RESEARCH-ARTICLE

DMSNet: A Lightweight and Efficient Facial Expression Recognition Model for IoT and WoT Applications

HEMRAJ SINGH, University of Petroleum and Energy Studies, Dehradun, UT, India

PRASHANTH BAITHI, Institute for Development and Research in Banking Technology India, Hyderabad, AP, India

ZARKA BASHIR, Indian Institute of Technology Hyderabad, Sangareddy, TG, India

Published: 08 May 2025

WWW '25: The ACM Web Conference 2025
April 28 - May 2, 2025
Sydney NSW, Australia

Conference Sponsors:
SIGWEB

https://doi.org/10.1145/3701716.3717576
ISBN: 9798400713316

DMSNet: A Lightweight and Efficient Facial Expression Recognition Model for IoT and WoT Applications

Hemraj Singh
hemraj.singh@ddn.upes.ac.in
The University of Petroleum and Energy Studies
Dehradun, Uttarakhand, India

Prashanth Baithi
baithi.prashanth.07@gmail.com
Institute for Development and Research in Banking Technology
Hyderabad, India

Zarka Bashir
cs21resch11010@iith.ac.in
Indian Institute of Technology
Hyderabad, Telangana, India

Abstract

Facial expression recognition (FER) plays a crucial role in computer vision, driving advancements in gesture recognition, patient monitoring, and human-robot interaction. Despite its potential, traditional FER methods struggle with geometric variations in facial expression features within static images, often leading to an imbalance between model performance and network complexity. This results in increased computational demands, hindering their deployment in resource-constrained environments such as the Internet of Things (IoT) or Web of Things (WoT) and mobile. To address these challenges, we propose the **D**eformable **M**ulti-**S**cale **Net**work **(DMSNet)**, a lightweight and efficient model specifically designed to capture multi-scale geometric variations in spatial expression features dynamically using depthwise separable convolution, deformable convolution and Receptive Field Blocks (RFBs) while minimizing parameters. It is highly suitable for real-time applications with just 5.6 million parameters, 100.8 million floating point operations, and 30 frame-per-second (FPS) inference speeds. Extensive experiments on five benchmark datasets demonstrate superior performance compared to state-of-the-art (SOTA) models for FER.

CCS Concepts

• **Information systems** → **Information extraction**; **Data streaming**.

Keywords

Facial expression recognition, geometric features, multi-scale features, multi-layer perceptron

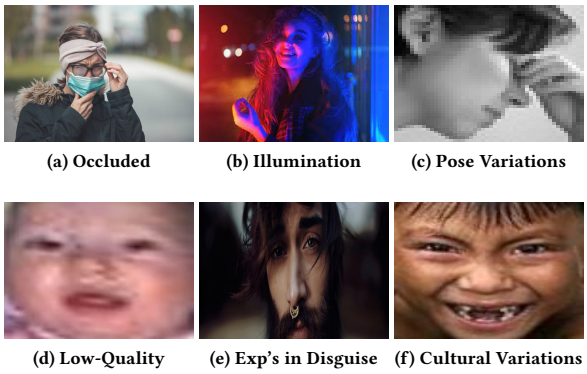
ACM Reference Format:
Hemraj Singh, Prashanth Baithi, and Zarka Bashir. 2025. DMSNet: A Lightweight and Efficient Facial Expression Recognition Model for IoT and WoT Applications. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3717576

WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717576

1 Introduction

The recognition of facial expressions, a highly effective and innate method for humans to communicate emotions, has become a focal point in the realm of computer vision. Automatic Facial Expression Recognition (FER) is gaining attention, particularly due to its diverse applications in areas such as medical diagnostics[4], driver fatigue monitoring [9], and human-computer interaction (HCI) [16]. FER primarily aims to categorize an image or video



(a) Occluded (b) Illumination (c) Pose Variations

(d) Low-Quality (e) Exp's in Disguise (f) Cultural Variations

Figure 1: The challenging scenarios of FER2013, RAF-DB, and CK+ datasets for facial expression recognition.

