DynaEFace: Dynamic Efficient Facial Expression Recognition in the wild

Lynda SAYOUD[1] and Slimane LARABI[1]

RIIMA Laboratory, Faculty of Computer Science, USTHB
Algiers, Algeria
{lsayoud, slarabi}@usthb.dz

Abstract. Dynamic Facial Expression Recognition (DFER) aims to recognize human emotions from unconstrained facial expression videos captured in real-world environments, where variations in pose, illumination, occlusion, and subtle temporal dynamics make the task highly challenging. Although recent Transformer-based approaches have improved temporal modeling, most supervised DFER methods still rely on generic face-recognition backbones, such as MS-Celeb-1M-pretrained ResNet models, which are not specifically designed to capture the fine-grained facial cues required for expression discrimination. To address these limitations, we propose DynaEFace, a lightweight and effective framework for DFER that combines expression-aware spatial representations with refined temporal modeling. Specifically, video frames are first processed using EfficientFace, a backbone tailored for facial expression recognition, and subsequently encoded using a temporal Transformer to model contextual dependencies across frames. To further enhance discriminative feature learning, we introduce a Post-Transformer Channel Gating (PTCG) module that adaptively recalibrates channel-wise responses on contextualized frame tokens. The refined temporal representations are then aggregated to obtain a complementary video-level representation for emotion recognition. Our experiments conducted on the challenging DFEW and FERV39k benchmark datasets show that the proposed method achieves competitive and robust performance while maintaining a lightweight architecture with only $5.87M$ parameters, demonstrating its potential for real-world DFER and HCI applications.

Keywords: Dynamic Facial Expression Recognition · EfficientFace · Channel Gating · Lightweight Deep Learning.

1 Introduction

Facial expressions are among the most immediate forms of human communication. They often reveal emotions before they are articulated in words [11]. Over the past decade, the automatic recognition of facial expressions, known as Facial Expression Recognition (FER), has become a prominent research area due to its diverse real-world applications, including Human-Computer Interaction (HCI) [8,14], clinical assessment [4,27,71], and driver monitoring systems

[65,72], among others [1,49]. Initially, research in this field focused on Static Facial Expression Recognition (SFER), which classifies individual images into basic emotional categories such as happiness, sadness, surprise, fear, disgust, anger, or neutral [75,33,59,67]. Although SFER methods have achieved strong performance on both controlled laboratory datasets (e.g. BU-3DFE [70]) and large-scale "in-the-wild" datasets (e.g. AffectNet [46]), their analysis is restricted to isolated frames. Prior studies [39,68] have shown that natural facial expressions are dynamic, evolving over time through continuous facial muscle movements. In order to capture these temporal dynamics, recent research has shifted toward Dynamic Facial Expression Recognition (DFER), which analyzes video sequences to track the full evolution of an emotion [48,64].

Early DFER studies were conducted under controlled laboratory conditions with frontal faces, uniform lighting, and limited occlusion [42,56,73]. While these settings simplified the task, they limited generalization to real-world scenarios. To address this gap, recent work introduced large-scale in-the-wild benchmarks such as DFEW [26] and FERV39K [62], which capture unconstrained variations in pose, illumination, occlusion, and head motion. Deep learning approaches initially explored CNN–RNN combinations and 3D CNNs to jointly model spatial and temporal cues, but their ability to capture long-range temporal dependencies remained limited. More recently, hybrid CNN–Transformer architectures [74,43,34,61] have emerged as a promising paradigm, where convolutional backbones extract frame-level features that are integrated by temporal Transformers to effectively model long-range dynamics. Beyond architectural advances, intensity-aware supervision schemes [32] have also been proposed to improve sensitivity to subtle expression variations in unconstrained settings. However, despite their strong recognition performance, many CNN–Transformer architectures achieve these gains through increasingly sophisticated temporal modules and additional attention mechanisms, resulting in higher computational complexity and larger parameter counts. In contrast, less attention has been given to the choice of the spatial backbone, with most supervised methods relying on generic CNN architectures, such as ResNet-18 [20] (11.18M parameters), pretrained on large-scale face recognition datasets such as MS-Celeb-1M [16]. Although such pretraining provides robust facial representations, these backbones are primarily designed for identity recognition rather than facial expression analysis. As a result, they may be less effective at capturing the subtle facial movements that distinguish visually similar expressions, such as fear and surprise. Beyond spatial representation learning, a further challenge lies in efficiently aggregating temporal information for video-level prediction. Most existing architectures obtain video-level representations from a single summary feature, such as a class token [43,44], global average pooling [61], or dynamically re-weighted aggregation strategies [34,58]. Although these approaches achieve strong performance, they often overlook fine temporal cues and do not explicitly refine feature channels after temporal encoding. As a result, the representations used for classification may still contain redundant or irrelevant information. Thus, there remains room to enhance the discriminative quality of temporal repre-

sentations while preserving computational efficiency. Therefore, there is a need for lightweight DFER architectures that achieve a balance between recognition accuracy and computational efficiency for deployment.

Motivated by this objective, we propose DynaEFace, a lightweight CNN–Transformer framework for in-the-wild DFER. To the best of our knowledge, this is the first framework that integrates a facial expression-oriented lightweight backbone into a CNN–Transformer pipeline for dynamic facial expression recognition. DynaEFace incorporates EfficientFace [75], originally designed for SFER, to learn discriminative facial expression representations while maintaining low computational complexity. To further improve temporal representation learning, we introduce a Post-Transformer Channel Gating (PTCG) module for channel-wise feature refinement after temporal encoding, together with a temporal attention pooling module that adaptively aggregates informative temporal features for video-level prediction. We evaluate DynaEFace on two large-scale benchmarks, DFEW [26] and FERV39K [62], under fully supervised settings. Experimental results demonstrate that the proposed framework achieves a favorable trade-off between recognition performance and computational efficiency, making it suitable for real-world deployment.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 reports experimental results and ablation studies. Finally, Section 5 concludes the paper.

