Facial Action Unit Detection using 3D Face Landmarks for Pain Detection

Kevin Feghoul, Mondher Bouazizi, Deise Santana

Facial Action Unit Detection using 3D Face Landmarks for Pain Detection

Kevin Feghoul[1], Mondher Bouazizi[2] and Deise Santana Maia[3]

***Abstract*— Automatic detection of facial action units (AUs) has recently gained attention for its applications in facial expression analysis. However, using AUs in research can be challenging since they are typically manually annotated, which can be time-consuming, repetitive, and error-prone. Advancements in automated AU detection can greatly reduce the time required for this task and improve the reliability of annotations for downstream tasks, such as pain detection. In this study, we present an efficient method for detecting AUs using only 3D face landmarks. Using the detected AUs, we trained state-of-the-art deep learning models to detect pain, which validates the effectiveness of the AU detection model. Our study also establishes a new benchmark for pain detection on the BP4D+ dataset, demonstrating an 11.13% improvement in F1-score and a 3.09% improvement in accuracy using a Transformer model compared to existing studies. Our results show that utilizing only eight predicted AUs still achieves competitive results when compared to using all 34 ground-truth AUs.**

I. Introduction

The Facial Action Coding System (FACS) [1] is a comprehensive taxonomy that objectively describes anatomical movements of the face. This coding system is composed of 32 non-overlapping fundamental actions of individual muscles or groups of muscles, known as action units (AUs). By combining AUs, any facial expression can potentially be encoded, allowing inference of an individual's emotional state. Each AU is identified by an ID and can characterize several emotions. For example, AU number 9, which represents nose wrinkle, can indicate emotions such as disgust or pain. Table I show different emotions and their corresponding AUs [2].

Automatic pain detection is of high interest due to its potential impact in various fields, such as medical diagnosis, remote monitoring, and robotics. With the recent advances in machine learning techniques, deep learning [3] is now the de facto choice for dealing with unstructured data, leading to significant breakthrough in computer vision [4], speech recognition [5], and natural language processing [6]. As such, deep learning is of great significance for the task of pain detection, as it enables the recognition of human pain through facial expressions and physiological data [7].

Privacy concerns are one of the main challenges faced by researchers conducting experiments on human subjects. The inability to share valuable and well-constructed datasets impedes progress in various fields, including pain and emotion recognition. In this study, we propose a novel approach to anonymize data specifically for pain detection. Our aim is to remove confidential information related to human faces from the dataset, while preserving critical features that enable the research community to utilize the dataset with minimal information loss.

In this study, we propose a novel method for extracting AUs from 3D face landmarks. To evaluate the effectiveness of our approach, we conducted experiments on the pain detection task using the extracted AUs to train Transformer [8] and LSTM [9] models. We assessed the performance of these models by comparing their results with the benchmark obtained from ground-truth AUs. All experiments were conducted on the BP4D+ [10] dataset.

The contributions of this work are fourfold and can be summarized as follows: (1) a method to extract AUs from 3D face landmarks; (2) the use of the extracted AUs for the task of pain detection; (3) the demonstration that the bare minimum number of AUs (*i.e*. 8 among 34 ground-truth AUs) extracted from the 3D face landmarks can be used for pain detection; (4) to the best of our knowledge, this is the first work to propose the utilization of a Transformer model for pain detection using AUs.

TABLE I
EMOTIONS AND THEIRS ASSOCIATED AUS

Emotion	AUs
Happiness	6, 7, 12, 25, 26
Sadness	1, 4, 6, 15, 17
Fear	1, 2, 4, 5, 7, 20, 25
Anger	4, 5, 17, 23, 24
Disgust	7, 9, 19, 25, 26
Pain	4, 6, 7, 9, 17, 18, 23, 24

II. Related Work

Our proposed framework is mainly related to the detection of AUs and their uses for facial expression analysis, thus we separated this section in two parts.

A. AUs detection

Traditional methods for AUs detection rely heavily on handcrafted features. For example, Valstar et al. [11] used a Support Vector Machine to recognize AUs and analyze their temporal behavior from face videos, utilizing a set of spatio-temporal features calculated from 20 facial fiducial points.

