Causal affect prediction model using a past facial image sequence

Geesung Oh
Graduate School of Automotive Engineering
Kookmin University, Seoul, Korea
gsethan17@kookmin.ac.kr

Euiseok Jeong
Graduate School of Automotive Engineering
Kookmin University, Seoul, Korea
euiseok_jeong@kookmin.ac.kr

Sejoon Lim
Department of Automobile and IT Convergence
Kookmin University, Seoul, Korea
lim@kookmin.ac.kr

Abstract

Among human affective behavior research, facial expression recognition research is improving in performance along with the development of deep learning. For improved performance, not only past images but also future images should be used along with corresponding facial images, but there are obstacles to the application of this technique to real-time environments. In this paper, we propose the causal affect prediction network (CAPNet), which uses only past facial images to predict corresponding affective valence and arousal. We train CAPNet to learn causal inference between past images and corresponding affective valence and arousal through supervised learning by pairing the sequence of past images with the current label using the Aff-Wild2 dataset. We show through experiments that the well-trained CAPNet outperforms the baseline of the second challenge of the Affective Behavior Analysis in-the-wild (ABAW2) Competition by predicting affective valence and arousal only with past facial images one-third of a second earlier. Therefore, in real-time application, CAPNet can reliably predict affective valence and arousal only with past data.

The code is publicly available.[1]

1. Introduction

Human affective behavior research is essential for the human-computer interaction (HCI) field and has been established for a long time. However, there are many obstacles for HCI systems used in real-world applications, such as in-the-wild or real-time tasks. To address these problems, Kollias et al. have been hosting the Affective Behavior Analysis in-the-wild (ABAW) Competition, which involves a variety of research activities, for two years [5, 6, 7, 8, 9, 11, 12, 30]. Most of the top-ranked teams in the competition of ABAW [5, 6] used convolutional neural networks (CNNs) with single facial images or sequences of such images. In cases where a single image was used, the captured image was inputted to be recognized, and even for teams that used image sequences, past or future images were used along with the image captured at that point [1, 14, 16]. Although these methods perform well with large-scale data in the wild, they encounter limitations when used in real time. Except in cases such as posterior analysis with recorded video, most real-world applications require future decisions only with past data in real time, such as facial expression recognition for interview assistant systems [21] or driver monitoring systems [19].

Hence, we propose a deep learning based network called causal affect prediction network (CAPNet), which enables affect prediction with only past data. CAPNet outputs the predicted affective state in the form of valence and arousal, which are the most popular continuous emotional representations proposed by Russell [22]. The input data require only a sequence of past facial images, not current or future ones. CAPNet consists of a modular architecture divided into a feature extractor and a causality extractor, which allows one to learn causality well from past facial images. The feature extractor is based on the facial expression recognition (FER) model proposed by Oh et al. [19], which was pretrained using the AffectNet dataset [17]. We fine-tuned this FER model with a pair of single images and corresponding labels of the Aff-Wild2 dataset [10] and used the feature extractor of the FER model as the feature extractor of CAPNet. The causality extractor consists of

[1] https://github.com/gsethan17/CAPNet_ABAW2021

long short-term memory (LSTM) and fully connected (FC) layers behind the feature extractor. We applied causal inference learning to perform supervised learning by pairing the sequence of past images with the current label using the Aff-Wild2 dataset [10].

The proposed network, CAPNet, can predict the affective state from only past facial images and was evaluated using the test datatset of Aff-Wild2 [10]. Announced results from the second challenge of ABAW (ABAW2) [5] showed that CAPNet outperforms baseline performance without current or future images. The proposed network achieved lower-quality than top-ranked scores but the predictions can be generated without having to wait for the current image. Nonetheless, CAPNet is the only one ranked on the leaderboard of ABAW2 using past data only. A considerable advantage of CAPNet is that it can be applied in real time by outputting reliable affective state one-third of a second ahead.

