Multi-scale Transformer-based Network for Emotion Recognition from Multi Physiological Signals

Tu Vu[†], Van Thong Huynh[†], Soo-Hyung Kim

Department of AI Convergence
Chonnam National University
Gwangju, South Korea
{tu369,vthuynh,shkim}@jnu.ac.kr

Abstract—This paper presents an efficient Multi-scale Transformer-based approach for the task of Emotion recognition from Physiological data, which has gained widespread attention in the research community due to the vast amount of information that can be extracted from these signals using modern sensors and machine learning techniques. Our approach involves applying a Multi-modal technique combined with scaling data to establish the relationship between internal body signals and human emotions. Additionally, we utilize Transformer and Gaussian Transformation techniques to improve signal encoding effectiveness and overall performance. Our model achieves decent results on the CASE dataset of the EPiC competition, with an RMSE score of 1.45.

Index Terms—Physiological signals, Deep Learning, Transformer, Multi-scale

I. Introduction

Recognizing emotions is a fundamental aspect of human communication, and the ability to accurately detect emotional states has significant impacts on a range of applications, from healthcare to human-computer interaction. Emotions are often reflected in physiological signals [1], facial [2], and speech [3]. Recently, the use of physiological signals for affective computing has gained considerable attention due to its potential to provide objective measures of emotional states in real-time [4].

Recently, there has been a growing interest in developing machine learning algorithms for affective computing using physiological signals [5]–[9]. These algorithms can be used to classify emotional states, predict changes in emotional states over time, or identify the specific features of physiological signals that are most informative for detecting emotional states. There has also been interested in developing wearable sensors that can capture physiological signals in real-world settings, such as in the workplace or in social situations [10].

The use of end-to-end deep learning architectures for physiological signals has the potential to simplify the development and deployment of an emotion recognition system [4]. By eliminating the need for preprocessing steps, these architectures can reduce the complexity and time required for system development, as well as improve the scalability and accuracy of the system. Moreover, end-to-end architectures can enable the development of systems that can process multiple physiological signals simultaneously, such as heart rate, respiration, and electrodermal activity, which can provide more comprehensive and accurate measures of emotional states.

Despite the potential benefits of end-to-end deep learning architectures for affective computing, there are still challenges that need to be addressed. One challenge is to develop architectures that can handle noisy and non-stationary physiological signals, which can be affected by movement artifacts, signal drift, and other sources of noise. Another challenge is to ensure that the learned features are interpretable and meaningful, which can help improve the transparency and explainability of the system.

In this paper, we propose an end-to-end Multi-scale architecture for continuous emotion regression with physiological signals. We evaluate the performance of the proposed architecture using EPiC 2023 competition benchmark, which contains data collected from experiments conducted in a laboratory setting.

II. Related Works

A. Continous emotion recognition from multimodal physiological signal

The utilization of physiological signals has been widely acknowledged as one of the most reliable data forms for affective science and affective computing. Despite the fact that individuals are capable of manipulating their physical signals such as facial expressions or speech, consciously controlling their internal state is quite a daunting task. Therefore, analysis of signals from the human body represents a dependable and robust approach for fully recognizing and comprehending an individual's emotional state [1], [11]. This reliability factor is especially crucial in medical applications, such as mental health treatment or mental illness diagnosis.

Recognizing affect from physiological data remains a significant challenge, not only during the data acquisition process but also in terms of emotion assessment. Laboratory-based research dominates the field of affective science due to the control it affords over experimental variables. Researchers can

carefully select and prepare emotional stimuli, and employ various sensor devices to trace and record a subject's emotional state with minimal unexpected event, interference [4]. However, most of these studies rely on discrete, indirect methods such as quizzes, surveys, or discrete emotion categories for emotion assessment, which overlook the time-varying nature of human emotional experience. Sharma et al. [12] introduced a novel Joystick-based Emotion Reporting Interface (JERI) to overcome a limitation in emotion assessment. JERI enables the simultaneous annotation of valence and arousal, allowing for moment-to-moment emotion assessment. The Continuously Annotated Signals of Emotion (CASE) dataset, acquired using JERI, provides additional information to researchers for identifying the timing of emotional triggers. Leveraging this dataset, the Emotion Physiology and Experience Collaboration (EPiC) Workshop organized a competition to model moment-to-moment, continuous ratings of valence and arousal from physiological signals. This competition serves as a valuable platform for evaluating the precision and temporal dynamics of emotion recognition models. Accordingly, we adopt the EPiC competition benchmark as our primary benchmark for evaluating our proposed emotion recognition model.