clip into various fundamental emotions, such as neutral, happiness, sadness, surprise, fear, disgust, anger, and sometimes disdain. FER is commonly divided into two main categories, static FER, and dynamic FER, based on whether the input is an image or a video. Additionally, it can be distinguished between laboratory-controlled [2] and in-the-wild conditions based on the context in which the recognition takes place. In recent years, Facial Expression Recognition (FER) has achieved remarkable results on controlled datasets like CK+ [17] and JAFFE [5], which feature frontal, unobstructed faces in laboratory settings. However, FER performance declines significantly when applied to real-world, in-the-wild conditions, where challenges such as illumination variations, occlusion, and pose changes are prevalent. Datasets like RAF-DB [17] expose these limitations, with FER models performing worse compared to laboratory-controlled scenarios [20]. Overcoming these challenges requires addressing issues like occlusion, pose variations, and illumination effects. Early approaches, such as [2], [13], and [14], attempted to manually reconstruct obstructed features to mitigate occlusion. Deep learning models, especially CNNs, have since become essential for tackling these issues in real-world settings,

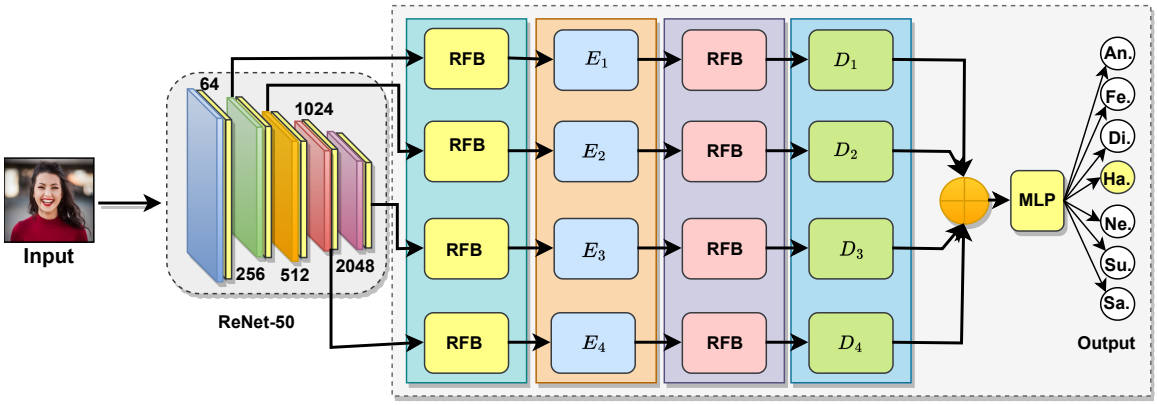


Figure 2: The architecture diagram of the proposed DMSNet. Where RFB is Receptive Field Blocks, $E_i, i = 1, 2, 3, 4$ is the encoder blocks, $D_i, i = 1, 2, 3, 4$ is the decoder blocks, $\oplus$ is the element-wise addition operation, and MLP is multi-layer perception layer. Where An is Angry, Sa is Sad, Ne is Neutral, Ha is Happy, Fe is Fear, Su is Surprise, and Di is Disgust.

with patch-based methods, often guided by facial landmarks [8], [9], offering effective solutions for occlusion and pose variation challenges.

Learning facial features from different perspectives has the potential to improve performance in scenarios with occlusion and pose variation. Psychological studies provide evidence that human face perception mechanisms extract information from both holistic and part-based information [5]. In response to this, [20] presents a Global Multi-Scale and Local Attention Network (MA-Net) designed to incorporate both global and local perspectives for capturing robust facial features. However, it faces limitations in extracting the geometric variation of multi-scale spatial features and dropping small expressions. To address the above-mentioned challenges, a **D**eformable **M**ulti-**S**cale **Net**work **(DMSNet)** is proposed, which extracts multi-scale geometry spatial features while effectively reduce the network's computational complexity using deformable convolution [3, 15], depth-wise separable convolution [7, 15], and Receptive Field Blocks (RFBs) [11]. The key contributions of our study are summarized as follows:

