Electronics Letters

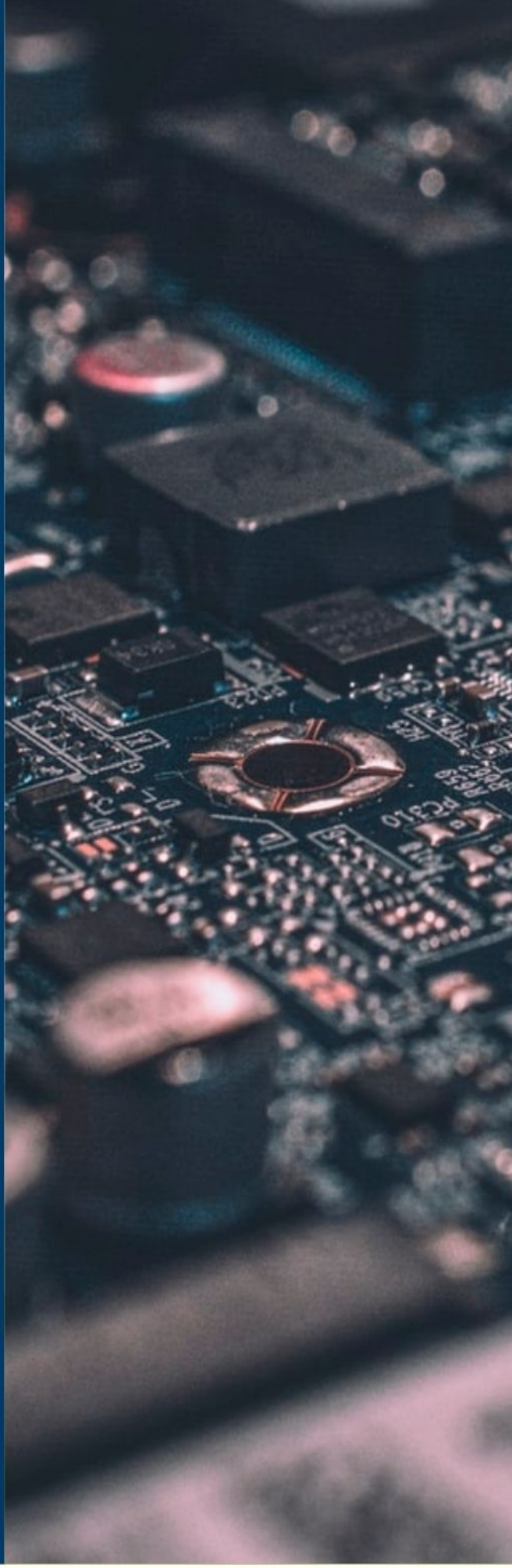
Special issue Call for Papers

Be Seen. Be Cited.
Submit your work to a new IET special issue

Connect with researchers and experts in your field and share knowledge.

Be part of the latest research trends, faster.

Read more



The Institution of Engineering and Technology

Check for updates

Depression recognition from facial videos: Preprocessing and scheduling choices hide the architectural contributions

Manuel Lage Cañellas,[1,✉] Constantino Álvarez Casado,[1] Le Nguyen,[1] and Miguel Bordallo López[1,2]

[1] *Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Oulu, Finland*

[2] *VTT Technical Research Centre of Finland, Oulu, Finland*

✉ E-mail: manuel.lage@oulu.fi

Deep learning models have been widely applied in video-based depression detection. It is observed that the diversity of preprocessing, data augmentation, and optimization techniques makes it difficult to fairly compare model architectures. In this study, the typical ResNet-50 model is enhanced by using specific face alignment methods, improved data augmentation, optimization, and scheduling techniques. The extensive experiments on two popular benchmark datasets (AVEC2013 and AVEC2014) obtained competitive results, compared to sophisticated spatio-temporal models for single streams. Moreover, the score-level fusion approach based on two texture streams outperformed the state-of-the-art methods. It achieved mean square errors of 5.82 and 5.50 on AVEC2013 and AVEC2014, respectively. These findings suggest that the preprocessing and training configurations result in noticeable improvements, which have been originally attributed to the network architectures.

