Temporal-Related Transformer with Depth Information for 4D Micro-Expression Recognition

Jinsheng Wei[1,2,*], Jialiang Sun[1,2], Guanming Lu[1,2,*] and Jingjie Yan[1,2]

[1]School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

[2]Jiangsu Key Laboratory of Intelligent Information Processing and Communication Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract

Micro-expressions can reflect real emotions, and recognizing micro-expressions has great potential in the fields of criminal investigation, security, and negotiation. Compared to 2D micro-expression images, 4D micro-expression data contains richer information, such as depth information. Thus, this paper explores the effectiveness of depth information in 4D micro-expression video and proposes a Depth Information Temporal Related Transformer (DITRTr) model. DITRTr model mines temporal depth information from 4D point clouds and maps it to the micro-expression category space by modelling the temporal correlation of depth information. Firstly, the model extracts the frontal depth image from each 3D point cloud frame; Then, the pre-trained convolutional neural networks (CNN) are used to extract spatial features of each depth image frame; Finally, a Transformer is adopted to model the correlation between the spatial features of different frames and extract spatiotemporal depth micro-expression features. The experimental results demonstrate the effectiveness of the proposed method, which ranked second in the 4DME Challenge at the 2025 IJCAI.

1. Introduction

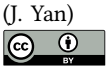
Micro-expressions are subtle facial expressions that appear when people try to hide their true emotions. Compared to ordinary facial expressions, micro-expressions have a shorter duration and lower intensity, which makes them difficult to detect with the naked eye. Psychological research [1, 2] has shown that when a person attempts to conceal their true emotions, their face will unconsciously reveal micro-expressions within a very short period of time, allowing people to interpret their true emotions by analysing these facial micro-expressions. Therefore, accurate analysis and understanding of micro-expressions are of great significance and widely applied in many important fields [3, 4], and micro-expression recognition technology meets the key technological needs of intelligent public and social security, national security, digital services, smart healthcare, and smart cities. For example, in public places, judging whether social personnel have negative emotions based on their micro-expressions can prevent the occurrence of violent safety incidents; In criminal investigation, judging whether a criminal is lying through their micro-expressions can help grasp the correct direction of criminal investigation and accelerate the process of handling criminal cases; In the negotiations, negotiators can judge the true emotions of the other party through their micro-expressions, thereby inferring their actual intentions and tendencies, which can assist negotiators in making more favorable decisions.

Although micro-expression analysis and understanding are of great significance, recognizing micro-expressions is a highly challenging task [5]. The main reason is that micro-expressions occur quickly, and the movement intensity of facial muscles is extremely subtle [6, 7]. Furthermore, due to the limitations of sensory organs, it is difficult for the human eye to detect the occurrence of micro-expressions and analyze the emotions represented by them. Therefore, how to fully extract micro-expression-related information from the whole face is crucial for the practical application of micro-expression recognition.

4DMR@IJCAI25: International IJCAI Workshop on 1st Challenge and Workshop for 4D Micro-Expression Recognition for Mind Reading, August 29, 2025, Guangzhou, China.

*Corresponding author.

weijs@njupt.edu.cn (J. Wei); 1024010316@njupt.edu.cn (J. Sun); lugm@njupt.edu.cn (G. Lu); yanjingjie@njupt.edu.cn (J. Yan)



Li et al. [8] explored 4-dimensional (4D, namely, 3D mesh + temporal changes) facial micro-expression information and collected a 4D micro-expression dataset (4DME). 4D micro-expression data contains richer clues and can provide more discriminative information for micro-expression recognition, such as depth information and different view information.

Compared to 2-dimensional (2D) images, depth information is one of the most distinctive features of 4D data. The movement of facial muscles can cause changes in depth information, and mining these changes can effectively extract micro-expression features. Therefore, this paper explores 4D micro-expression recognition from the perspective of depth information. That is, depth information is mined from 4D micro-expression data to obtain continuous depth image frames.

Micro-expressions are a dynamic process, and learning the temporal correlation between consecutive deep image frames can effectively extract dynamic features. Furthermore, CNN can learn spatial depth features from depth images, but cannot establish temporal correlation between continuous depth images. Transformer can effectively model the correlation between different tokens. Therefore, this paper adopts CNN and Transformer to model the spatial features of a single frame depth image and the temporal correlation features between consecutive frames, respectively.

