

Article

Lightweight Dynamic Perception Facial Expression Recognition for Power Business Hall Scenarios

Yuemin Qiu ¹, Gang Li ¹, Yun Chen ¹, Yanli Yang ¹, Zhen Zhong ¹, Haoran Huang ^{2,*} and Yuming Bo ²

¹ Nanjing Suyi Industrial Co., Ltd., Nanjing 210008, China; qiuyuemin@jit.edu.cn (Y.Q.); gangli@jit.edu.cn (G.L.); yunchen@jit.edu.cn (Y.C.); 05414@jit.edu.cn (Y.Y.); 2115011047@jit.edu.cn (Z.Z.)

² School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China; byuming@mail.njust.edu.cn

* Correspondence: 319110020173@njust.edu.cn

Abstract

Aiming at the problems of low facial expression recognition accuracy, high computational complexity, and difficulty in edge deployment caused by variable illumination, diverse head poses, severe partial occlusions, and the long-tail distribution of expression categories in complex power business hall scenarios, this paper presents an engineering-oriented lightweight dynamic perception facial expression recognition method for complex power business hall scenarios. The proposed method adopts an end-to-end joint face detection and expression classification framework built upon the anchor-free CenterNet architecture. In terms of technical implementation, this paper primarily focuses on the integration and adaptation of existing techniques to meet scenario-specific requirements; integrates a specially designed Multi-Scale Fusion Deformable Large Kernel Attention (MSF-DLKA) module to enhance multi-scale perception of subtle expression deformations under varying poses; designs a Task-Aware Dynamic Detection Head (TADDH) with decoupled spatial-channel attention to separately adapt the feature requirements of localization and classification subtasks; directly employs the off-the-shelf Label-Distribution-Aware Margin loss (LDAM Loss) to alleviate class imbalance; and combines three established compression techniques—structured pruning, quantization-aware training, and knowledge distillation—into a progressive lightweight pipeline for edge deployment. Experiments on two public micro-expression datasets, CASME II and SAMM, as well as a self-built electric power business hall scenario dataset (PBHD), show that the proposed method achieves mAP@0.5 of 93.2%, 93.9%, and 91.7%, and macro-F1 of 0.929, 0.936, and 0.914 on the three datasets, respectively, while using only 13.5 M parameters and attaining a single-image inference time of 28.9 ms on the NVIDIA Jetson AGX Orin. The paper provides a feasible solution for real-time expression perception in resource-constrained scenarios such as power business halls.

Academic Editor: Eva Cernadas

Received: 4 June 2026

Revised: 30 July 2026

Accepted: 1 August 2026

Published: 3 August 2026

Keywords: facial expression recognition; power business hall; dynamic perception; attention mechanism; lightweight network

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Facial expressions, as the most natural and direct non-verbal signals of humans, carry rich emotional information and psychological states, playing an irreplaceable role in interpersonal interaction and emotion understanding [1]. Facial Expression Recognition

(FER) technology has become a research hotspot in computer vision and pattern recognition, showing broad application prospects in smart healthcare, human–computer interaction, public safety, and service domains [2,3]. As typical public service places, power business halls have high customer flow and frequent interactions. Achieving efficient and accurate facial expression recognition in such complex scenarios holds important theoretical value and practical significance for improving the intelligence level of services and optimizing customer experience.

Existing FER methods are mainly divided into traditional methods based on hand-crafted features and deep learning-based methods. Traditional methods mainly rely on manually designed feature descriptors such as LBP, HOG, SIFT to extract corresponding features from images, and then combine classifiers like SVM or AdaBoost to achieve expression classification [4]. However, the feature representation capability of such methods is limited, and their robustness to interfering factors such as illumination changes, head poses, and local occlusions is insufficient, making it difficult to meet actual deployment requirements in real-world scenarios.

In recent years, with the rapid development of computer vision and artificial intelligence technologies, deep learning-based methods have made remarkable progress. Researchers have explored the FER task from various perspectives [5]. In terms of feature extraction and network architecture, Lan et al. [6] introduced dense connections to enhance feature propagation and fusion and used the Gini index to dynamically allocate multi-branch decision weights, solving the problem of insufficient utilization of multi-scale features. Yu et al. [7] separated channel attention and spatial attention and introduced a soft threshold mechanism to suppress noise, improving the network's ability to learn channel features and cross-channel correlations. Zheng et al. [8] fused a bilinear mixed attention module and designed an adaptive pooling weight algorithm to alleviate feature loss caused by traditional pooling. Luo [9] proposed a model based on a feature-enhanced convolutional network, combined with a multi-task convolutional neural network for face alignment, effectively improving the accuracy of expression recognition in dynamic environments. For fine-grained expression recognition, Liang et al. [10] proposed a multi-scale action unit-based network, MSAU-Net, providing a new benchmark for fine-grained FER. Chen et al. [11] filtered key frames via frequency amplitude, and fused self-attention with graph convolutional networks to capture subtle facial muscle movements from both local nodes and holistic parts. Wang et al. [12] designed a lightweight micro-attention mechanism that guides the network to focus on facial action unit regions without significantly increasing parameters. He et al. [13] addressed occlusion and pose variation in complex environments by using multi-scale feature fusion and feature decoupling to capture discriminative subtle differences, providing a new solution for fine-grained facial expression recognition. In terms of robustness for expression recognition, Huang et al. [14] proposed a hybrid augmentation strategy FERMix and mid-level representation learning, improving occlusion robustness while maintaining model efficiency. Liu et al. [15] generated pose-invariant expression images via a geometry-aware conditional network and used non-local operations to capture long-range dependencies, effectively distinguishing confusable expressions. Zhang et al. [16] used facial landmarks as geometric guidance to disentangle identity, pose, and expression via generative adversarial networks, improving recognition robustness under pose variations. Zhu et al. [17] proposed a dual-path stacked attention network that segments key facial regions and combines global and local features, effectively suppressing identity-related intra-class differences and further improving recognition accuracy on datasets.

For few-shot and model lightweighting, Fang et al. [18] proposed a deep joint learning framework ACNN-EC that generates pseudo-labels via expression-guided clustering to alleviate the few-shot problem. Talukder et al. [19] proposed a lightweight CNN

expression recognition system. Chang et al. [20] introduced the Mish activation function, CA attention mechanism, and CARAFE upsampling operator to improve detection accuracy while maintaining lightweightness. Xue et al. [21], based on MobileNetV3, introduced a feature pyramid module for multi-scale feature fusion and drew on the DenseNet structure to enhance feature reuse, improving expression recognition accuracy while keeping the model lightweight. Ge [22] constructed an intelligent emotion recognition system based on machine learning and clustering algorithms using an improved minimum redundancy maximum relevance algorithm and a hybrid neural network. Jiang et al. [23] systematically reviewed new advances in lightweight networks for facial expression recognition, deeply analyzing the application potential of MobileNet, ShuffleNet, and various model compression methods in FER tasks, highlighting the importance of efficient FER in resource-constrained scenarios. Lin et al. [24] introduced an ultra-lightweight dual attention module into ShuffleNetV2, combining channel and spatial attention, and optimized feature propagation via shortcut connections, further improving the facial expression recognition performance of lightweight models. Karani et al. [25] designed a lightweight model FER-BHARAT tailored for the facial expression characteristics of the Indian population, achieving superior accuracy and efficiency over baseline models and large pre-trained models on a culturally specific dataset.

In addition, Wan et al. [26] combined motion magnification with Transformer to enhance the extraction of motion amplitude representations in micro-expressions. Ye et al. [27] proposed explicit visual center encoding and implicit sparse attention center loss to cope with the challenges of inter-class similarity and intra-class diversity. Xia et al. [28] proposed a relation-aware convolutional neural network that adaptively mines the intrinsic correlations between regions such as eyes and mouth and expression categories. Chen et al. [29] reconstructed the SSD structure and incorporated center loss optimization to improve the robustness and discrimination of expression detection. Wang et al. [30] introduced the visual foundation model DINOv2 into FER tasks for the first time, using its frozen weights to extract highly generalizable image features and complementing facial detail features extracted by IR50 via bidirectional attention. Liu et al. [31] constructed a channel-spatial multi-scale enhancement structure to capture holistic multi-scale semantics and designed an efficient channel-spatial attention mechanism to obtain region-wise dual-pooling salient semantics, further validating the effectiveness of multi-scale features and attention mechanisms in improving expression recognition robustness in natural scenes. Hou et al. [32] introduced facial expression recognition technology into online teaching quality assessment, providing teachers with feedback by analyzing students' classroom emotional states in real time. Gloor et al. [33] found in video conferencing scenarios that speakers who elicit diverse emotional fluctuations from listeners receive higher presentation scores, providing useful references for human–computer interaction and customer emotion perception in power business halls. Zhao et al. [34], from a hardware implementation perspective, proposed an expression recognition method based on a memristor convolutional neural network, mapping network weights to memristor conductance values, offering a new approach for efficient deployment of expression recognition models on edge devices.

