Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

A Review of Dynamic Facial Expression Recognition: Methods, Datasets and Directions

Xuantao Nie NIE_XT@SHU.EDU.CN
School of Computer Engineering and Science, Shanghai University, Shanghai 200044, China

Zixiang Fei[*] ZXFEI@SHU.EDU.CN
School of Computer Engineering and Science, Shanghai University, Shanghai 200044, China

Wenju Zhou ZHOUWENJU@SHU.EDU.CN
School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200044, China

Minrui Fei MRFEI@STAFF.SHU.EDU.CN
School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200044, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Dynamic facial expression recognition (DFER) has emerged as an essential area of research in computer vision, with implications in human-computer interaction, psychological analysis, and security. Although image-based static facial expression recognition (SFER) is well-developed, DFER captures temporal dynamics, remains less explored. This paper comprehensively reviews DFER, focusing on feature extraction methods from traditional handcrafted features to advanced deep learning techniques, analyzing performance metrics, and examining publicly available datasets with their comparative characteristics. We discuss specific challenges faced by DFER systems such as occlusion, pose variations, and temporal alignment. Finally, we explore promising applications in healthcare and human-computer interaction, providing concrete implementation strategies and future research directions.

Keywords: Dynamic expressions, Feature extraction, Deep learning, Datasets, Challenges

1. Introduction

Facial expressions serve as one of the most fundamental channels for human emotional communication. Their importance spans evolutionary psychology, social interaction, and affective computing (Krumhuber et al., 2023). Traditional facial expression recognition (FER) research is foundational to understanding emotional expressions. Ekman and Oster (1979) identified six basic emotions—happiness, anger, sadness, surprise, fear, and disgust—linked universally to specific facial muscle movements. Standard FER tasks typically classify expressions into seven categories, adding a neutral state to these six basic emotions. Driven by AI advancements, FER has experienced rapid growth with applications in medical diagnosis (Zhao et al., 2016) and psychological research Rudovic et al. (2018).

While SFER focuses on analyzing isolated frames, DFER emphasizes the temporal evolution of expressions. DFER systems analyze sequences spanning 0.5-4 seconds, capturing micro-temporal features including onset (expression initiation), apex (peak intensity), and offset (expression decay) phases (Ćimić et al., 2021). These spatiotemporal patterns—ranging from rapid micro-expressions to prolonged affective displays—reflect the natural progression of emotions in real-world interactions.

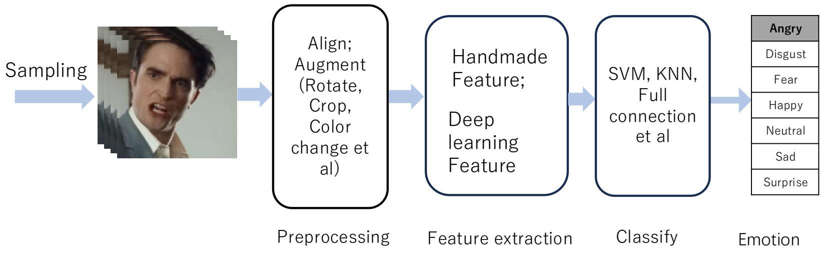

Figure 1: Workflow of DFER showing the progression from video input through preprocessing, feature extraction, and classification stages

DFER holds significant advantages over SFER by preserving the sequential structure of emotional expressions, which is essential for disambiguating subtle or context-dependent affective states. By capturing temporal coherence and anatomical constraints of facial behavior, DFER provides a more ecologically valid representation of emotion, bridging the gap between laboratory-controlled stimuli and natural human communication.

This review comprehensively examines DFER, analyzing feature extraction methods (both traditional and deep learning-based), evaluating challenges specific to dynamic recognition, comparing available datasets, discussing evaluation metrics, and exploring practical applications with implementation strategies.