- A novel lightweight Deformable Multi-Scale Network (DMSNet) is proposed, which features four encoder blocks, four decoder blocks, and four Receptive Field Blocks (RFBs) [11].
- It extracts local salient features and excels in capturing global salient patterns due to that it effectively mitigates challenges such as occlusion and poses variations encountered in real-world scenarios.
- DMSNet is the first model to dynamically extract multi-scale geometric spatial information, reducing the vulnerability of deeper convolution layers to occlusion and pose variations.
- Extensive experiments validate the effectiveness of the proposed model, highlighting its suitability for mobile applications. Our work addresses key challenges in facial expression recognition, excelling in occlusion and pose variations within complex real-world scenarios.

2 DMSNet: Proposed Architecture

The input frames $I_k, k = 1, 2, 3, ..., T$ is passed to the backbone network (ResNet-50) [6], which extracts the low-level backbone spatial features, where T is the total samples and k is batch size. As shown in *Fig.* 2, the proposed DMSNet model is a combination of four RFB blocks ($RFB_i, i = 1, 2, 3, 4$) with feature dimensions 256, 512, 1024, and 2048. These feature dimensions are connected with four encoder blocks ($E_i, i = 1, 2, 3, 4$), and each encoder block is a combination of DeCNet, SCNet, batch normalization layer (BN), and non-linearity operation (ReLU) with the same dimension of RFB. The encoder blocks ($E_i, i = 1, 2, 3, 4$) are connected with four RFB blocks and followed by four decoder blocks ($D_i, i = 1, 2, 3, 4$) and each decoder block is a combination of DeCNet with 3×3 filters, SCNet with 3×3 filters, and transpose convolution layer (TConv2d) 1×1 filters followed by ReLU with the same dimension of RFB. The decoder block outputs are fused to generate the generalized representation of multi-scale geometric spatial features. Further, the MLP is used to classify the objects and predict the correct expression.

2.1 Geometric Multi-Scale Feature Extraction Network

In the proposed DMSNet model, the first input is passed to the ResNet-50 backbone [6] network, which extracts the low-level backbone spatial features with dimensions (64,256,512,1024, 2048) and generates backbone spatial feature maps (X_k). The process is given in *Eq.* 1.

$$X_k^b = \text{ResNet-50}(X_k) \tag{1}$$

The top four blocks of the ResNet-50 are used with dimensions (256, 512, 1024, 2048) to connect the four RFB blocks ($\text{RFB}_i, i = 1, 2, 3, 4$), which generates the multi-scale spatial features. The RFB improves the network's performance by leveraging different scales of features for both input and output spaces, making it more effective at capturing the complex structure of the input data and handling

Table 1: Performance comparison of the proposed method (DMSNet) and twelve SOTA models. The top three results are shown in red > green > blue.

Models yr., cit.	Backbone Network	#Params (M)	FLOPs (M)	Speed (FPS)	FER2013 Acc.(%)	CK+ Acc.(%)	RAF-DB Acc.(%)	MUG Acc.(%)	JAFFE Acc.(%)
RAF-DB$_{20}$ [17]	ResNet-50	99.19	323.2	14.93	58.95	75.72	66.82	76.99	61.50
MA-Net$_{21}$ [20]	ResNet-50	45.91	234.8	19.98	**75.90**	84.59	75.39	85.34	85.42
DACL$_{21}$ [5]	ResNet-18	**11.91**	178.8	13.82	71.53	84.60	87.78	92.32	83.46
KTP$_{21}$ [9]	ResNet-50	89.93	345.5	18.89	69.50	83.23	67.84	76.99	84.45
Deep-Emotion$_{21}$ [12]	ViT	**0.66**	**168.5**	0.71	70.02	98.00	86.11	**94.46**	**92.80**
RUL$_{21}$ [19]	ResNet-18	14.88	181.1	16.35	73.75	89.69	**88.98**	82.64	84.32
FER-VT$_{22}$ [8]	ViT	24.08	282.2	23.09	74.28	88.26	73.14	77.83	85.90
Ad-Corre$_{22}$ [4]	ResNet-50	55.69	528.9	10.34	72.03	86.96	67.90	79.12	83.44
DLAF$_{22}$ [2]	ResNet-50	94.92	356.6	14.89	**75.82**	82.93	71.92	80.34	85.49
MCA-Net$_{23}$ [14]	ResNet-18	95.62	331.6	19.35	72.70	**98.80**	85.90	89.56	**93.33**
LDLVA$_{23}$ [10]	ResNet-50	85.66	287.3	24.32	72.45	89.78	86.99	91.23	85.54
MFER$_{24}$ [18]	ViT	115.69	189.2	22.87	73.49	87.70	87.90	90.29	86.50
DMSNet (Heavy)	ResNet-50	24.00	**140.2**	**25.00**	**76.34**	**100.0**	**95.56**	**94.34**	**94.34**
DMSNet (Light)	VGG-16	**5.6**	**100.8**	**30.00**	74.54	**98.96**	**91.84**	**92.53**	87.84