Despite the aforementioned advances, existing methods still share several systemic limitations: most research focuses on general FER or lab-controlled micro-expression recognition, with scarce dedicated studies for public service scenarios (e.g., power business halls, banks, government centers) where illumination variations, pose diversity, and occlusions are under-modeled. Although attention mechanisms and multi-scale fusion are widely used, few methods systematically address the long-tail class distribution—neutral/happy samples dominate while negative expressions (anger, disgust) are rare, causing poor performance on critical emotions. Recent works on long-tail losses, meta-learning

few-shot FER, and hard example mining offer valuable references, yet their integration into lightweight real-time frameworks for service scenarios remains unexplored. Moreover, existing lightweight networks reduce parameters but often sacrifice accuracy, failing to balance high precision with efficient edge inference. The power business hall scenario presents four main challenges: (1) variable illumination conditions from different lighting sources and times of day; (2) diverse head poses as customers naturally interact with service staff; (3) long-tail distribution of expression categories, where neutral and happy expressions dominate while negative expressions are rarely observed; and (4) the limited computing resources of edge devices limit the real-time response capability of expression recognition, which further affects service efficiency and customer experience. These challenges collectively motivate our work.

To address the above issues, this paper takes the Anchor-Free object detection framework CenterNet [35] as the foundation and proposes a lightweight dynamic perception expression recognition method for power business hall scenarios. The main contribution of this paper lies in the scenario-driven integration and adaptation of multiple existing techniques to address the multifaceted challenges of complex power business hall environments. Specifically, the engineering contributions include: (1) the design of a Multi-Scale Fusion Deformable Large Kernel Attention (MSF-DLKA) module, which enhances the perception of subtle expression deformations under varying poses through multi-scale offset generation and grouped deformable convolution; (2) the design of a Task-Aware Dynamic Detection Head (TADDH) with decoupled spatial-channel attention, which assigns independent attention parameters to localization and classification subtasks to mitigate the inherent conflict between the two tasks; (3) the direct adoption of the off-the-shelf Label-Distribution-Aware Margin (LDAM) loss [36] to alleviate long-tail class distribution without introducing extra parameters or structural modifications; and (4) the combination of three established compression techniques—structured pruning, quantization-aware training, and knowledge distillation—into a progressive engineering pipeline to meet edge-device deployment constraints. Rather than proposing theoretical innovations in network algorithms, this paper presents an engineering integration solution for power business hall scenarios, built entirely upon existing techniques. Experimental results on two public datasets, CASME II [37] and SAMM [38], and a self-built power business hall dataset demonstrate that the proposed method achieves competitive detection accuracy with substantially reduced computational complexity, offering a feasible technical solution for real-time emotion perception in power business halls.

2. Methods

2.1. Method Overview

The proposed method formulates facial expression recognition as a joint face detection and expression classification task. Given an input image, the network simultaneously predicts bounding boxes for all detectable faces and assigns each detected face an expression category label from the predefined set (Neutral, Happy, Disgust, Anger for the PBHD, or the corresponding categories for CASME II and SAMM). The overall structure of the lightweight dynamic perception expression recognition method for power business hall scenarios proposed in this paper is shown in Figure 1. The network uses CenterNet as the base detection framework. The backbone network embeds the MSF-DLKA module to enhance multi-scale deformation perception capability. The detection head adopts a TADDH, which achieves dynamic focusing of localization and classification features through parallel decoupled regression and classification branches and spatial-channel dual attention. The loss function introduces the off-the-shelf LDAM Loss to alleviate the long-tail class distribution problem. Finally, progressive engineering lightweight

compression is applied via structured channel pruning, quantization-aware training, and knowledge distillation-assisted fine-tuning to ensure efficient inference on edge devices.

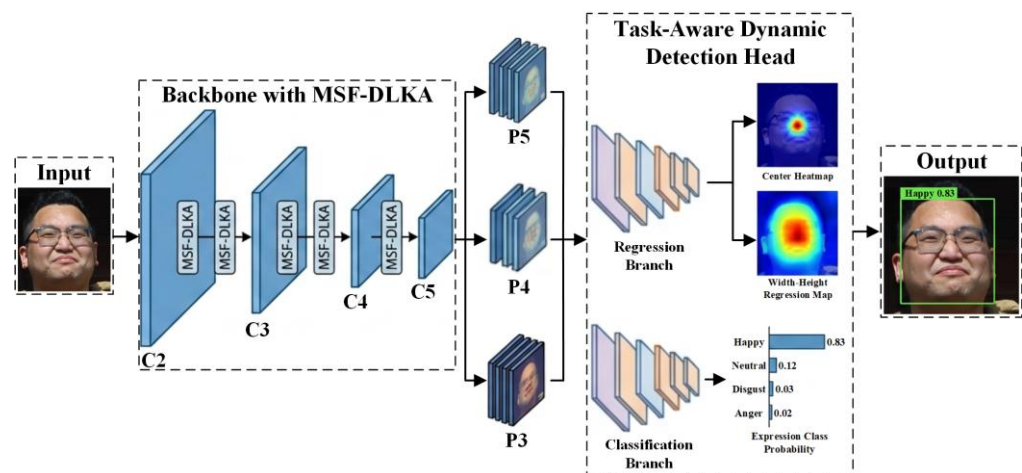


Figure 1. Overall structure of the network.

2.2. MSF-DLKA Module

In facial expression recognition, expression changes are usually caused by tiny movements of facial muscles, with key information concentrated in local areas such as eye corners, eyebrows, mouth corners, and nasal wings. Although traditional convolutional neural networks extract features effectively through hierarchical architectures, the fixed geometric structure of standard convolution kernels restricts their ability to accommodate the diverse poses and non-rigid deformations induced by facial expressions. Although the deformable large kernel attention module combines deformable convolution with large kernel attention to enhance modeling of local deformations and global context to some extent, it still has the following limitations: the offset field is generated by single-scale convolution with a single receptive field, making it difficult to capture fine deformations in multi-scale regions; the attention weights are generated statically, lacking dynamic adaptability to input content; and the two cascaded deformable convolution modules introduce some computational redundancy.

To enhance the model's adaptability to subtle expression changes while maintaining lightweightness, this paper designs the MSF-DLKA module, which integrates multi-scale offset field generation, grouped deformable convolution, and dynamic weight modulation. The design rationale of MSF-DLKA is as follows: Multi-scale offset generation with dilation rates 1, 3, and 5 enables the capture of both local fine-grained deformations and larger-scale pose variations simultaneously. The dilation rates of 1, 3, and 5 are empirically chosen as a heuristic configuration to cover three complementary receptive field scales: dilation = 1 captures local fine details (e.g., eye corner wrinkles and subtle mouth movements); dilation = 3 captures medium-range contextual information (e.g., mouth shape changes and nasolabial folds); and dilation = 5 captures global facial structure and overall pose variations. This combination outperformed alternatives such as 1, 2, 4 and 1, 3, 6 in preliminary validation and is thus adopted as the default configuration. We acknowledge that this set may not be globally optimal, yet it provides a reasonable multi-scale coverage for the current scenario. The grouped deformable convolution with channel shuffle reduces computational cost while maintaining cross-group information flow. Dynamic weight modulation allows content-adaptive feature enhancement, making the attention weights sample-dependent rather than fixed. The residual connection preserves the original feature information and facilitates gradient flow during training. The structure of MSF-DLKA is shown in Figure 2, and its overall computation flow is as follows:

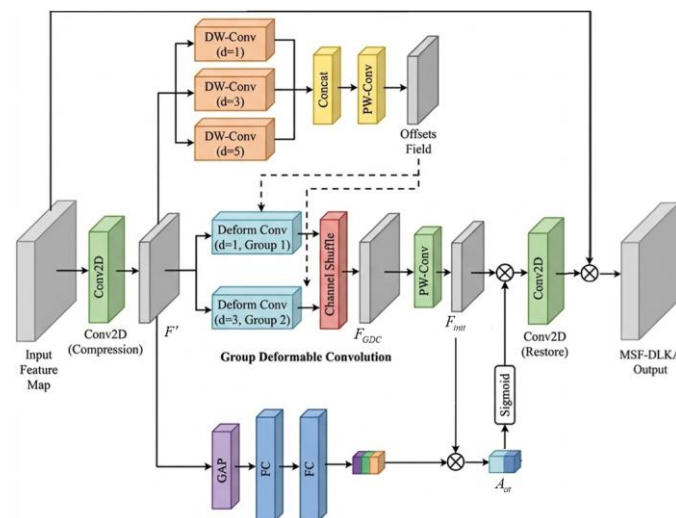


Figure 2. Structure of the MSF-DLKA module.

First, the input feature map passes through a standard convolution layer (Conv2D) for channel compression and preliminary feature transformation, obtaining an intermediate feature representation F' . This intermediate feature representation serves as the basis for subsequent attention computation and is also used to generate multi-scale offset fields.

To perceive expression deformations in regions of different scales, this paper designs a multi-scale offset field generation branch. This branch feeds the input feature map into three depthwise separable convolutions (DW-Conv) with different dilation rates (dilation rates = 1, 3, 5) to extract multi-scale context information. The outputs of the three branches are concatenated along the channel dimension and then fused via pointwise convolution (PW-Conv) to generate the final offset field.

Subsequently, the intermediate feature F' is fed into a grouped deformable convolution module. This module evenly divides the intermediate features into two groups along the channel dimension, performing deformable convolutions (Deform Conv) with different dilation rates (dilation rates = 1 and 3) on the two groups. The outputs of the two groups are subjected to channel shuffle to facilitate information exchange. The channel grouping strategy reduces the number of parameters by half compared to using a single large deformable convolution, while the different dilation rates allow the two groups to focus on different scales. The channel shuffle operation enables information exchange between groups without introducing additional computation. This design significantly reduces the parameter count and computational cost while maintaining multi-scale perception capability. The output feature map F_{GDC} of the grouped deformable convolution integrates multi-scale deformation information and further passes through a pointwise convolution to generate an initial attention feature map F_{init} .