2. Background and Challenges

2.1. Biological Basis of Dynamic Expressions

DFER is rooted in the Facial Action Coding System (FACS) developed by Friesen and Ekman (1978), which catalogs facial movements based on Action Units (AUs) corresponding to specific facial muscle contractions. Dynamic expressions are characterized not only by AU combinations but by their temporal evolution.

Micro-expressions (lasting less than a second) and macro-expressions have distinct neural signatures and recognition requirements. Micro-expressions typically activate more primitive emotional neural circuits and require high temporal resolution for detection, while macro-expressions engage both emotional and cognitive pathways and are easier to detect due to their prolonged visibility (Nia et al., 2024; Zhao et al., 2023).

2.2. Workflow of DFER

The DFER process typically follows a sequential pipeline as illustrated in Figure 1. The workflow consists of: image acquisition and sampling from facial expression videos or image sequences; preprocessing via alignment, normalization, and augmentation; facial expression feature extraction using attention mechanisms, RNNs, or 3D CNNs; and single or mixed emotion classification.

2.3. Specific Challenges in DFER

DFER faces several unique challenges beyond those encountered in SFER:

Temporal Alignment: While SFER analyzes single frames with fixed spatial features, DFER must process temporal sequences that vary significantly in duration and intensity across individuals. This creates unique difficulties in standardizing temporal dynamics and distinguishing between brief micro-expressions and sustained emotional displays (Zhao and Liu, 2021). The onset-apex-offset pattern critical to DFER has no counterpart in static approaches.

Head Pose Variations and Occlusions: In SFER, occlusion in a single frame may render recognition impossible. DFER, however, must maintain temporal coherence while tracking partial occlusions across multiple frames, potentially using information from unoccluded frames to complement occluded ones (Jiang et al., 2020). This temporal continuity requirement represents a distinctly different challenge from static recognition.

Multi-Expression Transitions: Real-world expressions frequently transition between emotional states or blend multiple emotions simultaneously, requiring systems to detect these transitions and compound expressions (Chen et al., 2024a).

3. Feature Extraction Methods

3.1. Handcrafted Methods

Traditional DFER approaches leverage engineered spatiotemporal features. Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) extends the LBP operator to three orthogonal planes, efficiently encoding appearance and motion patterns while maintaining robustness to illumination changes. Histogram of Oriented Gradients (HOG) captures gradient information across frames, effectively detecting edge patterns despite sensitivity to background noise (Polikovsky et al., 2009). Optical flow techniques directly measure facial movements but are computationally intensive and sensitive to illumination variations.

These methods excel in controlled environments but lack adaptability to unconstrained scenarios and require manual re-engineering for new variations compared to the deep learning methods. Table 1 presents a comparative analysis of traditional handcrafted and deep learning approaches for DFER.

Table 1: Comparison Between Handcrafted and Deep Learning Methods

Aspect	**Handcrafted**	**Deep Learning**
Interpretability	High	Lower
Data requirements	Less demanding	Requires large datasets
Computational cost	Lower for inference	Higher for training
Generalization	Limited to designed features	Better in varied conditions
Performance ceiling	Lower in unconstrained settings	Higher with sufficient data

3.2. Deep Learning Methods

3.2.1. 2DCNN-RNN MODELS

Early DFER methods used CNNs and RNNs for spatiotemporal feature extraction. The CNN-LSTM architecture (Tong et al., 2020) extracts spatial features via CNNs and models temporal dependencies through LSTMs. While effective for long-range dependencies, this approach may suffer from vanishing gradients over extended sequences.