The primary contribution of this paper is as follows:

- We propose CAPNet, which uses the order of the input data and the causal structure. CAPNet reliably predicts the affective state early with only past facial images through causal inference learning between the past images current affective state.

The rest of this paper is organized as follows. Section 2 introduces the related work of the human affective behavior research and causality learning. Section 3 discusses the details of proposed CAPNet for predicting the human affective state with only past facial images. Section 4 presents the details of implementation and analyzes the evaluation results announced by ABAW2. Section 5 concludes this work and describes further work.

2. Related work

2.1. ABAW

Human affective behavior research has been established for a long time and recently produced promising results with developments in deep learning. A contribution to this achievement is the ABAW Competition, hosted by Kollias et al. [6, 5]. ABAW consists of three challenges on the same dataset, Aff-Wild2 [10]: dimensional affect recognition (in terms of valence and arousal), categorical affect classification (in terms of the seven basic emotions), and 12 facial action unit detection. Most of the top-ranked teams in the first challenge of the Affective Behavior Analysis in-the-wild (ABAW1) Competition, which was held in conjunction with FG2020, proposed deep learning based multitask models that output the three challenges at once [1, 14]. For the input data, the corresponding image is basically used, and additional (previous or post) images are used to further leverage temporal information [1, 14, 16]. Kuhnke et al. were ranked a runner-up in the valence-arousal recognition challenge using not only visual data but also audio data as input [14]. Notably, the top-ranked model in the same challenge used an additional dataset, AFEW-VA [13], along with several methodologies to increase performance. Through ABAW1, human affective behavior recognition performance has been enhanced for in-the-wild environments. However, there are no results considering real-time environments.

2.2. Causality learning

Causal learning has been generally overlooked because of the success of machine learning and deep learning. Nonetheless Schölkopf et al. stated that causal inference can improve machine learning and deep learning [23]. They said that most of the current success of data-driven models is only the result of large-scale pattern recognition in formalized large data. Hence, causal inference learning is required to solve the problem with generalization outside formalized data. The use of structural causal models, such as modular architectures may enable the development of causality machine learning or deep learning models [23]. Entangled recurrent neural networks (E-RNN), proposed by Yoon et al. [29], are structural causal models.

Causality learning can also be achieved through sequential input data. It is mainly utilized in generative models such as PixelCNN [25] and WaveNet [20]. PixelCNN [25] first defines the order of each pixel in the image. The pixel values are inputted in the defined order, and PixelCNN generates the next pixel. WaveNet [20] deals with audio data that are made over time. Therefore, the audio data are inputted with a time sequence, and WaveNet generates the next audio data. In these processes, the causality between the previous data and the next data is learned.

Causal inference is also being considered in recent facial expression studies [18, 24]. Therefore, causal learning should also be considered in deep learning based facial expression recognition studies.

3. Methodologies

We propose CAPNet, which predicts human affective state indicators with valence and arousal values in past facial images. Through supervised learning using chronologically constructed facial images and corresponding affective labels, CAPNet learns the causal inference between past facial expressions and the current affective state. Multiple past images, past images are created in a single sequence and used as input data. Additional details about the input sequence are presented in Section 3.1. Each input image is first extracted into a feature vector through a feature extractor, which is imported from the driver's real emotion recognizer (DRER) [19]. Details about the feature extractor are discussed in Section 3.2. The extracted feature vectors are sequentially fed into a causality extractor, which inte-