Introduction: Depression is a common mental health disorder that negatively affects an individual's well-being [1]. Long-term medical depression can lead to severe complications, both at the psychological and physiological levels. Several studies suggest depression as a trigger for other diseases such as cardiovascular disease, osteoporosis, aging, pathological cognitive changes, Alzheimer's disease, and other dementias, and even an increase in the risk of early mortality [2].

Systems that automatically recognize depression are desirable because of their potential objectivity, speed, and reliability to avoid such an impact on a patient's health and well-being. In the last decade, many approaches based on classical statistical machine learning algorithms have been proposed to recognize signs of depression from facial videos, speech, and text data [3] to help physicians make decisions.

While the most novel architectures have shown noticeable improvements in the accuracy of depression recognition models, most previous work does not discuss or experiment with substantial components of the machine learning pipeline, such as preprocessing, face alignment, scheduling or optimization. Based on these shortcomings, in this article, we propose creating deep learning models for automatic depression screening using only static textural features extracted from facial video frames. In this context, we experiment with a set of changes that improve the results using this kind of architecture. Our main approach can be summarized as follows:

- We introduce a set of 2D-CNN models based on the ResNet-50 architecture [4] trained using only static textural information from video frames by applying two different face alignment techniques, and evaluate their impact in the final results.
- We explore novel training optimization and scheduling schemes to further improve the results of previous similar approaches that are based on spatial information.
- We propose the use of a fusion score approach to regress depression levels using different textural-based models that are shown to be complementary depending on the face alignment.
- We train and validate the models on the AVEC2013 [5] and AVEC2014 databases [6], and show how this simple approach can obtain comparable results to sophisticated spatio-temporal models, while the score-level fusion of both streams models outperforms state-of-the-art methods in the literature.

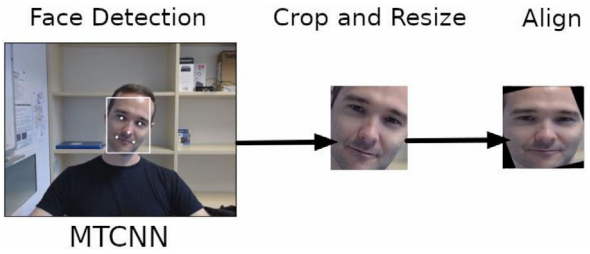


Fig. 1 *Incorrect alignment process. Cropping the facial region before the rotation deletes textural information within the facial boundaries, depicted as black triangles. Using a bigger bounding box will result in the need of a new MTCNN detection.*

Based on these findings, we conclude that relatively small changes in the pre-processing and training process could result in noticeable differences in the performance of the models, which could be hiding the contributions previously attributed to the differences in the neural network architectures.

Related work: In the last years, computer vision has been proposed as a powerful tool to diagnose clinical depression, as it shows promising performance in recognizing and analyzing facial expressions, a trait that is known to be deeply connected with depression. Several studies suggest that depression affects facial expressions in the following way: depressed people generally show a decrease in the intensity of positive emotions, a reduction of the number of smiles, an increase in the intensity of negative emotions, a reduction of the number of eye contacts, a reduction of the number of blinks, a reduction of the number of head turns and an increase in the number of head nods [7]. Many studies in the computer vision community have proposed automatic depression detection (ADD) methods based on facial expression analysis (FEA) in both static and video scenarios [8]. In particular, to recognize depressed expressions in static images, methods based on deep learning models have been proposed to extract embedding vectors from faces and classify images into depression or control classes [9]. Other studies propose a similar approach with static images, but focus on facial multi-regions instead of the entire face [10]. More recently, most studies have focused on exploiting the spatio-temporal information by leveraging the facial interframe information from videos. Some studies proposed 3D-CNN architectures to extract spatio-temporal features from short video clips, while others proposed temporal pooling techniques to capture and encode the dynamic information of video clips into an image map and train 2D-CNNs [11].