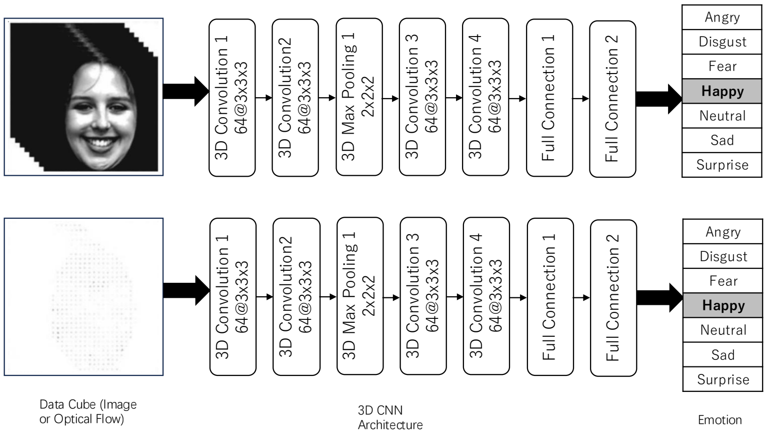

Figure 2: 3DCNN architecture for DFER, capturing spatiotemporal features through 3D convolutions. Figure reproduced based on Zhao et al. (2018)

Yu et al. (2020) introduced a multitask global-local network that focuses on both local facial regions and the whole face. The main advantage of 2DCNN-RNN models is their ability to leverage pre-trained image models, though they process spatial and temporal information separately, potentially missing joint patterns.

3.2.2. 3DCNN MODELS

3D convolutional networks capture spatial and temporal features simultaneously. Unlike 2D CNNs, 3D CNNs analyze entire video volumes, enabling end-to-end learning from raw video inputs. Zhao et al. (2018) demonstrated how 3D CNNs effectively capture local motion patterns crucial for expression recognition, though they require substantial training data and computational resources.

The key advantage of 3D CNNs is their ability to directly learn spatiotemporal features without separating spatial and temporal processing. This enables the detection of subtle motion patterns crucial for expression recognition. However, their increased parameters compared to 2D CNNs lead to higher computational costs and a greater susceptibility to overfitting, particularly on smaller datasets.

3.2.3. TRANSFORMER-BASED MODELS

Attention mechanisms have recently gained popularity in DFER due to their ability to focus on relevant facial regions and temporal segments. Zhao and Liu (2021) proposed Former-DFER, consisting of a convolutional spatial transformer and a temporal transformer. The spatial transformer guides learning of occlusion and pose-robust facial features, while the temporal transformer captures contextual information across frames. This architecture demonstrates superior performance in unconstrained environments with varying head poses and partial occlusions.

Transformer-based models offer several advantages for DFER: they effectively handle variable-length sequences, capture long-range dependencies without suffering from vanishing gradients, and

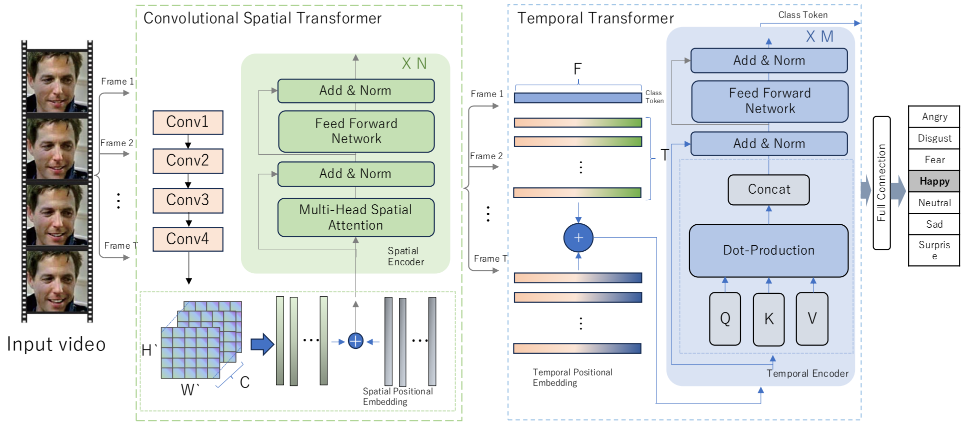

Figure 3: Transformer-based architecture for DFER, utilizing spatial and temporal attention mechanisms. Figure reproduced based on Former-DFER (Zhao and Liu, 2021)

can attend to the most informative facial regions across time. The primary limitations include increased computational requirements and the need for larger training datasets compared to CNN-based approaches.