Besides, it is claimed that a single physiological signal are relatively difficult to precisely reflect human's emotional changes. Hence, recently, there have been many researches focusing on detecting human's emotion on Multimodal Physiological Signal. There are many type of physiological signal used in these studies. While some study record heart related signals such as electrocardiographic (ECG) [13]–[15], blood volume pulse (BVP) [16], [17], others use electrical activity of the brain (Electroencephalogram/EEG) [18], [19] or muscle's electrical reaction (Electromyogram/EMG) [20], [21]. Furthermore, somes even employ skin temperature (SKT) [21], skin sweat glands (EDA) [18], [20] and the depth and rate of breathing (Respiratory/RSP) [20]. Based on signal modal provide by EPiC competition, our method utilized 8 different type of signals including ECG, BVP, EMG_CORU (corrugator supercilii), EMG_TRAP (trapezius), EMG_ZYGO (zygomaticus major), GRS, RSP and SKT.

B. Transformer based method for in multimodal emotion recognition from physiological signal

Similar to other emotion recognition problems that involve physical signals, Affective Computing in physiological data has witnessed extensive adoption of Machine Learning techniques, particularly Deep Learning methodologies. In their work, Dominguez et al [5] employed various conventional Machine Learning techniques, including Gaussian naive Bayes, k-Nearest Neighbours, and Support Vector Machines, for Valence-Arousal estimation. However, these approaches are heavily dependent on the quality of handcrafted feature selection and feature extraction processes. To overcome this challenge, other studies [6]–[8] proposed the use of Deep Learning techniques for an end-to-end approach, where the model learns to extract features automatically without the need for pre-designed feature descriptors.

With the advancement of Deep Learning, various state-of-the-art techniques have been employed for Affective Analysis from Physiological signals. Santamaria et al. [7] used Convolutional Neural Networks (CNNs) with 1D Convolution layers for Emotion Detection, while Harper et al. [8] combined CNNs with frequently used Recurrent Neural Networks (RNNs) for Emotion Recognition from ECG signals. Since their introduction in 2016, Transformers [22] have emerged as the preferred models in the field of Deep Learning. Their robust performance in Natural Language Processing, a type of data that shares some similar characteristics with time-series data, has demonstrated the potential of Transformers when applied to time-series signals. As a result, recent research in the Time-Series domain has utilized Transformers as the core module in their model architecture [23]–[25]. For physiological signals, some studies have proposed using Transformers and their variants for detecting emotions [6], [26]–[28]. In the works of Vazquez et al. [6], [26], they focused on applying pretrained Transformers for Multimodal signal processing. However, this is still a very basic application of Transformer modules. Wu et al. [27] and Yang et al. [28] proposed using more advanced techniques of Transformer-based models, which are self-supervised and Convolution-augmented Transformer for Single and Multimodal signal processing. Although these studies have demonstrated the effectiveness of Transformers for physiological signals, they often feed the model with fixed original size signals, which may lead to the loss of global feature information. To address this issue, we propose a new Multi-scale Transformer-based architecture for Multimodal Emotion Recognition.

III. PROPOSED APPROACH

A. Problem definition

The Emotion Recognition in Multimodal Physiological Signal problem takes as input 8 physiological signals, namely ECG, BVP, EMG_CORU, EMG_TRAP, EMG_ZYGO, GRS, RSP and SKT, extracted from human subjects during emotion-inducing stimuli. This denoted as 8 sequence with L length. In Affective computing field, the Emotion Recognition problem objective varies relying on indicated the Emotional Models. In the scope of this study, following the use of SAM (Self-Assessment Manikin) [29] model of the CASE dataset, the problem objective is Valence-Arousal (V-A) estimated value. The V-A score consists of two continuous floating-point numbers ranging from 0.5 to 9.5. A value of 0.5 denotes the most negative valence or lowest arousal, 5 indicates neutral valence or arousal, and 9.5 indicates the most positive valence or highest arousal.

