

Article

A Lightweight Hybrid CNN-ViT Network for Weed Recognition in Paddy Fields

Tonglai Liu ¹, Yixuan Wang ², Chengcheng Yang ^{1,3,*}, Youliu Zhang ⁴ and Wanzhen Zhang ^{1,*}¹ College of Artificial Intelligence, Zhongkai University of Agriculture and Engineering, Guangzhou 510550, China; tonglailiu@zhku.edu.cn² Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong SAR 999077, China³ School of Computer Science and Engineering, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau SAR 999078, China⁴ College of Engineering, South China Agricultural University, Guangzhou 510642, China

* Correspondence: yangchengcheng@zhku.edu.cn (C.Y.); wanzhenzhang@zhku.edu.cn (W.Z.)

Abstract

Accurate identification of weed species is a fundamental task for promoting efficient farmland management. Existing recognition approaches are typically based on either conventional Convolutional Neural Networks (CNNs) or the more recent Vision Transformers (ViTs). CNNs demonstrate strong capability in capturing local spatial patterns, yet they are often limited in modeling long-range dependencies. In contrast, ViTs can effectively capture global contextual information through self-attention, but they may neglect fine-grained local features. These inherent shortcomings restrict the recognition performance of current models. To overcome these limitations, we propose a lightweight hybrid architecture, termed **RepEfficientViT**, which integrates convolutional operations with Transformer-based self-attention. This design enables the simultaneous aggregation of both local details and global dependencies. Furthermore, we employ a structural re-parameterization strategy to enhance the representational capacity of convolutional layers without introducing additional parameters or computational overhead. Experimental evaluations reveal that RepEfficientViT consistently surpasses state-of-the-art CNN and Transformer baselines. Specifically, the model achieves an accuracy of 94.77%, a precision of 94.75%, a recall of 94.93%, and an F1-score of 94.84%. In terms of efficiency, RepEfficientViT requires only 223.54 M FLOPs and 1.34 M parameters, while attaining an inference latency of merely 25.13 ms on CPU devices. These results demonstrate that the proposed model is well-suited for deployment in edge-computing scenarios subject to stringent computational and storage constraints.

Keywords: weed recognition; EfficientViT; RepMBConv; re-parameterization**MSC:** 68T07; 68U10; 92B05

Academic Editor: Ioannis Tsoulos

Received: 17 August 2025

Revised: 1 September 2025

Accepted: 4 September 2025

Published: 8 September 2025

Citation: Liu, T.; Wang, Y.; Yang, C.; Zhang, Y.; Zhang, W. A Lightweight Hybrid CNN-ViT Network for Weed Recognition in Paddy Fields.

Mathematics **2025**, *13*, 2899. <https://doi.org/10.3390/math13172899>

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

During rice cultivation, weeds and seedlings compete intensely for essential resources such as nutrients, water, and space, which seriously hinders the healthy growth of seedlings and poses a direct risk to both yield and grain quality [1]. In addition, weeds create favorable conditions for the proliferation of pests and pathogens [2], often triggering large-scale infestations that further jeopardize rice production. At present, chemical herbicides remain

the dominant approach for weed suppression. Nevertheless, excessive and improper use of these chemicals can harm rice plants, pollute soil and water, and leave harmful residues in agricultural products [3]. By contrast, if herbicides could be applied in a targeted and precise manner while maintaining effective weed control, the overall chemical usage could be markedly reduced [4]. Consequently, the reliable identification of weed species in rice paddies is vital for guiding the rational selection of herbicides and enabling precision management strategies. This underscores the necessity and urgency of advancing efficient and accurate automatic weed recognition technologies for rice agriculture.

Computer vision methods, particularly Convolutional Neural Networks (CNNs), have achieved remarkable success in image analysis tasks, mainly due to their strong ability to capture local features and structural representations. By leveraging local receptive fields together with a weight-sharing mechanism, CNNs can efficiently extract detailed patterns such as edges and textures, which are highly relevant for distinguishing weed species with diverse morphological characteristics. For example, Lam et al. [5] employed the VGG-16 architecture [6] and reported a 92.1% recognition accuracy for *Rumex obtusifolius* in grassland scenarios. Similarly, Nasiri et al. [7] applied U-Net [8], a deep encoder–decoder framework, to achieve fine-grained pixel-level segmentation of sugar beet, weeds, and soil. Their study revealed that integrating well-balanced image datasets with a tailored loss function can substantially enhance segmentation performance. Moreover, Sharpe et al. [9] compared three representative architectures—VGGNet, GoogLeNet [10], and DetectNet—using both full images and locally cropped images. The findings indicated that DetectNet obtained higher accuracy on cropped local patches for weed detection and classification in strawberry crops. In another comparative study, Osorio et al. [11] examined YOLOv3 and Mask R-CNN, both based on Regional Proposal Networks (RPNs), against a Support Vector Machine (SVM). Their results showed that RPN-based models achieved 94% accuracy, outperforming SVM at 88%, thus highlighting the superiority of deep detection frameworks. Despite these advances, CNNs are inherently constrained in modeling long-range dependencies and global contextual relationships. Consequently, their recognition accuracy may decline in complex agricultural scenes, particularly when weeds exhibit overlapping morphologies or appear against cluttered backgrounds.

With the advantage of the self-attention mechanism, Transformer-based architectures exhibit unique strengths in sequential data modeling and global context integration, particularly excelling at capturing long-range dependencies. This property makes them naturally suitable for understanding the spatial distribution and interrelationships of weeds in field imagery. In weed recognition research, Espejo-Garcia et al. [12] reported 98.61% accuracy on the DeepWeed dataset by applying the Swin Transformer [13] with transfer learning. Their approach innovatively replaced the conventional Softmax classification head with a Support Vector Machine (SVM), leading to performance gains. Likewise, Castellano et al. [14] introduced a lightweight Vision Transformer for crop–weed discrimination in high-resolution UAV imagery. By leveraging RGB-based feature transfer in multispectral scenarios with the WeedMap dataset, their model achieved 99.18% accuracy, surpassing state-of-the-art methods. In another study, Jiang et al. [15] evaluated three Transformer variants—Swin Transformer, SegFormer [16], and Segmenter [17]—on a dedicated weed dataset. Their experiments revealed that SegFormer provided the best balance, reaching 75.18% mean accuracy (mAcc) and 65.74% mean Intersection-over-Union (mIoU). Furthermore, Wang et al. [18] proposed a fine-grained weed recognition approach by combining Swin Transformer with a two-stage transfer learning strategy. On a private dataset, their method achieved over 99% across accuracy, precision, recall, and F1-score, highlighting its effectiveness in handling data scarcity while improving recognition robustness. Despite these advantages, Transformers may underperform in capturing fine-grained local visual

cues, which are indispensable for differentiating morphologically similar weed species in rice fields.