To enhance the adaptability of attention weights to input content, this paper also introduces a dynamic weight modulation mechanism, enabling the model to adaptively enhance or suppress responses in different regions according to the expression characteristics of the current sample, thereby improving discriminability. This mechanism performs global average pooling (GAP) on the intermediate features to obtain a global semantic vector, which is then passed through two fully connected layers (FC) to generate a set of dynamic modulation coefficients. These coefficients are used to perform element-wise scaling on the initial attention weight map, yielding the final attention weight A_{at} .

Finally, the attention weight A_{at} is normalized by Sigmoid and then multiplied element-wise with the initial attention feature map F_{init} to enhance key features and

suppress redundant features. The result is then passed through a standard convolution for channel restoration and added to the module input via a residual connection, forming the complete output of the MSF-DLKA module.

In summary, through the organic combination of multi-scale offset generation, grouped deformable convolution, and dynamic weight modulation, the MSF-DLKA module enhances the perception of subtle expression features while maintaining low parameter count and computational complexity, thereby providing more discriminative feature representations for subsequent detection heads. It should be noted that the operations in this module are all implemented based on existing techniques. The contribution of this paper lies in the scenario-oriented configuration and adaptation of these operations for multi-scale deformation perception in facial expression recognition, rather than in proposing new convolutional operators or attention mechanisms.

2.3. TADDH

In power business hall scenarios, face region localization and expression category classification constitute two highly related yet heterogeneous subtasks. The localization task relies on spatial detail information to accurately regress bounding boxes, while expression classification requires high-level semantic features to capture subtle patterns of muscle deformation. Traditional detection heads usually use a single shared feature for joint prediction, which not only struggles to simultaneously meet the requirements of both tasks at the feature abstraction level, but also easily causes task conflicts under complex backgrounds, occlusions, and various poses, leading to accumulation of localization bias and classification errors. To this end, this paper designs a TADDH. The core design principle is task-specific feature recalibration through decoupled branches, each equipped with independently parameterized spatial and channel attention mechanisms. Unlike traditional detection heads, TADDH assigns independent attention parameters to the regression and classification branches, allowing each to learn its own optimal feature representation. This decoupling is a task-specific engineering adaptation of existing spatial-channel attention mechanisms rather than a novel attention formulation, but it effectively resolves the inherent conflict between the two tasks in our scenario.

The overall structure of TADDH is shown in Figure 3. It consists of two parallel task-aware branches, performing bounding box regression and facial expression classification, respectively. The two branches share the feature maps enhanced by the backbone network and the MSF-DLKA module, but their internal attention parameters are completely independent, achieving partial decoupling of localization and classification features. In the regression branch, the model needs to output a center point heatmap and a width-height regression map, requiring fine representation of the spatial position and scale of the target. To this end, this branch first performs average pooling along the channel dimension on the input feature map, compressing it into a single-channel spatial descriptor. Descriptors from different levels are concatenated and passed through a 7×7 large convolution kernel and Sigmoid activation to generate a spatial selective attention map. The use of a 7×7 large convolution kernel in this branch is a heuristic design choice intended to help capture the overall spatial continuity of the face region and suppress interference from background clutter and local occlusions. In preliminary experiments, we compared 5×5 , 7×7 , and 9×9 kernels; 7×7 achieved the highest mAP@0.5 on the validation set with moderate computational cost. The feature map weighted by spatial attention then enters a channel adaptive attention module, which obtains a channel descriptor via global average pooling, and then generates a channel weight vector through two 1×1 convolutions and a Sigmoid function, adaptively enhancing channels beneficial for boundary regression, such as those rich in edge and contour information, while suppressing channels related to texture or semantics that are irrelevant to localization. Spatial and channel attention are cascaded in

a residual manner, enabling the regression branch to dynamically focus on spatial regions and key feature channels of the target while retaining original feature information.

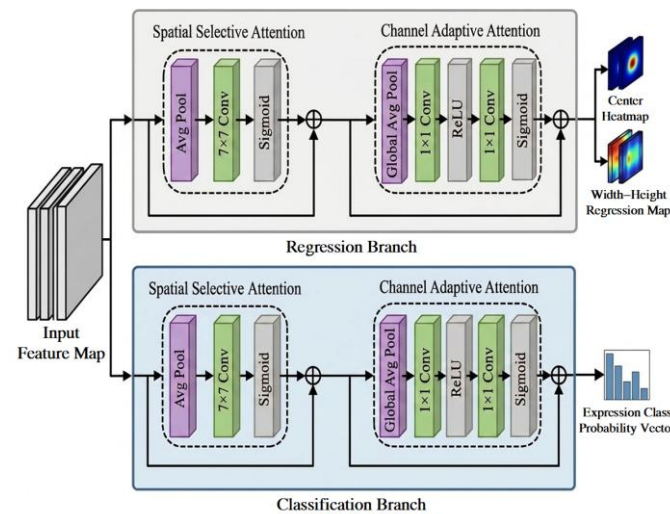


Figure 3. Structure of the TADDH.

The classification branch has the same structure as the regression branch, but its parameters are independent. The expression classification task requires the model to capture subtle expression deformations, such as mouth corner lifting, frown lines, and nasal wing dilation, from local details. Such discriminative information is often scattered in a few feature channels and is easily disturbed by individual differences and illumination changes. The spatial selective attention of the classification branch uses the same large convolution kernel mechanism to learn a spatial attention map centered on key facial areas, helping the model focus attention on the most discriminative areas for expressions, such as eyes, mouth, and nasal wings. The subsequent channel adaptive attention module dynamically increases the weights of sensitive channels encoding subtle deformation information according to the expression characteristics of the input sample, while suppressing channels sensitive to identity, illumination, and other irrelevant factors, enabling the classification branch to obtain more discriminative and invariant feature representations.

The outputs are respectively mapped to their final task spaces via separate convolutional layers: the regression branch outputs the center point heatmap and width-height regression map, while the classification branch outputs a probability vector for each center point over the expression categories. Given the inherent differences in feature utilization between regression and classification subtasks, TADDH assigns targeted feature recalibration capabilities to the two subtasks through independent spatial and channel attention mechanisms, effectively reducing negative interference between tasks and endowing the detection head with content-adaptive dynamic perception. This enables simultaneous high-precision face localization and expression classification in complex and variable power business hall scenarios. It should be clarified that the spatial and channel attention mechanisms themselves are established techniques. The contribution of this work lies in their parallel decoupled integration within the detection head and the assignment of independent learnable parameters for localization and classification tasks respectively.

2.4. Loss Function Optimization Based on LDAM Loss

In facial expression recognition tasks, imbalanced class distribution and varying sample difficulty are often key bottlenecks limiting model generalization. Taking the typical application scenario of power business halls as an example, common positive emotional states such as neutral and polite smiles occupy the vast majority of the sample space, while

negative emotional states such as anger are rare, showing a significant long-tail distribution. Moreover, interfering factors such as illumination changes, facial occlusions, head poses, and individual differences further amplify the sample difficulty variation within the same expression category, making some hard samples more prone to misclassification. The traditional cross-entropy loss function assigns the same weight to all training instances, failing to distinguish sample importance, which tends to bias the model toward predicting all samples as the majority positive emotions, thereby severely impairing recognition ability for minority classes such as negative emotions.

To alleviate the long-tail distribution problem without introducing additional parameters or modifying the network architecture, this paper directly adopts the off-the-shelf LDAM Loss as our classification loss function. The core idea of LDAM Loss is to introduce different classification margins for different classes based on the number of samples in each class. For tail classes with few samples, a larger classification margin is given, forcing the model to learn more discriminative features, thereby enhancing classification ability for minority classes. We emphasize that this loss function is adopted as an engineering plug-in without any structural novelty; its integration into our framework serves as a practical solution to the class imbalance issue inherent in power business hall scenarios, and it works synergistically with the other proposed components.

In the original classification task, the standard cross-entropy loss can be expressed as:

$$L_{CE} = -\sum_{c=1}^C y_c \log(p_c) \quad (1)$$

where C is the number of classes, y_c is the true label of the sample, and p_c is the probability predicted by the model for that class. In the case of class imbalance, the loss from majority class samples dominates the gradient update, leading to insufficient learning of minority classes.

On this basis, this paper introduces LDAM Loss, which is a modified cross-entropy loss:

$$L_{LDAM} = -\sum_{c=1}^C y_c \log\left(\frac{e^{\tilde{z}_c - \Delta_c}}{e^{\tilde{z}_c - \Delta_c} + \sum_{j \neq c} e^{\tilde{z}_j}}\right) \quad (2)$$

where $\tilde{z}_c$ is the logits value output by the model before Softmax, and Δ_c is the class-dependent margin introduced for class c . The calculation of Δ_c is as follows:

$$\Delta_c = \frac{\lambda}{n_c^{1/4}} \quad (3)$$

where n_c is the number of samples of class c in the training set, and λ is a hyperparameter controlling the margin strength. Classes with fewer samples have larger margins, forcing the model to output higher logits values for minority class samples. We set the hyperparameter to 0.3 based on preliminary experiments on the validation set, aiming to balance the trade-off between majority-class accuracy and minority-class recall. This value is chosen as a heuristic starting point rather than a rigorously optimized one, and we recognize that different datasets or tasks may benefit from alternative settings. The LDAM Loss described in this section is adopted exactly in its original form, without any modification to its mathematical formulation or optimization strategy.