Proposed methodology: In this paper, we propose using preprocessed RGB images to learn discriminative representations to determine the level of depression of a person. In contrast to most of the works in the literature, our work focuses on techniques for preprocessing and data augmentation of the input data, and keeps the same backbone network, a pretrained ResNet-50 architecture.

Preprocessing and face alignment: The first step of our method consists in segmenting and aligning the facial regions of every video frame. We use a state-of-the-art face detector, a Multi-task Cascade Convolutional Network (MTCNN) [12] based in deep learning, which provides also face alignment by detecting five fiducial landmarks in the eyes, nose and mouth. As seen in the related literature, [11, 13], typical face preprocessing is based on cropping, aligning, and rescaling the images. However, the order of cropping and aligning faces has a direct impact on the resultant image. Cropping the facial region before rotation and alignment deletes textural regions on the border. This can be depicted as black triangles, as shown in Figure 1 discarded facial alignment. Using a bigger bounding box would still result in the need of a new MTCNN detection.

As our model is only based on texture information, we instead apply a preprocessing scheme that conserves all textural information within the facial boundaries. In this context, we align and rescale the images in two different manners, pose independent and pose dependent, as shown in Figure 2.

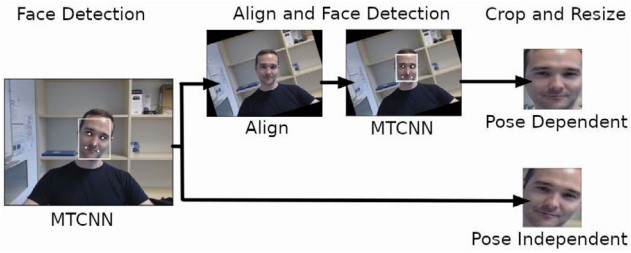


Fig. 2 *Two different face alignment processes. Pose dependent alignment detects the face and landmarks, rotates the image aligning the eyes horizontally and crops the face boundaries with a new face detection. Pose independent alignment detects the face and landmarks and crops the face boundaries directly.*

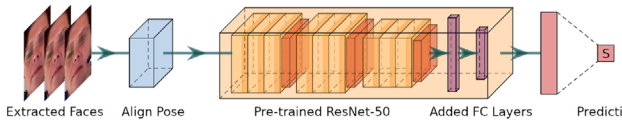


Fig. 3 *Proposed Architecture of a ResNet-50 architecture followed by two additional fully connected layers of 512 and 128 neurons.*

Pose dependent face alignment rotates the face based on the position of the eyes, which are horizontally aligned, subsequently applying a new face detection from where the aligned face is cropped and scaled, providing an input that varies the texture with the facial expression. Pose independent face alignment simply crops the face based only on the facial boundaries provided by MTCNN and then scales the image, providing an input that varies the texture with different head poses. These two types of alignments could give complementary information of the texture of the face.

Model architecture: We use a ResNet-50 architecture [4] as the backbone, followed by two additional fully connected (FC) layers of 512 neurons and a regression layer of 128 neurons to estimate the level of depression as shown in Figure 3. To take advantage of pre-learned features and transfer learning, we initialize the network with pre-trained weights based on the InsightFace framework [14]. In order to favor that new low-level textural features can be learnt from our specific data, we keep unfrozen all the layers of the architecture. The prediction S of each video is then calculated as the average of the predictions of all individual images.

Data augmentation: Since our proposed approach is based on the complementarity of information of both pose dependent and pose independent faces, we do not add any rotation of the faces for data augmentation. We propose instead random horizontal flips for each face, and changes in brightness, contrast and saturation. In contrast to other works in the literature, we do not include vertical flips nor add extra images to the dataset. We add these data augmentation techniques to our previous pose dependent and pose independent streams.