Overall, the contributions of this paper are as follows:

- 1) This paper explores the effectiveness of depth information in 4D micro-expression data for micro-expression recognition, proposes a Depth Information Temporal Related Transformer (DITRTr) model, and extracts discriminative depth spatiotemporal micro-expression features;
- 2) In the DITRTr, CNN and Transformer are introduced to extract deep spatial features and time-dependent features, respectively, effectively representing the dynamic depth information of micro-expressions;
- 3) The experimental results indicate that the proposed method is effective and ranked second in the 1st 4D micro-expression recognition (4DMR) challenge at IJCAI 2025.

2. Related Work

This paper focuses on micro-expression recognition based on a 4D micro-expression dataset. Therefore, related works are introduced from two aspects, namely, the micro-expression dataset and the recognition algorithm.

2.1. Micro-Expression Dataset

Early micro-expression research relied almost exclusively on 2D image or video datasets, such as CASME II [9], SAMM[10] and SMIC[11], which record subtle facial movements at high frame-rates but discard any steric information about the face. Based on these datasets, a large number of 2D image/video-based micro-expression recognition methods [12, 13, 14, 15, 16] have been proposed, greatly promoting the development of micro-expression recognition. However, the lack of depth cues makes these datasets sensitive to viewpoint changes and illumination, limiting their reliability in unconstrained criminal-investigation or surveillance scenarios.

To alleviate these weaknesses, Li et al. [8] collected a 4D micro-expression (4DME) dataset. 4DME is the first spontaneous 4D micro-expression dataset where every frame contains a high-resolution 3D mesh face. Nevertheless, harnessing 4D data for micro-expression recognition is non-trivial, and the sheer volume of 3D point clouds, irregular mesh topologies and subtle depth variations pose new feature-extraction challenges.

Consequently, this paper investigates how temporal depth cues can be exploited for 4D micro-expression recognition.

2.2. Recognition Algorithm

At present, there is a lack of work in 4D micro expression recognition, while there are some explorations in the field of ordinary 3D/4D expression recognition.

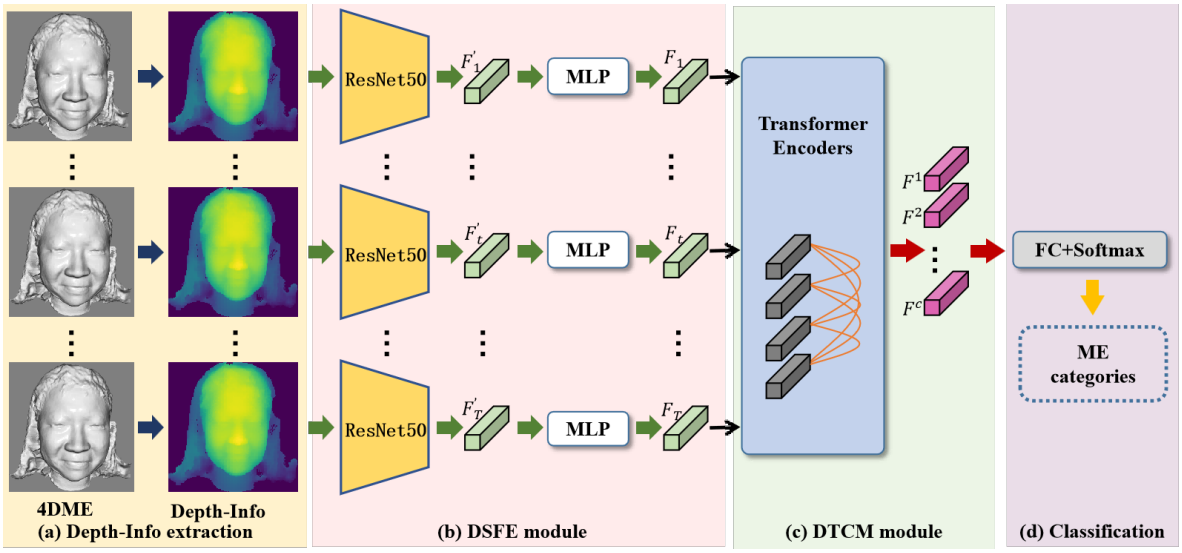


Figure 1: The framework of the proposed DITRTr. DSFE and DTCM modules extract the spatial features and the temporal correlation between depth image frames from the depth image, respectively.

Li et al.[17] presented a comprehensive survey on 3D face recognition methods, covering both traditional and modern methods. Traditional methods mainly extract distinctive facial features for matching, while modern methods rely on deep learning for end-to-end recognition. They also reviewed challenges such as pose, illumination, and expression variations.