2.5. Lightweight Strategy

In actual deployment scenarios of power business halls, the expression recognition algorithm needs to process multiple video streams in real time on edge devices while running in parallel with other business systems. Although the aforementioned MSF-DLKA module and TADDH effectively enhance the model's ability to perceive subtle expression deformations and decouple localization and classification, the parameter count and computational complexity of the model are still relatively high. Direct deployment on edge

computing devices (such as embedded GPUs, ARM processors, or Jetson series modules) would easily lead to increased inference latency and reduced system throughput, severely compromising the practical value of the technology and customer service experience. Therefore, further lightweight compression of the model is required to reduce its parameter count and computational complexity to meet actual deployment needs.

This paper integrates three well-established model compression techniques—structured channel pruning, quantization-aware training, and knowledge distillation—into a progressive engineering pipeline. This strategy is progressive in the sense that each compression step builds upon the previous one: pruning first reduces the model width, quantization reduces numerical precision, and knowledge distillation compensates for the accumulated accuracy loss. We emphasize that our contribution here lies in the sequential integration strategy tailored for this specific application and deployment hardware, rather than the invention of any individual compression algorithm. The relationships among the components are shown in Figure 4.

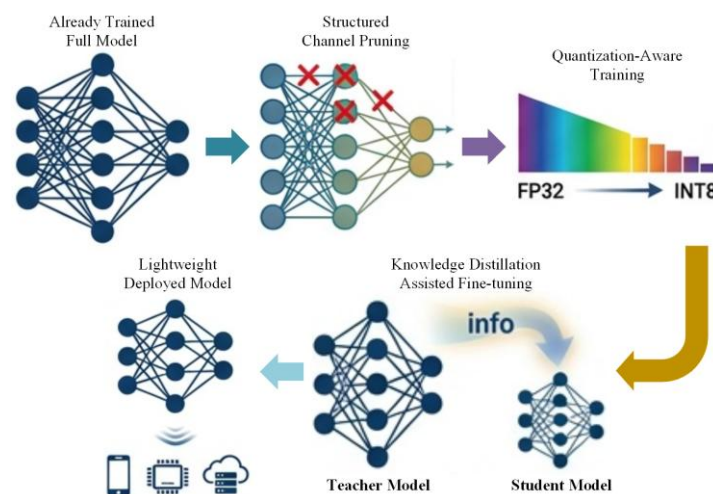


Figure 4. Flowchart of the lightweight strategy.

Structured channel pruning is the first step in the compression process. Although the network is designed with computational efficiency in mind, there is still a certain degree of channel redundancy in each convolutional layer; i.e., some feature maps contribute little to final expression recognition and can even be regarded as noise. The goal of structured channel pruning is to systematically identify and remove these redundant channels, thereby reducing parameter count and computational complexity, without introducing sparse matrix operations, which is very friendly to general hardware. This paper adopts a pruning strategy based on the scaling factor of the Batch Normalization (BN) layer. In convolutional neural networks, the BN layer usually follows the convolution layer, and its operation can be expressed as:

$$y = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \quad (4)$$

where γ and β are the learnable scaling factor and bias, respectively.

The magnitude of the scaling factor γ directly reflects the importance of the corresponding channel. If the γ value of a channel approaches zero, the output of that channel is nearly constant and has almost no influence on subsequent computations. The specific pruning process is as follows: first, perform sparse fine-tuning on the converged complete model, adding an L1 regularization term on γ to the original loss function:

$$L_{CLS} = L_{LDAM} + \omega \sum_{\gamma \in \Gamma} |\gamma| \quad (5)$$

where Γ is the set of scaling factors of all BN layers in the network and ω is the sparsity coefficient. We set the sparsity coefficient to 1×10^{-4} and the pruning ratio to 50%, which were determined empirically based on preliminary trials. We emphasize that these values are heuristic choices and have not been systematically optimized; alternative configurations may yield different trade-offs.

After several epochs of training, the scaling factors γ corresponding to unimportant channels will approach zero. Then, a global pruning ratio is set, all γ values are counted, and a threshold is calculated according to this ratio. Convolution kernels corresponding to channels below the threshold, along with their subsequent BN layers and activation layers, are removed. Because this is structured pruning with whole-channel removal, the input and output dimensions between network layers remain compatible after pruning, eliminating the need for sparse matrix operations, which is very hardware-friendly. Finally, the pruned model is fine-tuned for a short period using the original training set to restore detection accuracy.

The pruned model still stores weights and activations as 32-bit floating point numbers (FP32). To further compress the model size and accelerate inference, this paper introduces Quantization-Aware Training (QAT). Quantization technology maps high-precision values to a lower bit width (e.g., 8-bit integer INT8), which can greatly reduce model storage and increase computation speed. However, traditional Post-Training Quantization (PTQ) directly maps parameters of a pre-trained model, ignoring the impact of quantization noise on the network output, easily causing a significant drop in accuracy. Therefore, this paper adopts quantization-aware training, embedding quantization operations as part of the network into the training process, allowing the model parameters to actively adapt to low-precision representation. The core steps of QAT are as follows: first, insert pseudo-quantization nodes for each tensor to be quantized in the computation graph of the model. These nodes simulate the behavior of low-precision quantization:

$$\hat{x} = \text{round}\left(\frac{\text{clip}(x, -2^{b-1}, 2^{b-1}-1)}{s}\right) \times s \quad (6)$$

where $b=8$ is the quantization bit width and s is a learnable scaling factor.

During forward propagation, $\hat{x}$ is used for computation; during backpropagation, the Straight-Through Estimator (STE) is used to approximate the gradient, enabling the network to perceive quantization errors. This paper uses symmetric 8-bit quantization (signed INT8), which avoids additional operations for zero points. The scaling factor s is initialized as $\max(|x|)/127$ and adaptively adjusted during training. Then, about 1/5 of the original training set is used to fine-tune the model with inserted pseudo-quantization nodes at a low learning rate. During training, weights are still stored and updated in FP32 format, but each forward pass simulates the INT8 effect through the pseudo-quantization nodes, gradually converging the network to an optimal solution under low precision. After training, the pseudo-quantization nodes are removed, and the weights and biases are quantized into true INT8 integers, while the scaling factors for activations of each layer are saved. The final model is exported in standard ONNX INT8 format. After QAT, the model size is compressed to 1/4 of the original, and inference speed is also improved.

Although pruning and quantization each have a small impact on accuracy, the cumulative accuracy loss from two consecutive compression operations may be significant, especially on tail classes (e.g., anger, disgust) of the power business hall dataset, where performance degradation is more sensitive. To further restore accuracy, this paper introduces knowledge distillation, using the original uncompressed model (FP32, full network) as the teacher network and the pruned + quantized lightweight model as the student network for auxiliary fine-tuning. The distillation loss function consists of two parts:

$$L_{\text{KD}} = aL_{\text{hard}} + bL_{\text{soft}} \quad (7)$$

where L_{hard} is the cross-entropy loss between the student network's predictions and the true labels; L_{soft} is the KL divergence loss between the outputs of the student network and the teacher network, used to make the student mimic the teacher's soft prediction distribution. The hyperparameters a and b are empirically set to 0.3 and 0.7 respectively, placing greater emphasis on knowledge transfer from the teacher network. These values were chosen based on common practice and preliminary validation, and we acknowledge that they may not be optimal for all scenarios. After distillation fine-tuning, the lightweight model can restore its detection accuracy while maintaining the efficiency benefits brought by pruning and quantization.

In summary, through the progressive post-training compression lightweight strategy of “structured pruning → quantization-aware training → knowledge distillation-assisted fine-tuning”, this paper significantly reduces the storage and computational overhead of the model without modifying the network architecture, providing a feasible lightweight solution for real-time expression perception in power business hall scenarios.

3. Results

3.1. Datasets

To comprehensively validate the effectiveness of the proposed method, three datasets were selected for experimental evaluation: two public micro-expression datasets, CASME II and SAMM, and one self-constructed Power Business Hall Dataset (PBHD).

CASME II: This dataset released by the Institute of Psychology, Chinese Academy of Sciences, in 2014 is one of the most widely used benchmark datasets in the field of micro-expression recognition. It contains 255 video clips from 26 subjects under controlled laboratory conditions, covering seven emotion categories: happiness, disgust, fear, repression, surprise, sadness, and others. The total number of frames varies across videos, with approximately 17,021 frames in total. To avoid data leakage, we adopted a subject-wise split strategy: subjects were divided into training, validation, and test sets at a ratio of approximately 8:1:1 based on subject identity, ensuring that frames from the same subject do not appear in both training and testing sets. The detailed subject IDs and video-level split lists are provided in the compressed file included in the Data Availability Statement.

SAMM: This dataset is a spontaneous micro-expression dataset with high diversity. It contains 159 video clips from 32 subjects of different ethnicities, genders, and ages. Emotion categories include happiness, disgust, surprise, fear, anger, sadness, and others. The total number of frames varies across videos, with approximately 11,812 frames in total. As with CASME II, we extracted frames and applied subject-wise splitting with a ratio of approximately 8:1:1 to prevent data leakage. The detailed subject IDs and video-level split lists are provided in the compressed file included in the Data Availability Statement.