Considering these observations, integrating the complementary strengths of CNNs for local feature representation with Transformers' ability in global dependency modeling becomes essential for accurate weed recognition in rice agriculture. To this end, we propose a lightweight hybrid framework, termed **RepEfficientViT**, which combines the benefits of both paradigms while maintaining low parameter counts and computational cost. The network is organized into four stages: the initial high-resolution layers (Stages 1 and 2) employ re-parameterized MBConv (RepMBConv) blocks to extract detailed local structural features of weed images, while the deeper layers (Stages 3 and 4) adopt EfficientViT blocks to capture global semantic dependencies. This design not only enables precise discrimination of visually similar weed species but also effectively models their spatial distribution under complex field conditions. Moreover, by incorporating structural re-parameterization [19] and EfficientViT modules, the framework achieves notable improvements in inference efficiency, making it highly suitable for real-world, lightweight deployment scenarios.

We conducted comprehensive experiments on the self-constructed weed dataset to validate the effectiveness of the proposed **RepEfficientViT** framework. The results indicate that our model attains competitive accuracy with minimal parameter size and computational burden, consistently surpassing conventional lightweight CNN- and Transformer-based baselines. Furthermore, ablation analyses together with Grad-CAM visualizations confirm that integrating RepMBConv and EfficientViT blocks provides significant performance improvements. The main contributions of this work are as follows:

- We introduce **RepEfficientViT**, a new hybrid CNN–ViT architecture tailored for lightweight yet highly accurate weed identification.
- By applying structural re-parameterization to convolutional layers, the proposed model achieves further acceleration in inference without additional parameters or complexity.
- Experimental results on the constructed weed dataset demonstrate that RepEfficientViT outperforms representative CNN and Transformer counterparts in both accuracy and efficiency.

Unlike prior hybrid CNN–Transformer models such as MobileFormer [20], EdgeNeXt [21], ConvNeXt-ViT families [22], and MobileViT2 [23], as well as ensemble-based approaches [24], our design is explicitly CPU oriented. We employ structural re-parameterization that collapses multi-branch convolutions into a single standard convolution at inference, and we perform aggressive down-sampling before self-attention so that the Transformer operates on compact token sequences without specialized kernels. This yields low FLOPs, a small parameter budget, and real-time single-image CPU latency without cluster-scale resources, making our model practical for low-power agricultural devices.

2. Materials and Methods

2.1. Dataset Construction

Between March and April 2023, we constructed a dedicated weed dataset by capturing images in paddy fields located in Guangdong Province, China. To increase the variability of the data, images were collected under different conditions, including multiple times of day (morning, noon, and late afternoon), diverse weather scenarios, and varied shooting angles, as illustrated in Figure 1. Several mobile devices were employed for data acquisition, such as the iPhone 11, iPhone 14 Pro, Realme Q3s, and Honor 30. In total, the dataset comprises 2220 images covering nine weed categories. Specifically, it contains 192 samples of barnyard grass (*Echinochloa crusgalli*), 355 of climbing seedbox (*Ludwigia prostrate*), 279

of alligator weed (*Alternanthera philoxeroides*), 301 of false daisy (*Eclipta prostrata*), 306 of goosegrass (*Eleusine indica*), 290 of red sprangletop (*Leptochloa chinensis*), 164 of hairy bittercress (*Cardamine hirsuta*), 179 of coco-grass (*Cyperus rotundus*), and 154 of threeleaf arrowhead (*Sagittaria trifolia*). The dataset was split into training, validation, and test subsets following a 6:2:2 ratio, as detailed in Table 1. For training, each image was randomly cropped and resized to 224×224 pixels, after which standard data augmentation techniques, including random horizontal and vertical flipping, were applied to further enrich the dataset. The dataset has been made publicly available at <https://github.com/Yangcc0819/RepEfficientViT> (accessed on 1 September 2025), to promote reproducibility and community benchmarking.



Figure 1. Images of weeds in rice fields.

Table 1. Number of images per class.

Weed Species	Training	Validation	Testing
Barnyard grass	116	38	38
Climbing seedbox	213	71	71
Alligator weed	169	55	55
False daisy	181	60	60
Goosegrass	184	61	61
Red sprangletop	174	58	58
Hairy bittercress	100	32	32
Coco-grass	109	35	35
Threeleaf arrowhead	94	30	30

In addition to the random 6:2:2 split used in the main experiments, we also performed a stricter device-based split, where images from one acquisition device were held out entirely for testing. This setting better reflects deployment scenarios in which the model may encounter data from unseen devices. However, the current dataset remains limited to nine weed species collected in Guangdong Province during March to April 2023. While this provides a representative benchmark for paddy fields in South China, it does not capture all weed species or seasonal variations. This limitation is acknowledged and further discussed in Section 4.