2 Related work

2.1 From Static to Dynamic FER

Static Facial Expression Recognition (SFER) focuses on classifying emotions from individual images and has been widely studied in affective computing. Early deep SFER approaches primarily relied on convolutional neural networks (CNNs), where methods such as RAN [60] and SCN [59] introduced attention mechanisms to improve robustness to occlusions and noisy labels. These models achieved strong performance on benchmarks like RAF-DB [36] and AffectNet [47]. However, many of these architectures are relatively computationally demanding, which can limit their deployment in real-world scenarios. This has motivated the development of more efficient models, including MobileNetV2 [52] and ShuffleNetV2 [45]. While these models are lightweight, they are not specifically designed for facial expression recognition and may be less effective on challenging in-the-wild data. To address this limitation, EfficientFace [75] introduces a lightweight architecture, with 1.28M parameters, specifically designed for facial expression recognition by combining local facial feature extraction with channel-spatial attention mechanisms. Its AffectNet-7 pretrained weights are specifically optimized for facial expression recognition. Despite these advances, all SFER approaches operate on individual frames and do not capture how facial expressions evolve over time. Dynamic Facial Expression Recognition (DFER) addresses this limitation by modeling facial expressions as temporal sequences in videos. Early approaches relied on handcrafted spatiotemporal descriptors,

such as LBP-TOP [12], STLMBP [24], and HOG-TOP [9], which capture local texture and motion patterns over time. However, these methods have limited capacity when applied to unconstrained, in-the-wild scenarios.

Subsequent works introduced deep learning models based on a CNN-RNN paradigm [6,29,30]. In these approaches, a 2D CNN (e.g., VGGNet [53] or ResNet [21]) extracts frame-level features, which are then passed to recurrent modules such as LSTMs [22] or GRUs [10] to model temporal dynamics. Li et al. [38] incorporate GCNs to model facial landmark relationships, while Baddar et al. [2] propose variational LSTMs robust to unseen variation modes. Despite these improvements, such methods typically separate spatial and temporal modeling, which limits their ability to capture joint spatiotemporal patterns. To overcome this, 3D convolutional networks have been explored to jointly learn spatial and temporal representations directly from video clips, as in C3D [15] and I3D [7]. For DFER tasks, Hasani et al. [19] proposed Enhanced Deep 3D Convolutional Neural Networks with Inception-ResNet layers, while Liong et al. [5] introduced 3D-CNNs augmented with PCA and TMPCA dimensionality reduction for video-based FER. Nevertheless, 3D CNNs remain computationally demanding and struggle with long-range dependencies, motivating subsequent research on Transformer-based architectures for modeling temporal dynamics more efficiently.

2.2 Transformer-Based Architectures in DFER

The success of Transformers in natural language processing [57] and their extension to computer vision [13] has motivated their adoption in DFER, where self-attention enables modeling of long-range temporal dependencies across video frames. Former-DFER [74] was among the first to introduce a Transformer-based framework for this task, employing a dual-branch design in which a Convolutional Spatial Transformer (CS-Former) extracts frame-level spatial features, while a Temporal Transformer (T-Former) captures temporal relationships across the sequence. Ma et al. [43] proposed STT, which consolidates spatial and temporal attention into a unified Transformer encoder. Li et al.[34] introduced NR-DFERNet, which incorporates a Dynamic Class Token to suppress noisy and emotionally irrelevant frame contributions while preserving discriminative expression cues. GCA+IAL [32] further addressed expression intensity variation by combining global convolution attention with an intensity-aware loss function. Wang et al. [58] proposed M3DFEL, which leverages Multi-Instance Learning to handle inexact labels and introduces a dynamic long-term aggregation module to balance short- and long-term temporal relationships. LOGO-Former [44] later improved efficiency by combining local and global attention strategies within a unified Transformer framework.

Despite their architectural diversity, most supervised CNN-Transformer DFER methods follow a similar design paradigm, where a generic CNN backbone extracts spatial features and a Transformer models temporal dependencies. While these architectures have significantly improved recognition performance, they often rely on increasingly sophisticated temporal modules and attention mechanisms, leading to higher computational complexity and increased model size.

In addition, most methods employ generic face-recognition backbones, such as ResNet-18 or ResNet-50, pretrained on MS-Celeb-1M, which are not specifically optimized for facial expression analysis. Consequently, the design of lightweight and expression-oriented spatial representations remains comparatively underexplored in supervised DFER.

Video-level prediction is another important component of Transformer-based DFER frameworks. Existing methods commonly rely on global aggregation strategies, such as CLS-token classification [43,44], mean pooling [61], or attention-based dynamic aggregation [34,58]. More generally, temporal attention mechanisms have been widely explored to emphasize informative frames within video sequences. In particular, additive attention [3] computes learnable importance weights over temporal representations, enabling the model to focus on the most informative temporal cues

2.3 Channel Recalibration and Gating Mechanisms

Channel recalibration mechanisms aim to enhance informative feature channels while maintaining low computational cost. Hu et al. [23] introduced Squeeze-and-Excitation (SE) networks, which capture global context through average pooling and generate channel-wise weights using a bottleneck MLP. Woo et al. [66] later extended this idea with the Convolutional Block Attention Module (CBAM), combining channel and spatial attention. Yang et al. [69] further improved channel modulation through Gated Channel Transformation (GCT), which employs a residual gating formulation to stabilize feature refinement.