B. Methodology

We construct a new Multi-scale architecture for Valence-Arousal estimation from 8 Physiological signals. Our architecture comprises of two core modules: Feature encoding module and Multi-scale Fusion module. Overall architecture is showed in Figure 1.
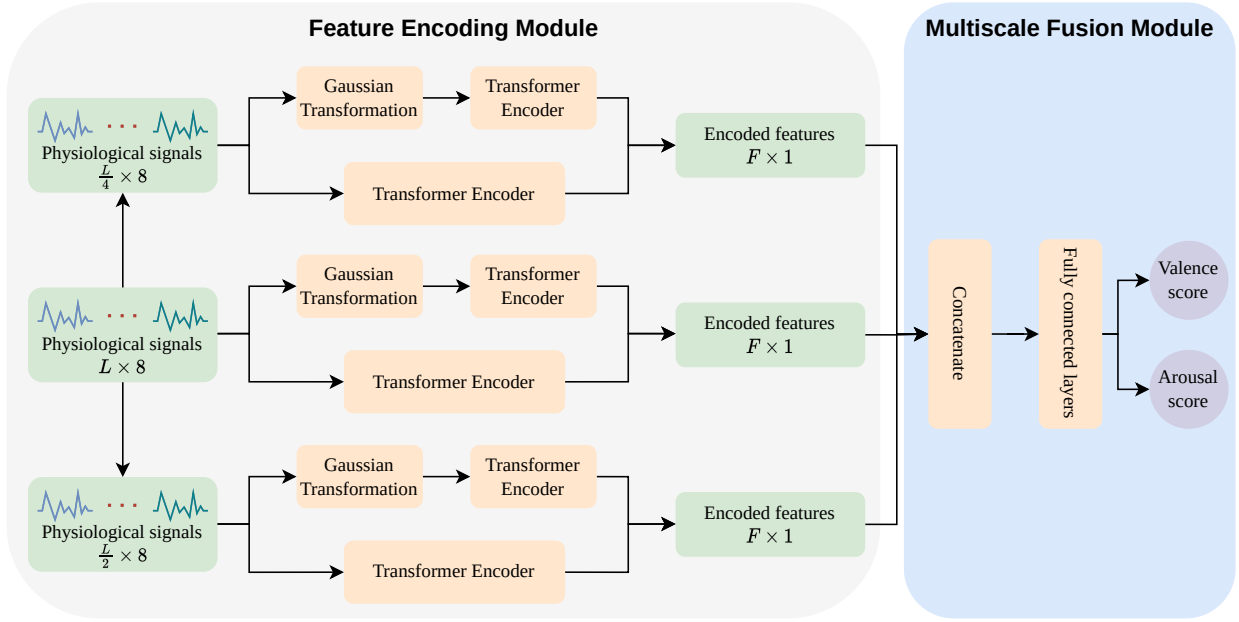
Fig. 1. An overview of our proposed architecture.

1) Feature encoding: To enable the feature encoding module to extract global features for the estimator and eliminate noise and interference information from the input, we employ 1-Dimensional Average Pooling to scale the 8 input signals into three different lengths: L, $L/2$, and $L/4$. This process helps to improve the model's ability to extract useful information and eliminate unwanted noise and interference.

Then, we simultaneously apply two kind of feature encoder which is Gaussian Transform [30] and Transformer Encoder [22]. The Transformer Encoder block is used Multi-Headed Self-Attention as it core mechanism. Given an input sequential signal $S \in R^{L \times C}$, where L represents the length of the signal sequence and $C = 8$ is the number of channels (signal modalities), we apply a Positional Encoding and Embedding layer to convert the raw input into a sequence of tokens. Subsequently, the tokens are fed into Transformer layers consisting of Multi-Headed Self-Attention (MSA) [22], Layer Normalization (LN), and Multi-Layer Perceptron (MLP) blocks. Each element is formalized in the following equations:

$$y^i = MSA(LN(x^i)) + x^i \qquad (1)$$
$$x^{l+i} = MLP(LN(y^i)) + y^i \qquad (2)$$

Here, i represents the index of the token, and x^i denotes the generated feature's token. It is worth noting that since the Multi-Headed Self-Attention mechanism allows multiple sequences to be processed in parallel, all 8 signal channels are fed into the Transformer Encoder at once.

