applied sciences

MDPI

Article

A Lightweight Teaching Assessment Framework Using Facial Expression Recognition for Online Courses

Jinfeng Wang *, Xiaomei Chen and Zicong Zhang

College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China; 13711117274@stu.scau.edu.cn (X.C.); zzc199788@126.com (Z.Z.)
* Correspondence: wangjinfeng@scau.edu.cn

Abstract

To ensure the effectiveness of online teaching, educators must understand students' learning progress. This study proposes LWKD-ViT, a framework designed to accurately capture students' emotions during online courses. The framework is built on a lightweight facial expression recognition (FER) model with modifications to the fusion block. In addition, knowledge distillation (KD) is integrated into the online course platform to enhance performance. The framework follows a defined process involving face detection, tracking, and clustering to extract facial sequences for each student. An improved model, MobileViT-Local, developed by the authors, extracts emotion features from individual frames of students' facial video streams for classification and prediction. Students' facial images are captured through their device cameras and analyzed in real time on their devices, eliminating the need to transmit videos to the teacher's computer or a remote server. To evaluate the performance of MobileViT-Local, comprehensive tests were conducted on benchmark datasets, including RAFD, RAF-DB, and FER2013, as well as a self-built dataset, SCAUOL. Experimental results demonstrate the model's competitive performance and superior efficiency. Due to the use of knowledge distillation, the proposed model achieves a prediction accuracy of 94.96%, surpassing other mainstream models. It also exhibits excellent performance, with optimal FLOPs of 0.265 G and a compact size of 4.96 M, while maintaining acceptable accuracy.

Keywords: facial expression recognition; online education; vision transformer; knowledge distillation

check for updates

Academic Editor: Pedro Couto

Received: 2 October 2025
Revised: 14 November 2025
Accepted: 19 November 2025
Published: 24 November 2025

Citation: Wang, J.; Chen, X.; Zhang, Z. A Lightweight Teaching Assessment Framework Using Facial Expression Recognition for Online Courses. *Appl. Sci.* **2025**, *15*, 12461. https://doi.org/10.3390/app152312461

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Since the onset of the COVID-19 pandemic, educational institutions have progressively adopted electronic learning technologies [1]. This widespread adoption can be primarily attributed to the unparalleled convenience of online education. During the global crisis, many schools postponed physical reopening and adopted online learning as their main instructional modality. At the same time, the number of Massive Open Online Courses (MOOCs) increased substantially [2]. Although the COVID-19 pandemic has been temporarily mitigated, the post-epidemic era requires greater emphasis on online education to ensure the continuity of educational services via the internet, particularly in anticipation of potential future public health crises. However, the effectiveness of online courses remains a subject of considerable debate. Compared with traditional classroom settings, online courses often lack timely and efficient channels for communication and feedback between students and instructors. Research indicates that facial expressions are among

the most powerful, innate, and universal means by which humans convey emotions and intentions, transcending cultural, ethnic, and gender boundaries [3,4]. Emotions play a fundamental and indispensable role in human experience. Understanding basic human behavior requires analyzing emotion-related data, including text, voice, and facial expressions [5]. The correlation between facial expressions and emotions is generally strong and consistent. Emotion analysis aims to recognize emotions in visual representations, track changes in emotions over time, classify related movements, and determine their orientation. The human face conveys abundant information that influences decision-making, with facial emotions closely linked to perceived engagement [6]. Identifying facial expressions in images provides a universal, non-verbal means of interpreting internal emotions through facial cues [7]. Accurately identifying facial emotions and applying this capability to online education holds significant research value but remains challenging. The main difficulty lies in processing facial features in real time while maintaining low computational demands.

Research has achieved rapid progress in facial expression recognition (FER), improving its accuracy and enabling real-time processing across a wide range of applications. However, the complexity and subtlety of facial expression variations pose numerous challenges. Current FER research remains somewhat limited, often focusing on specific methodologies while neglecting model deployment on mobile platforms such as smartphones and tablets [8]. Most current facial expression recognition systems rely heavily on large-scale deep convolutional neural networks [9,10] and multimodal features that combine audio, facial expressions, and body gestures [11–13]. Consequently, these models are not well-suited for real-world scenarios requiring real-time processing on devices with limited computational resources.

Moreover, to safeguard student privacy, an optimal solution is to process facial videos on each device and upload only processed data, thereby eliminating the need to transmit raw footage captured during the learning process [14].

This study presents an improved lightweight framework named MobileViT-Local, specifically designed for facial expression classification. This innovative framework modifies the fusion module in MobileViT and integrates a pre-trained deep neural network for knowledge distillation. These enhancements improve classification performance with only a slight increase in parameters, enabling implementation in online learning software on laptops with limited GPU capacity and memory, as well as on smartphones and tablets. The key contributions of this research are summarized as follows:

(1) A comprehensive, real-time emotion classification system is developed for online education, utilizing facial patterns for video-based analysis. The system predicts students' emotions locally on their personal devices, thereby safeguarding privacy and autonomy. The resulting emotion feature vectors are then securely transmitted to the instructor's device to facilitate classification of emotion trends across the entire student population. Moreover, the framework's seamless integration features allow effortless incorporation into various online course platforms and conferencing software, enhancing emotional intelligence and interactivity in virtual learning environments.

(2) To extract emotion features from facial images, this study proposes an enhanced lightweight facial expression recognition (FER) model based on the MobileViT architecture. By adjusting the input connections within the fusion module, the model acquires more representative features. Experimental results demonstrate a substantial improvement in prediction and classification accuracy on benchmark datasets.

(3) To further improve prediction performance while minimally increasing memory usage and parameters, knowledge distillation techniques are incorporated into MobileViT-Local. Compared with mainstream baseline models, the proposed model achieves competitive results.
(4) A customized emotion dataset is developed to support emotion recognition research in online learning environments. It includes facial recognition data collected from students during real learning sessions, followed by essential data preprocessing and augmentation.

The rest of this article is organized as follows. Section 2 reviews related works; Section 3 describes the proposed model; Section 4 presents the experimental data, setup, and comparisons with existing methods; Section 5 discusses the model's effectiveness; and Section 6 concludes this paper.

2. Related Work

Facial expressions are an indispensable aspect of human communication and represent one of the primary means by which individuals convey emotions. Even without verbal communication, information transmitted through non-verbal cues such as facial expressions can still be mutually understood. These cues play a vital role in facilitating and maintaining interpersonal relationships.

In pioneering research on facial expressions, Ekman and Friesen posited that, despite cultural differences, humans can universally recognize a set of fundamental emotions [15,16]. They categorized prototypical facial expressions into six types: anger, disgust, fear, happiness, sadness, and surprise. Subsequent studies by Ekman and Heider [17–19], as well as by Matsumoto [20], provided further evidence for a seventh universally recognized expression: contempt. Consequently, most public facial expression datasets include neutral expressions along with these six basic expressions, resulting in seven standard emotion labels, as shown in Figure 1. Electrodermal Activity (EDA) is widely used in emotion evaluation due to its sensitivity to sympathetic nervous system activity. A novel graph signal-processing approach [21] was proposed to address the nonlinear complexity of EDA for emotion recognition. To capture dynamic nonlinear variations in EDA, data were decomposed into phasic and tonic components, and features from a transition network were employed for classification [22]. Another study [23] proposed a deep learning-based autoencoder method for extracting nonlinear features from EDA signals. These EDA-based approaches can accurately discriminate complex emotions.

Over the past few decades, online education has achieved notable growth in universities and training institutions [24], creating new opportunities for applying facial expression recognition (FER) technology. Although online courses differ significantly from traditional face-to-face instruction and have been criticized for lacking rigor and direct communication—leading to skepticism among faculty [25,26]—several studies have posited that learning outcomes in online environments can be comparable to those in traditional classrooms [27,28], except in disciplines requiring precise skills or tactile engagement [29]. Despite ongoing debate, the growing popularity of online education, which provides greater convenience and flexibility for a broad range of learners, demonstrates significant potential for future expansion. Consequently, accurately monitoring students' learning progress in online settings is essential for the continued advancement of this educational model.