[1]Kevin Feghoul is with University of Lille, Inserm, CHU Lille, UMR-S1172 - LilNCog, UMR 9189 CRIStAL, F-59000 Lille, France, email: kevin.feghoul@univ-lille.fr

[2]Mondher Bouazizi is with Faculty of Science and Technology, Keio University, Japan, email: mondher.bouazizi@keio.jp

[3]Deise Santana Maia is with University of Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France, email: deise.santanamaia@univ-lille.fr

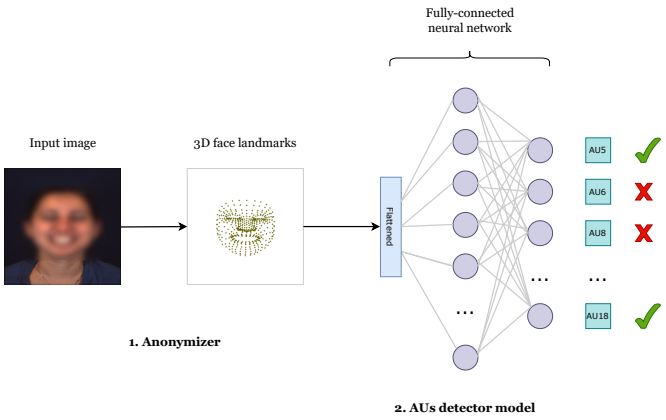

Fig. 1. A flowchart of the AUs detector: upon generating the 3D face landmarks, a fully-connected neural network with multiple outputs is used to detect the AUs.

Baltrušaitis et al. in [12] presents a real-time facial AUs intensity estimation and occurrence detection framework based on the combination of appearance features (Histogram of Oriented Gradients) and geometry features (shape parameters and landmark locations).

Recent advancements in deep learning have yielded impressive results in detecting AUs. Li et al. [13] utilized a Gated Graph Neural Network (GGNN) in a multi-scale CNN framework to integrate semantic relationships among AUs. In [14], Zhang et al. proposed a model called Multi-Head Fused Transformer that uses both RGB and depth images to learn discriminative AUs features representations. In their work, Jacob et al. [15] utilized image features and attention maps to feed different action unit branches, where discriminative feature embeddings were extracted using a novel loss function. Next, to capture the complex relationships between the different AUs, a Transformer encoder is employed.

Unlike deep learning models that preprocess entire images, our approach employs a straightforward fully-connected neural network to map 478 facial landmarks to 8 AUs. Our method stands out for its efficiency in terms of computational complexity and time.

B. AUs for facial expression analysis

FACS has been widely used for facial expression analysis in various applications. Darzi et al. [16] used AUs to evaluate the intensity of symptoms of OCD and depression in individuals undergoing deep brain stimulation. Other studies have explored the use of AUs for detecting stress [17] and pain. Hinduja et al. [7] trained a Random Forest model to recognize pain by combining AUs with physiological data. In [18], Meawad et al. proposed an approach for detecting pain in sequences of spontaneous facial expressions, based on extracted landmarks from a mobile device. For comprehensive surveys on automatic pain detection from facial expressions, the readers may refer to [19], [20].

As we could see, creating effective tools for detecting AUs can have a wide range of benefits in fields such as medicine, psychology, and affective computing.

III. Proposed Approach

Before presenting our proposed approach, we will introduce the method we used to extract the 3D face landmarks, as well as the dataset on which we conducted our experiments.

A. Face mesh

The face mesh is a 3D model of the human face. This type of model is commonly used in applications involving 3D modeling or augmented reality [21]. In our current work, we will be using a pre-trained neural network (NN) to extract the face mesh [22]. This NN has been trained to identify the x and y coordinates of the different landmarks as well as to estimate the z coordinate. The model is designed to predict the positions of 468 landmarks spread out across the facial surface, with an additional 10 landmarks allocated for the iris. Overall, the face mesh consists of 478 points, as illustrated in Fig. 2.

By reducing a face image to its face mesh, we can significantly decrease the size of the information while preserving most of the information, as we will demonstrate later. Additionally, converting a face image into a face mesh can help safeguard the privacy of the individuals participating in any study in which their faces are visible. In the present work, we only used the 3D face landmarks from the face mesh.

B. Dataset

The BP4D+ dataset includes a collection of data for each frame, including a 3D facial model, an RGB image, a thermal image, and eight physiological signals. For our study, we focused on analyzing the 2D RGB image data. From this data, we extracted the 3D face landmarks as described previously. The BP4D+ dataset consists of 140 participants, including 82 females and 58 males, with ages ranging from 18 to 66 years old. The dataset was designed to elicit a wide range of authentic emotions, with each participant performing a set of 10 activities. However, FACS experts annotated AUs for occurrence and intensity for four emotion elicitation tasks (happiness, embarrassment, fear, and pain). Nevertheless, regarding these four emotions, only the most expressive segments were annotated (roughly 20 seconds on average). We have defined a binary classification task for pain by considering pain sequences as the positive class and the remaining three emotion sequences as the negative class.