Self-constructed Dataset (PBHD): This dataset was collected via Hikvision DS-V122 cameras deployed in several power business halls within Nanjing. The study received approval from the institutional review board. The subjects include 47 male and female customers of various age groups, but are predominantly young people. All subjects were fully informed about the purpose of data collection, agreed to the use of their images for research purposes, and provided informed consent prior to data collection. For the example images used in this paper, additional written consent was obtained from the respective subjects for public display. This dataset comprises 15,339 images with a resolution of 640 × 480 pixels, which are divided into training, validation, and test sets in an 8:1:1 ratio using the same sampling method as the two datasets mentioned above to prevent data leakage. Expression labels were defined based on service interaction contexts as four categories:

neutral, happy, disgust, and anger, aligning more closely with emotional states in power business halls. The counts are 9056 for neutral, 4371 for happy, 1504 for disgust, and 408 for anger, which clearly demonstrates a long-tail distribution where the neutral and happy classes dominate. Furthermore, due to the rarity of the ‘anger’ category, we employed an emotion-guided facial image generation method based on the latent diffusion model to produce 1000 additional anger images exclusively for the training set. All generated images were manually filtered to remove poor-quality or ambiguous samples and were used only in the training set to avoid evaluation bias. All comparison methods in our experiments are trained on the same augmented dataset containing these generated samples, ensuring a fair evaluation. This generation method is available at <https://github.com/ComeRochas/GuidedEmotionGeneration> (accessed on 31 July 2026).

3.2. Experimental Setup

The training device used in this paper has a CPU of I9-14900K 3.2 GHz, 128 GB RAM, and a GPU of NVIDIA RTX5880 Ada with 48 GB VRAM, running Windows 11, with deep learning framework PyTorch 2.1.0+cu118, Python 3.9.19. The test device is an NVIDIA Jetson AGX Orin, equipped with a 12-core Arm® Cortex®-A78AE v8.2 64-bit CPU, 64 GB RAM, and a GPU with 2048 NVIDIA Ampere architecture cores and 64 Tensor cores, running Ubuntu 20.04.6, with PyTorch 2.2.0, Python 3.8.20.

All input images were resized to 512×512 pixels before training. The input resolution of 512×512 was selected to balance the need for sufficient spatial detail for subtle expression recognition and the computational constraints of edge deployment. This resolution is consistent with the input size used in the original CenterNet and provides a reasonable trade-off between accuracy and efficiency. To enhance model robustness, all models were trained using Mosaic data augmentation, color jittering, random horizontal flipping, and random rotation. The optimizer was stochastic gradient descent with an initial learning rate of 0.001, momentum of 0.937, weight decay of 0.0005, and a cosine annealing schedule reducing the learning rate to 0.00001 over 500 epochs. The batch size was 128, chosen to balance GPU memory and gradient stability. Training converged steadily after approximately 350 epochs, with validation loss plateauing and no significant overfitting observed.

All experiments were conducted under deterministic training conditions with fixed random seeds (seed = 42) to ensure reproducibility. The evaluation follows the standard object detection benchmarking protocol; under identical experimental settings, repeated runs yield consistent mAP values with negligible variance, as the evaluation protocol does not involve stochastic sampling during inference.

3.3. Evaluation Metrics

Since the proposed method performs joint face detection and expression classification, we adopt mAP@0.5 and mAP[0.5:0.95] as the primary evaluation metric, following the standard practice for object detection. The calculation method for mAP is as follows:

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

$$P_{\text{interp}}(R) = \max_{\tilde{R} \geq R} P(\tilde{R}) \quad (10)$$

$$AP = \sum_{i=0}^{n-1} (R_{i+1} - R_i) P_{\text{interp}}(R_{i+1}) \quad (11)$$

$$mAP = \frac{\sum_{i=1}^K AP_i}{K} \quad (12)$$

where TP represents the number of correctly predicted positive samples, FP represents the number of incorrectly predicted positive samples, FN represents the number of samples in which positive samples were incorrectly predicted as negative samples, n is the number of sample classes, and AP is the Average Precision for a single class.

The macro-F1 is defined as the harmonic mean of Precision and Recall:

$$F1_i = \frac{2 \times (P_i \times R_i)}{P_i + R_i} \quad (13)$$

$$\text{macro-F1} = \frac{\sum_{i=1}^K F1_i}{K} \quad (14)$$

For quantitative evaluation of detection efficiency, we also use the total number of model parameters (Para), computational complexity (Floating Point Operations, FLOPs), and average inference time per image on the test set (Time) as efficiency metrics. The average inference time is calculated over more than 1000 test images without any warm-up, and includes all steps from input image resizing to final output. Since mAP is a deterministic metric calculated based on model predictions and ground-truth annotations, repeated evaluations on the same test set under fixed model weights will produce the same results. This paper will not report the standard deviation of repeated training runs.

3.4. Ablation Experiments

To validate the effectiveness of each improvement strategy, a series of ablation experiments were designed and conducted on the self-built power business hall dataset. The experiments used the CenterNet model with the ResNet-50 backbone as the baseline. The control variable method was employed to sequentially add the MSF-DLKA module, TADDH, LDAM Loss, and the lightweight strategy, evaluating the impact of each module. The ablation experiment results are shown in Table 1. In the table, CenterNet represents the baseline model, CenterNet+M represents the model with only MSF-DLKA added, CenterNet+T represents the model with only TADDH added, CenterNet+LD represents the model with only LDAM Loss introduced, CenterNet+LI represents the model with only the lightweight strategy applied, and Proposed represents the final model with all improvement strategies. To validate the complementarity of different modules, we also conducted experiments on multi-module combinations, with results shown in Table 2.

Table 1. Ablation results of different improvement strategies.

Model	mAP@0.5 (%)	mAP[0.5:0.95]	Para (M)	FLOPs (G)	Time (ms)
CenterNet	82.4	64.2	44.5	29.4	42.5
CenterNet+M	87.6	68.4	48.0	32.2	46.3
CenterNet+T	86.9	67.8	45.7	30.8	44.6
CenterNet+LD	84.8	66.1	44.5	29.4	42.5
CenterNet+LI	81.5	63.5	12.3	17.5	25.1
Proposed	91.7	71.8	13.5	19.6	28.9

Table 2. Ablation results of multi-module combinations on the PBHD.

MSF-DLKA	TADDH	LDAM	Lightweight	mAP@0.5 (%)	mAP[0.5:0.95]
×	×	×	×	82.4	64.2
√	×	×	×	87.6	68.4
√	√	×	×	91.2	71.4
√	√	√	×	92.3	72.2
√	√	×	√	90.1	70.3
√	×	√	√	88.8	69.1
×	√	√	√	89.2	69.5
√	√	√	√	91.7	71.8

From the results in Table 1, it can be seen that each improvement strategy proposed in this paper has a corresponding effect. First, adding the MSF-DLKA module greatly improves accuracy with only a slight increase in parameter count and computation, indicating that the module effectively enhances feature representation capability through multi-scale deformation perception. Second, adding TADDH also improves accuracy by about 4.5% with only 1.2 M additional parameters, validating the advantage of the task-aware dynamic detection head in decoupling localization and classification tasks. Third, introducing LDAM Loss improves accuracy by about 2.4% without increasing parameter count and computational complexity, showing that this loss function effectively alleviates the long-tail distribution of expression categories. After applying the lightweight strategy, although accuracy drops slightly, the parameter count is reduced by about 32.2 M, computational complexity is reduced by about 11.9 GFLOPs, and inference time is shortened to 25.1 ms, laying a good foundation for edge deployment. Finally, from the effect of the final model, it can be seen that each improvement strategy plays a positive role in its respective dimension, and the modules have good complementarity and synergy, achieving synergistic improvement of accuracy and efficiency.

From the results in Table 2, it can be seen that all components work well. To quantify the synergy among modules, we compute the marginal gains from pairwise additions. On the PBHD, MSF-DLKA alone improves mAP@0.5 by 5.2% over the baseline, while TADDH alone provides a 4.5% gain. When MSF-DLKA and TADDH are combined, the mAP@0.5 reaches 91.2%, which is 8.8% higher than the baseline. This combined gain exceeds the gain of either module used individually and is close to the additive sum, indicating that the two modules largely complement each other with only minor overlap. Adding LDAM Loss on top of MSF-DLKA+TADDH further lifts the accuracy to 92.3%, confirming its effectiveness in handling the long-tail distribution without extra computational cost. Finally, the progressive lightweight pipeline reduces the model size by 72.6% (from 49.3 M to 13.5 M parameters) and FLOPs by 42.0%, while the mAP@0.5 drops only slightly from 92.3% to 91.7%. These quantitative results clearly demonstrate the complementary nature of the proposed modules and the acceptable accuracy–efficiency trade-off achieved by the lightweight strategy. Moreover, the reverse ablation results can be readily derived from Table 2 by comparing rows with and without each individual component, which further supports the independent contribution of each module.

To verify the consistency of module contributions across different datasets, we additionally conducted ablation experiments on the CASME II and SAMM datasets following the same conditions. The results are shown in Table 3. It can be seen that the contribution of each module is almost consistent across all three datasets.

Table 3. Ablation results across multiple datasets.

Model	CASME II	SAMM	PBHD
CenterNet	84.7	84.9	82.4
CenterNet+M	88.9	89.2	87.6
CenterNet+T	88.1	88.3	86.9
CenterNet+LD	86.5	86.6	84.8
CenterNet+LI	83.8	84.1	81.5
Proposed	93.2	93.9	91.7

To further illustrate the contribution of each compression step in detail, we provide a detailed step-by-step analysis in Table 4. It can be seen that structured pruning reduces the number of parameters and FLOPs by removing redundant channels, while quantization-aware training further decreases model storage and accelerates inference with almost no change to the number of parameters. Knowledge distillation subsequently compensates for the accuracy degradation caused by the previous two compression operations, ultimately yielding a lightweight model with balanced performance.

Table 4. Ablation results of the lightweight strategy on the PBHD.