3.3. Emerging Trends

3.3.1. SFER to DFER

Recent approaches like S2D (Chen et al., 2024b) leverage pre-trained static recognition models with additional dynamic information, circumventing limited DFER dataset availability while potentially inheriting SFER biases.

3.3.2. Multimodal DFER

Multimodal fusion approaches integrate multiple data sources such as facial expressions, voice, physiological signals, and text descriptions to improve recognition accuracy. Systems combining facial cues and vocal intonation achieve superior performance in real-world contexts where facial expressions alone are insufficient. Chen et al. (2024a) introduced FineCLIPER, a multimodal fine-grained framework that extends category labels to text descriptions and hierarchically mines useful clues (video frames, facial segmentation masks, keypoints, and cross-frame facial change descriptions). Such multimodal approaches promise to advance DFER in challenging real-world scenarios where a single modality is insufficient.

3.4. Comparison and evaluation metrics

Table 2 presents a comparative analysis of key deep learning DFER methods. We report accuracy (A), unweighted average recall (UAR), and weighted average recall (WAR). Standard evaluation metrics for DFER include accuracy, F1-score, and confusion matrices for balanced datasets. For imbalanced datasets common in real-world scenarios, UAR and WAR are preferred as they account

Table 2: Summary of Deep Learning methods of DFER

Scheme	Tech.	Result	Advantage	Disadvantage
Cross-LayerLSTM (Tong et al., 2020)	2DCNN-RNN	A=97.47% on CK+	Spatial-temporal modeling via cross-layer LSTM	Complex, high computation cost
3D CNN (Zhao et al., 2018)	3DCNN	A=98.47% on CK+	compact and efficient	Limited to supervised learning; may miss high-level semantics
Former-DFER (Zhao and Liu, 2021)	Transformer	UAR=53.69%, WAR=65.70% on DFEW; 31.16% and 43.27% on MAFW	Robust to occlusion and pose	High data, computation needs
S2D (Chen et al., 2024b)	SFER to DFER	UAR=65.45%, WAR=76.03% on DFEW; 43.40%, 57.37% on MAFW; 43.97%, 52.56% on FERV39k	Leverages SFER knowledge; adapts image models for videos	Relies on external landmark detectors; limited by SFER data
FineCLIPER (Chen et al., 2024a)	Multi-modal	UAR=65.98%, WAR=76.21% on DFEW; 45.01%, 56.91% on MAFW; 47.89%, 56.41% on FERV39k	Hierarchical multi-modal approach; efficient fine-tuning	Complex design; high computational cost; depends on pre-trained models

for class imbalance. Cross-validation strategies vary across datasets, with subject-independent protocols being essential to assess generalization capabilities. In table 2, while laboratory datasets like CK+ yield high accuracy scores, performance decreases significantly on in-the-wild datasets such as DFEW, demonstrating the substantial challenges in unconstrained environments.

4. Datasets and Evaluation

DFER heavily relies on robust datasets and standardized evaluation protocols. Current datasets vary significantly in size, diversity, collection environment, and annotation methods, presenting different challenges for model development and evaluation.

Table 3: Comparison of DFER Datasets

Dataset	Sample Size	Resolution	Annotation Method	Cultural Bias
CK+	593	640×480	8 labels performed by subjects	Western
DFEW	16,372	Varied	10 labels by 12 experts	Western
MAFW	10,045	Varied	11 labels by 11 annotators	Mixed
FERV39k	38,935	Varied	Automated + manual curation	Mixed

CK+ (Lucey et al., 2010): The Extended Cohn-Kanade(CK+) contains 593 sequences from 123 subjects under controlled laboratory conditions. While it offers high-quality, posed expressions with standardized illumination and frontal poses, its limited scale and diversity restrict its applicability to real-world scenarios. CK+ serves primarily as a baseline dataset for initial model validation.