C. Overall system description

In Fig. 1 and 2, we show the flowchart of our proposed system. Our system is composed of three main parts:

1) An anonymizer: this component refers to the transformation of images in our dataset into 3D face landmarks. In brief, it generates a set of 478 landmark coordinates over time. Such point coordinates are very useful in extracting information allowing for pain detection, as we will demonstrate, yet, they do not allow identifying the identity of the subjects.

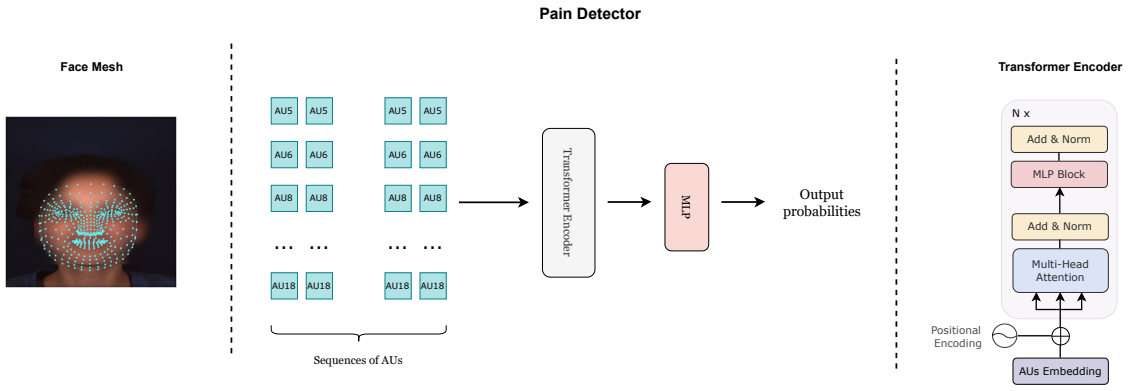

Fig. 2. A flowchart of the pain detector: the AUs detected from the face mesh (on the left) is processed through a Transformer encoder to identify the class (pain).

TABLE II
THE INDIVIDUAL AND AVERAGE F1-SCORE OF THE DETECTED ACTION UNIT.

Method	AU5	AU6	AU8	AU9	AU10	AU12	AU14	AU18	AVG
FCN	82.87	86.27	89.10	77.43	87.83	80.10	75.27	55.13	79.25

2) An AU detector: this component simply allows extracting the main AUs from a given set of 3D face landmarks. Conventionally, NNs are trained to identify face meshes from full images, making use of all embedded information. By creating NNs that extract the AUs from the 3D face landmarks, we demonstrate the usefulness of the first component, and that it is still possible to derive relevant information from a condensed and small-size vector such as the set of 3D landmarks.
3) A pain detector: this component relies on the detected AUs to identify the pain expressed by the subjects. While the accuracy of the AU detection is nowhere near perfect, we show that it is still possible to identify pain with an accuracy nearly similar to that when using the ground-truth AUs. We also demonstrate that for the pain detection task, very few AUs are required to achieve an accuracy relatively close to that when using all the AUs.

D. Detailed system description

1) The AUs detector: Conventionally, AUs are extracted from the images of human faces (as shown in Fig 2). However, the transformations we have applied that allowed us to remove the identify-related information of the participants led to a transformation of the input format itself. In the current work, we use a neural network that follows the structure of a typical fully connected neural network. In Fig. 1, we show the structure of the proposed network. The network is composed of one dense layer with 128 neurons. The input layer's shape follows the size of the input vector generated for the face landmarks, *i.e.* a shape of $3 \times N_s$, with N_s referencing to the number of landmarks (478 in our case). The results of the detection will be shown later in Section IV.

2) The pain detector: In this study, we propose using advanced deep learning models to detect pain by utilizing the AUs extracted from our AU detector approach as well as all ground-truth AUs. We believe that tailored neural networks will enhance performance and enable real-time operation. Deep learning is particularly well-suited for processing sequential data and eliminates the need for feature engineering, which is often required in traditional machine learning algorithms. We employed both Transformer and LSTM and compared their performance.