Although these approaches were originally developed for convolutional feature maps, Vision Transformers [13] generally lack explicit channel-wise recalibration after self-attention encoding. As a result, the adaptation of channel recalibration mechanisms to Transformer representations remains comparatively underexplored in DFER.

3 Proposed Method

Overview

As illustrated in Figure 1, a facial expression clip is dynamically sampled from the input video and processed by EfficientFace [75] to extract frame-level features. The extracted features are projected into a shared embedding space of dimension D_{emb} to form frame tokens, which are combined with temporal positional embeddings and prepended with a learnable classification token before being fed into a Transformer encoder to capture temporal dependencies across frames and produce contextualized tokens. The resulting frame tokens are subsequently refined by the proposed Post-Transformer Channel Gating (PTCG) module, which performs residual channel-wise recalibration by adaptively re-weighting feature channels according to the global temporal context, allowing expression-relevant channels to be emphasized while reducing the influence of redundant or weakly

informative feature responses. The refined frame tokens are then aggregated using temporal attention pooling to obtain a compact video-level representation. Finally, the pooled representation is fused with the classification token through element-wise addition and fed into a two-layer multilayer perceptron (MLP) for emotion classification.

Input Clips

Our method takes as input a video clip $X \in \mathbb{R}^{T\times H\times W\times 3}$ consisting of T RGB frames with spatial resolution $H \times W$. Following prior work, the input clip is dynamically sampled from the original video. During training, each video is divided into U uniform temporal segments, from which V frames are randomly sampled. During testing, the video is similarly partitioned, while frames are sampled from the mid of each segment. This yields a total of $T = U \times V$ frames for both training and inference.

Each sampled frame x_t is independently processed by EfficientFace [75]. To the best of our knowledge, this is the first work to leverage a facial-expression-specialized backbone for DFER. EfficientFace partitions the input face into four spatial regions processed by depthwise convolutions to capture localized expression cues, while its Modulator employs dual channel and spatial pathways to enhance globally informative features. For each frame x_t, the backbone produces a compact feature vector $f_t \in \mathbb{R}^{D_b}$, which is projected into a shared embedding space to obtain the frame embedding $e_t \in \mathbb{R}^{D_{emb}}$. The resulting frame embeddings are combined with temporal positional embeddings and prepended with a classification token before being fed into the temporal Transformer encoder:

$$H = \text{Transformer}\left([e_{\text{cls}}, e_1, \ldots, e_T]\right) = [z_{\text{cls}}, h_1, \ldots, h_T] \in \mathbb{R}^{(T+1)\times D_{emb}}. \quad (1)$$

where $z_{\text{cls}} \in \mathbb{R}^{D_{emb}}$ denotes the final classification token representation and $h_t \in \mathbb{R}^{D_{emb}}$ denotes the contextualized representation of frame t.

Post-Transformer Channel Gating (PTCG)

The contextualized frame representations $h_t \in \mathbb{R}^{D_{emb}}$ are subsequently refined using a Post-Transformer Channel Gating (PTCG) module that performs adaptive channel-wise recalibration on each frame token. While the Transformer encoder captures temporal dependencies across frames, PTCG further enhances discriminative channel responses by computing a channel gating vector:

$$g_t = \sigma\left(W_2\, \phi\left(W_1\, \text{LN}(h_t)\right)\right), \quad (2)$$

where $\phi(\cdot)$ and $\sigma(\cdot)$ denote the GELU and sigmoid activation functions, respectively. The resulting gating vector is then applied through residual channel-wise modulation:

$$\hat{h}_t = \text{LN}\left(h_t + \gamma(h_t \odot g_t)\right), \quad (3)$$

where γ is a fixed scaling factor set to 0.5 to control the strength of residual channel modulation and keep the update near-identity.

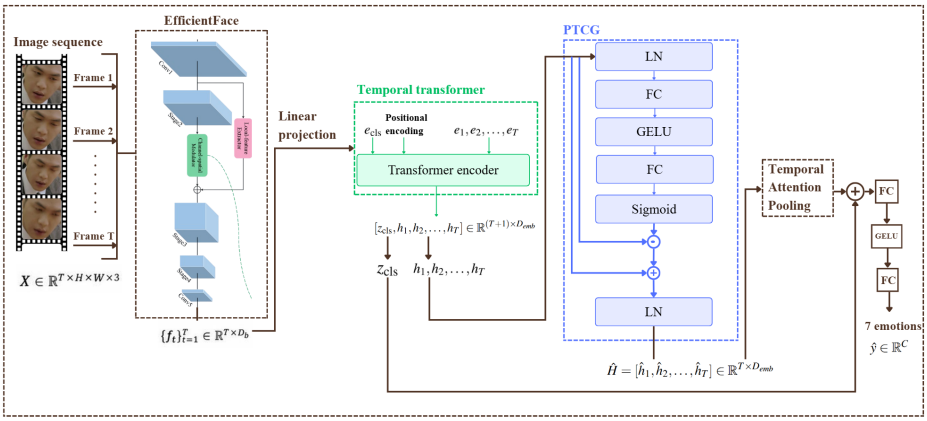

Fig. 1. An overview of the proposed DynaEface framework.