2.2. Hybrid CNN–Transformer and Ensemble Approaches

Recent studies have explored hybrid architectures that integrate convolutional networks and transformers. **MobileFormer** builds a two-stream design with MobileNet for local features and a lightweight Transformer for global modeling, connected by bidirectional bridges. Although effective, this requires additional cross-token fusion modules that increase parameters and inference cost, and it is optimized mainly for GPU parallelism. **EdgeNeXt** introduces hardware-aware attention blocks and grouped convolutions to balance accuracy and throughput on GPUs/TPUs, but still depends on windowed attention and specialized kernels that are less suitable for CPUs. **ConvNeXt-ViT families** unify ConvNeXt blocks with ViT-style attention, often leveraging ImageNet pretraining and large training resources; although strong in accuracy, these models are parameter heavy (tens of millions) and unsuitable for low-power deployment. **MobileViT2** employs separable self-attention combined with convolutional priors, improving efficiency on mobile GPUs, but still requires relatively complex attention operations that limit CPU efficiency. **Ensemble-based approaches**, such as Nanni, train multiple CNNs and Transformers with different optimizers and aggregate them to improve accuracy, significantly increasing training and storage costs. In contrast, our method targets CPU-side deployment with a single lightweight hybrid: (i) shallow CNNs are re-parameterized so that multi-branch convolutions collapse into single-path kernels at inference, removing runtime overhead; (ii) aggressive down-sampling is performed before attention, reducing token sequence length and avoiding heavy windowed attention; (iii) the model remains extremely compact (1.34 M parameters, 223.54 M FLOPs), enabling real-time inference on commodity CPUs without large-scale pretraining or ensembles.

2.3. Design of RepEfficientViT

2.3.1. Algorithm Design Overview

To achieve a balance between lightweight design and high recognition accuracy for diverse paddy field weeds, we propose a novel recognition framework termed **RepEfficientViT**, which integrates re-parameterized convolution with the EfficientViT architecture. As illustrated in Figure 2, the network begins with two convolutional layers with a stride of 2, followed by four hierarchical stages, each incorporating progressive downsampling to generate multi-scale feature representations. In the shallow stages, RepMBConv blocks are employed to process high-resolution weed images and capture fine-grained local details. In the deeper stages, EfficientViT blocks are utilized to perform global spatial reasoning over the feature maps. This design exploits the efficiency of convolution for handling high-resolution features at the earlier layers, while leveraging the self-attention mechanism at lower resolutions in the later layers to capture long-range dependencies. By combining these complementary strengths, the model reduces overall computational cost while retaining robust recognition capability.

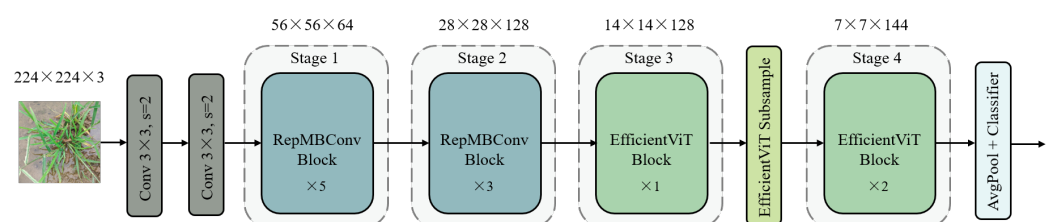


Figure 2. The overall architecture of our proposed RepEfficientViT.

2.3.2. RepMBConv Block

The MBConv module [25] is designed to enhance feature representation while keeping computational cost low. It begins with a 1×1 pointwise convolution that expands the

channel dimension, thereby improving the expressive capacity of the network. This is followed by a 3×3 depthwise (DW) convolution, which performs spatial filtering in a computationally efficient manner. To further strengthen feature discrimination, a Squeeze-and-Excitation (SE) attention mechanism is incorporated, allowing the model to emphasize informative channels while attenuating irrelevant ones. Finally, another 1×1 pointwise convolution reduces the expanded channels back to their original size. Throughout this process, residual connections are employed to facilitate stable gradient propagation and to preserve spatial consistency across layers.

We extend the MBConv block by introducing structural re-parameterization to improve performance while avoiding additional inference complexity. During training, a multi-branch structure is adopted, which is later equivalently transformed into a single branch at inference, as illustrated in Figure 3. This strategy ensures strong representational capacity during training while significantly accelerating inference. The detailed re-parameterization process is depicted in Figure 4.

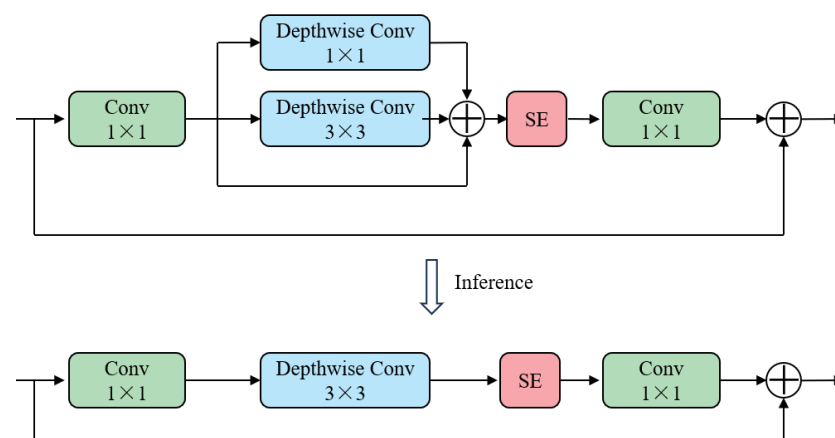


Figure 3. Detailed structure of the RepMBConv block.