Stage	mAP@0.5 (%)	mAP[0.5:0.95]	Para (M)	FLOPs (G)	Time (ms)
Original	82.4	64.2	44.5	29.4	42.5
Pruning	80.3	62.1	21.9	20.6	31.2
QAT	79.9	61.9	21.6	19.9	28.9
Distillation	81.5	63.5	12.3	17.5	25.1

3.5. Experimental Results and Analysis

To further verify the effectiveness of the proposed method, comprehensive comparisons were conducted with CenterNet, CenterNet with a MobileNetV3 backbone (CenterNet+MV3), CenterNet with a ShuffleNetV2 backbone (CenterNet+SV2), YOLOv11s [39], RT-DETR [40], and D-FINE [41] on the two public micro-expression datasets CASME II and SAMM, as well as the self-built power business hall dataset PBHD. Among these, MobileNetV3 and ShuffleNetV2 are widely adopted lightweight backbones in facial expression recognition tasks. Therefore, our comparison covers both general-purpose object detectors and lightweight FER-specific backbones, providing a comprehensive baseline for evaluating the proposed method. For fair comparison, all methods were evaluated using the same training and test sets. The experimental results are shown in Tables 5–7.

Table 5. Comparison results with other methods on the CASME II dataset.

Model	mAP@0.5 (%)	mAP[0.5:0.95]	macro-F1	Para (M)	FLOPs (G)	Time (ms)
CenterNet	84.7	66.0	0.844	44.5	29.4	42.5
CenterNet+MV3	80.9	62.7	0.805	8.3	20.9	35.4
CenterNet+SV2	79.6	60.9	0.788	7.7	19.7	33.8
YOLOv11s	88.7	69.1	0.884	9.4	21.5	34.1
RT-DETR	87.4	68.3	0.871	20.2	58.6	48.6
D-FINE	92.5	72.1	0.922	16.5	51.2	43.7
Proposed	93.2	72.9	0.929	13.5	19.6	28.9

Table 6. Comparison results with other methods on the SAMM dataset.

Model	mAP@0.5 (%)	mAP[0.5:0.95]	macro-F1	Para (M)	FLOPs (G)	Time (ms)
CenterNet	84.9	66.2	0.846	44.5	29.4	42.5
Center-Net+MV3	81.1	63.2	0.807	8.3	20.9	35.4
Center-Net+SV2	79.9	61.5	0.792	7.7	19.7	33.8
YOLOv11s	89.1	69.7	0.888	9.4	21.5	34.1
RT-DETR	87.8	68.5	0.875	20.2	58.6	48.6
D-FINE	92.8	72.6	0.925	16.5	51.2	43.7
Proposed	93.9	73.4	0.936	13.5	19.6	28.9

Table 7. Comparison results with other methods on the PBHD.

Model	mAP@0.5 (%)	mAP[0.5:0.95]	macro-F1	Para (M)	FLOPs (G)	Time (ms)
CenterNet	82.4	64.2	0.821	44.5	29.4	42.5
Center-Net+MV3	78.7	61.3	0.783	8.3	20.9	35.4
Center-Net+SV2	77.2	59.2	0.762	7.7	19.7	33.8
YOLOv11s	86.2	67.6	0.859	9.4	21.5	34.1
RT-DETR	84.9	66.1	0.846	20.2	58.6	48.6
D-FINE	91.2	71.2	0.909	16.5	51.2	43.7
Proposed	91.7	71.8	0.914	13.5	19.6	28.9

From the results in Tables 5–7, the proposed method achieves competitive detection performance across all three datasets. Compared with the lightweight FER baselines, the proposed method achieved better detection performance with only a slight increase in parameters, confirming the effectiveness of the proposed modules for facial expression recognition. On the CASME II dataset, the proposed method achieves an mAP@0.5 of 93.2% and mAP[0.5:0.95] of 72.9%, outperforming CenterNet+MV3, CenterNet+SV2, YOLOv11s, and RT-DETR, and is slightly higher than D-FINE. More importantly, it uses only 13.5 M parameters, 19.6 GFLOPs, and 28.9 ms inference time, which are substantially lower than those of D-FINE, demonstrating a much better trade-off between accuracy and efficiency. On the SAMM dataset, the proposed method achieves an mAP@0.5 of 93.9% and mAP[0.5:0.95] of 73.4%, again surpassing all compared methods with similar efficiency advantages. On the PBHD, the proposed method still reaches 91.7% mAP@0.5 and 71.8% mAP[0.5:0.95], also demonstrating the best overall performance.

To further verify the performance of the proposed method for each category, we present the per-class precision, recall, macro-F1, mAP@0.5 and mAP[0.5:0.95] on the PBHD test set in Table 8 and the corresponding confusion matrix in Figure 5.

Table 8. The performance of the proposed method for each category on the PBHD.

Category	Precision (%)	Recall (%)	macro-F1	mAP@0.5 (%)	mAP[0.5:0.95]
Neutral	94.4	93.9	0.941	94.7	74.3
Happy	92.0	92.7	0.923	93.1	72.9
Disgust	88.4	90.1	0.892	89.8	70.2
Anger	89.6	87.1	0.883	88.9	69.6

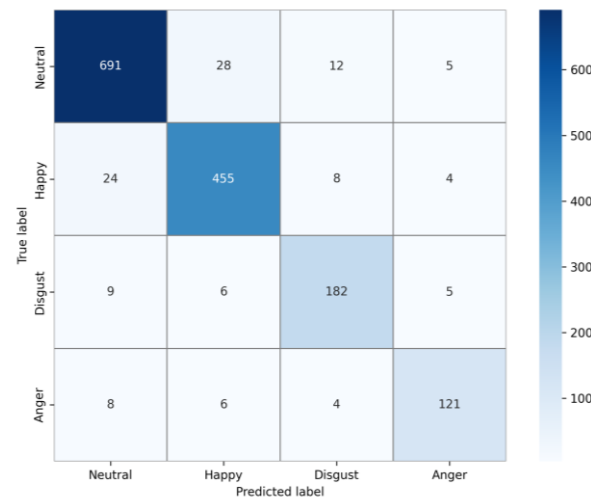


Figure 5. Confusion matrix of the proposed method on the PBHD.

These results further demonstrate that the proposed method achieves relatively balanced performance across all four expression categories. The neutral and happy categories obtain the highest precision and recall, benefiting from abundant training samples and relatively clear facial muscle patterns. For the minority classes, disgust and anger achieve mAP@0.5 of 89.8% and 88.9% and $\text{mAP}[0.5:0.95]$ of 70.2% and 69.6%, respectively, which are considerably competitive despite the limited real samples. The success on these tail classes can be attributed to the synergistic effect of our three key designs. First, the LDAM Loss directly counteracts the long-tail distribution by imposing larger margins on under-represented classes, effectively forcing the network to produce more discriminative logits for disgust and anger. Second, the MSF-DLKA module enhances sensitivity to subtle facial deformations—such as nasal-wing wrinkling and lip raising for disgust, versus eyebrow lowering and mouth tightening for anger—which are the primary differentiating cues. Third, the TADDH decouples classification from localization, allowing the classification branch to focus on high-level semantic features that are more robust to pose and illumination variations, thereby reducing the interference of environmental factors that often confound rare expressions.

To further demonstrate the advantage of the proposed end-to-end joint detection and classification framework over a conventional two-stage pipeline, we constructed a baseline consisting of YOLOv11s as a face detector (trained to localize faces without expression labels) and MobileNetV3 as an expression classifier (trained on cropped face images). The two stages are executed sequentially: the detector first outputs bounding boxes, which are then cropped and fed into the classifier for expression prediction. The results are summarized in Table 9, where the mAP@0.5 (%) and $\text{mAP}[0.5:0.95]$ are computed using the same standard as in previous tables, and the Para (M), FLOPs (G), and Time (ms) are the cumulative totals of the two-stage pipeline (i.e., the sum of the detector's and classifier's respective costs). The results in Table 9 show that the proposed end-to-end joint detection and classification framework achieves 91.7% mAP@0.5 , 71.8% $\text{mAP}[0.5:0.95]$, and 0.914 macro F1, substantially outperforming the two-stage pipeline across all accuracy metrics. This consistent superiority in both accuracy and efficiency validates the advantage of integrating face detection and expression classification into a unified framework, which eliminates the redundant feature extraction and propagation inherent in cascaded pipelines, and is particularly beneficial for resource-constrained edge deployment in power business hall scenarios. In summary, the proposed method achieves a good balance between accuracy and efficiency, possesses strong scene adaptability and engineering application value,

and provides a feasible solution for real-time expression perception in resource-constrained scenarios such as power business halls.

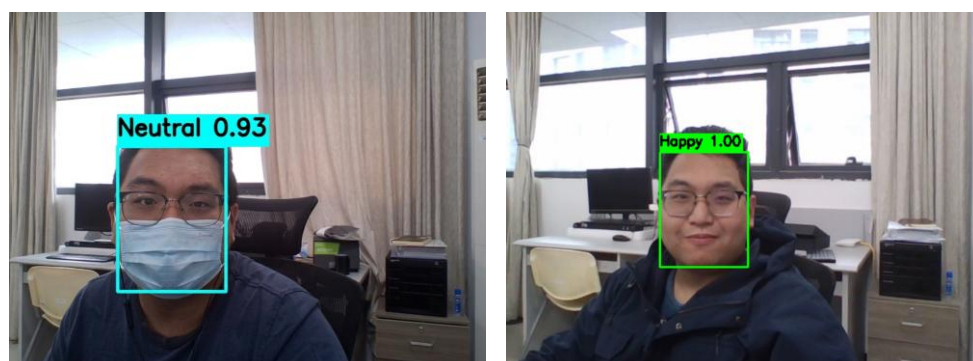
Table 9. Comparison of end-to-end and two-stage pipelines on the PBHD.