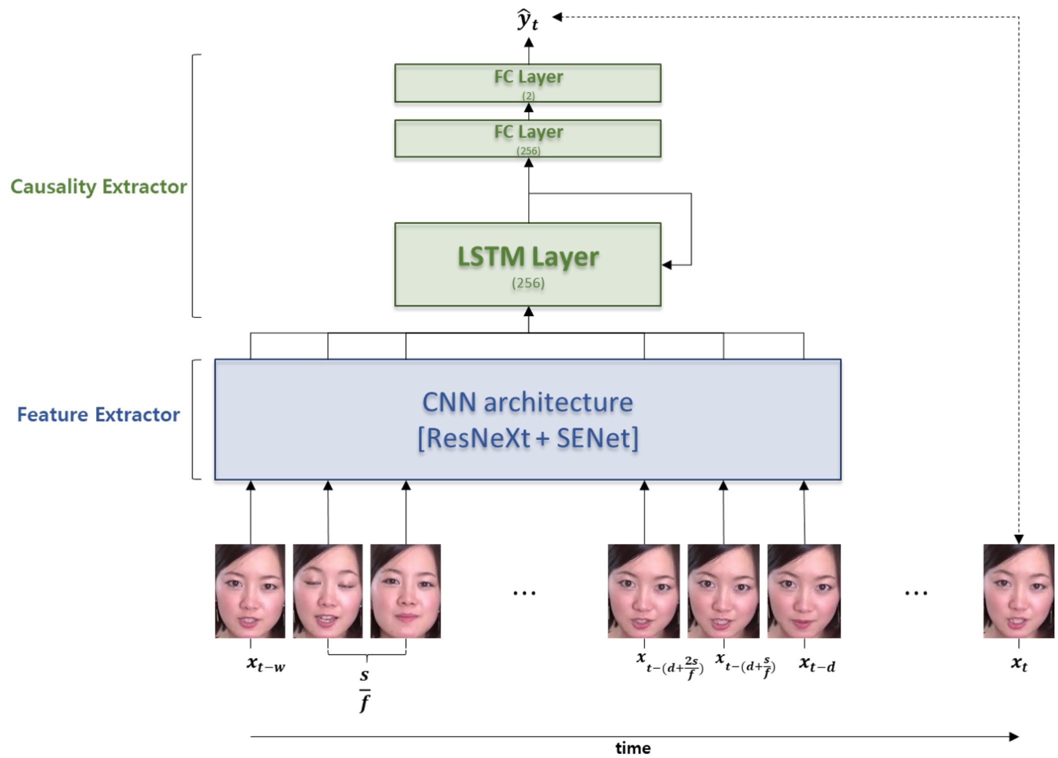

Figure 1. Overall pipeline of CAPNet–the network predicts the current affective state in terms of valence and arousal, $\hat{y}_t$, with only past facial images.

grates causal inference from the feature vectors. The way by which the causality extractor integrates the causal information is described in Section 3.3. Figure 1 shows the overall pipeline of CAPNet.

3.1. Input sequence

Ordering of past images, proposed by Oord et el. [20, 25], can be conducted to learn causal inference from past facial images. Even though the order of past images should be determined, most of the images we use are made over time; thus, we only need to organize the order of the input data according to an intrinsic sequence of time steps. Therefore, causal inference is trained on the basis of the sequence of chronological past images and the corresponding labels. CAPNet trains the distribution of valence and arousal over the past facial images of humans under the following conditional distribution:

$$p(y_t) = p(y_t | x_{t-(d+\frac{n}{f})}, \; n = f \times (w-d), \ldots, 2s, s, 0) \quad (1)$$

where y_t is the valence and arousal values of the human at time t, x_t is the facial expression image of the person captured at time t, d is the time duration for prediction, f is the number of images captured per second (frame rate), w is the window size (in seconds), and s is the number of stride images. To predict y_t, CAPNet needs facial images at intervals of s from time $t-w$ to $t-d$.

3.2. Feature extractor

The CNN architecture we use is based on the FER model of DRER [19], which uses ResNeXt [28] and SENet [3] to recognize driver valence and arousal from a single image of the driver's face during driving. The structure of the FER model consists of a feature extractor, which has a CNN structure, and an FC layer, as shown in Figure 2. Using the FER model as it is, we can output the valence and arousal of the subject by inputting a single image. Furthermore, the top layer of the FER model (FC layer) can be removed, and the causality extractor can be connected to process a sequence of facial images, as shown in Figure 1.