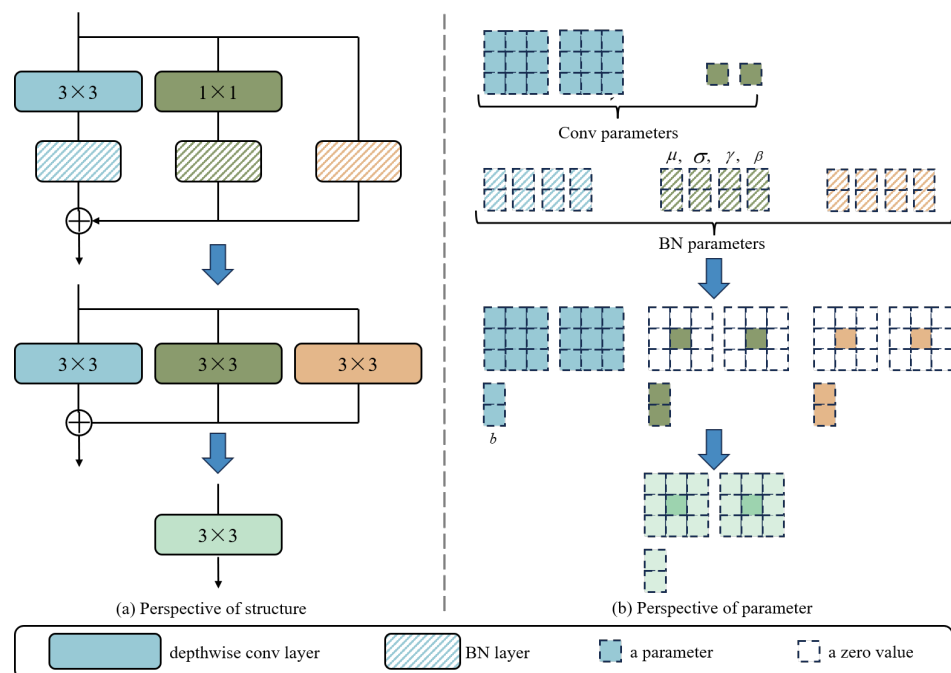


Figure 4. Structural re-parameterization of a RepMBConv block.

For the 3×3 depthwise convolution with Batch Normalization (BN), we denote its weight parameters as $W^{(3)} \in \mathbb{R}^{C_2 \times C_1 \times 3 \times 3}$, where C_1 and C_2 represent the input and

output channels, respectively. The BN parameters include the running mean $\mu^{(3)}$, standard deviation $\sigma^{(3)}$, as well as the learnable scaling and bias terms $\gamma^{(3)}$ and $\beta^{(3)}$.

Similarly, for the branch of depth convolution 1×1 followed by BN, the weight matrix is expressed as $W^{(1)} \in \mathbb{R}^{C_2 \times C_1 \times 1 \times 1}$, with its BN parameters denoted as $\mu^{(1)}$, $\sigma^{(1)}$, $\gamma^{(1)}$, and $\beta^{(1)}$.

Finally, the identity branch, also equipped with BN, is parameterized by $\mu^{(0)}$, $\sigma^{(0)}$, $\gamma^{(0)}$, and $\beta^{(0)}$. Given an input feature map X , the aggregated output Y of the three branches can be formulated as

$$Y = \gamma^{(3)} \frac{W^{(3)} * X - \mu^{(3)}}{\sigma^{(3)}} + \beta^{(3)} + \gamma^{(1)} \frac{W^{(1)} * X - \mu^{(1)}}{\sigma^{(1)}} + \beta^{(1)} + \gamma^{(0)} \frac{X - \mu^{(0)}}{\sigma^{(0)}} + \beta^{(0)} \quad (1)$$

To fuse the three branches into a single branch, we next pad the 1×1 convolution, where only the center position has a value and the other positions are zero. For the identity branch, we equate it to a 3×3 convolution kernel, where only the center position has a value of 1, and the other positions are 0. Consequently, for any branch i , the output Y_i is:

$$Y_i = W_i X \frac{\gamma_i}{\sigma_i} + (\beta_i - \frac{\mu_i \gamma_i}{\sigma_i}) \quad (2)$$

Since μ_i , σ_i , γ_i and β_i are constants, we can smoothly fuse these parameters into the convolution kernel. The weights W'_i and bias b_i of the convolution kernel for this branch are respectively:

$$W'_i = W_i \frac{\gamma_i}{\sigma_i}, \quad b_i = \beta_i - \frac{\mu_i \gamma_i}{\sigma_i} \quad (3)$$

After convolutional padding and BN layer fusion, we obtain three 3×3 depthwise convolutions with the same number of channels. By summing these three depthwise convolutions, we achieve a single depthwise convolution that is rich in information.

2.3.3. EfficientViT Block

To overcome the limitations of conventional Vision Transformers [26], such as redundant computation, poor memory utilization, and heavy parameterization, the EfficientViT block [27] introduces three key components: a memory-efficient sandwich layout, a cascaded group attention mechanism, and a parameter reallocation strategy, as illustrated in Figure 5. These modules collectively improve efficiency in terms of memory cost, computational throughput, and parameter utilization.

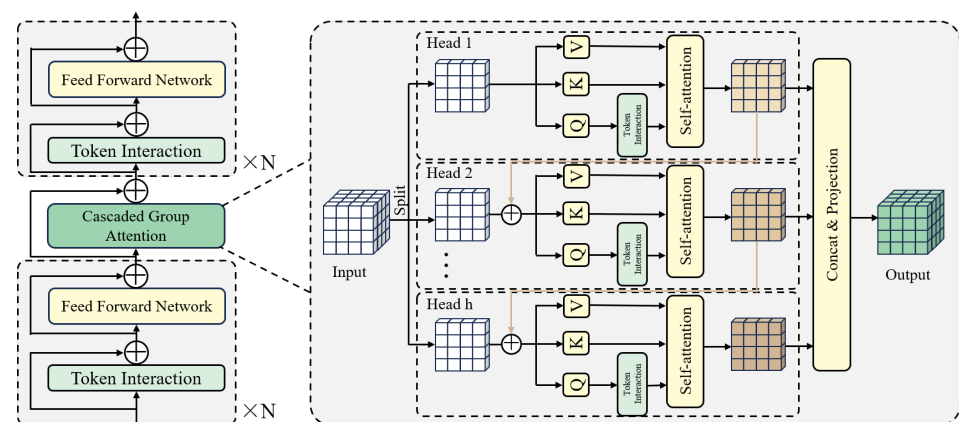


Figure 5. Detailed structure of the EfficientViT block.

Within the sandwich layout, the number of memory-intensive self-attention layers is reduced, while additional lightweight feed-forward network (FFN) layers are inserted

to facilitate inter-channel information flow. This design alleviates the high memory and time consumption typically caused by self-attention and leverages the efficiency of FFNs to enable fast communication across feature channels.

The cascaded group attention mechanism partitions the feature map into subsets assigned to different attention heads, resembling the operation of depthwise convolution. Each head independently performs its attention computation, and the cascading design not only diversifies the distribution of attention but also deepens the network without introducing extra parameters. Consequently, the model gains stronger representational capacity and improved ability to capture complex spatial dependencies.

Finally, EfficientViT employs a parameter reallocation strategy that adaptively adjusts channel widths across different modules. Specifically, channels in critical components are expanded to capture richer feature representations, while those in less significant modules are reduced to eliminate redundancy. This fine-grained allocation accelerates inference, reduces unnecessary overhead, and enhances the overall efficiency of the model.