LSTM is a type of recurrent neural network (RNN) that handles sequential data by storing and retrieving information over time. It uses a series of gates to selectively update and forget information at each time step, allowing it to capture long-term dependencies in the input sequence. On the other hand, the Transformer is a type of neural network model that has become the go-to method for handling natural language processing tasks. Unlike LSTMs, it does not use recurrent connections, but instead uses self-attention mechanisms to model the relationships between different positions in the input sequence. This allows it to capture dependencies between distant positions in the sequence, which can be difficult for LSTMs. In this work, we utilize only the encoder component of the Transformer model, as shown in Fig. 2.

IV. RESULTS AND DISCUSSION

A. Evaluation framework

1) AUs detection: To validate our AU detection approach, we employed a subject-independent 3-fold cross-validation strategy. Each fold consisted of extracted 3D face landmarks from distinct subjects, which ensured that the model was

trained on one set of subjects and evaluated on another set of subjects to ensure generalizability. In terms of performance metrics, we utilized the F1-score to account for imbalanced class distribution.

2) Pain detection: Following prior works [7], we utilized a subject-independent 10-fold cross-validation strategy, where each fold comprised distinct subjects whose AUs were extracted using the AUs detection model. For evaluation, we used the accuracy and F1-score.

B. Implementation details

1) AUs detection: A total of 197,782 frames have been used for the training and testing. More precisely, we have used approximately 66% of the frames for training and the remaining ones for testing. The FCN model is composed of one hidden layer with 128 neurons. As for hyperparameters, we fixed the batch size to 64, the maximum number of epochs to 500, and the learning rate was set to 0.01.

2) Pain detection: During the training process, we used a fixed timestamp of 350 frames, which corresponds to approximately 14 seconds of data. To determine the optimal parameters for our models, we employed a grid-search strategy. For the Transformer model, we set the dimension of the linear projection layer to 1024, the number of multi-head attention to four, the number of encoder layers to two, and the learning rate to 10^{-5}. Regarding the LSTM model which is composed of two layers, we fixed the hidden dimension to 512, and the learning rate to 10^{-3}. Both models were trained with a batch size of 16 and a maximum number of epochs of 150.

All models (for AUs and pain detection) were trained using the Adam optimizer [23], with an exponential decay rate for the first and second-moment estimates fixed at 0.9 and 0.999, respectively. The whole pipeline was implemented using the PyTorch framework [24].

C. Results

1) Action units detection: In Table II, we show the performance of our proposed method for AUs detection. As previously stated, since numerous AUs are absent in the majority of the frames, most of the 34 AUs has not been detected. However, since those infrequent AUs represent only a very small portion of the ground-truth, it has not impacted the overall performance of the downstream detection task, as we demonstrate later on. Therefore, we present in Table II the 8 most present AUs, which also happens to be the most useful in our study. The detection accuracy ranges from 55.13% for AU18 to 89.1% for AU8, with an overall average F1-score equal to 79.25%. The main reason behind the poor detection accuracy of AU18 is its low presence in the dataset. As opposed to the other 7 AUs identified here, AU18 is present in only 14.76% of the frames in our dataset. That being said, the performance is overall good for our method which relies on a limited number of face landmarks, as opposed to other deep learning models which process the whole images. More importantly, as we will demonstrate, the detected AUs from the human face landmarks can be used to perform classification tasks such as pain detection. We will present the performance of the pain detection task using three different settings, which are:

- **8 predicted AUs (8AUP)**, in which the top 8 predicted AUs (see Table III) are used for pain detection
- **8 ground-truth AUs (8AUG)**, in which the ground truth of the same 8AUP are used for pain detection
- **All ground-truth AUs (All-AUG)**, in which all the 34 ground truth AUs are used for pain detection

TABLE III
TOP 8 DETECTED AUS

AU number	FACS name
AU 5	Upper Lid Raiser
AU 6	Cheek Raiser
AU 8	Lips toward each other
AU 9	Nose wrinkler
AU 10	Upper lip raiser
AU 12	Lip corner puller
AU 14	Dimpler
AU 18	Lip pucker

D. Pain detection

In Table IV, we show the performance for the task of pain detection. We report the overall accuracy and F1-score for each of the three experiments. As we can see, the results of training Transformer and LSTM models with 8AUP and 8AUG show a marginal difference. For Transformer and LSTM models, the difference in F1-score is 0.26% and 0.09% in favor of 8AUG, respectively. When comparing 8AUP with All-AUG, we can see a small difference of 2.43% and 1.13% in terms of F1-score, and 1.30% and 0.58% in terms of accuracy in favor of All-AUG for Transformer and LSTM, respectively.