Training: On the AVEC2014 database, we process every frame of each video for a total of 304929 facial frames. As AVEC2013 database is nine times bigger than AVEC2014, in order to keep a similar amount of data, and similarly to previous works, we decide to process 1 frame every 9. The cost of this regression model is calculated as the L1 norm, the mean average error function (MAE) that aims at minimizing the error for the predictions of each individual frame. In contrast with other previous publications, we include a state-of-the-art optimizer, RAdam [15], enhanced with an implementation of Lookahead optimization [16]. We use a ReduceLROnPlateau optimizer that reduces our initial learning rate of 3.0×10^{-4} [17] when learning has stopped improving.

Experimental analysis and results:

Datasets: The performance of our proposed method has been evaluated on two publicly available databases: Audio/Visual Emotion Challenge

Table 1. Experimental results for AVEC2013

Streams	MAE	RMSE
Pose independent	6.16	8.99
Pose dependent	6.02	**8.23**
Fused	**5.82**	8.41

Table 2. Experimental results for AVEC2014. Comparison for Separated and Joint tasks

	Separated tasks		Joint tasks	
Streams	MAE	RMSE	MAE	RMSE
Pose independent	6.00	8.08	5.94	7.88
Pose dependent	5.61	7.52	5.59	7.34
Fused	**5.53**	7.42	**5.50**	7.29

AVEC2013 [5] and AVEC2014 [6] depression sub-challenge datasets. Both are derived from a subset of the audio-visual depressive language corpus (AViD-Corpus). AVEC2013 is constituted of one task while AVEC2014 is distributed into two different subsets: Northwind and Freeform, where the same subject performs two tasks. Each task is segmented into three partitions: training, development, and test, each of them with 50 videos. Every video contains a record of a subject speaking in front of a camera. The videos are labeled with a Beck-Depression Inventory BD-II [1] score indicating a depression level that ranges between 0 and 63. According to the BD-II score, the severity of depression can be classified into four levels: minimal (0-13), mild (14-19), moderate (20-28), and severe (29-63).

Experimental setup: The proposed methodologies are evaluated using only the static texture features extracted from both benchmark dataset videos. The results across different models are compared against other state-of-the-art models based on visual information, including static and spatio-temporal models. We provide results for both for individual streams and for their complementary score-level fusion. The experiments are performed using a computer that integrates an AMD Ryzen 7 5800 8-Core processor and an NVIDIA GeForce RTX 3060 running on Linux. We used Python 3.8 as the programming language with PyTorch 1.11.0 framework.

Protocol and performance metrics: To evaluate the performance of these models and make a fair comparison with the state-of-the-art methods, we provide the two most common metrics in the automatic depression assessment literature, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The overall predicted depression score for each input video is obtained by averaging the estimation scores for all its frames.

Experimental results: The performance and validity of the proposed modality is evaluated through a series of experiments in the benchmark databases. Table 1 shows the results of both pose dependent and pose independent streams and the fusion of them, for AVEC2013. In Table 2, we provide the results for AVEC2014, with two different ways for measuring the performance: separated and joint tasks. Separated task evaluation considers each video as an independent task. MAE is calculated over the mean of 100 predictions, one per each video. Joint task evaluation fuses each Northwind and Freeform video done by the same subject by averaging the predictions of both tasks. The MAE is then obtained as the mean over the 50 predictions corresponding to each subject. For the two tables, it can be seen that the average error, in terms of MAE, the error is smaller when streams are fused. These results show that the information contained in pose-dependent and pose-independent streams might be complementary. Hence, the exploration of different preprocessed spatial streams is an effective way to utilize the static information.