variable-length sequences. The process is given in *Eq.* 2.

$$X_k^{ms} = \sum_{i=i}^{n} RFB_i(X_k^b) \tag{2}$$

Next, these multi-scale spatial features (X_k^{ms}) are passed to four encoder blocks ($E_i, i = 1, 2, 3, 4$), which extract the multi-scale geometric features from a variable-length sequence and produce a fixed-length vector representation. It performs by enabling it to adapt to the structure of the input data and capture complex features more effectively. This feature is particularly beneficial for handling irregular shapes and varying sizes within the input data, as the deformable encoder incorporates deformable convolutions that can adjust to the input data's structure. By doing so, the encoder can extract more geometric variation of multi-scale spatial feature maps as follows given *Eq.* 3.

$$X_k^e = \sum_{i=i}^{n} E_i(X_k^{ms}) \tag{3}$$

Further, these multi-scale encoder geometric feature maps (X_k^e) passed the RFB block to extract the multi-scale spatial feature maps and enhance it. The procedure is given in *Eq.* 4.

$$X_k^{mse} = \sum_{i=i}^{n} RFB_i(X_k^e) \tag{4}$$

These enhance multi-scale geometric features (X_k^{mse}) are decoded by decoder blocks ($D_i, i = 1, 2, 3, 4$). The decoder decodes multi-scale geometric characterized by irregular shapes, such as images with diverse sizes or objects featuring arbitrary boundaries. This capability is achieved by incorporating deformable convolutions, enabling the decoder to adjust to the structural intricacies of the images dynamically. Consequently, the architecture gains the capacity to learn and adapt its convolution operations based on the inherent structure of the images. The process is given in *Eq.* 5.

$$X_k^d = \sum_{i=i}^{n} RFB_i(X_k^{mse}) \tag{5}$$

Further, the output of four decoder blocks (X_k^d) is fused together using element-wise addition operation ($\oplus$) to generate the generalized representation of multi-scale geometric features X_K^f. The process is given in *Eq.* 6.

$$X_k^f = \sum_{i=i}^{k} (X_k^d) \tag{6}$$

These fused multi-scale geometric spatial (X_k^f) are passed to the Multi-layer Perceptron (MLP), which predicts and classifies the facial expression of the object efficiently. The process is given in *Eq.* 7.

$$\text{Out} = MLP\left(X_k^f\right) \tag{7}$$

Loss Function: The Adam optimizer is used to minimize the Cross-Entropy loss [2], which quantifies the difference between predicted and actual class label distributions. This widely-used loss function is crucial for multi-class classification tasks, penalizing deviations between predicted and true class probabilities. The process is given in *Eq.*8.

$$L_{\text{Soft-max}}(X, Y) = -\frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{\exp(X_{i,Y_i})}{\sum_{j=1}^{C} \exp(X_{i,j})}\right) \tag{8}$$

Where, N is the number of samples, C is the number of classes, X is the predicted logits, and Y is the true class labels.