In addition, 8AUP outperforms the Random Forest (RF) model proposed in [7], who is the only direct comparison to our work. Our approach results in a significant improvement in F1-score and accuracy, by 8.7% and 1.79% when using Transformer, and by 8.1% and 2.14% when using LSTM, compared to the Random Forest model. When using all-AUG, the F1-score and accuracy are improved by 11.13% and 2.72% when using Transformer, and by 9.23% and 2.14% when using LSTM, compared to the Random Forest model.

The confusion matrix as shown in Table V summarizes the classification performance for the pain detection task using 8AUP. Overall, we can see that we have a well-balanced confusion matrix indicating that the pain detection model is performing well for both classes (pain *vs* no pain), and is not biased towards one class or the other.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we first proposed an approach for AUs detection using only 3D face landmarks. We achieved an overall F1-score of 79.25% when using 8 among all the ground truth AUs. Then, to further demonstrate the effectiveness of our

TABLE IV

PAIN DETECTION PERFORMANCE USING 8AUP AS WELL AS THE GROUND TRUTH ONES.

	8AUP		8AUG		All-AUG	
Method	F1	Acc	F1	Acc	F1	Acc
RF [7]	-	-	-	-	73.40	89.02
LSTM	81.50	91.16	81.59	91.03	82.63	91.74
Transformer	82.10	90.81	82.36	90.89	**84.53**	**92.11**

TABLE V

SUM OF ALL THE CONFUSION MATRICES FOR THE TASK OF PAIN DETECTION OVER THE 10-FOLDS USING THE 8 PREDICTED AUS (8AUP).

Class	**Classified as**	
	Others	Pain
Others	**387**	22
Pain	29	**117**

method, we trained a Transformer and LSTM models for the task of pain detection using solely the 8 extracted AUs from our AUs detection model. Those results are then compared to the ones obtained from ground-truth AUs. The results from the 8 predicted AUs are similar to the ones from the 8 ground truth AUs, which confirms the relevance of our AUs detection model. Nevertheless, we observe an acceptable drop in performance when compared to results from all the available AUs. Moreover, the experimental results show that our framework, even only using the 8 predicted AUs, outperforms the state-of-the-art existing approach for the task of pain detection on the challenging BP4D+ dataset.

In a future study, we will work toward (1) the estimation of AUs intensity using 3D face landmarks, and; (2) on the design of a more sophisticated model for both AUs detection and AUs intensity estimation in a multi-task setting, to reach higher performance and efficiency.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[2] Cordaro, "Universals and cultural variations in expression in five cultures," *Unpublished doctoral dissertation, University of California, Berkeley*, 2013.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[5] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[7] S. Hinduja, S. Canavan, and G. Kaur, "Multimodal fusion of physiological signals and facial action units for pain recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 577–581.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.

[11] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, pp. 149–149.

[12] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–6.

[13] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8594–8601.

[14] X. Zhang and L. Yin, "Multi-modal learning for au detection based on multi-head fused transformers," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8.

[15] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7680–7689.

[16] A. Darzi, N. R. Provenza, L. A. Jeni, D. A. Borton, S. A. Sheth, W. K. Goodman, and J. F. Cohn, "Facial action units and head dynamics in longitudinal interviews reveal ocd and depression severity and dbs energy," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–6.

[17] G. Giannakakis, M. R. Koujan, A. Roussos, and K. Marias, "Automatic stress analysis from facial videos based on deep facial action units recognition," *Pattern Analysis and Applications*, pp. 1–15, 2021.

[18] F. Meawad, S.-Y. Yang, and F. L. Loy, "Automatic detection of pain from spontaneous facial expressions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 397–401.

[19] Z. Chen, R. Ansari, and D. Wilkie, "Automated pain detection from facial expressions using facs: A review," *arXiv preprint arXiv:1811.07988*, 2018.

[20] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 1815–1831, 2019.

[21] V. Kitanovski and E. Izquierdo, "3d tracking of facial features for augmented reality applications," in *WIAMIS 2011: 12th International Workshop on Image Analysis for Multimedia Interactive Services, Delft, The Netherlands, April 13-15, 2011*. TU Delft; EWI; MM; PRB, 2011.

[22] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," *arXiv preprint arXiv:1907.06724*, 2019.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.