DFEW (Jiang et al., 2020): Dynamic Facial Expression in-the-Wild(DFEW) comprises 16,372 video clips extracted from movies, featuring varied lighting, occlusions, and natural head movements. Each clip has been annotated ten times by 12 expert annotators, providing robust expression labels. The dataset's in-the-wild nature makes it particularly valuable for evaluating models intended for real-world deployment.

MAFW (Liu et al., 2022): Multi-modal Affective Facial in-the-Wild(MAFW) contains 10,045 clips collected from diverse geographical regions. Its unique contribution lies in its multi-cultural representation and comprehensive annotation scheme, including single expression labels, multiple expression labels, and bilingual emotional descriptive text. This dataset specifically addresses compound expressions, where multiple emotions are expressed simultaneously.

FERV39k (Wang et al., 2022): FERV39k is one of the largest DFER datasets, comprising 38,935 annotated video clips categorized into 22 fine-grained scenes spanning four distinct scenario types. Its organization by interaction context (including routine daily activities, varying levels of interactivity such as passive viewing versus dynamic participation, and atypical events), enables context-specific model development and evaluation.

5. Applications and Implementation Strategies

5.1. Healthcare Integration

DFER systems offer significant potential in mental health monitoring and neurological disorder detection. Implementation involves:

Continuous monitoring: Capturing facial videos during therapeutic sessions or daily activities through smartphone cameras.

Temporal analysis: Analyzing expression dynamics to identify micro-expressions and emotional transitions that may indicate underlying psychological states.

DFER demonstrates particular promise in detecting early signs of neurodegenerative diseases. Studies show that specific deficits in dynamic expression recognition correlate with cognitive impairment in conditions like depression and Alzheimer's disease (Fei et al., 2019; Gu et al., 2024). Subtle changes in expression dynamics—including reduced intensity, delayed onset, or premature offset—can serve as biomarkers preceding other cognitive symptoms (Chen et al., 2023). By implementing longitudinal tracking of these expression parameters, healthcare systems can potentially identify at-risk individuals before traditional diagnostic criteria are met.

5.2. Advanced Human-Computer Interaction

DFER technology enables more natural and responsive human-computer interactions through emotion-aware systems. Implementation strategies include:

Real-time adaptive interfaces: Systems that modify their behavior based on detected emotional states, providing additional support when frustration is detected or simplifying interfaces when confusion is observed.

Social robotics: Robots equipped with DFER capabilities can recognize user emotions and adapt their responses accordingly, creating more natural and empathetic interactions (Jiang et al., 2023).

Practical implementations include emotion-controlled music players (Umer et al., 2023) and VR/AR systems that dynamically adjust to maintain optimal engagement or therapeutic effectiveness. In educational applications, this allows for personalized learning experiences that adapt to students' emotional states, potentially improving learning outcomes.

6. Conclusion

This review has examined the evolution of DFER from handcrafted features to sophisticated deep learning architectures, highlighting the unique challenges and opportunities in capturing the temporal dynamics of facial expressions. We have analyzed how different methodologies address specific challenges like occlusions, pose variations, and temporal alignment, while comparing their performance across standardized metrics and diverse datasets.

Emerging trends such as multimodal fusion and transfer learning from static to dynamic recognition demonstrate promising directions for improving DFER robustness and generalization. Their integration into human-computer interaction and healthcare monitoring highlights DFER's practical applications beyond academia.

Future research should focus on developing lightweight architectures for real-time applications, addressing cultural and individual variability in expression dynamics, and improving generalization across diverse populations. As DFER technologies mature, their integration into everyday applications will likely accelerate, enabling more intuitive, responsive, and empathetic technological interactions.

Acknowledgments

This study is supported by the Natural Science Foundation of China (NSFC) under Grant 62203290.