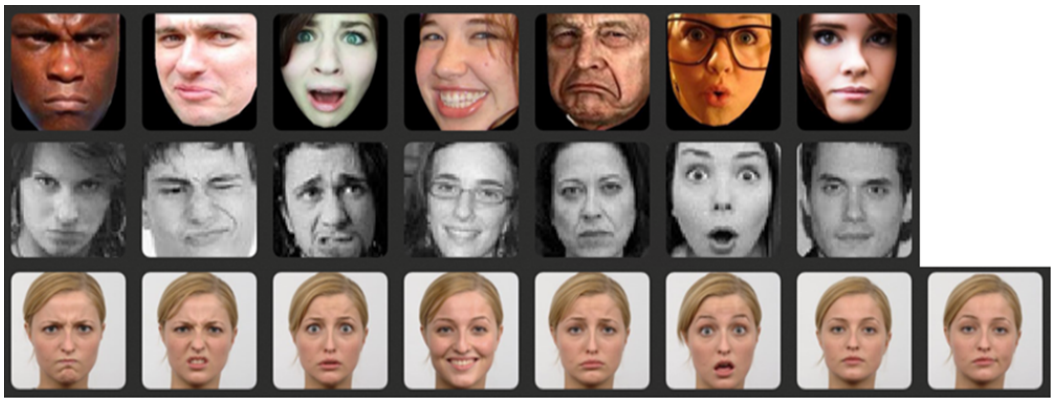


Figure 1. Prototypical facial expressions in the FER2013 [30], RAF-DB [31,32], and RAFD [33] datasets. The expressions—anger, disgust, fear, happiness, sadness, surprise, neutral, and contempt—are shown from left to right.

The transformer architecture, originally proposed by Vaswani [34], revolutionized machine translation and has since gained substantial prominence in the field of Natural Language Processing (NLP). Over the past three years, the Vision Transformer (ViT) [35] has represented a pioneering step in applying transformers to computer vision tasks. Specifically, ViT has demonstrated remarkable performance in large-scale image recognition, achieving accuracy comparable to that of convolutional neural networks (CNNs) while eliminating the need for image-specific inductive biases on large datasets such as JFT-300M.

Touvron [36] further advanced this field by introducing the Data-efficient Image Transformer (DeiT), which employs a novel training technique to reduce reliance on extensive pre-training data, thereby enhancing performance. Notably, this approach also introduced a teacher–student distillation strategy for auxiliary representation learning within the transformer framework. Experimental results showed that CNNs outperformed transformers when used as teacher models. At the same time, hybrid Vision Transformer architectures emerged, characterized by dynamic attention to neighboring elements and improved local feature extraction through local attention mechanisms. For example, the Swin Transformer [37] adopted a sliding-window approach along the spatial dimension to capture both global and boundary features, while VOLO [38] introduced an outlook attention mechanism, similar to patch-wise dynamic convolution, that performs three operations—unfolding, linear-weights attention, and refolding—to capture finer-grained features.

A mobile-friendly, lightweight hybrid architecture, MobileViT [39], was later introduced as a novel method for capturing global representations. Standard convolution typically involves three steps: unfolding, localized computation, and folding. The MobileViT block replaces the localized computation step with global processing using transformers, thereby combining the advantages of CNNs and ViTs. This hybrid approach enables MobileViT to learn effective representations with fewer parameters and simplified training protocols, such as basic data augmentation techniques. These innovations facilitate the development of lightweight transformer-based networks.

Prior to this, numerous mainstream lightweight CNN-based networks had been developed, including MobileNet [40], ShuffleNet [41], and EfficientNet [42]. Although CNNs have been extensively studied and have achieved promising results across various visual tasks, transformer-based architectures, owing to their superior generalization capability, have the potential to surpass CNNs in many transfer learning applications. Since the introduction of ViT and its proven effectiveness in computer vision (CV) tasks, Vision Transformers have attracted significant attention, challenging the long-standing dominance of CNNs in the field.

On the student side, learners can launch an application on their personal devices, as shown in Figure 2. This application provides real-time emotion analysis results during the learning process without requiring the recording or uploading of personal facial videos. The implementation process is described below.

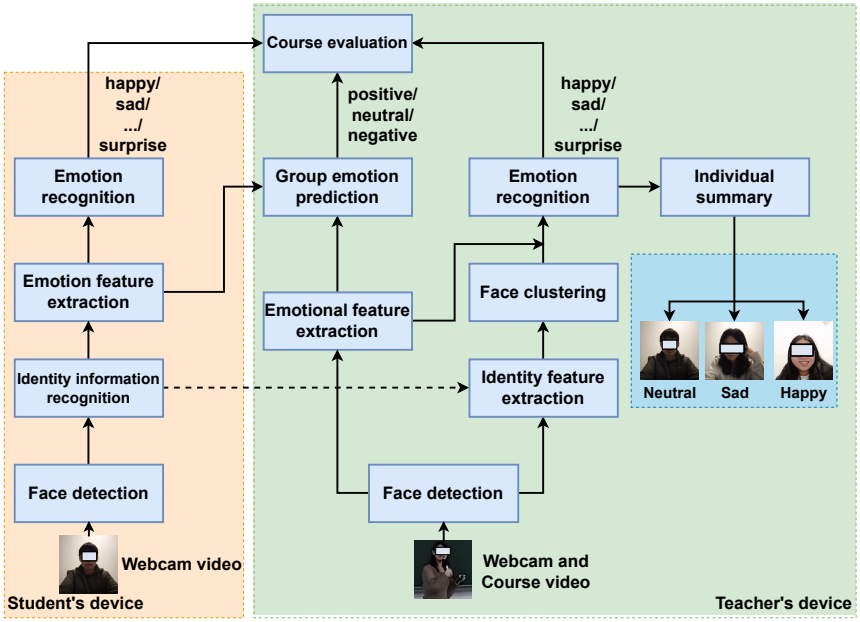


Figure 2. Online course framework in which solid arrows represent procedure direction and dash arrow represents the similar module.

Step 1—Image Acquisition and Preprocessing. The camera captures real-time video frames of students, and each frame undergoes systematic preprocessing:

(1) Face detection: A lightweight CNN-based detector identifies facial regions and generates bounding boxes around each face.
(2) Alignment: Key facial landmarks (eyes, nose, and mouth) are detected, and an affine transformation aligns faces to a standardized frontal pose.
(3) Rotation correction: Faces tilted beyond ±5° are rotated to ensure vertical alignment of the nasal bridge.
(4) Resizing: Aligned faces are resized to a fixed resolution (e.g., 112 × 112 pixels) using bilinear interpolation, normalizing input dimensions for downstream modules.

Step 2—Identity Verification and Feature Extraction. Aligned face images are processed through an identity verification pipeline:

(1) Feature embedding: A lightweight face recognition model generates 128-dimensional feature vectors encoding identity-specific attributes.
(2) Database matching: These features are compared with a pre-registered student database using cosine similarity, and matches exceeding a threshold authenticate identity.

Step 3—Emotion Feature Extraction. Verified face images are then fed into a specialized emotion analysis network:

(1) Network architecture: Pre-trained on datasets, the network extracts spatiotemporal features. Bottleneck layers reduce dimensionality while preserving emotion-relevant patterns such as eyebrow furrows and lip curvature.
(2) Feature representation: The network outputs feature vectors capturing probabilities of action unit activations.

Step 4—Temporal Emotion Classification. Sequential analysis of 3–5-s video segments enhances classification robustness:

(1) Feature aggregation: Emotion features across frames are averaged or processed through a lightweight LSTM to model temporal dynamics.
(2) Classification: A shallow fully connected layer maps aggregated features to seven emotion classes (anger, disgust, fear, happiness, sadness, surprise, and neutral) using a softmax function.

Step 5—Deployment and Reporting. Results are optimized for edge deployment:

(1) Data transmission: Packets containing emotion labels, timestamps, and confidence scores are transmitted to the teacher dashboard.
(2) Edge optimization: Quantization and model pruning reduce network size to an optimal level.
(3) Teacher interface: Real-time heatmaps display the emotion distribution across the class, with alerts triggered for persistent negative states.

On the teacher side, once the video streams captured during an online course are obtained, a series of subsequent processing steps is performed independently in offline mode. The detailed operational procedures are described below.

Step 1—Facial Feature Extraction via Deep Neural Network Modeling. Video streams from online course sessions undergo frame-by-frame processing. The key steps are as follows:

(1) Localization: Bounding-box regression identifies participants' facial regions.
(2) Feature encoding: High-dimensional embeddings are extracted through triplet-loss optimization, capturing invariant facial attributes across pose variations.
(3) Computational workflow: Inference is parallelized on GPU-accelerated cloud instances to reduce per-frame latency.

Step 2—Cohort-Level Affective State Analysis. Emotion descriptors collected from distributed learner devices are aggregated through a federated framework. The analytical pipeline performs the following operations:

(1) Integrates sparse facial expression indicators (e.g., eyebrow movements, lip curvature) across temporal windows.
(2) Applies attention mechanisms to weight salient behavioral cues.
(3) Generates time-synchronized engagement metrics using recurrent neural architectures.
(4) Produces continuous emotion-tendency estimates reflecting collective responses to instructional content.