3 Experiment Setting

Table 2: Ablation studies for the components setting of DMSNet.

Modules	Component Setting RFB	ResNet-50	MLP	FER2013 Acc.%	CK+ Acc.%	RAF-DB Acc.%	MUG Acc.%	JAFFE Acc.%
Conv2d	✓	✓	✓	0.708	0.779	0.768	0.793	0.757
DeCNet	✓	✓	✓	0.725	0.924	0.763	0.891	0.786
SCNet	✓	✓	✓	0.738	0.898	0.863	0.906	0.766
DeCNet + SCNet	✓	✓	✓	**0.745**	**0.989**	**0.918**	**0.925**	**0.878**

Experimental Setup and Datasets

The experiment is performed on Ubuntu 18.04.6 LTS operation system, which has one NVIDIA GPU Quadro P5000/PCIe/SSE2 with 16 GB, 32 GB RAM, and 1 TB Hard Disk. For experiments, we installed Anaconda version 4.12.0, Python Version 3.6, and PyTorch Version 1.7.0, and the GPU is connected with NVIDIA 450 Driver and communicated with CUDA 11.4 and CuDNN 10.2. For the training and testing of the proposed model, we use lr=0.001, batch size=32, Adam optimizer, and cross-entropy loss function. For FER, five benchmark datasets are used: FER2013 [9] with 28,709 training, 3,589 validation, and 3,589 test images; RAF-DB [17] with 12,271 training and 3,068 test images; MUG [1] with 1,259 training and 315 test images; CK+ [17] with 785 training and 196 test images; and JAFFE [5] with 171 training and 42 test images. These datasets collectively serve to evaluate and benchmark FER models.

Table 3: Ablation studies for the components setting of DMSNet.

	# Params	FLOPs	Speed	FER2013	CK+	RAF-DB	MUG	JAFFE
Modules	(M)	(M)	(FPS)	Acc.%	Acc.%	Acc.%	Acc.%	Acc.%
Conv2d	43.0	242.5	12	0.708	0.779	0.768	0.793	0.757
DeCNet	33.4	223.6	17	0.725	0.924	0.763	0.891	0.786
SCNet	23.2	200.4	18	0.738	0.898	0.863	0.906	0.766
DeCNet + SCNet	5.6	100.8	30	**0.745**	**0.989**	**0.918**	**0.925**	**0.878**

3.1 Comparative Analysis

The proposed model DMSNet is compared with twelve state-of-the-art (SOTA) models on five publicly available datasets. It gives 74.54% in FER2013 fourth, 98.96% in CK+ first, 91.84% in RAF-DB first, 92.53% in MUG second, and 87.84% in JAFFE third as shown in *Table* 1. Further, compared with network complexity, our proposed model takes 5.6 million parameters and gives the best prediction result in comparison to Deep-Emotion [12] models, floating point operation (FLOPs) 100.8 million less to all SOTA models and testing speed (inference) is 30 FPS is more to all SOTA models. The results show its capability to be used in mobile applications.

3.2 Ablation Study

The ablation study of the proposed DMSNet model is detailed in *Table* 2, illustrating the effectiveness of each component setting. DMSNet is structured in parallel, densely connected to the ResNet-50 Backbone Network, featuring four encoders and decoders. This parallel architecture proves more effective, surpassing hierarchical visual connection processing. Notably, our proposed model incorporates an attention-based multi-scale geometric spatial and temporal channels mechanism, a pioneering approach to our knowledge. The results indicate that utilizing these attention-based channels outperforms hierarchical visual perception learning. *Table* 2 demonstrates a gradual performance increase and computational complexity decreases as DeCNet and SCNet are used in the proposed model encoder and decoder blocks. Additionally, the comparison between simple Conv2d and both DeCNet and SCNet highlights the proposed solution's inferiority compared to the baseline. *Table* 3 reveals that as components are integrated into the model, the number of parameters and FLOPs decreases due to the feature map dimension reduction. Including, DeCNet and SCNet effectively balance parameters and maintain performance (accuracy).