3. Results and Discussion

3.1. Experiment Setup

In this work, the weed dataset described in Section 3.1 was employed to train and evaluate multiple recognition models. To ensure comparability across different approaches, all models were optimized using AdamW [28] with a weight decay of 0.05 and an initial learning rate of 1×10^{-5} . The training objective was defined as the cross-entropy loss, which measures the discrepancy between predicted labels and ground truth. Each model was trained for 150 epochs with a batch size of 24, and input images were resized to 224×224 pixels. After training, the model parameters corresponding to the highest validation accuracy were selected for testing. Performance was assessed using classification accuracy on the test set as the unified evaluation metric. The detailed experimental settings are summarized in Table 2.

Table 2. Experimental conditions.

Name	Parameter
CPU	Intel Core i7-11700 (Intel Corporation, Santa Clara, CA, USA)
GPU	NVIDIA GeForce RTX 3090 (NVIDIA Corporation, Santa Clara, CA, USA)
System	Ubuntu 20.04
Programming Language	Python 3.7.13
Deep Learning Framework	PyTorch 1.12.1

In addition to the standard random 6:2:2 split used in the main experiments, we also conducted a device-split evaluation. In this setting, all images from one acquisition device were excluded from training and used solely for testing, while the remaining devices provided the training and validation data. This supplementary protocol enables us to assess the robustness of RepEfficientViT and baseline models under heterogeneous device conditions, thereby offering a closer approximation to real deployment scenarios.

To evaluate the effectiveness of the proposed **RepEfficientViT** model, we adopted two complementary approaches: quantitative analysis and Gradient-weighted Class Activation Mapping (Grad-CAM) [29].

For the quantitative analysis, model performance on the test set was examined using multiple indicators, including FLOPs (floating-point operations), number of parameters (Params), inference latency, accuracy, precision, recall, and the F1-score. Specifically, FLOPs provide a measure of computational demand, while Params indicate the overall model complexity in terms of learnable weights. Inference time was obtained through a customized

PyTorch script. The model was first loaded on the designated CPU, and a randomly generated tensor with dimensions $[1, 3, 224, 224]$ was used to simulate a single RGB image of resolution 224×224 . To minimize startup fluctuations, we performed 50 warm-up forward passes before formal measurements. Subsequently, 300 inference runs were executed, and the average execution time (in milliseconds) was reported as the final inference speed.

In addition, accuracy, precision, recall, and F1-score were used as classification metrics. Their definitions are given as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

where TP, TN, FP, and FN denote the numbers of True Positive, True Negative, False Positive, and False Negative samples, respectively.

Grad-CAM is a widely used visualization approach that provides interpretability for deep neural networks by revealing their decision-making process. It highlights the regions in the input image that have the greatest influence on the prediction of a specific class, thereby offering intuitive insights into the model's reasoning. The method operates by tracing the gradients flowing into the last convolutional layers, where the magnitude of the gradients reflects the relative importance of spatial locations. Based on this information, Grad-CAM constructs a heatmap that quantifies the contribution of each pixel to the target prediction. The resulting visualization effectively illustrates the areas of the image that attract the model's attention, thus clarifying which features are most critical to its classification decision.

3.2. Model Evaluating

To assess the effectiveness of the proposed **RepEfficientViT**, we compared its performance against several representative lightweight models. The baselines included CNN-based networks: EfficientNet [30], MobileNetV2 [31], and ShuffleNet v2 [32], Transformer-based architectures: EfficientViT-M1, LightViT-T [33], LVT [34], along with our proposed hybrid model. The results are summarized in Table 3. Among these models, RepEfficientViT demonstrates superior overall performance with fewer parameters and reduced computational cost. Concretely, it achieves 94.09% precision, 94.07% precision, 94.03% recall, and 94.05% F1 score, with an inference time of only 34.86 ms. In comparison, MobileNetV2, ShuffleNet v2, and LVT yield lower classification accuracy, though ShuffleNet v2 and LVT provide faster inference (22.12 ms and 35.91 ms, respectively). While EfficientNet, EfficientViT-M1, and LightViT-T achieve relatively competitive accuracy and speed, RepEfficientViT outperforms them when considering the balance across all metrics.

As shown in Table 4, the device-split protocol holds out all images from one acquisition device as the test set, while the remaining devices provide training and validation data. Unlike the random 6:2:2 split, this setup creates an imbalanced ratio depending on device data volume and poses a more challenging scenario for generalization to unseen devices. Under this stricter evaluation, RepEfficientViT still maintains over 91% accuracy and F1-score, consistently outperforming baseline CNN models. This demonstrates both the robustness of our design and its suitability for real deployment conditions. Although the dataset is geographically and temporally constrained (nine species collected in Guangdong Province

during March–April 2023), the device-split evaluation provides supporting evidence that RepEfficientViT generalizes well to heterogeneous acquisition conditions. This suggests that the proposed approach has the potential to adapt to broader deployment scenarios.

Table 3. Performance comparison of different models on the weed dataset.

Model	FLOPs	Params	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Time (ms)
EfficientNet-B0	412.56 M	4.02 M	91.59	91.87	92.01	91.94	50.03
MobileNetV2	326.28 M	2.24 M	75.45	79.70	76.68	78.16	116.65
ShuffleNet V2	151.69 M	1.26 M	80.68	86.86	80.62	81.72	22.12
EfficientViT-M1	164.22 M	2.77 M	92.27	92.39	92.72	92.55	26.35
LightViT-T	738.74 M	8.09 M	93.18	93.01	93.50	93.25	44.08
LVT	731.63 M	3.42 M	85.45	85.93	85.49	85.71	35.91
RepEfficientViT	223.54M	1.34 M	94.77	94.75	94.93	94.84	25.13

Table 4. Performance of RepEfficientViT under random and device-split protocols. The random split follows a 6:2:2 ratio, while the device-split holds out all images from one acquisition device as the test set, providing a stricter evaluation of generalization.