Step 3—Identity-Associated Affective Profiling. Individual learner analysis combines verification and clustering techniques:

(1) Biometric matching: Facial embeddings are compared against registered identity databases using metric learning, with adaptive thresholds to accommodate occlusion scenarios.
(2) Temporal clustering: Unsupervised association algorithms link discontinuous facial appearances into identity-continuous sequences.
(3) Personalized baseline modeling: Individual neutral-expression benchmarks are established through historical analysis.
(4) Adaptive recognition: Emotion classification adapts to individual expressiveness patterns via transfer learning approaches.

Step 4—Multimodal Course Evaluation Framework. The course assessment integrates macro- and micro-level affective indicators:

(1) Macro-analysis: Correlates cohort engagement trajectories with pedagogical events.
(2) Micro-analysis: Identifies learning anomalies through longitudinal emotion tracking.
(3) Synthesis framework: Cross-references group-level engagement metrics with personalized emotion profiles to generate instructional-effectiveness indices, personalized intervention triggers, and content-optimization recommendations.

This enables teachers to promptly interpret students' emotions during learning and identify course sections that students find challenging. Meanwhile, a personalized summary of each student's performance can be generated based on emotions identified through facial clustering, allowing teachers to detect periods of negative affect and provide targeted support and feedback.

To optimize emotion evaluation in online courses, the proposed framework operates across two device platforms. This design relieves students from the burden of recording and uploading their own videos, significantly reducing potential privacy risks. Instead, students contribute to the course evaluation process by submitting only their facial features, thereby protecting their personal data. Furthermore, since many online course and conference platforms cannot capture clear facial images of all participants in a single frame, a selective-display strategy is recommended. Specifically, only students who have consented to share their facial data are included in the evaluation process, ensuring accurate emotion assessment during online sessions. This approach enhances the effectiveness of the evaluation while maintaining a high level of respect for students' privacy.

3. Proposed Model

This study presents an enhanced lightweight Vision Transformer model, MobileViT-Local, derived from the MobileViT framework. Figure 3 illustrates the original MobileViT architecture, which primarily comprises a convolution block, an MV2 block, a MobileViT block, a global pooling layer, and a fully connected layer. The MV2 block, originating from the lightweight CNN MobileNet v2, is an inverted residual block composed mainly of 1 × 1 convolutional layers and 3 × 3 depthwise convolution units. Modules marked with "↓2" indicate a stride value of 2, representing a downsampling operation.

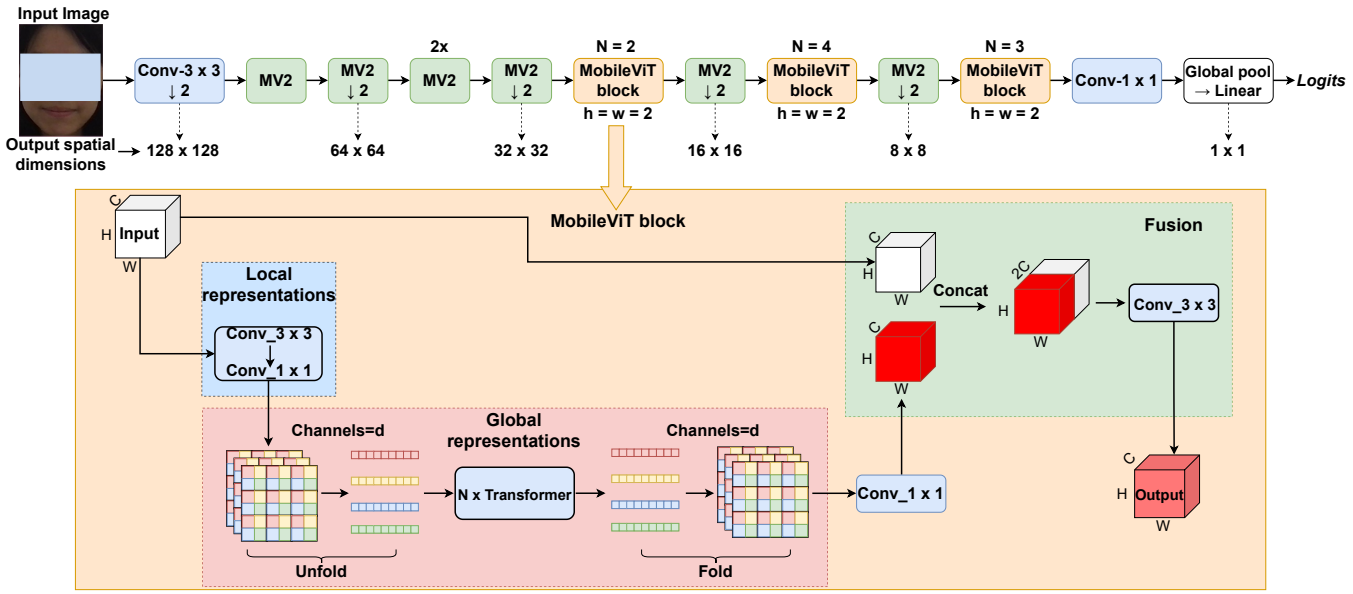


Figure 3. Original MobileViT block.

3.1. Lightweight Vision Transformer

The MobileViT block comprises several key components: convolutional layers, transformer modules, residual connections, and a fusion block. First, a 3 × 3 convolution captures local feature representations from the input feature map. Next, a 1 × 1 convolution adjusts the channel dimensionality to match the transformer module's input requirements. The transformer module then applies an "unfold–transform–fold" procedure to extract global feature representations. A subsequent 1 × 1 convolution restores the feature map's original dimensionality, and a shortcut branch combines the modified feature map with the original input. This fusion, implemented via a 1 × 1 convolutional layer, produces the final global feature output. Finally, a global pooling layer and a fully connected layer generate the prediction logits.

The MobileViT framework employs the "unfold–transform–fold" strategy to minimize parameters and computational complexity. In computing token self-attention, MobileViT does not evaluate individual tokens; instead, it focuses on tokens occupying the same spatial positions across patches, thereby reducing computational overhead. Recognizing the redundancy and correlation among adjacent pixels, MobileViT groups these pixels to lower computation while maintaining accuracy. Furthermore, by applying the convolutional operation before the transformer module, local image representations are directly utilized in self-attention for global representation learning. This approach eliminates the need for exhaustive pixel-level processing to generate global representations.

The MobileViT-Local variant has been optimized with a primary focus on enhancing its fusion module. As shown in Figure 4, the updated fusion process concatenates the local representation—rather than the original input features—with the global representation. This modification is based on the observation that the locally derived representation exhibits a stronger correlation with the global context than the original input features. By leveraging the local representation outputs, the model can better recognize fundamental low-level characteristics of the input image, thereby reducing redundant information in the raw data. In essence, this refinement enriches the image features integrated with the global representation, resulting in a more comprehensive and efficient feature representation.

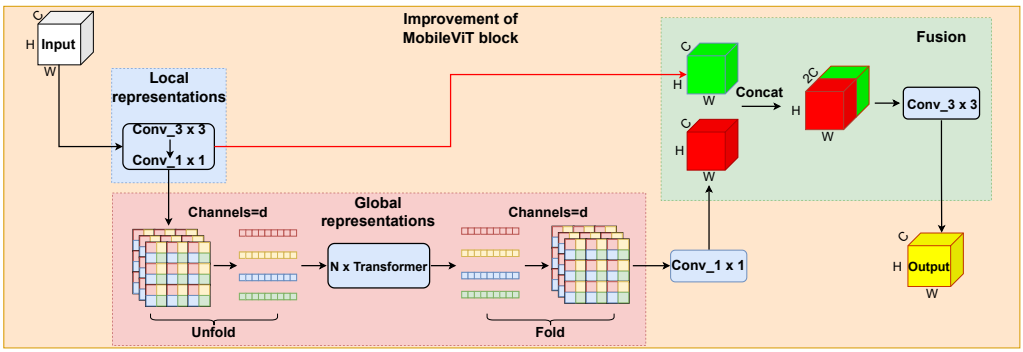


Figure 4. Improved MobileViT-Local block.

3.2. MobileViT-Local Through Distillation

This section introduces the integration of knowledge distillation (KD) into MobileViT-Local to enhance its performance without altering the model's scale or parameter count. To achieve this, a powerful image feature extraction model—such as a CNN, transformer, or hybrid architecture—is employed as the teacher model. This teacher serves as a reference for MobileViT-Local, enabling knowledge transfer from teacher to student and improving the latter's performance.

The soft target method can enhance model generalization by capturing inter-class relationships [43]. These soft targets represent the probabilities of an input belonging to

different classes, estimated using the softmax function, as shown in Equation (1). A temperature factor τ adjusts the influence of each soft target:

$$p(z_i, \tau) = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} \quad (1)$$

where z_i denotes the logit for the i-th class, τ is the temperature factor, and j represents all classes to be predicted.