Temporal Attention Pooling

After channel-wise refinement, the sequence of refined frame tokens $\hat{H}$ is aggregated into a compact video representation using a temporal attention pooling mechanism inspired by Bahdanau additive attention [3].

For each frame token, a scalar relevance score is computed:

$$e_t = v^\top \tanh(W_a \hat{h}_t), \tag{4}$$

where W_a and v are learnable parameters. These scores are then normalized across the temporal dimension using a softmax function, producing attention weights:

$$w_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}. \tag{5}$$

The final video representation is obtained by aggregating the refined frame tokens using the computed weights:

$$p = \sum_{t=1}^{T} w_t \, \hat{h}_t, \qquad p \in \mathbb{R}^{D_{emb}}. \tag{6}$$

This pooled representation p is then fused with the global representation $z_{\text{cls}} \in \mathbb{R}^{D_{emb}}$, which encodes global information from the Transformer via element-wise addition followed by layer normalization:

$$z = \text{LN}(z_{\text{cls}} + p), \qquad z \in \mathbb{R}^{D_{emb}}. \tag{7}$$

The final prediction $\hat{y} \in \mathbb{R}^{C}$ is obtained using a two-layer MLP classifier.

4 Experiments and Analysis

We conduct experiments on two widely used in-the-wild DFER datasets, DFEW [26] and FERV39K [62]. We first introduce the datasets and provide implementation details. We then perform ablation studies to evaluate the contribution of each component. Finally, we compare our proposed DynaEFace with several state-of-the-art methods and analyze the results to validate its effectiveness.

4.1 Datasets

DFEW [26] consists of 16,372 in-the-wild video clips collected from over 1,500 movies worldwide. These clips feature various challenging conditions, such as extreme illumination, occlusions, and pose changes. Each clip is annotated multiple times by professional annotators and assigned to one of seven basic expressions: happiness, sadness, neutral, anger, surprise, disgust, and fear. From these, 12,059 clips are single-labeled and split into five non-overlapping subsets. We adopt 5-fold cross-validation as the evaluation protocol, sequentially using one subset for testing and the remaining four for training. Similar to other FER datasets, DFEW is characterized by an imbalanced class distribution.

FERV39K [62] is currently the largest publicly available in-the-wild DFER dataset, containing 38,935 video clips from four main scenarios, further subdivided into 22 fine-grained scenes (e.g., daily life, talk shows, business, and crime). Each clip is independently annotated by 30 professional annotators to ensure high-quality labels and assigned to one of the seven primary expressions, consistent with DFEW. The official benchmark provides fixed, non-overlapping splits: 80% training (31,088 clips) and 20% testing (7,847 clips), which we adopt for all experiments.

4.2 Implementation details

Training Settings. For both DFEW and FERV39K, we use the pre-cropped face regions provided by the dataset authors. All faces are resized to 224×224 pixels to match the input resolution of EfficientFace. During training, standard data augmentation strategies, including random cropping, horizontal flipping, and dataset-specific appearance transformations, are applied to improve generalization. In addition, the EfficientFace backbone is progressively unfrozen during training to better adapt the spatial representations to the temporal modeling stage, with Stage 4 unfrozen at epoch 5, followed by Stage 3 and the Modulator block at epoch 10. For temporal sampling, the dynamically sampled sequence length is set to $T = 16$. Specifically, we use $U = 16$ and $V = 1$ for FERV39K, and $U = 8$ and $V = 2$ for DFEW. The EfficientFace backbone produces frame-level features of dimension $D_b = 1024$, which are projected into an embedding space of dimension $D_{emb} = 256$. The temporal Transformer consists of 4 layers with 8 attention heads. All models are trained for 50 epochs with a batch size of 8

using the AdamW [41] optimizer with a weight decay of 1×10^{-2}. Two parameter groups are employed: the temporal head, including the always-trainable conv5 layer, uses a learning rate of 1×10^{-4}, while the unfrozen backbone parameters are updated with a lower learning rate of 1×10^{-5}. The learning rate is decayed by a factor of 0.1 at epochs 20 and 35 using a MultiStepLR scheduler. Gradient clipping is applied with a maximum ℓ_2 norm of 1.0. We fix the PTCG residual gate scale to $\gamma = 0.5$ to preserve stable residual modulation during channel recalibration. All experiments are implemented in PyTorch [50] and trained on a single NVIDIA GeForce RTX 4070 GPU.

Evaluation Metrics. Following standard DFER evaluation protocols, we report Unweighted Average Recall (UAR), defined as the average recall across all classes, and Weighted Average Recall (WAR), corresponding to the overall classification accuracy [74,34,43,44,32].

4.3 Ablation Studies

To better understand the contribution of each design choice, we conduct three complementary ablation studies on the FERV39K and DFEW benchmarks. We first evaluate the contribution of the proposed architectural components. We then analyze the impact of the spatial backbone on both recognition performance and computational efficiency. Finally, we compare the proposed Post-Transformer Channel Gating (PTCG) module with representative channel recalibration mechanisms.