Protocol	Acc (%)	F1 (%)
Random split (6:2:2)	94.8	94.8
Device-split	91.3	91.0

To further emphasize the distinction between our approach and recent CNN–Transformer hybrids or ensemble-based strategies, we extend the comparison to include MobileFormer, EdgeNeXt, ConvNeXt-Tiny, MobileViT2, as well as a CNN–Transformer ensemble. The corresponding results are summarized in Table 5 under a unified 224×224 evaluation setting. While ensemble methods and ConvNeXt-based models deliver strong accuracy, they demand substantial computational resources and memory, rendering them impractical for low-power devices. MobileFormer, EdgeNeXt, and MobileViT2 offer a more favorable balance, but their designs remain primarily optimized for GPU/TPU acceleration. In contrast, our RepEfficientViT achieves the most balanced trade-off: it attains comparable or superior accuracy (94.77% Acc, 94.84% F1) while maintaining the smallest parameter count (1.34 M), the lowest FLOPs (223.54 M), and the lowest CPU latency (25.1 ms). These results underscore its suitability for real-time inference on resource-constrained agricultural devices.

Table 5. Extended comparison with recent hybrid CNN–Transformer and ensemble approaches under a unified 224×224 setting. Results are reported as average performance over 3 runs. CPU latency is measured with batch = 1 on an Intel i7-11700 CPU in single-thread mode.

Model	Params (M)	FLOPs (M)	Acc/F1 (%)	Latency (ms)
MobileFormer-294M	3.5	294	93.12/93.05	42.7
EdgeNeXt-XXS	2.3	280	93.84/93.71	36.9
ConvNeXt-Tiny	28.6	4500	94.10/94.02	87.5
MobileViT2-XXS	2.0	230	93.56/93.48	31.2
RepEfficientViT (ours)	1.34	223.54	94.77/94.84	25.1

To further examine training dynamics, Figure 6 presents accuracy curves for both training and validation. As shown in Figure 6a, all evaluated models converge within approximately 120 epochs. In general, CNN-based models lag slightly behind Transformer-based models in final accuracy. Interestingly, the training trajectory of LightViT-T closely resembles that of RepEfficientViT; however, as shown in Figure 6b, our model delivers better validation results relative to training accuracy, highlighting its stronger generalization ability.

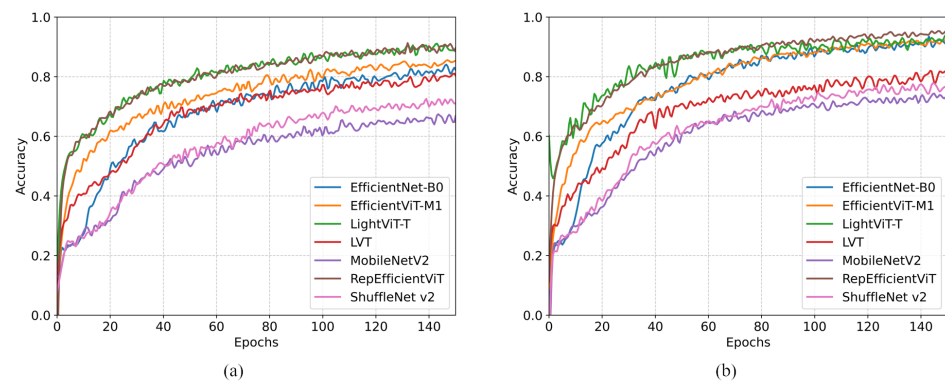


Figure 6. Comparison of training and validation accuracy across various models: (a) the training accuracy graph; (b) the validation accuracy graph.

In addition, Figure 7 provides the confusion matrix of RepEfficientViT on the weed dataset. Rows correspond to ground-truth labels, while columns indicate predicted labels, with each cell showing the number of samples classified into the corresponding category. The visualization reveals that the model attains consistently high accuracy across most weed species. Nonetheless, certain misclassifications occur between morphologically or chromatically similar weeds, such as barnyard grass, red sprangletop, and goosegrass (similar shape), as well as climbing seedbox and alligator weed (similar color). These confusions illustrate the inherent difficulty of distinguishing species with overlapping visual characteristics, even for advanced recognition models.

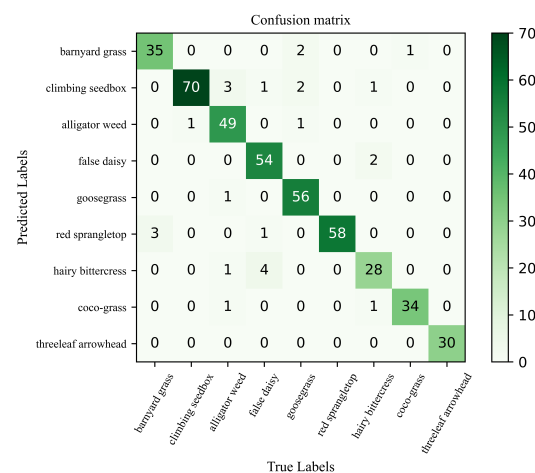


Figure 7. Confusion matrix for weed classification with RepEfficientViT network model.

3.3. Ablation Studies

To comprehensively evaluate the contribution of individual components in the proposed **RepEfficientViT**, we conducted a series of ablation experiments, with results summarized in Table 6. In the baseline EfficientViT design, the network performs $16\times$ downsampling and partitions the model into three stages with stacked EfficientViT blocks. Although this configuration improves computational efficiency, the aggressive downsampling inevitably leads to substantial loss of fine-grained features.

To mitigate this issue, we modified the structure by eliminating the $16\times$ downsampling layer and the third stage of EfficientViT. Instead, two convolutional downsampling layers were introduced at the beginning of the network, followed by two additional stages composed of RepMBConv blocks. This variant, denoted as MBConv + EfficientViT, enhances local feature representation through convolutional operations and thus improves classification accuracy. However, the increased reliance on convolution also results in slower inference due to higher computational cost.

Table 6. Comparison of hybrid structures.