Tian et al. [18] designed a novel deep feature fusion convolution neural network (CNN) for 3D facial expression recognition (FER). They represented each 3D face scan as 2D facial attribute maps (including depth, normal, and shape index values) and then combined different facial attribute maps to learn facial representations by fine-tuning a pre-trained deep feature fusion CNN subnet. Global Average Pooling was used to reduce overfitting.

Bouzid et al. [19] presented a method for dynamic facial expression recognition based on 4D facial expression data. Their approach directly extracted spatio-temporal information from the 3D mesh sequences. Every mesh in the sequences was fed into a spatial auto-encoder using spatial convolutions to extract spatial embeddings, and then a temporal transformer processed the sequence of embeddings for facial expression classification.

Trimech et al. [20] aimed to achieve the task by using point cloud-based DNNs. They enlarged the dataset by generating synthetic 3D facial expressions and applied a level curve-based sampling strategy to obtain discriminative point-based representations of 3D faces, achieving promising results.

Kalapala et al.[21] proposed direct classification of normalised and flattened 3D facial landmarks reconstructed from 2D images. They used a pre-trained convolutional Face Alignment Network (FAN) for 3D projection of 2D facial landmarks and tried different classifiers in the spherical coordinate system.

The above methods explore 3D/4D expression recognition. Unlike these methods, this paper proposes a Depth Information Temporal Related Transformer to achieve 4D micro-expression recognition tasks

3. Method

As shown in Figure 1, this paper proposes a novel DITRTr that includes depth information extraction, depth spatial features extraction (DSFE) and depth temporal correlation modelling (DTCM) module.

3.1. Depth Information Extraction

Depth information plays a crucial role in capturing subtle facial muscle movements. Depth maps are extracted from the 3D meshes in order to obtain the depth information invariant to illumination and texture, which is essential for accurately representing the minor and rapid facial movements

characteristic of micro-expressions. First, the vertices of meshes are parsed to obtain the point cloud data:

$$P = \{p_i | p_i = (x_i, y_i, z_i), i = 1, \ldots, N\}. \tag{1}$$

To convert the point cloud into a structured depth map representation, we adopt a distance mapping strategy. The 2D plane is subdivided into a uniform grid with a resolution of $n_d \times n_d$, and the Z-coordinate (depth) of the nearest point is assigned to each grid cell. The grid step sizes along the X and Y axes are computed as:

$$step_x = \frac{X_{max} - X_{min}}{n_d - 1}, \quad step_y = \frac{Y_{max} - Y_{min}}{n_d - 1}, \tag{2}$$

where $X_{max}, X_{mim}, Y_{max}, Y_{min}$ represent the range of the point cloud along the X and Y dimensions, respectively. Finally, each depth map $D \in \mathbb{R}^{n_d \times n_d}$ is constructed by projecting the point cloud onto the grid and assigning: $D(i, j) = z_k$, where (i, j) is the grid cell index corresponding to point p_k. TFollowing this procedure, we generated the frontal-view depth maps for each ME sequence s, denoted as $D_s \in \mathbb{R}^{N_f \times n_d \times n_d}$, where N_f indicates the the number of normalized frames in the sequence.

3.2. DSFE Module

To extract discriminative features from the depth maps of micro-expression video sequences, we employ a ResNet18 as a backbone that has proven effectiveness in visual feature extraction tasks. However, depth maps possess fundamentally different structural and distributional properties compared to RGB images. Therefore, directly applying a ResNet18 pretrained on RGB images may result in suboptimal feature representations for depth inputs. To address this challenge, we fine-tune the ResNet18 model to better adapt to the characteristics of depth maps.

Specifically, the original ResNet18 first convolutional layer, which expects a three-channel input, is modified to accept single-channel depth maps. The weights of the new convolutional layer are initialized by averaging the pretrained RGB weights along the channel dimension to retain transferable low-level features. Formally, given a depth map input $D \in \mathbb{R}^{N_f \times H \times W}$, the modified ResNet18 produces an output feature vector $\mathrm{F}_{res} \in \mathbb{R}^{N_f \times d_{res}}$ after global average pooling:

$$F_{res} = ResNet18_{mod}(D). \tag{3}$$

To further adapt the feature representations for micro-expression recognition, we introduce a MLP adapter that projects the features into a lower-dimensional space suitable for the subsequent Transformer-based temporal modeling. The adapter operation is defined as:

$$F_b = MLP(F_{res}), \quad F_b \in \mathbb{R}^{N_f \times d_{model}}, \tag{4}$$

where d_{model} denotes the dimension in the proposed DTCM Module. By fine-tuning the modified ResNet18 and training the MLP adapter jointly, the model learns the feature representations tailored to the depth properties, thereby enhancing its capability to capture subtle micro-expression cues effectively.