Model	mAP@0.5 (%)	mAP[0.5:0.95]	macro-F1	Para (M)	FLOPs (G)	Time (ms)
two-stage pipeline	85.1	66.4	0.843	13.6	21.8	39.2
Proposed	91.7	71.8	0.914	13.5	19.6	28.9

To more intuitively validate the detection performance of the proposed method, the detection results on the PBHD and camera are visualized. The results are shown in Figures 6 and 7, where the colored labels indicate the predicted category, and the numbers after the labels represent the confidence scores. It can be seen that the proposed method accurately locates face regions and effectively recognizes expression categories such as neutral, happy, disgust, and anger. Even under complex conditions such as illumination changes and partial mask occlusion, the model still outputs high-confidence classification results, and the bounding boxes tightly fit the facial regions without missed detections or false detections.



Figure 6. Schematic diagram of detection results on the PBHD.



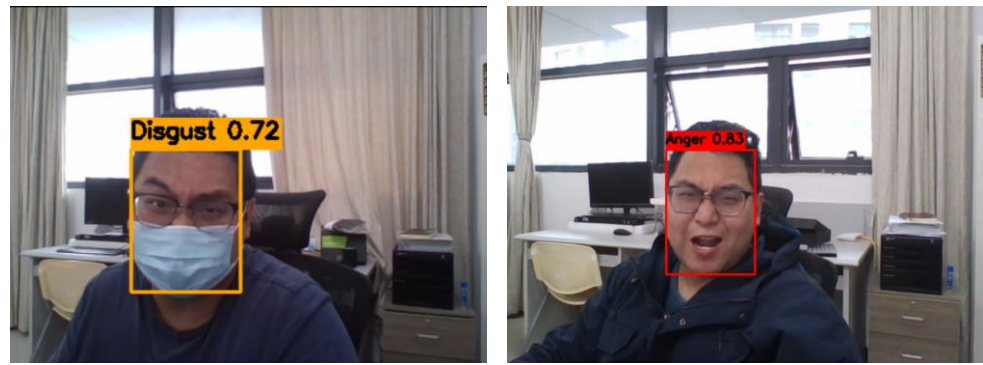


Figure 7. Schematic diagram of real-time detection results on camera.

To further analyze the limitations of the proposed method, we present representative detection results on profile faces and typical failure cases in Figure 8. The figure shows that while the model maintains reasonable localization for profile faces, the classification confidence decreases noticeably for extreme yaw angles, and occluded regions may lead to missed detections or misclassifications. These failure cases primarily arise from the lack of such hard examples in the training set and the inherent difficulty of extracting discriminative expression features from limited visible facial areas.

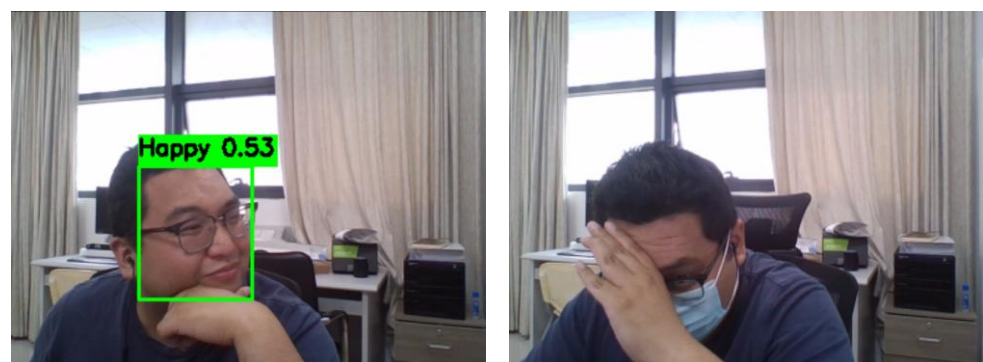


Figure 8. Schematic diagram of detection results on profile faces and typical failure cases.

4. Discussion

This paper addresses the typical public service scenario of power business halls and proposes a lightweight dynamic perception facial expression recognition method, aiming to solve the problems of low recognition accuracy, high computational overhead, and difficulty in edge deployment in complex scenarios. The method is based on the anchor-free CenterNet detection framework, incorporates an MSF-DLKA module, designs a TADDH, introduces LDAM Loss, and combines a progressive lightweight strategy. The following sections discuss the overall performance, contributions of each module, practical application potential, and limitations.

From the perspective of overall performance, the proposed method achieves synergistic improvement in both accuracy and efficiency on all three datasets. From the perspective of each module's contribution, the ablation experiments further confirm the rationale behind each improvement. The MSF-DLKA module enhances the perception of key expression regions such as the eyes and mouth through multi-scale offset field generation, grouped deformable convolution, and dynamic weight modulation, enabling the model to more accurately capture subtle deformations caused by expressions. The TADDH decouples localization and classification features via spatial and channel dual-path attention, allowing the regression branch to focus on spatial details and the classification branch on high-level semantics. The introduction of LDAM Loss, without adding

any parameters or computational cost, alleviates the long-tail distribution of expression categories in power business hall scenarios by assigning larger margins to minority classes. Although the progressive lightweight strategy causes a slight drop in the original model's accuracy, it reduces the parameter count by about 32.2 M, lowers computational complexity by 11.9 GFLOPs, and shortens inference time to 25.1 ms, laying a solid foundation for edge deployment. A good balance between accuracy and efficiency has been achieved though the synergistic effect of all modules.

From the perspective of practical application, the proposed method demonstrates strong deployability in power business halls. On NVIDIA Jetson AGX Orin, the inference time per image is only 28.9 ms, meeting the real-time processing requirements of multiple video streams. In large service halls with dense customer traffic, system scalability can be achieved through batch inference, parallel processing across multiple edge devices, or load balancing via a lightweight central server. The model's compact size enables efficient parallel processing, and the overall system throughput can be further improved via GPU resource pooling and dynamic task scheduling. By recognizing customers' facial expressions during service interactions, the system assists staff in promptly perceiving emotional changes, thereby optimizing service approaches and improving the customer experience. Moreover, the lightweight characteristics make it easy to integrate into existing security or service terminal devices, effectively reducing hardware upgrade costs.

The proposed method still has several limitations. First, the self-built PBHD suffers from scarce anger and disgust samples. Although we augmented the training set with 1000 synthetic anger images generated by a latent diffusion model, synthetic data cannot fully emulate real-world variability, particularly in terms of illumination, texture fidelity, and natural facial dynamics. This introduces a mild distribution shift between the training and test domains, since the test set contains only real captured images. Moreover, the PBHD was collected from a single geographic region and predominantly from young customers, which may introduce demographic bias and limit the generalizability of the model to other populations, age groups, or service environments. The overlapping action units and pose variations further blur class boundaries. Consequently, recognition performance on these minority classes remains suboptimal, and cross-scenario adaptability requires further verification. Future work could employ meta-learning strategies, such as contrastive learning or prototypical networks, to enforce compact intra-class and distinct inter-class representations, thereby boosting robustness under limited labeled data. Additionally, we plan to collect real facial data from power business halls in other regions to construct an external test set for evaluating cross-scene adaptability once such data becomes available. Second, the current single-image framework ignores temporal continuity; in dynamic service interactions, fleeting expression changes may be missed. Third, reliance on only visual cues limits performance when facial details are obscured; multimodal signals (speech, text) could offer complementary information. Future work will target these issues through targeted hard example mining, semi-supervised learning, temporal modeling, and lightweight multi-modal fusion.

5. Conclusions

This paper presents a lightweight dynamic perception facial expression recognition method for power business hall scenarios, with a focus on engineering deployability. Built upon the CenterNet detection framework, the method integrates an MSF-DLKA module and a TADDH to address multi-scale deformation and task-conflict challenges, directly adopts the off-the-shelf LDAM loss to alleviate class imbalance, and combines three established compression techniques into a progressive pipeline for edge deployment. It should be clearly stated that the core contribution of this work is not fundamental algorithmic innovation, but rather the integration and engineering adaptation of existing

techniques for complex power business hall scenarios. Experiments on CASME II, SAMM, and our self-built PBHD demonstrate that the proposed method achieves competitive performance, with mAP@0.5 of 93.2%, 93.9%, and 91.7% and macro-F1 of 0.929, 0.936, and 0.914, respectively. Meanwhile, it maintains only 13.5 M parameters and 28.9 ms inference time on the NVIDIA Jetson AGX Orin, offering a viable solution for edge deployment.

Nevertheless, several limitations should be acknowledged. The PBHD was collected from a single geographic region and predominantly from young customers, which may introduce demographic bias and affect generalizability to other populations or service settings. Synthetic anger samples, generated via a latent diffusion model, were added only to the training set to alleviate class imbalance; however, they cannot fully capture the natural variability of real facial expressions in terms of illumination, texture, and dynamic appearance, leading to a mild domain shift between training and test sets. This may slightly weaken the objectivity of the reported evaluation metrics. Privacy concerns were alleviated through anonymization and encrypted storage, but real-world deployment must adhere to strict data protection regulations. Failure cases mainly involve extreme profile views and severe occlusions, where facial details are insufficient. Generalization to other service environments would likely require additional fine-tuning with domain-specific data. Due to current practical constraints, we are unable to obtain real facial data from power business halls in other regions; cross-region validation is therefore left as an important direction for future work.