4 Conclusion and Future Work

This paper addresses the challenge of facial expression recognition in unconstrained scenarios by proposing a novel lightweight Deformable Multi-Scale Network (DMSNet). DMSNet integrates receptive field blocks (RFBs), deformable convolutions, and depth-wise separable convolutions to efficiently extract multi-scale geometric spatial features while minimizing network complexity. The RFBs capture diverse spatial features, deformable convolutions extract geometric information, and depth-wise separable convolutions reduce computational overhead. Encoder and decoder blocks enhance feature quality and preserve the localization of multi-scale receptive fields. Extensive experiments validate the efficacy of each module, demonstrating the model's potential in advancing FER. Future work will focus on developing a real-time, mobile-friendly FER model optimized for edge devices.

References

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. 2010. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, Desenzano del Garda, Italy, 1–4.

[2] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. 2022. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics* 18, 8 (2022), 5619–5627.

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. IEEE, Honolulu, Hawaii, 764–773.

[4] Ali Pourramezan Fard and Mohammad H Mahoor. 2022. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access* 10 (2022), 26756–26768.

[5] Amir Hossein Farzaneh and Xiaojun Qi. 2021. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. IEEE, Virtual, 2402–2411.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Las Vegas, Nevada, 770–778.

[7] Andrew G Howard. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *Proceedings of the IEEE conference on computer vision and pattern recognition* 783 (2017), 654–765.

[8] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. 2021. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences* 580 (2021), 35–54.

[9] Mukku Nisanth Kartheek, Rapolu Madhuri, Munaga VNK Prasad, and Raju Bhukya. 2021. Knight tour patterns: Novel handcrafted feature descriptors for facial expression recognition. In *Computer Analysis of Images and Patterns: 19th International Conference, CAIP 2021, Virtual Event, September 28–30, 2021, Proceedings, Part II 19*. Springer, Virtual Event, 210–219.

[10] Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, and Anh Nguyen. 2023. Uncertainty-aware label distribution learning for facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, Vancouver Canada, 6088–6097.

[11] Songtao Liu, Di Huang, et al. 2018. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, Munich, Germany, 385–400.

[12] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. 2021. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* 21, 9 (2021), 3046.

[13] Trinh Thi Doan Pham, Sesong Kim, Yucheng Lu, Seung-Won Jung, and Chee-Sun Won. 2019. Facial action units-based image retrieval for facial expression recognition. *IEEE Access* 7 (2019), 5200–5207.

[14] Tongping Shen and Huanqing Xu. 2023. Facial Expression Recognition Based on Multi-Channel Attention Residual Network. *CMES-Computer Modeling in Engineering & Sciences* 135, 1 (2023), 539–560.

[15] Hemraj Singh, Mridula Verma, and Ramalingaswamy Cheruku. 2023. DSNet: Efficient Lightweight Model for Video Salient Object Detection for IoT and WoT

Applications. In *Companion Proceedings of the ACM Web Conference 2023*. ACM, Virtual, 1286–1295.

[16] Hemraj Singh, Mridula Verma, and Ramalingaswamy Cheruku. 2024. Dsfnet: video salient object detection using a novel lightweight deformable separable fusion network. *IEEE Transactions on Instrumentation and Measurement* 73 (2024), 1557–9662.

[17] Yuan Xie, Tianshui Chen, Tao Pu, Hefeng Wu, and Liang Lin. 2020. Adversarial graph representation adaptation for cross-domain facial expression recognition. In *Proceedings of the 28th ACM international conference on Multimedia*. ACM, New York, United States, 1255–1264.

[18] Jie Xu, Yang Li, Guanci Yang, Ling He, and Kexin Luo. 2024. Multiscale facial expression recognition based on dynamic global and static local attention. *IEEE Transactions on Affective Computing* 18 (2024), 1–14.

[19] Yuhang Zhang, Chengrui Wang, and Weihong Deng. 2021. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems* 34 (2021), 17616–17627.

[20] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing* 30 (2021), 6544–6556.