3.3. DTCM Module

To effectively capture the temporal dynamics information in depth image sequences, we adopt a Transformer-based architecture as the depth temporal correlation modeling module. The self-attention mechanism in Transformer enables the model to learn global dependencies across depth image frames, which is crucial for identifying subtle and rapid facial muscle movements.

Given an input depth feature sequence F_b, we prepend c learnable class tokens to the input sequence before feeding it into the DTCM Module. Each class token is designed to represent a specific micro-expression category, enabling the model to independently learn discriminative features for each class. Formally, let $C \in \mathbb{R}^{c \times d_{model}}$ denote the class tokens. The concatenated Transformer input is thus:

$$X = \mathrm{concat}(C, F_b) \in \mathbb{R}^{(c+N_f) \times d_{model}}. \tag{5}$$

The self-attention mechanism within the Transformer encoder then performs information integration across both the temporal frame tokens and the class tokens. Specifically, the self-attention output is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{6}$$

where $Q, K, V \in \mathbb{R}^{(c+N_f)\times d_{model}}$ represent the query, key, and value matrices derived from X. This formulation enables bidirectional interactions between each class token and the temporal feature tokens, allowing each class token to aggregate relevant information from all frames in the sequence. After Transformer encoding, the output representations corresponding to the first c class tokens are extracted and passed through a classification head to produce the final multi-label predictions:

$$Y = \sigma(WF_c + b), \tag{7}$$

where $F_c \in \mathbb{R}^{(c+N_f)\times d_{model}}$ denotes the encoded class token features, and σ is the sigmoid activation function producing the probability scores for each micro-expression class.

By assigning each expression class a dedicated token, the model is encouraged to learn class-specific representations, thereby enhancing its ability to distinguish between subtle and co-occurring micro-expressions in a multi-label classification setting.

3.4. Loss Function

To optimize the proposed model for multi-label recognition, we adopt the binary cross-entropy loss (BCE Loss), incorporating class-specific positive weights to address class imbalance in the dataset. In micro-expression recognition, certain classes are significantly underrepresented compared to others, which can bias the model towards majority classes if standard loss functions are used without reweighting. Let $\hat{y}_{i,j}$ denote the predicted logit for the j-th class of the i-th sample, and $y_{i,j} \in \{0, 1\}$ denote the corresponding ground truth label. The weighted binary cross-entropy loss for a single sample is defined as:

$$L_i = -\sum_{j=1}^{c} \left[w_j y_{i,j} \log \sigma(\hat{y}_{i,j}) + (1 - y_{i,j}) \log(1 - \sigma(\hat{y}_{i,j})) \right], \tag{8}$$

where c is the total number of classes, and w_j is the positive weight for class j, computed as:

$$w_j = \frac{N_{neg,j}}{N_{pos,j} + \epsilon}, \tag{9}$$

where $N_{pos,j}$ and $N_{neg,j}$ represent the number of positive and negative samples for class j, respectively, and is a small constant added for numerical stability.

4. Experiment

4.1. Dataset

We conduct our experiments on the 2025 4DMR challenge dataset, a high-resolution 4D micro-expression dataset. The dataset consists of 100 4D micro-expression samples, each annotated with multiple expression category labels across five fine-grained emotions.

Each micro-expression sequence is captured as a dynamic mesh sequence, where every frame is represented by a 3D mesh file in OBJ format, encoding the detailed facial geometry at that moment. Since the number of frames per sequence varies across samples, temporal normalization is required for batch-wise processing. To this end, we analyzed the temporal distribution of the sequences and found that the average number of frames across the 100 samples is 18.81, while the 90th percentile is 26.10. Balancing computational efficiency and temporal coverage, we normalized all sequences to a fixed length of 20 frames. For sequences exceeding 20 frames, we retained only the first 20 and discarded

Table 1
4DMR challenge competition results on IJCAI 2025

Rank	Team	Macro F1
1	Red-Green-Blue	0.536
2	WS(Our)	0.518
3	Infinite Messtropy	0.476

the remaining ones; for shorter sequences, we applied zero-padding at the end to reach the required frame count. To facilitate model training and hyperparameter tuning, 100 samples in the dataset were randomly divided into a training set of 90 samples and a validation set of 10 samples.