Model	FLOPs	Params	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Inference Time (ms)
EfficientViT-M1	164.22 M	2.77 M	92.27	92.39	92.72	92.55	26.35
MBConv + EfficientViT	226.35 M	1.34 M	93.41	93.69	93.54	93.61	34.96
MBConv (Three Branches) + EfficientViT	229.30 M	1.35 M	94.77	94.75	94.93	94.84	36.37
RepEfficientViT	223.54 M	1.34 M	94.77	94.75	94.93	94.84	25.13

Building upon this variant, we further extended the MBConv blocks by incorporating three parallel branches, referred to as MBConv(Three Branches) + EfficientViT. This design improves recognition accuracy by leveraging richer multi-branch feature extraction, though at the expense of a minor increase in inference latency.

Finally, in our proposed RepEfficientViT, these three-branch MBConv blocks are merged into a single branch at inference time via structural re-parameterization. This re-parameterization enables the model to maintain the accuracy gains achieved during training while significantly reducing inference overhead. As a result, RepEfficientViT attains higher accuracy, faster inference, and improved convergence speed compared to its ablated counterparts.

3.4. Visualization of Grad-CAM

To further examine the quality of visual representations learned by different models, we applied Grad-CAM to generate activation heatmaps for each stage of **RepEfficientViT**, EfficientNet-B0, EfficientViT-M1, and ShuffleNet v2, as illustrated in Figure 8. These visualizations highlight the image regions most relevant to the network predictions. For instance, EfficientNet-B0 is effective at capturing fine-grained details such as weed leaves, while ShuffleNet v2 demonstrates strong responses for goosegrass and climbing seedbox but exhibits difficulty in accurately identifying features of alligator weed and false daisy. In contrast, EfficientViT-M1, lacking CNN components, produces comparatively coarse feature maps.

RepEfficientViT addresses these limitations by leveraging both convolutional and Transformer-based components, enabling it to extract detailed local patterns while simultaneously capturing global contextual cues. The stage-wise activation maps further reveal that our model is capable of focusing on intricate details during the shallow layers and progressively attending to semantically meaningful regions in deeper layers. This hierarchical attention to both low-level and high-level features confirms that RepEfficientViT achieves more robust and comprehensive feature representations than other mainstream lightweight networks.

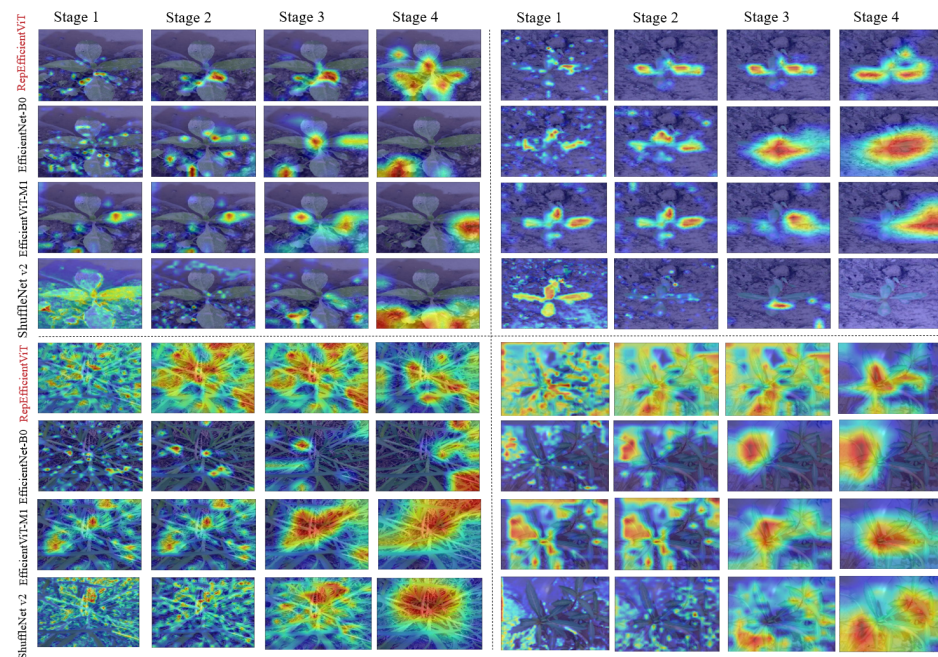


Figure 8. Grad-CAM visualization of the models trained on weed dataset. Red and yellow regions indicate areas of high model attention, while green and blue regions represent low attention.

4. Conclusions

In this work, we presented **RepEfficientViT**, a lightweight hybrid CNN–ViT architecture tailored for weed recognition. The framework is organized into four sequential stages: the first two stages employ lightweight convolutional layers to simultaneously reduce input resolution and capture fine-grained local structures, while the latter two stages utilize an efficient Transformer to model long-range dependencies and global contextual information. Furthermore, structural re-parameterization is applied within the convolutional layers to accelerate inference without compromising representational capacity.

Extensive experiments demonstrate that RepEfficientViT effectively balances accuracy and efficiency by jointly leveraging local and global features. Specifically, the model attains 94.77% accuracy, 94.75% precision, 94.93% recall, and 94.84% F1-score, with only 223.54 M FLOPs and 1.34 M parameters. These results consistently outperform conventional CNN-based and Transformer-based counterparts, providing a promising lightweight solution for weed image recognition.

Compared with prior hybrid CNN–Transformer models (e.g., MobileFormer, EdgeNeXt, ConvNeXt-ViT, MobileViT2) and ensemble strategies, our method emphasizes CPU efficiency and single-model compactness. This enables real-time inference on resource-limited agricultural devices without cluster-scale training or multi-model deployments.

In addition to random splits, we also conducted a device-split evaluation to simulate deployment on unseen acquisition devices. Although accuracy dropped compared with the random split, RepEfficientViT remained robust and outperformed lightweight CNN baselines. This supplementary analysis further confirms the practical applicability of our model in heterogeneous deployment scenarios.

We also recognize that our dataset is restricted to nine weed species collected during March–April 2023 in Guangdong Province. This inevitably limits the generalizability to other regions or seasons. Nevertheless, the use of multiple devices and field locations, together with the supplementary device-split evaluation, demonstrates encouraging robustness. In future work, we plan to extend the dataset to additional provinces, different

seasons, and more weed species to further validate the scalability and generalizability of the proposed approach.