3.3. Soft Distillation

The student model aims to approximate the teacher model's output by minimizing the difference between their softmax distributions, measured by the Kullback–Leibler (KL) divergence. Let Z_t and Z_s denote the logits of the teacher and student models, respectively. A temperature parameter τ regulates the distillation process, while λ balances the KL divergence loss and the cross-entropy (CE) loss based on the ground-truth labels y. ψ represents the softmax function that converts logits into probabilities. The overall objective loss function is defined in Equation (2):

$$\begin{aligned} \mathcal{L}_{\text{Total}} &= (1-\lambda)\mathcal{L}_{\text{CE}}(\psi(Z_s), y) \\ &\quad + \lambda\tau^2 \text{KL}(\psi(Z_s/\tau), \psi(Z_t/\tau)) \end{aligned} \quad (2)$$

where Z_t and Z_s represent the logits of the teacher and student models, respectively. τ is the temperature coefficient, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss measuring the difference between the predicted probability $\psi(Z_s)$ and the true label y. λ is a hyperparameter that balances the KL divergence and cross-entropy losses over the true label y. The softmax function ψ is applied here, and $\mathcal{L}_{\text{Total}}$ denotes the total loss function, which serves as the objective function for model training.

As shown in Figure 5, the process begins by converting the input image into a vector representation through embedding. This vector is then simultaneously input to both the teacher and student models. The teacher model's prediction is scaled using the temperature parameter and processed by a softmax function to yield a smoother probability distribution (the soft target) within the range [0, 1]. The temperature value determines the distribution's shape: a higher value results in a more uniform distribution, whereas a lower value increases the risk of misclassification and may introduce spurious fluctuations. For complex classification or detection tasks, the temperature coefficient is typically set to one, emphasizing accurate teacher predictions.

The true class label, or hard target, is typically expressed as a one-hot vector. The total loss function combines the cross-entropy losses for both soft and hard targets—namely, the knowledge distillation (KD) loss and the standard cross-entropy (CE) loss. As the weight of the soft target increases, the learning process increasingly relies on the teacher's guidance. Initially, this accelerates the student model's learning on simpler cases; however, as training progresses, the influence of soft targets should gradually decrease, allowing the true labels to strengthen recognition of more complex samples.

The teacher network, possessing superior predictive capability compared with the student network, is not strictly constrained by its dimensions. Its high inference accuracy provides a favorable learning environment for the student, facilitating optimal knowledge transfer. Thus, a large and complex neural network serves as the teacher, transferring its expertise to the MobileViT-Local student model through knowledge distillation, ensuring efficient knowledge transfer and enhanced student performance.

To address the limited availability of publicly accessible datasets specifically designed for online learning scenarios, this study also developed an emotion dataset that accurately captured real-world online learning dynamics. The goal was to reproduce the genuine en-

gagement exhibited by students in online educational environments. Moreover, the dataset incorporated preprocessing procedures to ensure robust support for subsequent experimental research on learner emotion recognition.

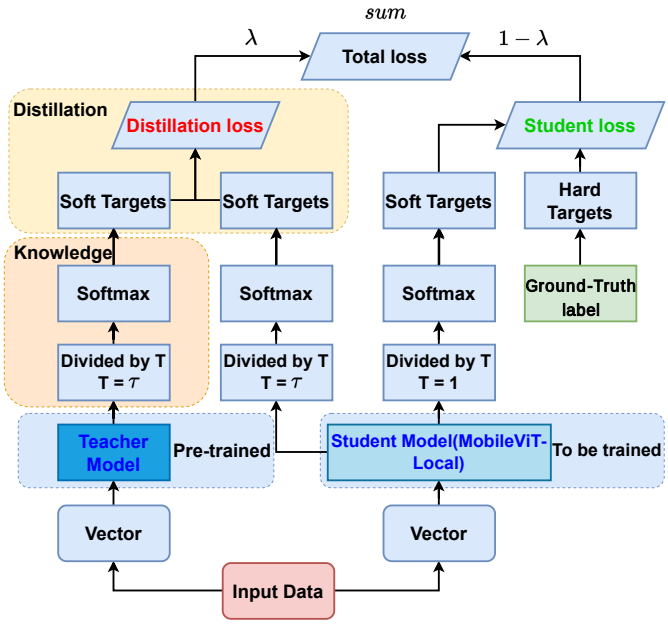


Figure 5. Knowledge distillation architecture.

For data collection, participants were selected from a cohort of undergraduate and postgraduate students, primarily in computer-related disciplines. A total of 26 subjects aged between 19 and 25 participated, comprising 18 males and 8 females.

Before data collection, each participant was placed in a quiet, isolated environment to replicate the authentic engagement typical of self-directed online learning. Capturing the subtle details of facial emotions required a camera setup that ensured each participant's facial expressions were clearly visible in the video recordings. During the learning sessions, participants occasionally shifted their gaze, turned their heads due to external stimuli, or covered their faces with their hands. Such instances reduced the accuracy of facial expression data and were therefore excluded from the emotion analysis dataset.

After the learning sessions, the recorded facial expressions were carefully examined. Eight distinct learning emotions were identified and labeled: joy, inquisitiveness, contempt, neutral, irritation, perplexity, tedium, and diversion, as illustrated in Figure 6. Following comprehensive data organization and analysis, 122 high-quality video segments were obtained. Each video, with a resolution of 1280 × 720, provided a vivid depiction of a wide range of learning emotions. This comprehensive dataset served as a critical foundation for in-depth analyses and provided deeper insights into the emotions of learners throughout the study period.

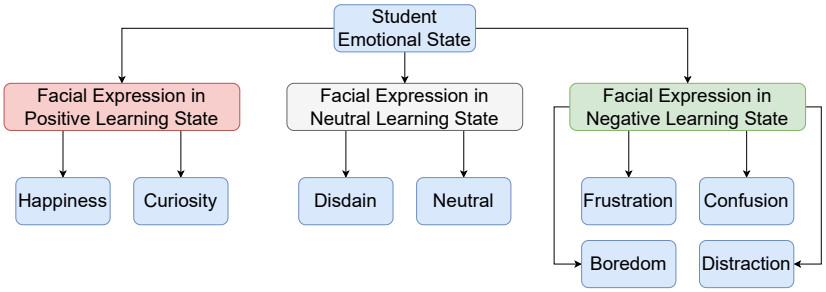


Figure 6. Relationship between a learner's emotions and facial expressions.

Subsequently, the FFMPEG framework was used to process the collected video streams. To generate image sequences, one frame was extracted every 48 frames. Figure 7 illustrates sequences of images extracted from videos depicting the online learning emotions of the selected participants.

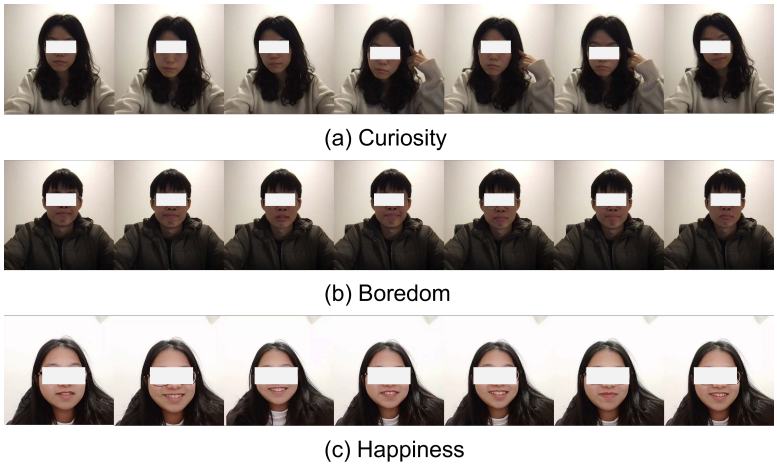


Figure 7. Image sequences of selected participants' emotions during online learning.

Finally, the YOLO-v7 algorithm [44], pre-trained on a large-scale facial recognition dataset, was applied to detect and locate faces in the images. This advanced object detection model effectively identified faces in each frame and provided the corresponding bounding-box coordinates. The positioning information of these bounding boxes was crucial for accurately extracting facial emotion regions from each frame. This targeted extraction ensured that only relevant facial data were analyzed, improving the accuracy of emotion recognition and enabling a more focused exploration of participants' responses during online learning. For example, the final facial images extracted from the sequence of learning emotion images corresponded to the bounding-box positions predicted by YOLO-v7.

4. Experimental Results

This section presents a series of analytical experiments conducted on publicly available facial emotion datasets, including FER2013 [30], RAF-DB [31,32], and RAFD [33]. In addition, experiments were extended to a custom dataset, SCAUOL, specifically designed for emotion analysis in online learning environments. First, we evaluate the emotion classification performance of MobileViT-Local and subsequently evaluate the effectiveness of knowledge distillation by introducing teacher networks with diverse architectures and assessing their impact on model efficiency and accuracy across these datasets.