Error distribution: To further analyze the performance of our best model we show the error distribution in AVEC2014 benchmark, and

1350911x, 2023, 20, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ell2.12992 by University Of Oulu, Wiley Online Library on [28/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

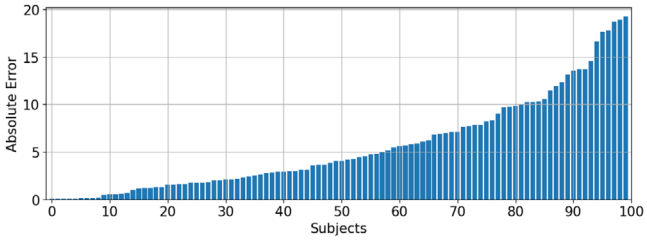


Fig. 4 *Absolute error distribution for each video in AVEC2014.*

Table 3. Comparison of methods for depression detection on AVEC2013 dataset

Methods	MAE	RMSE
ResNet-50+Pool (Zhou et al. [18])	6.37	8.43
MTB-DFE (Xu et al. [19])	6.31	8.20
Four DCNN (Zhou et al. [10])	6.20	8.28
MDN-100 (Melo et al. [13])	6.14	7.62
MTB-DFE + SEG (Xu et al. [19])	6.05	7.92
2xResNet-50 (Melo et al. [11])	5.96	7.97
MTB-DFE + SPG (Xu et al. [19])	5.95	7.57
Fusion of texture streams (ours)	**5.82**	8.41

Table 4. Comparison of methods for depression detection on AVEC2014 dataset

Methods	MAE	RMSE
ResNet-50+Pool (Zhou et al. [18])	6.37	8.43
MTB-DFE (Xu et al. [19])	6.30	7.83
MTB-DFE + SPG (Xu et al. [19])	6.24	7.65
Four DCNN (Zhou et al. [10])	6.21	8.39
2xResNet-50 (Melo et al. [11])	6.20	7.94
MDN-152 (Melo et al. [13])	6.06	7.65
MTB-DFE + SPG (Xu et al. [19])	5.86	7.18
Fusion of texture streams (ours)	**5.50**	7.29

show it in Figure 4. The figure shows the absolute error for each of the 100 test videos ordered from the smallest to the largest. We can observe that the error distributions of our approach shows a relatively small overall error. Our textural model shows a higher portion of videos with a very small error (that would not result in a misclassification diagnosis). However, a small portion of the videos, still show significantly high errors.

Comparison with state-of-the-art: Tables 3 and 4, show the evaluation of the performance of the proposed approach using different alignment and the improved pretraining, optimization and scheduling techniques applied to facial videos for both AVEC2013 and AVEC2014. We make the comparison with the result that we obtained by fusing the streams. The comparison shows how using only textural information and a well-known deep learning architecture, it is possible to obtain state-of-the-art results comparable with the most sophisticated novel architectures that exploit spatio-temporal information.

In order to show that our approach is usable for other recent architectures, we train both Vision Transformer (ViT) [20] and Regularized Network (RegNet) [21] models. Table 5 shows the results without preprocessing (Raw), with preprocessing and the stream fusion. Although these complementary architectures do not improve our main results using ResNet, the experiments support our claim that fused and preprocessed streams show generally better results.

Although our results are based only on static information, they obtain the best results in terms of MAE and very competitive results

Table 5. MAE comparison for SOTA architectures in AVEC2014.

Methods	Raw	Pose Independent	Pose Dependent	Fused
ViT	7.76	7.56	7.33	**6.89**
RegNet	7.80	6.71	6.87	**6.35**

in terms of RMSE. However, since we use a very simple architecture, it is possible to argue that similar preprocessing and optimization techniques applied to more sophisticated spatio-temporal models would result in an even better performance. Hence, we argue that more studies with standardized preprocessing and optimization are needed to fairly compare the performance across different neural network architectures.