Contribution of the proposed components. As shown in Table 1, the baseline in Setting (a) consists of the EfficientFace backbone, the temporal Transformer encoder, and the CLS-token classifier. When ATP is introduced alone in Setting (b), the performance remains nearly unchanged: on DFEW, UAR increases only marginally from 58.67% to 58.71%, while on FERV39k it slightly decreases from 38.83% to 38.39%. This suggests that applying adaptive temporal aggregation directly to Transformer features provides limited benefit. In contrast, introducing PTCG with average pooling in Setting (c) improves performance on FERV39k to 38.70% UAR and 48.72% WAR, indicating that channel-wise recalibration enhances the discriminative quality of temporal representations. However, the use of average pooling leads to slightly lower performance on DFEW, which suggests that fixed aggregation is less effective than adaptive temporal weighting. Finally, combining PTCG with ATP in Setting (d) yields the best overall results on both datasets, achieving 39.72% UAR and 48.86% WAR on FERV39k, as well as 59.10% UAR on DFEW. Although Setting (b) obtains a marginally higher WAR on DFEW (70.62% vs. 70.53%), the full model consistently improves UAR, which is more reliable under class imbalance. Overall, these results show that PTCG and ATP are complementary: PTCG refines channel-level feature responses, while ATP selectively aggregates the most informative temporal cues, leading to more discriminative and robust video-level

Table 1. Assessing the contribution of each component of the proposed model on **FERV39k** and **DFEW**. ATP denotes Temporal Attention Pooling. **Bold** indicates the best results.

Setting	PTCG	ATP	FERV39k		DFEW	
			UAR	WAR	UAR	WAR
a (Baseline)			38.83	48.68	58.67	70.53
b		✓	38.39	48.69	58.71	**70.62**
c	✓	Avg	38.70	48.72	58.53	70.51
d (Ours)	✓	✓	**39.72**	**48.86**	**59.10**	70.53

Table 2. Comparison of different backbones integrated into the proposed framework on **FERV39k** and **DFEW**. **Bold** indicates the best results.

Backbone	FERV39k		DFEW		GFLOPs	Params (M)
	UAR	WAR	UAR	WAR		
ResNet-18 (MS-Celeb-1M)	38.77	**49.01**	56.44	69.06	58.503	15.648
EfficientFace (AffectNet-7)	**39.72**	48.86	**59.10**	**70.53**	**5.229**	**5.871**

Backbone analysis. To assess the impact of the spatial backbone, we replace EfficientFace with the widely used ResNet-18 pretrained on MS-Celeb-1M while keeping all other components of the proposed framework unchanged. As reported in Table 2, EfficientFace achieves higher UAR on both FERV39K (39.72% vs. 38.77%) and DFEW (59.10% vs. 56.44%), while also improving WAR on DFEW (70.53% vs. 69.06%). Although ResNet-18 achieves a marginally higher WAR on FERV39K (49.01% vs. 48.86%), EfficientFace consistently yields better class-balanced performance across both benchmarks. Furthermore, it requires only 5.229 GFLOPs and 5.871M parameters, compared with 58.503 GFLOPs and 15.648M parameters for ResNet-18, corresponding to more than an 11-fold reduction in computational cost and approximately a 2.7-fold reduction in parameter count. These results demonstrate that an expression-oriented lightweight backbone offers a more favorable trade-off between recognition performance and computational efficiency, making it well suited for practical DFER deployment.

Channel recalibration analysis. To evaluate the effectiveness of the proposed PTCG module, we compare it with two representative channel recalibration mechanisms, namely SE [23] and GCT [69]. For a fair comparison, each module is inserted at the same post-Transformer position while keeping the remainder of the framework unchanged. As shown in Table 3, SE consistently yields the lowest performance across both datasets, suggesting that a conventional channel attention mechanism is less effective in this setting. Compared with GCT, PTCG achieves the highest UAR on both FERV39K (39.72% vs. 39.66%) and DFEW (59.10% vs. 58.87%). Although GCT obtains slightly higher WAR on both datasets (48.94% vs. 48.86% on FERV39K and 70.85% vs. 70.53% on DFEW), PTCG consistently provides better class-balanced recognition performance, as reflected by UAR. These results suggest that residual channel-

Table 3. Ablation study of channel recalibration modules on FERV39K and DFEW datasets.

Module	FERV39K		DFEW	
	UAR (%)	WAR (%)	UAR (%)	WAR (%)
SE [23]	38.42	48.80	58.04	70.47
GCT [69]	39.66	**48.94**	58.87	**70.85**
PTCG (ours)	**39.72**	48.86	**59.10**	70.53

wise recalibration applied after temporal encoding is more effective for refining Transformer representations than directly applying existing channel recalibration mechanisms originally developed for convolutional feature maps.

4.4 Comparison with State-of-the-Art

Results on DFEW. The comparison results on the DFEW benchmark under the standard 5-fold cross-validation protocol are reported in Table 4. The proposed DynaEFace achieves the best overall performance among all compared methods, obtaining 59.10% UAR and 70.53% WAR with only 5.23 GFLOPs and 5.87M parameters. In terms of UAR, it surpasses M3DFEL [58], AEN [31], and Dual-STI [35] by 3.00%, 2.44%, and 2.21%, respectively. It also achieves the highest WAR, outperforming MSSTNet$_{T=8}$ [61], Dual-STI [35], and NSNPNet [17]. Although M3DFEL [58] requires only 1.65 GFLOPs, its performance remains substantially lower in both UAR and WAR. Regarding per-class accuracy, the proposed method achieves the best results on the Happy and Sad categories, reaching 90.59% and 71.44%, respectively. It also achieves competitive results on the Fear and Disgust categories, with accuracies of 35.92% and 12.41%, respectively. These categories are particularly challenging because they contain fewer training samples in DFEW. Overall, the results indicate that the proposed method generalizes effectively across both frequent and underrepresented expressions.