The Gaussian Transform [30] is traditionally employed to kernelize linear models by nonlinearly transforming input features via a single layer and subsequently training a linear model on top of the transformed features. However, in the context of deep learning architectures, random features can

also be leveraged, given their ability to perform dimensionality reduction or approximate certain functions via random projections. As a non-parametric technique, this transformation maps input data to a more compressed representation that exclude noise information while still enabling the computationally efficient processing. Such a technique may serve as a valuable supplement to Transformer Encoder architectures, compensating for any missing information.

2) Multi-scale Fusion: From the feature extracted from the Feature Encoder Module at different scale, we fuse them by using Concatenation operation. The concatenated features are then fed through a series of Fully-connected layers (FCN) for the estimation of the 2 Valence and Arousal scores. The Rectified Linear Unit (ReLU) activation function is chosen for its ability to introduce non-linearity into the model, thus contributing to the accuracy of the score estimation. The effectiveness of this approach lies in its ability to efficiently estimate the desired scores, while maintaining a simple and straightforward architecture.

IV. EXPERIMENTAL AND RESULTS

A. Dataset

Following the Emotion Physiology and Experience Collaboration (EPiC) Workshop, we evaluate our model using the Continuously Annotated Signals of Emotion (CASE) dataset [12]. The dataset contains data from several physiological sensors and continuous annotations of emotion. This data was acquired from 30 subjects while they watched several video-stimuli and simultaneously reported their emotional experience using JERI. The devices used include sensors for Electrocardiograph (ECG), Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Respiration (RSP), Skin Temperature (SKT), and Electromyography (EMG). These sensors returns 8 type of

Scenarios type	Arousal		Valence	
	RMSE	STD	RMSE	STD
Across-time scenario	1.503	1.083	1.639	1.2246
Across-subject scenario	1.336	0.2276	1.345	0.0576
Across-elicitor scenario	1.509	0.6771	1.514	0.6054
Across-version scenario	1.369	0.2525	1.352	0.1919

Physiological signals: ECG, BVP, EMG_CORU, EMG_TRAP, EMG_ZYGO, GRS, RSP and SKT. The emotional stimuli consisted of 11 videos, ranging in duration from 120 to 197 seconds. The annotation and physiological data were collected at a sampling rate of 20 Hz and 1000 Hz, respectively. The initial Valence-Arousal score range was set at [-26225, 26225].

In EPiC Competition, the organizers restructure the data and scale the annotation value into [0.5, 9.5] range. The reconstructed dataset has 4 scenarios for 4 different evaluation approach: Across-time scenario, Across-subject scenario, Across-elicitor scenario and Across-version scenario.

- Across-time scenario: Each sample represents a single person watching a single video, and the training and test sets are divided based on time. Specifically, the earlier parts of the video are used for training, while the later parts are reserved for testing.
- Across-subject scenario: Participants are randomly assigned to groups, and all samples from a given group belong to either the train or test set depending on the fold.
- Across-elicitor scenario: Each subject has two samples (videos) per quadrant in the arousal-valence space. For each fold, both samples related to a given quadrant are excluded, resulting in four folds, with one quadrant excluded in each fold.
- Across-version scenario: Each subject has two samples per quadrant in the arousal-valence space. In this scenario, one sample is used for training the model, and the other sample is used for testing, resulting in two folds.

B. Experiments setup

Our networks were implemented using the Tensorflow framework. We trained our models using the AdamW optimizer [31] with a learning rate of 0.001 and the Cosine annealing warm restarts scheduler [32] over 10 epochs. The MSE loss function was used to optimize the network. RMSE is the metrics for evaluation stage. The sequence length was set to 2048. We utilized 4 Transformer layers for the Transformer Encoder, with each Attention module containing 4 heads. The hidden dimension of the Transformer was set to 1024. All training and testing processes were conducted on a GTX 3090 GPU.

C. Results

Table I presents the results of our model on the test set in terms of scenario-level evaluation. Overall, the final RMSE

score for both Valence and Arousal estimation task we gain is 1.45. Our model achieved the best performance in the Across-subject scenario, with an Arousal score of 1.336 and a Valence score of 1.345, with relatively low standard deviation of 0.2276 and 0.0576 for Arousal and Valence, respectively. These results suggest that our model can effectively generalize to new subjects and accurately capture the emotion change after fully viewing the entire video-viewing process. Meanwhile, the relatively low performance in the Across-elicitor scenario, with scores of 1.509 and 1.514 in Arousal and Valence, respectively, suggests that our model did not perform well in inferring emotional states that were not seen during training, given the previously learned specific emotional states.