Future work will further investigate adaptive perception strategies that dynamically adjust multi-scale receptive fields and attention modulation according to environmental changes, as well as online self-calibration and incremental learning mechanisms. Deployment-oriented optimizations, including dynamic inference scheduling, lightweight model updating, and resource-aware task distribution across edge devices, will also be explored to enhance robustness and maintain real-time performance under varying operational conditions.

Author Contributions: Conceptualization, Y.Q. and H.H.; methodology, Y.Q. and H.H.; software, Y.Y. and Z.Z.; formal analysis, G.L. and Y.C.; investigation, G.L. and Y.C.; data curation, Y.Y. and Z.Z.; writing—original draft preparation, Y.Q. and H.H.; writing—review and editing, Y.Q., G.L. and Y.B.; supervision, Y.B.; project administration H.H. and Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the State Grid Jiangsu Electric Power Co., Ltd. Provincial Management Industry Science and Technology Project, grant number JCNJ2025007.

Informed Consent Statement: All subjects provided written informed consent for the collection and research use of their images. Additional written consent was obtained from the respective subjects for the example images shown in this paper.

Data Availability Statement: The public dataset CASME II is available at https://drive.google.com/file/d/1ndwMCNDeyJMJz1uUIREAVaE2YnlzA7_u/view?usp=sharing (accessed on 31 July 2026). The public dataset SAMM is available at <https://drive.google.com/file/d/1akXOWVmYz3fy18Q1BCU-bNaeoOToFnhb/view?usp=sharing> (accessed on 31 July 2026). The PBHD is not publicly released due to institutional data governance policies and privacy considerations.

Conflicts of Interest: The authors declare no conflicts of interest. Yuemin Qiu, Gang Li, Yun Chen, Yanli Yang, and Zhen Zhong were employed by Nanjing Suyi Industrial Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Revina, I.M.; Emmanuel, W.R.S. A Survey on human face expression recognition techniques. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 619–628.
- Runzer-Colmenares, F.M.; Cahuapaza-Gutierrez, N.L.; Calderon-Hernandez, C.C.; Loret de Mola, C. Artificial intelligence-enabled facial expression analysis for mental health assessment in older adults: A systematic review and research agenda. *Future Internet* **2025**, *17*, 541–557.
- Shuai, T.; Beng, S.; Khalid, F.B.; Rahmat, R.W.B.O.K. Advances in facial micro-expression detection and recognition: A comprehensive review. *Information* **2025**, *16*, 876.
- Guo, X.; Zhang, Y.D.; Lu, S.Y.; Lu, Z.H. Facial expression recognition: A review. *Multimed. Tools Appl.* **2024**, *83*, 23689–23735.
- Sarvakar, K.; Rana, K. Revolutionizing facial emotion recognition: In-depth analysis of cutting-edge models, methodologies, and datasets. *Discov. Artif. Intell.* **2025**, *5*, 388.
- Lan, Z.J.; Wang, L.; Huang, Y. An expression recognition algorithm based on improved Dense-HRNet and gini-index dynamic weighted decision. *Telecommun. Eng.* **2022**, *62*, 1683–1690.
- Yu, J.F.; Li, Z.K.; Chen, T. Facial expression recognition based on separate hybrid attention mechanism. *Telecommun. Eng.* **2022**, *62*, 1207–1214.
- Zheng, R.; Fu, W.H.; Du, J.X.; Wei, S.F. Application of adaptive convolutional neural network in facial expression recognition. *Telecommun. Eng.* **2023**, *63*, 314–322.
- Luo, M.W. Facial expression recognition based on FECN under artificial intelligence. *Discov. Artif. Intell.* **2025**, *5*, 420.
- Liang, L.Q.; Lang, C.Y.; Li, Y.D.; Feng, S.H.; Zhao, J. Fine-grained facial expression recognition in the wild. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 482–494.
- Chen, Y.; Wu, L.C.; Wang, C. A micro-expression recognition method based on multi-level information fusion network. *Acta Autom. Sin.* **2024**, *50*, 1445–1457.
- Wang, C.Y.; Peng, M.; Bi, T.; Chen, T. Micro-attention for micro-expression recognition. *Neurocomputing* **2020**, *410*, 354–362.
- He, Y.J.; Han, Y.G.; Zhang, H.Y. FFDNet: Fine-grained facial expression recognition in challenging environments. *Appl. Res. Comput.* **2024**, *41*, 1578–1584.
- Huang, Y.S.; Peng, J.J.; Zhang, W.Q.; Zhao, T.; Chen, G.; Tan, S.H.; Yi, F.; Wang, L. FERMixNet: An occlusion robust facial expression recognition model with facial mixing augmentation and mid-level representation learning. *IEEE Trans. Affect. Comput.* **2025**, *16*, 639–654.
- Liu, T.; Li, J.; Wu, J.; Du, B.; Wan, J.; Chang, J. Confusable facial expression recognition with geometry-aware conditional network. *Pattern Recognit.* **2024**, *148*, 110174.
- Zhang, F.F.; Zhang, T.Z.; Mao, Q.R.; Xu, C.S. Geometry guided pose-invariant facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4445–4460.
- Zhu, H.T.; Xu, H.H.; Ma, X.J.; Bian, M.J. Facial expression recognition using dual path feature fusion and stacked attention. *Future Internet* **2022**, *14*, 258–275.
- Fang, B.; Zhao, Y.J.; Han, G.X.; He, J.H. Expression-guided deep joint learning for facial expression recognition. *Sensors* **2023**, *23*, 7148.
- Talukder, A.; Ghosh, S. Facial image expression recognition and prediction system. *Sci. Rep.* **2024**, *14*, 27760.
- Chang, X.H.; Wang, J.R. Expression recognition algorithm based on optimized YOLOv7-Tiny. *J. Test Meas. Technol.* **2025**, *39*, 113–120.
- Xue, Z.C.; Yilihamu, Y.E.M.M.T.; Yan, T.X. Facial expression recognition based on MobileNetV3 multi-scale feature fusion. *Electron. Meas. Technol.* **2023**, *46*, 38–44.
- Ge, X.X. Emotion recognition intelligent system based on machine learning and clustering algorithm. *Discov. Artif. Intell.* **2026**, *6*, 129.
- Jiang, B.; Cui, X.M.; Jiang, H.B.; Ding, H.Q.; Yuan, J.L. New advances in lightweight networks for facial expression recognition. *Appl. Res. Comput.* **2024**, *41*, 663–670.
- Lin, E.H.; Wang, F.; Tan, X.L. Facial expression recognition of ShuffleNetV2 combined with ultra-lightweight dual attention modules. *Electron. Meas. Technol.* **2024**, *47*, 168–174.
- Karani, R.; Jani, J.; Desai, S. FER-BHARAT: A lightweight deep learning network for efficient unimodal facial emotion recognition in Indian context. *Discov. Artif. Intell.* **2024**, *4*, 35.
- Wan, C.; Zhang, W.B.; Chen, Y.D.; Song, L.L.; Cheng, P. ARTNet for micro-expression recognition. *Sensors* **2026**, *26*, 247.

27. Ye, J.Y.; Yu, Y.H.; Wang, Q.X.; Liu, G.L.; Li, W.T.; Zeng, A.; Zhang, Y.Q.; Liu, Y.; Zheng, Y.S. CmdVIT: A voluntary facial expression recognition model for complex mental disorders. *IEEE Trans. Image Process.* **2025**, *34*, 3013–3024.
28. Xia, Y.F.; Yu, H.; Wang, X.; Jian, M.W.; Wang, F.Y. Relation-aware facial expression recognition. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 1143–1154.
29. Chen, P.A.; Deng, Q. Expression detection algorithm based on SSD network model reconstruction. *Comput. Sci.* **2025**, *52*, 508–513.
30. Wang, H.J.; Deng, Y.J.; Liu, T.T.; Yang, Z. Hybrid feature facial expression recognition model based on DINO prior. *Comput. Eng.* **2025**, *51*, 284–294.
31. Liu, J.; Zhang, M.Y.; Hu, M.; Huang, Z.; Jiang, J.L. Expression recognition network of channel-spatial multi-scale enhancement and dual-pooling attention. *Appl. Res. Comput.* **2025**, *42*, 3182–3191.
32. Hou, C.B.; Ai, J.J.; Lin, Y.; Guan, C.Y.; Li, J.W.; Zhu, W.Y. Evaluation of online teaching quality based on facial expression recognition. *Future Internet* **2022**, *14*, 177–189.
33. Rößler, J.; Sun, J.; Gloor, P. Reducing videoconferencing fatigue through facial emotion recognition. *Future Internet* **2021**, *13*, 126–141.
34. Zhao, Y.B.; Jiang, W.; Meng, R.Y.; Li, Y.N. Facial expression recognition based on memristor convolutional neural network. *Electron. Meas. Technol.* **2022**, *45*, 93–101.
35. Zhou, X.Y.; Wang, D.Q.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
36. Zhang, Y.F.; Kang, B.Y.; Hooi, B.; Yan, S.C.; Feng, J.S. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10795–10816.
37. Yan, W.J.; Li, X.B.; Wang, S.J.; Zhao, G.Y.; Liu, Y.J.; Chen, Y.H.; Fu, X.L. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **2014**, *9*, e86041.
38. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. SAMM: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.* **2016**, *9*, 116–129.
39. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
40. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974.
41. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. *arXiv* **2024**, arXiv:2410.13842.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.