4.1. Dataset Description

The datasets used in the experiments are described as follows:

(1) FER2013 [30]: Introduced during the Representation Learning Challenge at ICML 2013, this large-scale, unconstrained dataset comprises 35,887 images. The images exhibit challenges such as occlusion, partial facial visibility due to cropping errors, and low resolution, closely reflecting the complexities of real-world scenarios.

(2) RAF-DB [31,32]: A widely used public facial emotion dataset consisting of 15,339 images depicting seven fundamental emotions sourced from the internet. These images represent a variety of ethnicities, genders, and age groups under natural conditions and are accompanied by high-quality annotations. Emotion labels were determined by multiple independent annotators to ensure precision and consistency.

(3) RAFD [33]: Compiled under controlled laboratory conditions, this dataset comprises 1607 high-resolution facial images from 67 subjects. Each subject's expressions were captured from multiple angles, including frontal, lateral, and 45-degree oblique views, making it an exceptional resource for facial expression recognition and emotion analysis.
(4) SCAUOL: This dataset comprises genuine experimental data collected by our institution from 26 participants aged between 19 and 25 years (18 males and 8 females). We identified eight distinct learning emotions: joy, inquisitiveness, contempt, neutral, irritation, perplexity, tedium, and diversion. Recordings were captured via computer cameras in participants' natural, self-directed learning environments. After the videos were recorded, Python 3.9 was used as the development environment, and the FFMPEG framework was employed to extract and store every 48th frame from the captured video stream as an image. Following frame extraction, emotion annotation, object detection, and cropping were performed, resulting in 1519 facial expression images. To address the small size of the dataset, a StyleGAN-based data augmentation method [45] was applied, expanding the dataset to 2519 images.

All datasets were divided into training, validation, and test sets in an 8:1:1 ratio. The experiments were conducted under controlled conditions to ensure consistency across trials. The computational setup comprised a Core i9-9900K processor operating at 3.6 GHz, 16 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. The model was implemented using the PyTorch 2.0 deep learning framework with CUDA version 11.3. During training, consistent hyperparameters were applied: a batch size of 16, an initial learning rate of 1×10^{-4}, a cosine annealing learning rate schedule, the Adam optimizer, 100 training epochs, and the same knowledge distillation parameters—a temperature coefficient of $\tau = 7.0$ and a balance coefficient factor of $\lambda = 0.7$.

4.2. Emotion Feature Extraction

This section evaluates the emotion recognition capability of the proposed model on three benchmark datasets: FER2013, RAFD, and RAF-DB. The analysis compares the performance of MobileViT-Local, which incorporates an enhanced fusion module, with the baseline MobileViT model. The results are systematically summarized in Table 1. Notably, the RAFD dataset incorporates an eighth emotion—contempt—in addition to the seven standard emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). This expansion enhances the comprehensiveness of the evaluation framework and demonstrates the robustness of both models in recognizing a wider range of emotions. For MobileViT-Local, various model sizes were examined—MobileViT-Local-S, MobileViT-Local-XS, and MobileViT-Local-XXS—where S, XS, and XXS denote small, extra small, and ultra-small versions, respectively. For instance, in the case of MobileViT-Local-S, integrating the improved fusion module increased the Top-1 classification accuracy from 77.05% to 77.61% on RAF-DB. On FER2013, the Top-1 accuracy rose from 60.30% to 60.41%, while on RAFD, a more pronounced improvement from 97.93% to 98.55% was observed.

MobileViT-Local combined with knowledge distillation was further validated using the RAF-DB, FER2013, RAFD, and self-constructed SCAUOL online learning emotion datasets. In general, robust image feature extraction models—such as CNNs, transformers, or hybrid architectures—serve effectively as teacher models. Based on the findings of DeiT [46], CNN-based networks tend to perform optimally as teacher models for transformers, likely due to the inductive bias introduced during distillation. Nevertheless, the experimental results indicated that teacher models sharing the same architecture as their student counterparts achieved the greatest distillation gains. This architectural similarity facilitated more efficient knowledge transfer. Consequently, VOLO was selected as the teacher network to investigate the transfer of knowledge from a deeper, larger model to a lightweight student model. To demonstrate the

effectiveness of knowledge distillation, MobileViT-Local was compared with several mainstream baseline models. The results are presented in Table 2, where MobileViT-Local-S with VOLO-D2* denotes the optimized configuration obtained after fine-tuning the balance coefficient λ and the distillation temperature τ. Detailed results from the ablation studies are presented in Tables 3–5.

Table 1. Comparison between the original MobileViT and the modified fusion architecture, MobileViT-Local, across the RAF-DB, FER2013, and RAFD datasets.

Model	Params	Dataset		
		RAF-DB	FER2013	RAFD
		Top-1 Accuracy		
MobileViT-XXS	1.268 M	73.89%	57.51%	95.45%
MobileViT-Local-XXS	1.296 M	74.51%	57.94%	95.86%
MobileViT-XS	2.309 M	75.36%	58.71%	97.10%
MobileViT-Local-XS	2.398 M	75.72%	58.92%	97.72%
MobileViT-S	5.566 M	77.05%	60.30%	97.93%
MobileViT-Local-S	5.797 M	77.61%	60.41%	98.55%

Table 2. Comparison between MobileViT-Local with KD and various CNN- and transformer-based models. The bold formatting represents the optimal results. * denotes the optimized configuration obtained after fine-tuning the balance coefficient λ and the distillation temperature τ. The darker color is, the greater comlexity is.

Model	Params	FLOPS	Memory	Model Size	Dataset			
					RAF-DB	FER2013	RAFD	SCAUOL
					Top-1 Accuracy			
EfficientNet-B0	5.247 M	0.386 G	77.99 MB	20.17 MB	74.28%	57.37%	95.45%	73.24%
MobileNet v3-Large 1.0	5.459 M	0.438 G	55.03 MB	20.92 MB	72.62%	57.57%	93.79%	74.55%
Deit-T	5.679 M	1.079 G	49.34 MB	21.67 MB	71.68%	-	-	44.21%
ShuffleNet v2 2x	7.394 M	0.598 G	**39.51 MB**	28.21 MB	75.13%	58.30%	96.07%	75.43%
EfficientNet-B1	7.732 M	0.570 G	109.23 MB	29.73 MB	73.47%	57.23%	94.20%	75.16%
MobileViT-XXS with VOLO-D2	**1.268 M**	**0.257 G**	53.65 MB	**4.85 MB**	77.64%	60.50%	95.65%	74.60%
MobileViT-Local-XXS with VOLO-D2	1.296 M	0.265 G	53.65 MB	4.96 MB	**78.75%**	**63.20%**	**96.69%**	**76.36%**
ResNet-18	11.690 M	1.824 G	**34.27 MB**	44.59 MB	79.40%	61.19%	98.14%	78.16%
ResNet-50	25.557 M	4.134 G	132.76 MB	97.49 MB	**80.15%**	61.35%	97.72%	82.97%
VOLO-D1	25.792 M	6.442 G	179.49 MB	98.39 MB	79.20%	60.62%	95.45%	**86.33%**
Swin-T	28.265 M	4.372 G	106.00 MB	107.82 MB	77.18%	-	97.10%	33.11%
ConvNeXt-T	28.566 M	4.456 G	145.73 MB	109.03 MB	73.63%	56.95%	94.00%	79.36%
EfficientNet-B5	30.217 M	2.357 G	274.87 MB	115.93 MB	75.07%	58.90%	94.82%	79.96%
EfficientNet-B6	42.816 M	3.360 G	351.93 MB	164.19 MB	75.29%	60.54%	95.24%	80.71%
ResNet-101	44.549 M	7.866 G	197.84 MB	169.94 MB	79.60%	61.88%	97.93%	83.38%
MobileViT-XS with VOLO-D2	**2.309 M**	**0.706 G**	135.32 MB	**8.84 MB**	78.75%	61.28%	97.72%	79.96%
MobileViT-Local-XS with VOLO-D2	2.398 M	0.728 G	135.32 MB	9.18 MB	79.79%	**63.91%**	**98.34%**	80.43%
ConvNeXt-S	50.180 M	8.684 G	233.58 MB	191.54 MB	74.97%	58.74%	92.75%	86.58%
VOLO-D2	57.559 M	13.508 G	300.40 MB	219.57 MB	79.40%	60.04%	97.72%	90.49%
ResNet-152	60.193 M	11.604 G	278.23 MB	229.62 MB	79.63%	61.91%	97.93%	87.42%
CoAtNet-2	73.391 M	15.866 G	589.92 MB	280.23 MB	**82.89%**	64.76%	98.96%	**92.33%**
RegNetY-16GF	83.467 M	15.912 G	322.94 MB	318.87 MB	79.24%	62.68%	97.52%	84.69%
MobileViT-S with VOLO-D2	**5.566 M**	**1.421 G**	**162.18 MB**	**21.28 MB**	81.03%	63.50%	98.96%	83.78%
MobileViT-Local-S with VOLO-D2	5.797 M	1.473 G	**162.18 MB**	22.16 MB	81.42%	65.19%	**99.17%**	85.33%
MobileViT-Local-S with VOLO-D2 *	5.797 M	1.473 G	**162.18 MB**	22.16 MB	81.91%	**66.20%**	**99.17%**	88.72%

Table 3. Ablation study of the effects of different distillation temperatures and balance coefficients on the RAF-DB dataset. The bold formatting represents the optimal results.