Conclusion: This paper experimented with a simple depression detection approach that leverages two complementary face alignment techniques to derive a set of deep models based on textural static information. The training process of the simple models was enhanced using data augmentation and recent optimization and scheduling schemes. Extensive experiments on the AVEC2013 and AVEC2014 benchmark datasets show that individual models trained this way have comparable results to sophisticated spatio-temporal models, while the score-level fusion of several streams outperforms more sophisticated state-of-the-art methods based on visual information. We argue that, at least for AVEC2013 and AVEC2014, the impact of using novel architectures in depression estimation from videos, can not be clearly distinguished from the contributions due to other processing components, and should be investigated in a more systematic manner.

Author contributions: Manuel Lage Cañellas: Conceptualization, data curation, formal analysis, investigation, software, validation, visualization, writing - original draft. Constantino Alvarez Casado: Investigation, validation, writing - original draft, writing - review and editing. Constantino Alvarez Le Nguyen: Supervision, validation, writing - review and editing. Miguel Bordallo López: Formal analysis, methodology, supervision, validation, writing - review and editing.

Conflict of interest statement: The authors declare no conflict of interest.

Data availability statement: Data sharing not applicable no new data generated.

© 2023 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
Received: *19 July 2023* Accepted: *13 October 2023*
doi: 10.1049/ell2.12992

References

1 Beck, A.T., Steer, R.A., Brown, G.K.: *BDI-II: Beck depression inventory*. Pearson, London (1996)
2 Verhoeven, J., et al.: Major depressive disorder and accelerated cellular aging: Results from a large psychiatric cohort study. *Mol. Psychiatry* **19**(8), 151 (2013)
3 Wu, P., et al.: Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Trans. Intell. Technol.* **8**(3), 701–711 (2023)
4 He, K., et al.: Deep residual learning for image recognition. In: CVPR 2016. IEEE, Piscataway (2016)
5 Valstar, M., et al.: Avec 2013: The continuous audio/visual emotion and depression recognition challenge. In: Proc. of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, (2013)
6 Valstar, M., et al.: Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. ACM, New York (2014)

1350911x, 2023, 20, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ell2.12992 by University Of Oulu, Wiley Online Library on [28/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

7 Fiquer, J., et al.: What is the nonverbal communication of depression? assessing expressive differences between depressive patients and healthy volunteers during clinical interviews. *J. Affect. Disord.* **238**, 636–644 (2018)
8 Yadav, U., Sharma, A.K.: Review on automated depression detection from audio visual clue using sentiment analysis. In: 2nd International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, Piscataway (2021)
9 de Melo, W.C., Granger, E., Hadid, A.: Depression detection based on deep distribution learning. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE, Piscataway (2019)
10 Zhou, X., et al.: Visually interpretable representation learning for depression recognition from facial images. *IEEE Trans. Affective Comput*. **11**(3), 542–552 (2020)
11 de Melo, W., Granger, E., Bordallo Lopez, M.: Encoding temporal information for automatic depression recognition from facial analysis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Piscataway (2020)
12 Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process Lett*. **23**(10), 1499–1503 (2016)
13 de Melo, W.C., Granger, E., Bordallo Lopez, M.: Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Trans. Affective Comput.* **14**(1), 578–590 (2023)
14 Guo, J., et al.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021)
15 Liu, L., et al.: On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations. ICML, San Diego (2020)
16 Zhang, M., et al.: Lookahead optimizer: k steps forward, 1 step back. In: Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA (2019)
17 Bermant, P.C.: Biocppnet: automatic bioacoustic source separation with deep neural networks. *Sci. Rep.* **11**, 23502 (2021)
18 Zhou, X., et al.: Learning content-adaptive feature pooling for facial depression recognition in videos. *Electron. Lett*. **55**(11), 648–650 (2019)
19 Xu, J., et al.: Two-stage temporal modelling framework for video-based depression recognition using graph representation. CoRR (2021)
20 Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint, arXiv:2010.11929 (2020)
21 Radosavovic, I., et al.: Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10428–10436. IEEE, Piscataway (2020)

1350911x, 2023, 20, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ell2.12992 by University Of Oulu, Wiley Online Library on [28/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License