4.2. Metrics

Following the requirements of the challenge competition, we adopt the F1-score as the primary evaluation metric to evaluate the performance of the model, which effectively balances precision and recall, especially in scenarios with class imbalance. The F1-score is defined as:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{10}$$

The macro-averaged F1-score, denoted as $F1_{macro}$, is computed to provide a more comprehensive evaluation across all classes. This metric gives equal weight to each class, regardless of its sample size, and is particularly suitable for micro-expression datasets with unbalanced class distributions.

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} F1_i, \tag{11}$$

where N is the total number of classes, and $F1_i$ is the F1-score calculated for the i^{th} class individually.

4.3. Implement Detail

The proposed model was trained using the Adam optimizer with a learning rate set to 1e-4 and a weight decay of 1e-4 to prevent overfitting. The batch size was set to 32, and the model was trained for 100 epochs to ensure convergence. For the input data, each sequence was normalized to 20 frames. Sequences with fewer than 20 frames were zero-padded, while those with more than 20 frames were truncated to maintain a consistent temporal dimension. The depth maps resolution n_d was set to 128. The feature dimension d_{model} of the transformer was set to 128.

4.4. Challenge Competition Results

Our method is compared with the first and third-ranked methods in terms of the 4DMR challenge competition on IJCAI 2025. The performance results are presented in Table 1. Our method secured the second rank, highlighting its strong competitiveness and overall effectiveness. Specifically, our method is 0.018 Macro F1-score lower than the first-ranked method and 0.042 Macro F1-score higher than the third-ranked method.

4.5. Ablation Study

We evaluate the DTCM Module by comparing the model's performance with and without this module. For the model without the DTCM Module, the temporal modeling and feature aggregation components were replaced by a simpler structure consisting of linear layers followed by average pooling, thereby removing the dedicated temporal–class token interaction mechanism introduced in DTCM. As shown in

Table 2
Performance with and without DTCM Module

Method	Macro F1(Val)	Macro F1(Test)
with DTCM	0.8311	0.518
without DTCM	0.7156	0.399

Table 2, the model incorporating the DTCM Module demonstrates significant improvements. Specifically, the model with the DTCM achieves a 0.518 Macro F1-score that is 0.119 higher than that of the model without DTCM.

It is worth noting that the model with DTCM Module achieves a 0.8311 Macro F1 score on the validation set and 0.518 on the test set. There is a significant gap in the results between the validation set and the test set, which may be due to the small number of samples in the validation set.

5. Discussion

In this work, we proposed a DITRTr model for 4D micro-expression recognition, which explicitly integrates depth information extracted from facial 3D mesh sequences. By fine-tuning the pretrained ResNet-18, the backbone network better accommodates single-channel depth maps. Furthermore, a set of learnable class tokens was introduced into the Transformer architecture to enhance the discriminability of temporal representations under a multi-label setting.

Our method achieves competitive performance on the 4DMR benchmark, achieving second place in the IJCAI 2025 challenge. These results underscore the effectiveness of incorporating depth spatial information and temporal dependencies for subtle facial movement analysis.

There are potential directions for further exploration. While our method extracts depth maps from raw mesh sequences and applies downsampling to produce fixed-resolution representations, this process may overlook fine-grained spatial cues that are critical for identifying subtle micro-expressions. More adaptive depth feature encoding strategies—such as region-aware sampling or attention-based depth refinement may offer enhanced sensitivity to critical facial muscle movements.

6. Conclusion

This paper presents a novel DITRTr model to recognize 4D micro-expressions, based on depth information. The proposed method designs the DSFE and DTCM modules to extract the spatial features and the temporal correlation between depth image frames from the depth image, respectively, which effectively captures subtle facial movements in facial 3D sequences (4D data). The experiments conducted on the 4DME dataset validate the performance of our method, which ranks second in the 4DMR IJCAI Workshop 2025 challenge.

Acknowledgments

This work was supported in part by the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (Grant No.24KJB520022), in part by the Nanjing Science and Technology Innovation Foundation for Overseas Students (Grant No. RK002NLX23004), in part by Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No.NY223030).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Haggard, Ernest A., Isaacs, Kenneth S., Micromonlentary facial expressions as indicators of ego mechanisms in psychotherapy, in: Methods of Research in Psychotherapy, Springer US, Boston, MA, 1966, pp. 154–165.