V. CONCLUSION

This paper proposes a new Multi-scale architecture for Multimodal Emotion Recognition from Physiological signals. Our approach involves encoding the signal with Transformer Encoder at multiple scales to capture both global and local features and obtain more informative representations. Our method achieved decent results on the test data of the EPiC Competition.

ETHICAL IMPACT STATEMENT

Our study leveraged deep learning methods for emotion estimation from physiological signals has ethical implications regarding potential bias, generalizability. The data used for model training was collected from a specific population, introducing the possibility of bias and limiting the generalizability of the models to other populations. Physiological signals can vary among individuals, affecting the accuracy of emotion estimation and further reducing generalizability. We acknowledge these limitations and recommend future research to address these issues and develop ethical guidelines to ensure appropriate use of the models.

REFERENCES

[1] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[3] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[4] S. Saganowski, B. Perz, A. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: a systematic literature review," *IEEE Transactions on Affective Computing*, 2022.

[5] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical signal processing and control*, vol. 55, p. 101646, 2020.

[6] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Emotion recognition with pre-trained transformers using multimodal signals," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, IEEE, 2022.

[7] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2018.

[8] R. Harper and J. Southern, "A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat," *IEEE transactions on affective computing*, vol. 13, no. 2, pp. 985–991, 2020.

[9] M. Algarni, F. Saeed, T. Al-Hadhrami, F. Ghabban, and M. Al-Sarem, "Deep learning-based approach for emotion recognition using electroencephalography (eeg) signals using bi-directional long short-term memory (bi-lstm)," *Sensors*, vol. 22, no. 8, p. 2976, 2022.

[10] S. Saganowski, M. Behnke, J. Komoszyńska, D. Kunc, B. Perz, and P. Kazienko, "A system for collecting emotionally annotated physiological signals in daily life using wearables," in *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–3, IEEE, 2021.

[11] Z. Ahmad and N. Khan, "A survey on physiological signal-based emotion recognition," *Bioengineering*, vol. 9, no. 11, p. 688, 2022.

[12] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific data*, vol. 6, no. 1, p. 196, 2019.

[13] L. Hu, J. Yang, M. Chen, Y. Qian, and J. J. Rodrigues, "Scai-svsc: Smart clothing for effective interaction with a sustainable vital sign collection," *Future Generation Computer Systems*, vol. 86, pp. 329–338, 2018.

[14] K. Rattanyu, M. Ohkura, and M. Mizukawa, "Emotion monitoring from physiological signals for service robots in the living space," in *ICCAS 2010*, pp. 580–583, IEEE, 2010.

[15] K. Rattanyu and M. Mizukawa, "Emotion recognition using biological signal in intelligent space," in *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part III 14*, pp. 586–592, Springer, 2011.

[16] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 346–355, IEEE, 2018.

[17] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion recognition using physiological signals: laboratory vs. wearable sensors," in *Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pp. 15–22, Springer, 2018.

[18] T. Martens, M. Niemann, and U. Dick, "Sensor measures of affective leaning," *Frontiers in Psychology*, vol. 11, p. 379, 2020.

[19] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," *IEEE Access*, vol. 6, pp. 49325–49338, 2018.

[20] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, "Multi-target affect detection in the wild: an exploratory study," in *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pp. 211–219, 2019.

[21] L. Romeo, A. Cavallo, L. Pepa, N. Bianchi-Berthouze, and M. Pontil, "Multiple instance learning for emotion recognition using physiological signals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 389–407, 2019.

[22] 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.

[23] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Advances in neural information processing systems*, vol. 32, 2019.

[24] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," *arXiv preprint arXiv:2001.08317*, 2020.

[25] Y. Li, X. Peng, J. Zhang, Z. Li, and M. Wen, "Dct-gan: dilated convolutional transformer-based gan for time series anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[26] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Transformer-based self-supervised learning for emotion recognition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2605–2612, IEEE, 2022.

[27] Y. Wu, M. Daoudi, and A. Amad, "Transformer-based self-supervised multimodal representation learning for wearable emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[28] K. Yang, B. Tag, Y. Gu, C. Wang, T. Dingler, G. Wadley, and J. Goncalves, "Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 562–570, 2022.

[29] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[30] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in neural information processing systems*, vol. 20, 2007.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[32] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.