Results on FERV39K. Table 5 reports the results on the challenging FERV39k benchmark, which is generally more difficult than DFEW due to higher intra-class variability and stronger class imbalance. Our method achieves a UAR of 39.72% and a WAR of 48.86%, outperforming several recent approaches, including Former-DFER [74], NR-DFERNet [34], GCA+IAL [32], Dual-STI [35], NSNPNet [17], and M3DFEL [58]. In particular, it improves over Former-DFER by 1.69% in WAR, and over M3DFEL by 3.78% in UAR and 1.19% in WAR. FERV39k exhibits a highly imbalanced class distribution, where Disgust and Fear account for only 5.89% and 5.4% of the sequences, respectively, which may explain the reduced performance on these categories. In addition, DynaEFace runs at an average of 4.73 ms per clip (0.30 ms per frame) across both benchmarks, confirming its efficiency and suitability for real-time deployment.

Table 4. Comparison with state-of-the-art methods on the DFEW benchmark. ⋆ Temporal interpolation-based sampling. ⋆⋆ Dynamic sampling-based methods. Bold indicates the best result in each column.

Method	Per-Emotion Accuracy (%)							Metrics (%)		GFLOPs	Params
	Hap.	Sad	Neu.	Ang.	Sur.	Dis.	Fear	UAR	WAR		
C3D⋆ [15]	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54	38.57	-
P3D⋆ [51]	74.85	43.40	54.18	60.42	50.99	0.69	23.28	43.97	54.47	-	-
R(2+1)D18⋆	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22	42.36	-
3D-ResNet18⋆ [18]	73.13	48.26	50.51	64.75	50.10	0.00	26.39	44.73	54.98	8.32	-
I3D-RGB⋆ [7]	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27	6.99	-
VGG11+LSTM⋆	76.89	37.65	58.04	60.70	43.70	0.00	19.73	42.39	53.70	31.65	-
ResNet18+LSTM⋆	78.00	40.65	53.77	56.83	45.00	4.14	21.62	42.86	53.08	7.78	31
3D-ResNet18⋆⋆ [18]	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27	8.32	-
ResNet18+GRU⋆⋆	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02	7.78	-
Former-DFER⋆⋆ [74]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70	9.11	146
STT⋆⋆ [43]	87.36	67.90	64.97	71.24	53.10	3.49	34.04	54.58	66.65	-	-
DPCNet⋆⋆ [63]	-	-	-	-	-	-	-	57.11	66.32	9.52	51
NR-DFERNet⋆⋆ [34]	88.47	64.84	70.03	75.09	61.60	0.00	19.43	54.21	68.19	6.33	-
GCA+IAL⋆⋆ [32]	87.95	67.21	70.10	**76.06**	**62.22**	0.00	26.44	55.71	69.24	-	-
M3DFEL⋆⋆ [58]	89.59	68.38	67.88	74.24	59.69	0.00	31.63	56.10	69.25	**1.65**	-
MIDAS⋆⋆ [28]	87.40	67.34	58.64	68.06	59.65	**28.69**	44.50	57.45	69.16	-	-
AEN⋆⋆ [31]	89.24	69.38	70.67	72.08	59.07	4.17	32.00	56.66	69.37	-	-
MSSTNet$^{\star\star}_{T=4}$ [61]	89.57	71.28	69.01	70.92	55.89	11.03	43.68	58.77	69.71	5.05	-
MSSTNet$^{\star\star}_{T=8}$ [61]	89.20	70.12	**72.35**	70.23	59.22	13.10	39.02	59.04	70.16	10.10	-
EST⋆⋆ [40]	-	-	-	-	-	-	-	53.94	65.85	-	-
MSCM⋆⋆ [37]	89.29	71.33	69.61	73.45	57.59	8.28	**39.91**	58.49	70.16	8.10	-
FreqHD [54]	-	-	-	-	-	-	-	44.24	54.98	-	-
DynaEFace (Ours)⋆⋆	**90.59**	**71.44**	70.48	72.02	60.86	12.41	35.92	**59.10**	**70.53**	5.23	**5.87**

4.5 Confusion Matrices

Figure 2 shows the confusion matrix of DynaEFace on FERV39k. The model achieves the highest recognition rates for Happy (68.91%), Neutral (54.29%), and Sad (52.55%), which are among the most frequent and visually distinguishable categories in the dataset. Angry is correctly classified in 49.70% of cases, although confusion with Neutral (22.53%) and Sad (13.92%) remains noticeable, likely due to the subtle and ambiguous nature of spontaneous anger expressions in unconstrained settings. Surprise remains difficult to recognize, with only 26.02% correct predictions and frequent misclassification as Angry (23.82%) and Neutral (20.22%). The most challenging categories are Disgust (12.63%) and Fear (13.92%), which are often confused with other negative expressions, particularly Angry and Neutral. These results indicate that the proposed framework effectively captures dominant expression patterns, while subtle and low-intensity negative emotions remain challenging under real-world conditions.

5 Conclusion

In this paper, we presented DynaEFace, a lightweight CNN–Transformer framework for dynamic facial expression recognition in the wild. By combining an expression-oriented lightweight backbone with the proposed Post-Transformer

Table 5. Comparison (%) with state-of-the-art methods on **FERV39k**. **Bold** denotes the best result.

