minimized. The pain scales used in [10] are VAS, OPI (the Observer Pain Intensity), SEN (the Sensory Scale), AFF (the Affective Motivational Scale). Our proposed model achieved a better performance. The comparison with our model is given in Table 2.

Transformer model, when trained with MAE loss, also achieved a superior performance compared to the previous studies, when the best published scores are considered. The performance of our model (with MAE loss) and its

TABLE II
PERFORMANCE COMPARISON OF OUR MODEL WITH EREKAT ET AL (MAE) [10]

Model	MAE (VAS)
CNN-GRU, 2 Labels (VAS+OPI) [10]	2.74
CNN-GRU, 4 Labels (VAS+OPI+AFF+SEN) [10]	2.38
Transformer	2.17

comparison with previous studies are given in Table 3.

TABLE III
PERFORMANCE COMPARISON OF OUR MODEL WITH PREVIOUS STUDIES

Model	MAE (VAS)
DeepFaceLift [8]	2.91[a]
CNN-GRU [10]	2.34[a]
Transformer	2.17

[a]The best MAE for each model.

D. Experiment with Proposed Custom Loss

We conducted the initial experiments with custom loss function proposed in section 3.C. Initial results show a worse results compared to the MAE loss but we are currently working on debugging the implementation. The current MAE of VAS scores is around 2.3 and example histogram of the results is given in Figure 3 to inform the reader about the current point.

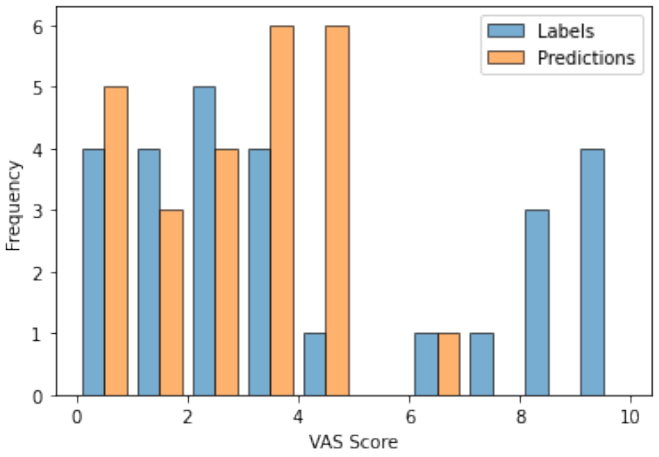

Fig. 3. Histogram of Transformer Model (with Custom Loss) Predictions and Labels (see section 3.C).

V. CONCLUSION

In this study, we proposed an attention-based learning framework for automated estimation of the self-reported pain scores from videos. We also aimed to provide the consistency between predictions and labels by introducing a custom loss function. Our initial results shows improvements in the VAS score estimation compared to the previous state of the art models. Therefore, our results show that transformer networks can be applied on video-based prediction problems and may achieve state of the art performance. Our proposed objective function still needs some corrections and the detailed analysis will be shared in the final report. We will also share the detailed comparison of training and inference time of transformer model with others. For the next steps of our work, we are planning to transformer our model into an end-to-end differentiable form (including the part of feature extraction with pre-trained network) to better utilize the frame features of videos.

REFERENCES

[1] Z. Hammal and J. F. Cohn, "Automatic, objective, and efficient measurement of pain using automated face analysis," Social and Interpersonal Dynamics in Pain, pp. 121–146, 2018.

[2] M. Matsangidou, A. Liampas, M. Pittara, C. S. Pattichi, and P. Zis, "Machine learning in pain medicine: An up-to-date systematic review," Pain and Therapy, vol. 10, no. 2, pp. 1067–1084, 2021.

[3] D. Liu, D. Cheng, T. T. Houle, L. Chen, W. Zhang, and H. Deng, "Machine learning methods for automatic pain assessment using facial expression information," Medicine, vol. 97, no. 49, 2018.

[4] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," Pain, vol. 139, no. 2, pp. 267–274, 2008.

[5] T. Hadjistavropoulos, K. Herr, K. M. Prkachin, K. D. Craig, S. J. Gibson, A. Lukas, and J. H. Smith, "Pain assessment in elderly adults with dementia," The Lancet Neurology, vol. 13, no. 12, pp. 1216–1227, 2014.

[6] M. A. Ferreira-Valente, J. L. Pais-Ribeiro, and M. P. Jensen. Validity of four pain intensity rating scales. Pain, 152(10):2399–2404, 2011.

[7] D. D. Price, P. A. McGrath, A. Rafii, and B. Buckingham. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain, 17(1):45–56, 1983.

[8] D. Liu, F. Peng, O. Rudovic, and R. W. Picard, "DeepFaceLIFT: Interpretable Personalized Models for Automatic Estimation of Self-Reported Pain" In Proceedings of the 1st IJCAI Workshop on Artificial Intelligence in Affective Computing (Proceedings of Machine Learning Research). pp. 1–16. 2017

[9] D. L. Martinez, O. Rudovic, and R. Picard, "Personalized automatic estimation of self-reported pain intensity from facial expressions," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[10] D. Erekat, Z. Hammal, M. Siddiqui, and H. Dibeklioğlu, "Enforcing multilabel consistency for automatic spatio-temporal assessment of shoulder pain intensity," Companion Publication of the 2020 International Conference on Multimodal Interaction, 2020.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin. "Attention is all you need". In NIPS, 2017.

[12] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on Transformer vs RNN in speech applications," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019.

[13] G. Tang, M. Müller, A. Rios, and R. Sennrich, "Why self-attention? A targeted evaluation of Neural Machine Translation Architectures," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[14] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, and M. Shah "Transformers in Vision: A Survey" unpublished.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". unpublished.

[16] Z. Wang, M. Wang. "Chi-square Loss for Softmax: an Echo of Neural Network Structure", unpublished.

[17] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-mcmaster shoulder pain expression archive database," Face and Gesture 2011, 2011.

[18] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, S. Crawcour, P. Werner, A. Al-Hamadi, and A. O. Andrade, "The Biovid Heat Pain Database data for the advancement and systematic validation of an automated pain recognition system," 2013 IEEE International Conference on Cybernetics (CYBCO), 2013.

[19] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: Requirements, challenges and the Multimodal Emopain Dataset," IEEE Transactions on Affective Computing, vol. 7, no. 4, pp. 435–451, 2016.

[20] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12, 2012.

[21] S. Rezaei, A. Moturu, S. Zhao, K. M. Prkachin, T. Hadjistavropoulos, and B. Taati, "Unobtrusive pain monitoring in older adults with dementia using pairwise and contrastive training," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1450–1462, 2021.

[22] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," Advances in Visual Computing, pp. 368–377, 2012.

[23] M. Monwar and S. Rezaei, "Pain recognition using artificial neural network," 2006 IEEE International Symposium on Signal Processing and Information Technology, 2006.

[24] J. Egede, M. Valstar, and B. Martinez, "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation," 2017 12th IEEE International Conference on Automatic Face amp; Gesture Recognition (FG 2017), 2017.

[25] J. Chen, Z. Chi, and H. Fu, "A new framework with multiple tasks for detecting and locating pain events in video," Computer Vision and Image Understanding, vol. 155, pp. 113–123, 2017.

[26] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic, "Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation," Computer Vision – ACCV 2016, pp. 171–186, 2017.

[27] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020.

[29] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal selfattention network for referring image segmentation," in CVPR, 2019.

[30] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in ICCV, 2019

[31] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.