References

Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multimodal fine-grained clip for dynamic facial expression recognition with adapters. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2301–2310. Association for Computing Machinery, Inc, 2024a.

Pan Chen, Hong Cai, Wei Bai, Zhaohui Su, Yi-Lang Tang, Gabor S Ungvari, Chee H Ng, Qinge Zhang, and Yu-Tao Xiang. Global prevalence of mild cognitive impairment among older adults living in nursing homes: a meta-analysis and systematic review of epidemiological surveys. *Translational psychiatry*, 13(1):88, 2023.

Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, 2024b.

Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual review of psychology*, 1979.

Zixiang Fei, Erfu Yang, David Day-Uei Li, Stephen Butler, Winifred Ijomah, and Huiyu Zhou. A survey on computer vision techniques for detecting facial features towards the early diagnosis of mild cognitive impairment in the elderly. *Systems Science & Control Engineering*, 7(1):252–263, 2019.

E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

Shanshan Gu, Xinlu Sun, Bin Chen, and Weijing Tao. Depression micro-expression recognition technology based on multimodal knowledge graphs. *Traitement du Signal*, 41(4), 2024.

Cheng-Shan Jiang, Zhen-Tao Liu, Min Wu, Jinhua She, and Wei-Hua Cao. Efficient facial expression recognition with representation reinforcement network and transfer self-training for human–machine interaction. *IEEE Transactions on Industrial Informatics*, 19(9):9943–9952, 2023.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.

Eva G Krumhuber, Lina I Skora, Harold CH Hill, and Karen Lander. The role of facial movements in emotion recognition. *Nature Reviews Psychology*, 2(5):283–296, 2023.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pages 24–32, 2022.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

Alireza Farrokhi Nia, Vanessa Tang, Valery Malyshau, Amit Barde, Gonzalo Maso Talou, and Mark Billinghurst. Fead: Introduction to the fnirs-eeg affective database-video stimuli. *IEEE Transactions on Affective Computing*, 2024.

Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *3rd international conference on imaging for crime detection and prevention (ICDP 2009)*, pages 1–6. IET, 2009.

Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):eaao6760, 2018.

Ying Tong, Rui Chen, and Ruiyu Liang. Unconstrained facial expression recognition based on feature enhanced cnn and cross-layer lstm. *IEICE TRANSACTIONS on Information and Systems*, 103(11):2403–2406, 2020.

Saiyed Umer, Ranjeet Kumar Rout, Shailendra Tiwari, Ahmad Ali AlZubi, Jazem Mutared Alanazi, and Kulakov Yurii. Human-computer interaction using deep fusion model-based facial expression recognition system. *CMES - Computer Modeling in Engineering and Sciences*, 135(2):1165–1185, 2023.

Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.

Mingjing Yu, Huicheng Zheng, Zhifeng Peng, Jiayu Dong, and Heran Du. Facial expression recognition based on a multi-task global-local network. *Pattern Recognition Letters*, 131:166–171, 2020.

Jianfeng Zhao, Xia Mao, and Jian Zhang. Learning deep facial expression features from image and optical flow sequences using 3d cnn. *The Visual Computer*, 34:1461–1475, 2018.

Q Zhao, X Zhang, G Chen, and J Zhang. Eeg and fnirs emotion recognition based on modal attention graph convolutional feature fusion. *J. Zhejiang Univ. Sci*, 57:1987–1997, 2023.

Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. Predicting personalized image emotion perceptions in social networks. *IEEE transactions on affective computing*, 9 (4):526–540, 2016.

Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1553–1561, 2021.

Goran Ćimić, Mladenka Tkalčić, Vana Vukić, Damir Mulc, Ena Španić, Marina Šagud, Francisco E. Olucha-Bordonau, Mario Vukšić, and Patrick R. Hof. Understanding emotions: Origins and roles of the amygdala. *Biomolecules*, 11, 2021.