[2] P. Ekman, Lie Catching and Microexpressions, in: C. Martin (Ed.), The Philosophy of Deception, Oxford University Press, 2009, pp. 118–136. URL: https://academic.oup.com/book/6899/chapter/151126752. doi:10.1093/acprof:oso/9780195327939.003.0008.

[3] G. Zhao, X. Li, Y. Li, M. Pietikäinen, Facial Micro-Expressions: An Overview, Proceedings of the IEEE 111 (2023) 1215–1235. URL: https://ieeexplore.ieee.org/document/10144523/. doi:10.1109/jproc.2023.3275192, publisher: Institute of Electrical and Electronics Engineers (IEEE).

[4] Y. Li, J. Wei, Y. Liu, J. Kauttonen, G. Zhao, Deep Learning for Micro-Expression Recognition: A Survey, IEEE Transactions on Affective Computing 13 (2022) 2028–2046. URL: https://ieeexplore.ieee.org/document/9915437/. doi:10.1109/TAFFC.2022.3205170.

[5] J. Wei, J. Sun, G. Lu, J. Yan, D. Zhang, Multi-information hierarchical fusion transformer with local alignment and global correlation for micro-expression recognition, in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 5873–5882.

[6] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–1. URL: https://ieeexplore.ieee.org/document/9382112/. doi:10.1109/TPAMI.2021.3067464.

[7] J. Wei, W. Peng, G. Lu, Y. Li, J. Yan, G. Zhao, Geometric Graph Representation With Learnable Graph Structure and Adaptive AU Constraint for Micro-Expression Recognition, IEEE Transactions on Affective Computing 15 (2024) 1343–1357. URL: https://ieeexplore.ieee.org/document/10345706/. doi:10.1109/TAFFC.2023.3340016.

[8] X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, M. Pantic, G. Zhao, 4dme: A spontaneous 4d micro-expression dataset with multimodalities, IEEE Transactions on Affective Computing 14 (2022) 3031–3047.

[9] W. Yan, X. Li, S. Wang, G. Zhao, Y. Liu, Y. Chen, X. Fu, Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, PloS one 9 (2014) e86041.

[10] A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE transactions on affective computing 9 (2016) 116–129.

[11] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

[12] Y. Wang, J. See, R. C.-W. Phan, Y.-H. Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: D. Cremers, I. Reid, H. Saito, M.-H. Yang (Eds.), Computer Vision – ACCV 2014, Springer International Publishing, Cham, 2015, pp. 525–537.

[13] J. Wei, G. Lu, J. Yan, H. Liu, Micro-expression recognition using local binary pattern from five intersecting planes, Multimedia Tools and Applications 81 (2022) 20643–20668. URL: https://link.springer.com/10.1007/s11042-022-12360-x. doi:10.1007/s11042-022-12360-x.

[14] M. Wei, X. Jiang, W. Zheng, Y. Zong, C. Lu, J. Liu, CMNet: Contrastive Magnification Network for Micro-Expression Recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 119–127. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25083. doi:10.1609/aaai.v37i1.25083.

[15] J. Wei, G. Lu, J. Yan, Y. Zong, Learning two groups of discriminative features for micro-expression recognition, Neurocomputing 479 (2022) 22–36. URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231221019433. doi:10.1016/j.neucom.2021.12.088.

[16] J. Yang, Z. Wu, R. Wu, Micro-expression recognition based on contextual transformer networks, The Visual Computer 41 (2025) 1527–1541. URL: https://link.springer.com/10.1007/s00371-024-03443-x. doi:10.1007/s00371-024-03443-x.

[17] M. Li, B. Huang, G. Tian, A comprehensive survey on 3d face recognition methods, Engineering Applications of Artificial Intelligence 110 (2022) 104669.
[18] K. Tian, L. Zeng, S. McGrath, Q. Yin, W. Wang, 3d facial expression recognition using deep feature fusion cnn, in: 2019 30th Irish Signals and Systems Conference (ISSC), IEEE, 2019, pp. 1–6.
[19] H. Bouzid, L. Ballihi, 3d facial expression recognition using spiral convolutions and transformers, in: 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), IEEE, 2023, pp. 1–7.
[20] I. H. Trimech, A. Maalej, N. E. Ben Amara, Facial expression recognition using 3d points aware deep neural network., Traitement du Signal 38 (2021).
[21] L. Kalapala, H. Yadav, H. Kharwar, S. Susan, Facial expression recognition from 3d facial landmarks reconstructed from images, in: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), IEEE, 2020, pp. 1–5.