Method	UAR	WAR	Method	UAR	WAR
C3D [15]	22.68	31.69	Former-DFER [74]	37.20	46.85
P3D [51]	23.20	33.39	LOGO-Former [44]	38.22	48.13
I3D-RGB [7]	30.17	38.78	NR-DFERNet [34]	33.99	45.97
R(2+1)D [55]	31.55	41.28	MIDAS [28]	39.20	47.37
3D-ResNet18 [18]	26.67	37.57	AEN [31]	38.18	47.88
VGG13+LSTM [62]	32.41	43.37	M3DFEL [58]	35.94	47.67
VGG16+LSTM [62]	30.43	41.47	GCA+IAL [32]	35.82	48.54
ResNet18+LSTM [62]	30.92	42.95	Dual-STI [35]	38.27	48.46
Two VGG13+LSTM [62]	32.79	44.54	NSNPNet [17]	37.80	46.40
Two ResNet18+LSTM [62]	31.28	43.20	SFT [25]	35.16	47.80
			FreqHD [54]	32.24	43.93
			DynaEFace (Ours)	**39.72**	**48.86**

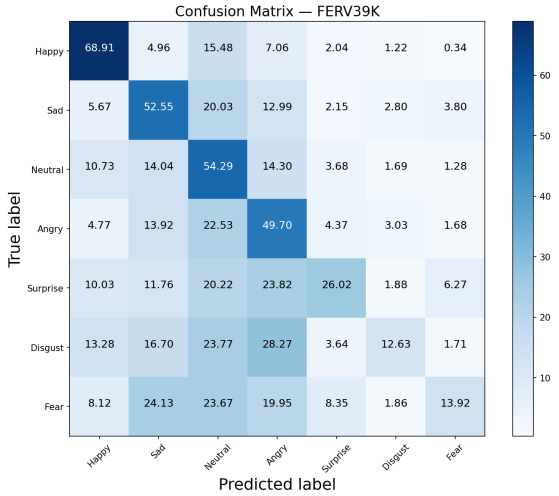

Fig. 2. Experimental results of confusion matrix on FERV39K.

Channel Gating module and temporal attention pooling, the proposed framework achieves an effective balance between recognition performance and computational efficiency. Experimental results on the DFEW and FERV39K benchmarks demonstrate competitive performance while requiring only 5.87M parameters and 5.23 GFLOPs. The ablation studies further validate the contribution of each proposed component. Overall, these results show that accurate in-the-wild DFER can be achieved without relying on computationally expensive architectures, making DynaEFace well suited for deployment in resource-constrained real-world applications. Future work will explore more efficient temporal modeling and lightweight multimodal extensions for robust affective computing.

References

1. Al-Modwahi, A.A.M., Sebetela, O., Batleng, L.N., Parhizkar, B., Lashkari, A.H.: Facial expression recognition intelligent security system for real time surveillance. In: Proc. of world congress in computer science, computer engineering, and applied computing (2012)
2. Baddar, W.J., Ro, Y.M.: Mode variational lstm robust to unseen modes of variation: Application to facial expression recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3215–3223 (2019)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Bishay, M., Palasek, P., Priebe, S., Patras, I.: Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. IEEE Transactions on Affective Computing **12**(4), 949–961 (2019)
5. Byeon, Y.H., Kwak, K.C.: Facial Expression Recognition Using 3D Convolutional Neural Network. International Journal of Advanced Computer Science and Applications **5**(12) (2014)
6. Cai, Y., Zheng, W., Zhang, T., Li, Q., Cui, Z., Ye, J.: Video based emotion recognition using cnn and brnn. In: Chinese conference on pattern recognition. pp. 679–691. Springer (2016)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
8. Chattopadhyay, J., Kundu, S., Chakraborty, A., Banerjee, J.S.: Facial expression recognition for human computer interaction. In: New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India, pp. 1181–1192. Springer (2020)
9. Chen, J., Chen, Z., Chi, Z., Fu, H.: Emotion recognition in the wild with feature fusion and multiple kernel learning. In: Proceedings of the 16th international conference on multimodal interaction. pp. 508–513 (2014)
10. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
11. Darwin, C.: The Expression of the Emotions in Man and Animals. John Murray, London (1872)
12. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International conference on multimodal interaction. pp. 509–516 (2013)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Duric, Z., Gray, W.D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M.J., Schunn, C., Wechsler, H.: Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proceedings of the IEEE **90**(7), 1272–1289 (2002)
15. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: Proceedings of the 18th ACM international conference on multimodal interaction. pp. 445–450 (2016)
16. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)

17. Han, Z., Meichen, X., Hong, P., Zhicai, L., Jun, G.: Nsnp-dfer: a nonlinear spiking neural p network for dynamic facial expression recognition. Computers and Electrical Engineering **115**, 109125 (2024)
18. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
19. Hasani, B., Mahoor, M.H.: Facial expression recognition using enhanced deep 3d convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 30–40 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
23. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
24. Huang, X., He, Q., Hong, X., Zhao, G., Pietikainen, M.: Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In: Proceedings of the 16th international conference on multimodal interaction. pp. 514–520 (2014)
25. Huang, Z., Zhu, Y., Li, H., Yang, D.: Dynamic facial expression recognition based on spatial key-points optimized region feature fusion and temporal self-attention. Engineering Applications of Artificial Intelligence **133**, 108535 (2024)
26. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2881–2889 (2020)
27. Jin, B., Qu, Y., Zhang, L., Gao, Z.: Diagnosing parkinson disease through facial expression recognition: video analysis. Journal of medical Internet research **22**(7), e18697 (2020)
28. Kawamura, R., Hayashi, H., Takemura, N., Nagahara, H.: Midas: Mixing ambiguous data with soft labels for dynamic facial expression recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6552–6562 (2024)
29. Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Transactions on Affective Computing **10**(2), 223–236 (2017)
30. Kuo, C.M., Lai, S.H., Sarkis, M.: A compact deep learning model for robust facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2121–2129 (2018)
31. Lee, B., Shin, H., Ku, B., Ko, H.: Frame level emotion guided dynamic facial expression recognition with emotion grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5681–5691 (June 2023)
32. Li, H., Niu, H., Zhu, Z., Zhao, F.: Intensity-aware loss for dynamic facial expression recognition in the wild. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 67–75 (2023)