3.3. Causality extractor

Following the proposal of Schölkopf et al. [23], we build the causality extractor as a modular structure so that CAPNet becomes a structural causal model. Since the causality extractor exists independently of the feature extractor, it can be trained individually for the causal inference. The causality extractor consists of one LSTM layer and two FC layers. The LSTM layer lies between the feature extractor and FC

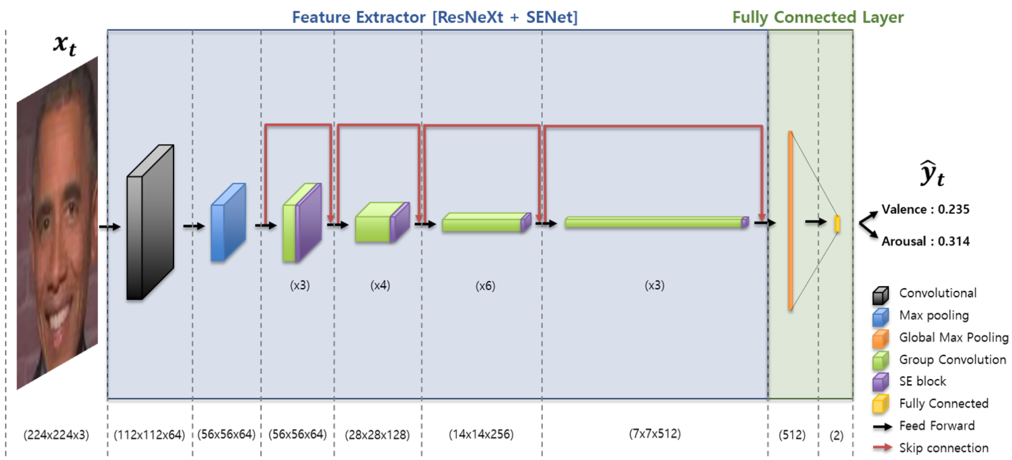

Figure 2. Overall pipeline of the FER model of DRER [19]–the FER model recognizes the current affective state in terms of valence and arousal, $\hat{y}_t$, with a single (current) facial image. The feature extractor of the FER model is utilized for CAPNet as is.

layers, and it is based on the LSTM network, proposed by Hochreiter and Schmidhuber [2]. The LSTM layer outputs a single hidden state for the input sequence of feature vectors generated by the sequence of input images. During the integration of sequential data into the single hidden state, the LSTM layer learn the causal inference between the past facial images and the affective state. The FC layers eventually convert the single hidden state to the predicted affective state. The last FC layer, which has two units and a tanh activation function, outputs a continuous two-dimensional value between -1 and 1 representing valence and arousal, respectively.

4. Experiments

We studied the proposed model, CAPNet, which consists of feature extractor and causality extractor. The backbone of feature extractor is the FER model [19] pretrained with AffectNet [17], one of the largest datasets composed of single images. We discarded the top layer of the FER model, then put on an LSTM and FC layers as the causality extractor. CAPNet configured in above step was trained using the Aff-Wild2 dataset [10]. Proposed CAPNet was evaluated by ABAW2 [5] because the label for test sets of Aff-Wild2 [10] is not disclosed. We submitted prediction results by CAPNet for the test sets of Aff-Wild2 [10] and the results announced from ABAW2 [5] are described in Section 4.3.

4.1. Data preprocessing

For training, we used cropped facial images, provided by ABAW2, and the valence-arousal label of each image was provided with 564 videos on the Aff-Wild2 dataset [10]. Except when the cropped images contained detection errors, we used the training set consisted of 1,573,844 images and the validation set of 334,865 images. These data pairs can be represented by Equation 2, where d and n are 0.