Dataset: RAF-DB	
Teacher Model: VOLO-D2	
Student Model: MobileViT-Local-S	
Distillation Temperature τ and Balance Coefficient λ	**Top-1 Accuracy**
temp = 10.0, alpha = 0.7	81.58%
temp = 9.0, alpha = 0.7	81.00%
temp = 8.0, alpha = 0.7	80.64%
temp = 7.0, alpha = 0.7	**81.91%**
temp = 6.0, alpha = 0.7	81.39%
temp = 5.0, alpha = 0.7	81.68%
temp = 4.0, alpha = 0.7	80.77%
temp = 7.0, alpha = 0.9	81.16%
temp = 7.0, alpha = 0.8	81.16%
temp = 7.0, alpha = 0.7	**81.91%**
temp = 7.0, alpha = 0.6	81.03%
temp = 7.0, alpha = 0.5	79.73%

Table 4. Ablation study of the effects of different distillation temperatures and balance coefficients on the FER2013 dataset. The bold formatting represents the optimal results.

Dataset: FER2013	
Teacher Model: VOLO-D2	
Student Model: MobileViT-Local-S	
Distillation Temperature τ and Balance Coefficient λ	**Top-1 Accuracy**
temp = 10.0, alpha = 0.7	65.51%
temp = 9.0, alpha = 0.7	**66.20%**
temp = 8.0, alpha = 0.7	65.24%
temp = 7.0, alpha = 0.7	**65.54%**
temp = 6.0, alpha = 0.7	65.35%
temp = 5.0, alpha = 0.7	65.43%
temp = 4.0, alpha = 0.7	65.18%
temp = 9.0, alpha = 0.9	66.15%
temp = 9.0, alpha = 0.8	65.68%
temp = 9.0, alpha = 0.7	**66.20%**
temp = 9.0, alpha = 0.6	65.26%
temp = 9.0, alpha = 0.5	64.00%

4.3. Ablation Study

To verify the effectiveness of knowledge distillation using teacher models with different architectures, tests were conducted on several networks, including RegNet (CNN-based), CoAtNet (a CNN–transformer hybrid), and VOLO (a transformer with local attention similar to MobileViT). The complete experimental results are presented in Table 6.

Table 5. Ablation study of the effects of different distillation temperatures and balance coefficients on the RAFD dataset. The bold formatting represents the optimal results.

Dataset: RAFD	
Teacher Model: VOLO-D2	
Student Model: MobileViT-Local-S	
Distillation Temperature τ and Balance Coefficient λ	**Top-1 Accuracy**
temp = 10.0, alpha = 0.7	97.93%
temp = 9.0, alpha = 0.7	**98.76%**
temp = 8.0, alpha = 0.7	98.55%
temp = 7.0, alpha = 0.7	**99.17%**
temp = 6.0, alpha = 0.7	98.96%
temp = 5.0, alpha = 0.7	98.55%
temp = 4.0, alpha = 0.7	98.34%
temp = 7.0, alpha = 0.9	98.76%
temp = 7.0, alpha = 0.8	98.55%
temp = 7.0, alpha = 0.7	**99.17%**
temp = 7.0, alpha = 0.6	98.14%
temp = 7.0, alpha = 0.5	98.76%

Table 6. Performance of MobileViT-Local with teacher models of different architectures. The bold formatting represents the optimal results.

Teacher Model	**Dataset**		
	RAF-DB	**FER2013**	**RAFD**
	Top-1 Accuracy		
RegNetY-16GF	79.24%	62.68%	97.52%
CoAtNet-2	**82.89%**	**64.76%**	**98.96%**
VOLO-D2	79.40%	62.13%	97.72%
Student Model with KD			
MobileViT-XXS	73.89%	57.51%	95.45%
MobileViT-XXS with RegNetY-16GF	76.17%	58.94%	**96.89%**
MobileViT-XXS with CoAtNet-2	77.05%	**61.47%**	96.07%
MobileViT-XXS with VOLO-D2	**77.64%**	60.50%	95.65%
MobileViT-Local-XXS	74.51%	57.94%	95.86%
MobileViT-Local-XXS with RegNetY-16GF	76.14%	59.51%	**97.52%**
MobileViT-Local-XXS with CoAtNet-2	77.31%	61.49%	97.10%
MobileViT-Local-XXS with VOLO-D2	**78.75%**	**63.20%**	96.69%
MobileViT-XS	75.36%	58.71%	97.10%
MobileViT-XS with RegNetY-16GF	77.41%	60.60%	97.31%
MobileViT-XS with CoAtNet-2	78.36%	**62.59%**	97.52%
MobileViT-XS with VOLO-D2	**78.75%**	61.28%	**97.72%**

Table 6. *Cont.*

Teacher Model	Dataset		
	RAF-DB	FER2013	RAFD
	Top-1 Accuracy		
MobileViT-Local-XS	75.72%	58.92%	97.72%
MobileViT-Local-XS with RegNetY-16GF	78.23%	60.71%	97.93%
MobileViT-Local-XS with CoAtNet-2	79.50%	62.70%	**98.34%**
MobileViT-Local-XS with VOLO-D2	**79.79%**	**63.91%**	**98.34%**
MobileViT-S	77.05%	60.30%	97.93%
MobileViT-S with RegNetY-16GF	78.78%	62.13%	98.14%
MobileViT-S with CoAtNet-2	79.34%	**63.90%**	98.14%
MobileViT-S with VOLO-D2	**81.03%**	63.50%	**98.96%**
MobileViT-Local-S	77.61%	60.41%	98.55%
MobileViT-Local-S with RegNetY-16GF	79.89%	62.15%	98.76%
MobileViT-Local-S with CoAtNet-2	80.64%	64.08%	98.55%
MobileViT-Local-S with VOLO-D2	**81.42%**	**65.19%**	**99.17%**

Analysis of the results across the three datasets showed that VOLO, when used as a teacher model, achieved the most substantial distillation gains due to its architectural similarity to the student model. Although CoAtNet's hybrid architecture demonstrated the highest prediction accuracy, the results indicate that the structure of an effective teacher model should closely align with the student model to ensure efficient knowledge transfer. The transformer's local attention mechanism—which integrates local image representations, retains the CNN's inductive bias, and captures global feature information—proved particularly effective for image feature extraction.

To further investigate the effects of different distillation temperatures (τ) and balance coefficients (λ) on distillation performance, additional experiments were conducted using VOLO as the teacher network and MobileViT-Local-S as the student network for 100 training epochs. The results indicated that the optimal configuration was achieved when the balance coefficient was 0.7 and the distillation temperature was either 7.0 or 9.0. Detailed experimental results are presented in Tables 3–5.

5. Analysis and Discussion

This section analyzes the performance of the proposed model based on the comparative results and addresses relevant ethical considerations.

5.1. Results Analysis

The evaluation of emotion recognition was conducted on three benchmark datasets to compare the performance of MobileViT-Local, equipped with an improved fusion module, against the baseline MobileViT model. As shown in Table 1, the modifications of the fusion module yielded consistent improvements in classification performance with minimal increase in the number of parameters, regardless of the model's scale. The classification confusion matrices for MobileViT-Local across the datasets are presented in Figures 8–10. Most emotion predictions demonstrate satisfactory accuracy, except for disgust expressions. This may be because certain disgust expressions are subtle and not easily distinguished from neutral expressions. Notably, the accuracy reported in Table 2 on the FER2013 dataset is generally low. However, the improved MobileViT-Local-S with VOLO-D2* model

achieves a comparatively high accuracy of 66.20%. The primary reason is data quality issues: we used raw data with low resolution and some incorrect annotations, which reduced overall accuracy. In future experiments, we plan to process all raw data to enhance classification precision.