33. Li, H., Sui, M., Zhao, F., Zha, Z., Wu, F.: Mvt: Mask vision transformer for facial expression recognition in the wild. arXiv preprint arXiv:2106.04520 (2021)
34. Li, H., Sui, M., Zhu, Z., et al.: Nr-dfernet: Noise-robust network for dynamic facial expression recognition. arXiv preprint arXiv:2206.04975 (2022)
35. Li, M., Zhang, X., Fan, C., Liao, T., Xiao, G.: Dual-sti: Dual-path spatial-temporal interaction learning for dynamic facial expression recognition. Information Sciences **678**, 120953 (2024)
36. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2852–2861 (2017)
37. Li, T., Chan, K.L., Tjahjadi, T.: Multi-scale correlation module for video-based facial expression recognition in the wild. Pattern Recognition **142**, 109691 (2023)
38. Liu, D., Zhang, H., Zhou, P.: Video-based facial expression recognition using graph convolutional networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 607–614. IEEE (2021)
39. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1749–1756 (2014)
40. Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., Zhan, Y.: Expression snippet transformer for robust video-based facial expression recognition. Pattern Recognition **138**, 109368 (2023)
41. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. International Conference on Learning Representations (ICLR) (2019), https://arxiv.org/abs/1711.05101
42. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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. pp. 94–101. IEEE (2010)
43. Ma, F., Sun, B., Li, S.: Spatio-temporal transformer for dynamic facial expression recognition in the wild. arXiv preprint arXiv:2205.04749 (2022)
44. Ma, F., Sun, B., Li, S.: Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
45. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)
46. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE transactions on affective computing **10**(1), 18–31 (2017)
47. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE transactions on affective computing **10**(1), 18–31 (2017)
48. Nie, X., Fei, Z., Zhou, W., Fei, M.: A review of dynamic facial expression recognition: Methods, datasets and directions. In: International Conference on Machine Learning and Intelligent Computing. pp. 171–180. PMLR (2025)
49. Pabba, C., Kumar, P.: An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Systems **39**(1), e12839 (2022)

50. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
51. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541 (2017)
52. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
54. Tao, Z., Wang, Y., Chen, Z., Wang, B., Yan, S., Jiang, K., Gao, S., Zhang, W.: Freq-hd: An interpretable frequency-based high-dynamics affective clip selection method for in-the-wild facial expression recognition in videos. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 843–852 (2023)
55. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
56. Valstar, M., Pantic, M., et al.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect. vol. 10, p. 65. Paris, France. (2010)
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
58. Wang, H., Li, B., Wu, S., Shen, S., Liu, F., Ding, S., Zhou, A.: Rethinking the learning paradigm for dynamic facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17958–17968 (2023)
59. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6897–6906 (2020)
60. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)
61. Wang, L., Kang, X., Ding, F., Nakagawa, S., Ren, F.: Msstnet: A multi-scale spatio-temporal cnn-transformer network for dynamic facial expression recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3015–3019. IEEE (2024)
62. Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., Zhang, W.: 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. pp. 20922–20931 (2022)
63. Wang, Y., Sun, Y., Song, W., Gao, S., Huang, Y., Chen, Z., Ge, W., Zhang, W.: Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In: Proceedings of the 30th ACM international conference on multimedia. pp. 101–110 (2022)
64. Wang, Y., Yan, S., Liu, Y., Song, W., Liu, J., Chang, Y., Mai, X., Hu, X., Zhang, W., Gan, Z.: A survey on facial expression recognition of static and dynamic emotions. arXiv preprint arXiv:2408.15777 (2024)

65. Wilhelm, T.: Towards facial expression analysis in a driver assistance system. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–4 (2019). https://doi.org/10.1109/FG.2019.8756565
66. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
67. Xue, F., Wang, Q., Guo, G.: Transfer: Learning relation-aware facial expression representations with transformers. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 3601–3610 (2021)
68. Yang, P., Liu, Q., Metaxas, D.N.: Boosting encoded dynamic features for facial expression recognition. Pattern Recognition Letters **30**(2), 132–139 (2009)
69. Yang, Z., Zhu, L., Wu, Y., Yang, Y.: Gated channel transformation for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11794–11803 (2020)
70. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)
71. Yolcu, G., Oztel, I., Kazan, S., Oz, C., Palaniappan, K., Lever, T.E., Bunyak, F.: Facial expression recognition for monitoring neurological disorders based on convolutional neural network. Multimedia Tools and Applications **78**(22), 31581–31603 (2019)
72. Zhang, Y., Hua, C.: Driver fatigue recognition based on facial expression analysis using local binary patterns. Optik **126**(23), 4501–4505 (2015)
73. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image and vision computing **29**(9), 607–619 (2011)
74. Zhao, Z., Liu, Q.: Former-dfer: Dynamic facial expression recognition transformer. In: Proceedings of the 29th ACM international conference on multimedia. pp. 1553–1561 (2021)
75. Zhao, Z., Liu, Q., Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3510–3519 (2021)