For CAPNet training, the image sequence is paired with sampling images and labeled as follows:

$$([x_{t-(d+\frac{n}{f})}], y_t) \tag{2}$$

where the set of x contains the cropped past facial images when n is intervals of s from $f \times (w-d)$ to 0 (i.e. $n = f \times (w-d), \ldots, 2s, s, 0$), y_t is the valence-arousal label at time t, d is the time duration for prediction, f is the video frame rate, w is the window size (in seconds), and s is the number of stride images. If the cropped facial image does not exist at the $t-(d+\frac{n}{f})$ point, the image will be sampled while navigating sequentially up to the $t-(d+\frac{n+s}{f})+1$ point except n is $f \times (w-d)$. All the videos provided by Aff-Wild2 [10] have a frame rate of 30 fps; for our experiments, we set d to one-third of a second, w to 3 s, s to 10 images. In this case, except when the value of x is insufficient or y_t is invalid, we used 1,512,258 training sets and 322,643 validation sets for CAPNet training.

4.2. Implementation

We trained CAPNet with the following parameters: the input image size was $224 \times 224 \times 3$; the mini-batch size was 128, and the optimizer we used was Adam [4], with a 0.00001 learning rate. Training was conducted repeatedly and was terminated if there was no improvement of the validation metric in next four consecutive epochs. The metric used was the concordance correlation coefficient (CCC) [15] value, same as that proposed in the ABAW2 competition [5] and the equation is as follows:

$$CCC = \frac{2k_{\hat{y}y}}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{3}$$

Model	window size (seconds)	CCC		
		Valence	Arousal	Mean
CAPNet	1	0.476	0.478	0.477
	2	0.498	**0.486**	0.492
	3	**0.510**	0.483	**0.497**

Table 1. Evaluation results using validation sets of Aff-Wild2.

Model	window size (seconds)	CCC			Remark
		Valence	Arousal	Mean	
CAPNet	1	0.2416	0.2323	0.2370	ours
	2	**0.2680**	**0.2561**	**0.2621**	
	3	0.2607	0.2502	0.2555	
VGG-FACE [5]	-	0.23	0.21	0.22	baseline of ABAW2

Table 2. Evaluation results using test sets of Aff-Wild2 with baseline of ABAW2 challenge [5].

where $\hat{y}$ and y are the model output and label, respectively; μ is each mean value; σ^2 is each variance; and k is the corresponding covariance value. The average of each CCC value (CCC_{mean}) for valence and arousal was used as the overall metric. For training, the loss function for backpropagation was set to $1 - CCC_{mean}$. For CAPNet training, additional data-related parameters were needed. The time duration for prediction (d) was wet to one-third of a second; the number of stride images (s) was set to 10 images; and the window size (w) was set to 1, 2, and 3 s. Additionally, we applied dropout to drop the inputs of the LSTM layer and the first FC layer at a rate of 0.2.

4.3. Results

Table 1 shows the results of our methods on the validation set. On the validation set, it performed well overall regardless of window size. The CCC values of valence and arousal were not biased. The CCC values of the valence and arousal showed the highest values when the window size was 3 s and 2 s, respectively. Mean CCC values of the valence and arousal showed the highest value when the window size was 3 s.

Table 2 presents the results of our methods on the ABAW2 challenge with the baseline [5]. First, all our proposed models achieved superior CCC values compared with the baseline and were listed on the leaderboard of ABAW2 challenge [5]. Moreover, the CCC values of valence and arousal were still not biased on the test set. The difference from the results evaluated in the validation set is that both of valence and arousal CCC values were highest when the window size was 2 s.

We compared the performance of our proposed models with other state-of-the-art models ranked on the leaderboard of ABAW2 competition[2]. Table 3 shows the comparison results. Our proposed CAPNet is a single task model that outputs only valence and arousal, while other state-of-the-art models are all multi task models that outputs emotion category or action unit with valence and arousal [26, 27, 31]. Netease Fuxi Virtual Human [31] achieves high performance by streaming hierarchical relationships between three emotional expressions, Morphoboid [26] uses the knowledge distillation technique. STAR [27] uses audio input data as well as the knowledge distillation technique. There are many differences between the state-of-the-art models and CAPNet, but we focused on input images. CAPNet uses only past images while other state-of-the-art models use either a single current image [31] or an image sequence containing the current image [26, 27] as input image. CAPNet is the only one listed on the leaderboard of ABAW2 using past data only.