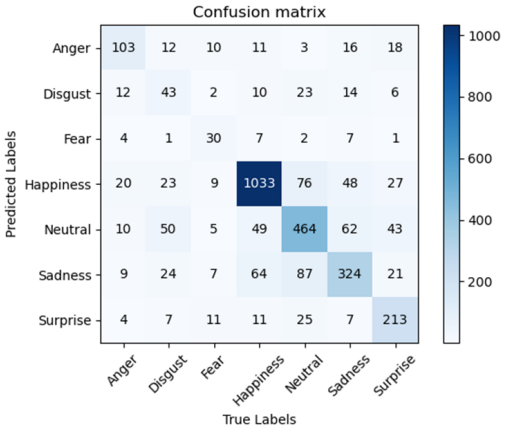


Figure 8. Confusion matrix of emotion recognition on the RAF-DB dataset.

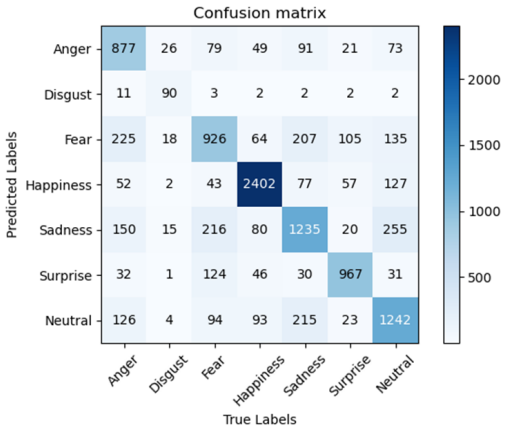


Figure 9. Confusion matrix of emotion recognition on the FER2013 dataset.

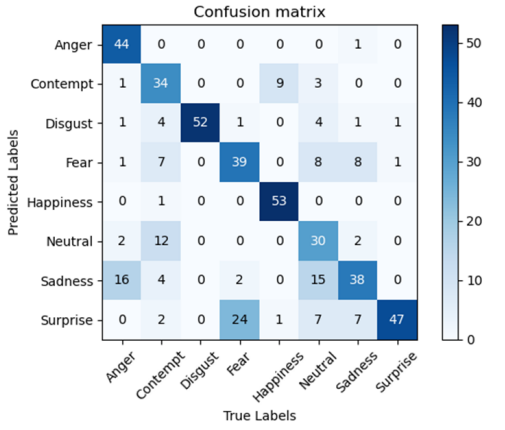


Figure 10. Confusion matrix of emotion recognition on the RAFD dataset.

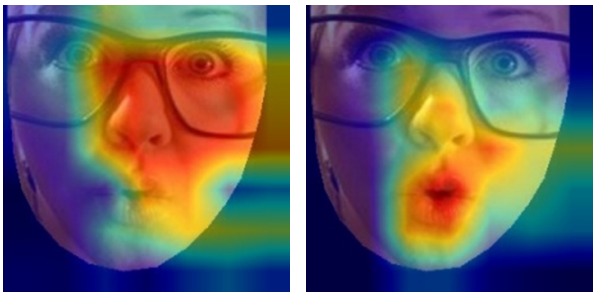
To better understand the input data processed by the student model, Grad-CAM visualization was used to assess the performance gains achieved through knowledge distillation. This technique demonstrates how the network extracts crucial information from input images.

For comparison, two sets of visualization graphs from cases selected from the RAF-DB dataset [31,32] are shown in Figures 11a,b and 12a,b. The model without knowledge distillation relies more on global search and attention during feature extraction. In contrast, the model trained with knowledge transfer from the teacher model focuses more precisely on key regions that distinguish different expressions. For example, in Figure 11, the "happiness" emotion is primarily expressed in the eyes. The left panel focuses only on the facial region without capturing essential details, whereas the right panel identifies precise facial expression features by leveraging the abundant "happiness" labels in the dataset. Similarly, for the "surprised" expression in Figure 12, the critical feature is mouth movement. The left panel captures only partial eye information and misses the crucial mouth activity, whereas the right panel accurately identifies mouth movement patterns by analyzing common characteristics of this emotion.



(**a**) (**b**)

Figure 11. Grad-CAM visualization of the happiness emotion from the RAF-DB dataset [31,32], predicted by MobileViT-Local (**a**) and MobileViT-Local with KD (**b**).



(**a**) (**b**)

Figure 12. Grad-CAM visualization of the surprise emotion from the RAF-DB dataset [31,32], predicted by MobileViT-Local (**a**) and MobileViT-Local with KD (**b**).

5.2. Limitations and Implications

To address the scarcity of emotion data in online learning, this study designed a data collection strategy, constructed an emotion dataset, and annotated eight labels specific to online learning environments. Comparative experiments with mainstream models demonstrated the proposed model's improved generalization and robustness and confirmed the dataset's effectiveness. The dataset also showed high usability, providing valuable support for future research.

Traditional frameworks for online learning emotion analysis rely on recording students' emotions during lectures for offline processing. However, this approach often discourages privacy-conscious students from sharing videos, limiting effective course evaluation. A potential solution is to process real-time emotion data locally on students' devices instead of uploading actual videos. Transmitting this information directly to teachers' devices

allows educators to gain intuitive insights into students' real-time emotional fluctuations during class.

We proposed a prototype system that processes real-time facial emotion features locally on students' personal devices and transmits the computed data to teachers' devices. Since most students use low-end laptops or mobile devices for online classes, implementing real-time emotion recognition on such hardware requires lightweight models that maintain high accuracy and efficiency. Traditional image recognition approaches typically rely on large-scale deep learning networks or multimodal feature fusion systems integrating audio, facial, and body pose data. Although these approaches improve accuracy in detecting students' emotions during classroom interactions, their computational demands make real-time processing on low-end devices impractical. To address this, we implemented an optimized lightweight model using knowledge distillation, combined with an enhanced self-constructed online learning emotion dataset and data augmentation techniques. The trained model was then embedded within the prototype system to perform real-time emotion recognition.

The prototype system comprises a student facial recognition login module and three core functional modules: (1) a student-side facial expression recognition and emotion analysis module, (2) a teacher-side online classroom student emotion recognition module, and (3) a teacher-side backend statistics module.

The student-side module features a login interface for facial recognition and identity authentication. It continuously monitors students' emotions and expressions in real time and transmits the results to the teacher.

The teacher-side modules enable viewing all authenticated students' facial images captured during online classes. The system detects and locates individual student faces in real time to analyze emotions and expressions. It collects students' emotion data and generates relationship diagrams illustrating how learning progress correlates with changes in emotion during class. Finally, it aggregates all students' emotion information, counts students exhibiting negative emotions, and produces comprehensive classroom performance reports.

However, several areas for improvement remain:

(1) The lightweight emotion recognition model based on visual self-attention and knowledge distillation addresses the challenge of running large networks on low-end devices, but it currently relies solely on single-modal facial expression data. Real classroom environments are more complex, and leveraging multimodal information could improve emotion recognition. Therefore, future work should explore lightweight multimodal recognition methods.
(2) The online learning emotion recognition model using generative adversarial data augmentation improved performance on the self-built dataset, but the limited original data means that generated samples may lack quality. Low-quality data could mislead model training. Future efforts should focus on collecting larger datasets and, for generative augmentation, combining large public facial emotion datasets with online learning-specific datasets to enhance data diversity and authenticity.

5.3. Ethical Considerations

In this study, we constructed a self-built dataset using data collected on a campus, contributed voluntarily by participants. All data were legally obtained or generated through observation without interfering with public behavior, in accordance with item 32 of the *Ethical Review Measures for Human Life Science and Medical Research* (National Health Commission of the People's Republic of China, effective 27 February 2023).

In future applications, real-time facial data captured using classroom equipment will remain within the local network, ensuring a closed system with minimal risk. All students

will be informed in advance and retain the right to opt out of participation. The informed consent process will include transparent explanations of the study's purpose, data collection procedures, potential risks and benefits, and participants' rights, including the right to withdraw. Parents or guardians will also be informed as appropriate, and consent will be obtained before any data collection begins.

6. Conclusions

In this study, we optimized the lightweight MobileViT model for resource-constrained environments. By incorporating knowledge distillation into the teacher model, we achieved substantial performance improvements while maintaining comparable parameter counts and memory usage. Building on this optimized network, we introduced a novel application in online education: an emotion analysis framework for online courses. This framework integrates smoothly with existing online learning tools, enabling rapid and precise evaluation of students' emotions. Through these evaluations, instructors can detect potential negative emotions or confusion among students, supporting improved curriculum design and enhancing the overall quality of online courses.

Leveraging the lightweight model, students only need to activate their device cameras to transmit real-time emotion information to teachers' devices. This approach simplifies the emotion analysis process while protecting the privacy of students who are reluctant to share facial videos. Teachers can record the entire online course for analysis, but students who consent to sharing facial images need only grant camera access, allowing instructors to monitor emotional responses in real time. Students who decline to share facial images can still receive an average emotion rating for the course based on individual predictions. Overall, the enhanced MobileViT-Local model yielded promising results on three benchmark emotion datasets as well as a customized online learning emotion dataset, all while maintaining low memory consumption and parameter count.