5. Conclusions

In this paper, we propose CAPNet, which outperforms the baseline (ABAW2 results) and is ranked the leaderboard of ABAW2 challenge [5]. CAPNet can predict current or future affective states with only a sequence of past facial images. Although CAPNet achieves lower-quality than the state-of-the-art models, the state-of-the-art models still need the current image, which hinders its use in real-time application. In other words, CAPNet can predict at the same time the event occurs or before the event occurs, even when it is applied to systems where facial images are inputted in real time. The evaluation results shows that CAPNet has sufficient causality learning to predict corresponding valence and arousal from past facial images. Hence, CAPNet predicts reliable affective state one-third of a second ahead.

Although the performance of CAPNet outperforms the baseline, CAPNet is slightly poorer than the state-of-the-art

[2]https://ibug.doc.ic.ac.uk/resources/iccv-2021-2nd-abaw/

Model	Input image			CCC			Place of ABAW2 [5]
	Type	Past	Current	Valence	Arousal	Mean	
Netease Fuxi Virtual Human [31]	single		✓	0.4859	0.4950	0.4905	2^{nd}
Morphoboid [26]	sequence	✓	✓	0.5051	0.4747	0.4899	3^{rd}
STAR [27]	sequence	✓	✓	0.4784	0.4975	0.4876	4^{th}
CAPNet [ours]	sequence	✓		0.2680	0.2561	0.2621	10^{th}

Table 3. Evaluation results using test sets of Aff-Wild2 with results of state-of-the-art models.

models. Thus, filling this gap is one of our future works. We are also considering the multi task model, the knowledge distillation technique, and use of past audio data. Additionally, we plan to apply CAPNet to real-time environments.

6. Acknowledgments

This research was supported by the Knowledge Service Industry Core Technology Development Program funded by the Ministry of Trade, Industry, and Energy of Korea (No. 20003519), the AI-based Autonomous Driving Computing Module Development and Service Demonstration Program funded by the Ministry of Trade, Industry, and Energy of Korea (No. 20005673), and the BK21 Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 5199990814084).

References

[1] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[6] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.

[7] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[8] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[9] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019.

[10] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[11] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[12] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[13] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[14] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020.

[15] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[16] I Li et al. Technical report for valence-arousal estimation on affwild2 dataset. *arXiv preprint arXiv:2105.01502*, 2021.

[17] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[18] Lea Müller, Maha Shadaydeh, Martin Thümmel, Thomas Kessler, Dana Schneider, and Joachim Denzler. Causal inference in nonverbal dyadic communication with relevant interval selection and granger causality. *arXiv preprint arXiv:1810.12171*, 2018.

[19] Geesung Oh, Junghwan Ryu, Euiseok Jeong, Ji Hyun Yang, Sungwook Hwang, Sangho Lee, and Sejoon Lim. Drer: Deep learning–based driver's real emotion recognizer. *Sensors*, 21(6):2166, 2021.

[20] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[21] Kalyani Pampattiwar, Subramanian Gopalakrishnan, Pranav Sainath, Somesh Banerjee, and Jayaraman Shankaran. Intelligent hiring with facial expression detection and report generation. *Journal of Applied Science and Computations*, 6(5):804–807, 2019.

[22] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[23] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[24] Maha Shadaydeh, Lea Müller, Dana Schneider, Martin Thümmel, Thomas Kessler, and Joachim Denzler. Analyzing the direction of emotional influence in nonverbal dyadic communication: A facial-expression study. *IEEE Access*, 9:73780–73790, 2021.

[25] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016.

[26] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition, 2021.

[27] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis, 2021.

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[29] Jinsung Yoon and Mihaela van der Schaar. E-rnn: Entangled recurrent neural networks for causal prediction. In *Proc. ICML Workshop Principled Approaches Deep Learn.*, pages 1–5, 2017.

[30] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017.

[31] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition, 2021.