Author Contributions: Conceptualization, J.W. and Z.Z.; methodology, J.W.; software, Z.Z.; validation, Z.Z.; formal analysis, J.W.; investigation, J.W.; resources, X.C.; data curation, X.C.; writing—original draft preparation, J.W.; writing—review and editing, J.W.; visualization, J.W.; supervision, J.W.; project administration, J.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Platforms and Projects of the Guangdong Provincial Education Department (2023ZDZX4002, 2021ZDZX1078).

Institutional Review Board Statement: According to Item 32 of <Ethical review measures for human life science and medical research> (National Health Commission of the People's Republic of China, effective 27 February 2023) (https://www.nhc.gov.cn/wjw/c100375/202302/902b4a1dc3af4aba862a6387e6e376dc.shtml (accessed on 10 November 2025)). This study qualified for exemption from ethical review due to the following: (1) This research exclusively involves the secondary analysis of publicly accessible, completely anonymized datasets. There is no potential for identifying individual participants or breaching privacy, as the data contains no personal identifiers and cannot be linked back to individuals. (2) This research constitutes a routine evaluation of standard educational practices and curriculum implementation conducted within the researcher's institution. It does not involve sensitive personal data collection, interventions outside normal practice, or procedures posing more than minimal risk. The findings are solely intended for internal quality improvement purposes. (3) This research utilizes pre-existing, archived records or biological specimens that have been irreversibly anonymized prior to the researcher accessing them. The researcher has no access to any key or code that could link the data/specimens back to identifiable individuals. The design and conduct of this study strictly adhere to the requirements stipulated for exempt research within the aforementioned regulation, particularly concerning respect for participant rights (as applicable to the data source), privacy protection, and data security.

Informed Consent Statement: Informed consent for participation was obtained from all subjects involved in the study.

Data Availability Statement: For more details and access to the code, please visit https://github.com/bess926/MobileViT accessed on 10 November 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, X.; Liu, T.; Wang, J.; Tian, J. Understanding learner continuance intention: A comparison of live video learning, pre-recorded video learning and hybrid video learning in COVID-19 pandemic. *Int. J. Hum.-Interact.* **2022**, *38*, 263–281. [CrossRef]
2. Shen, J.; Yang, H.; Li, J.; Cheng, Z. Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimed. Syst.* **2022**, *28*, 469–478. [CrossRef] [PubMed]
3. Darwin, C.; Prodger, P. *The Expression of the Emotions in Man and Animals*; Oxford University Press: Cary, NC, USA, 1998.
4. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115. [CrossRef]
5. Walker, S.A.; Double, K.S.; Kunst, H.; Zhang, M.; MacCann, C. Emotional intelligence and attachment in adulthood: A meta-analysis. *Personal. Individ. Differ.* **2022**, *184*, 111174. [CrossRef]
6. Whitehill, J.; Serpell, Z.; Lin, Y.C.; Foster, A.; Movellan, J.R. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Trans. Affect. Comput.* **2014**, *5*, 86–98. [CrossRef]
7. Dujaili, M.J.A. Survey on facial expressions recognition: Databases, features and classification schemes. *Multimed. Tools Appl.* **2023**, *83*, 7457–7478. [CrossRef]
8. Sajjad, M.; Ullah, F.U.M.; Ullah, M.; Christodoulou, G.; Cheikh, F.A.; Hijji, M.; Muhammad, K.; Rodrigues, J.J. A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. *Alex. Eng. J.* **2023**, *68*, 817–840. [CrossRef]
9. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: London, UK, 2016.
10. Liu, T.; Wang, J.; Yang, B.; Wang, X. NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom. *Neurocomputing* **2021**, *436*, 210–220. [CrossRef]
11. Zhu, B.; Lan, X.; Guo, X.; Barner, K.E.; Boncelet, C. Multi-rate attention based gru model for engagement prediction. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 25–29 October 2020; pp. 841–848.
12. Liu, C.; Jiang, W.; Wang, M.; Tang, T. Group level audio-video emotion recognition using hybrid networks. In Proceedings of the 2020 International Conference on Multimodal Interaction, Online, 25–29 October 2020; pp. 807–812.
13. Liu, T.; Wang, J.; Yang, B.; Wang, X. Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom. *Infrared Phys. Technol.* **2021**, *112*, 103594. [CrossRef]
14. Savchenko, A.V.; Demochkin, K.V.; Grechikhin, I.S. Preference prediction based on a photo gallery analysis with scene recognition and object detection. *Pattern Recognit.* **2022**, *121*, 108248. [CrossRef]
15. Ekman, P.; Friesen, W. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
16. Ekman, P. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychol. Bull.* **1994**, *115*, 268–287. [CrossRef]
17. Ekman, P.; Friesen, W.V. A new pan-cultural facial expression of emotion. *Motiv. Emot.* **1986**, *10*, 159–168. [CrossRef]
18. Ekman, P.; Friesen, W.V. Who knows what about contempt: A reply to Izard and Haynes. *Motiv. Emot.* **1988**, *12*, 17–22. [CrossRef]
19. Ekman, P.; Heider, K.G. The universality of a contempt expression: A replication. *Motiv. Emot.* **1988**, *12*, 303–308. [CrossRef]
20. Matsumoto, D. More evidence for the universality of a contempt expression. *Motiv. Emot.* **1992**, *16*, 363–368. [CrossRef]
21. Mercado-Diaz, L.R.; Veeranki, Y.R.; Marmolejo-Ramos, F.; Posada-Quintero, H.F. EDA-Graph: Graph Signal Processing of Electrodermal Activity for Emotional States Detection. *J. Biomed. Health Inform. J-BHI* **2024**, *28*, 14. [CrossRef]
22. Veeranki, Y.R.; Posada-Quintero, H.F.; Swaminathan, R. Transition Network-Based Analysis of Electrodermal Activity Signals for Emotion Recognition. *Innov. Res. Biomed. Eng. IRBM* **2024**, *45*, 100849. [CrossRef]
23. Veeranki, Y.R.; Mercado-Diaz, L.R.; Posada-Quintero, H.F. Autoencoder Based Nonlinear Feature Extraction from EDA Signals for Emotion Recognition. In Proceedings of the 2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Eindhoven, The Netherlands, 26–28 June 2024; pp. 1–5.
24. Allen, I.E.; Seaman, J. *Digital Compass Learning: Distance Education Enrollment Report 2017*; Babson Survey Research Group: Boston, MA, USA, 2017.
25. Shea, P.; Bidjerano, T.; Vickers, J. *Faculty Attitudes Toward Online Learning: Failures and Successes*; SUNY Research Network: Albany, NY, USA, 2016.
26. Lederman, D. *Conflicted Views of Technology: A Survey of Faculty Attitudes*; Inside Higher Ed: Washington, DC, USA, 2018.

27. Cason, C.L.; Stiller, J. Performance outcomes of an online first aid and CPR course for laypersons. *Health Educ. J.* **2011**, *70*, 458–467. [CrossRef]
28. Maloney, S.; Storr, M.; Paynter, S.; Morgan, P.; Ilic, D. Investigating the efficacy of practical skill teaching: A pilot-study comparing three educational methods. *Adv. Health Sci. Educ.* **2013**, *18*, 71–80. [CrossRef] [PubMed]
29. Dolan, E.; Hancock, E.; Wareing, A. An evaluation of online learning to teach practical competencies in undergraduate health science students. *Internet High. Educ.* **2015**, *24*, 21–25. [CrossRef]
30. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Republic of Korea, 3–7 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; Proceedings, Part III 20, pp. 117–124.
31. Li, S.; Deng, W.; Du, J. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: New York, NY, USA, 2017; pp. 2584–2593.
32. Li, S.; Deng, W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Trans. Image Process.* **2018**, *28*, 356–370. [CrossRef]
33. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
35. 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* **2020**, arXiv:2010.11929.
36. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 18–24 July 2021; pp. 10347–10357.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 10–17 October 2021; pp. 10012–10022.
38. Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; Yan, S. Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6575–6586. [CrossRef] [PubMed]
39. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
40. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
41. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
42. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
43. Yu, J.; Markov, K.; Matsui, T. Articulatory and spectrum features integration using generalized distillation framework. In Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Salerno, Italy, 13–16 September 2016; IEEE: New York, NY, USA, 2016; pp. 1–6.
44. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
45. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4217–4228. [CrossRef]
46. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Hervé, J. Training data-efficient image transformers & distillation through attention. *arXiv* **2020**. [CrossRef]

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.