Author Contributions: Conceptualization, T.L.; Methodology, T.L.; Software, T.L. and Y.W.; Validation, Y.W.; Formal analysis, Y.W. and Y.Z.; Writing—original draft, T.L.; Writing—review and editing, C.Y., Y.Z. and W.Z.; Supervision, C.Y.; Project administration, C.Y.; Funding acquisition, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011230, the Science and Technology Program of Guangzhou under Grant 2023E04J0037, the Heyuan Social Science and Agriculture Project under Grant 2023015, the Science and Technology Planning Project of Yunfu under Grant 2023020205, the Key Construction Discipline Research Ability Enhancement Project of Guangdong Province under Grant 2022ZDJS022, and the Guangdong Province Science and Technology Innovation Strategy Special Fund (University Student Science and Technology Innovation Cultivation) Project for 2024 under Grant pdjh2024a199.

Data Availability Statement: The data presented in this study are openly available in GitHub at <https://github.com/Yangcc0819/RepEfficientViT> (accessed on 10 August 2025).

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

References

- Hosoya, K.; Sugiyama, S. Weed communities and their negative impact on rice yield in no-input paddy fields in the northern part of Japan. *Biol. Agric. Hortic.* **2017**, *33*, 215–224. [\[CrossRef\]](#)
- Patterson, D.T.; Westbrook, J.K.; Joyce, R.J.V.; Lingren, P.D.; Rogasik, J. Weeds, insects, and diseases. *Clim. Change* **1999**, *43*, 711–727. [\[CrossRef\]](#)
- Mahzabin, I.A.; Rahman, M.R. Environmental and health hazard of herbicides used in Asian rice farming: A review. *Fundam. Appl. Agric.* **2017**, *2*, 277–284.
- Li, J.; Zhang, W.; Zhou, H.; Yu, C.; Li, Q. Weed detection in soybean fields using improved YOLOv7 and evaluating herbicide reduction efficacy. *Front. Plant Sci.* **2024**, *14*, 1284338. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lam, O.H.Y.; Dogotari, M.; Prüm, M.; Vithlani, H.N.; Roers, C.; Melville, B.; Zimmer, F.; Becker, R. An open source workflow for weed mapping in native grassland using unmanned aerial vehicle: Using *Rumex obtusifolius* as a case study. *Eur. J. Remote Sens.* **2021**, *54* (Suppl. S1), 71–88. [\[CrossRef\]](#)
- Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [\[CrossRef\]](#)
- Nasiri, A.; Omid, M.; Taheri-Garavand, A.; Jafari, A. Deep learning-based precision agriculture through weed recognition in sugar beet fields. *Sustain. Comput. Inform. Syst.* **2022**, *35*, 100759. [\[CrossRef\]](#)
- Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. [\[CrossRef\]](#)
- Sharpe, S.M.; Schumann, A.W.; Boyd, N.S. Detection of Carolina geranium (*Geranium carolinianum*) growing in competition with strawberry using convolutional neural networks. *Weed Sci.* **2019**, *67*, 239–245. [\[CrossRef\]](#)
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–9. [\[CrossRef\]](#)
- Osorio, K.; Puerto, A.; Pedraza, C.; Jamaica, D.; Rodríguez, L. A deep learning approach for weed detection in lettuce crops using multispectral images. *AgriEngineering* **2020**, *2*, 471–488. [\[CrossRef\]](#)
- Espejo-Garcia, B.; Panoutsopoulos, H.; Anastasiou, E.; Rodríguez-Rigueiro, F.J.; Fountas, S. Top-tuning on transformers and data augmentation transferring for boosting the performance of weed identification. *Comput. Electron. Agric.* **2023**, *211*, 108055. [\[CrossRef\]](#)
- 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 (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022. [\[CrossRef\]](#)
- Castellano, G.; De Marinis, P.; Vessio, G. Weed mapping in multispectral drone imagery using lightweight vision transformers. *Neurocomputing* **2023**, *562*, 126914. [\[CrossRef\]](#)
- Jiang, K.; Afzaal, U.; Lee, J. Transformer-based weed segmentation for grass management. *Sensors* **2022**, *23*, 65. [\[CrossRef\]](#) [\[PubMed\]](#)

16. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2021**, *34*, 12077–12090. [\[CrossRef\]](#)
17. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 7262–7272. [\[CrossRef\]](#)
18. Wang, Y.; Zhang, S.; Dai, B.; Yang, S.; Song, H. Fine-grained weed recognition using Swin Transformer and two-stage transfer learning. *Front. Plant Sci.* **2023**, *14*, 1134932. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 13733–13742. [\[CrossRef\]](#)
20. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. MobileFormer: Bridging MobileNet and Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 5270–5279. [\[CrossRef\]](#)
21. Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S.W.; Anwer, R.M.; Khan, F.S. EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 3–20. [\[CrossRef\]](#)
22. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 11976–11986. [\[CrossRef\]](#)
23. Mehta, S.; Rastegari, M. Separable Self-Attention for Mobile Vision Transformers. *arXiv* **2022**, arXiv:2206.02680. [\[CrossRef\]](#)
24. Nanni, L.; Loreggia, A.; Barcellona, L.; Ghidoni, S. Building Ensemble of Deep Networks: Convolutional Networks and Transformers. *IEEE Access* **2023**, *11*, 124962–124974. [\[CrossRef\]](#)
25. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1314–1324. [\[CrossRef\]](#)
26. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [\[CrossRef\]](#)
27. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. EfficientViT: Memory efficient vision transformer with cascaded group attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 14420–14430. [\[CrossRef\]](#)
28. Loshchilov, I. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101. [\[CrossRef\]](#)
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 618–626. [\[CrossRef\]](#)
30. Tan, M. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946. [\[CrossRef\]](#)
31. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520. [\[CrossRef\]](#)
32. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Munich, Germany, 2018; pp. 116–131. [\[CrossRef\]](#)
33. Huang, T.; Huang, L.; You, S.; Wang, F.; Qian, C.; Xu, C. LightViT: Towards light-weight convolution-free vision transformers. *arXiv* **2022**, arXiv:2207.05557. [\[CrossRef\]](#)
34. Yang, C.; Wang, Y.; Zhang, J.; Zhang, H.; Wei, Z.; Lin, Z.; Yuille, A. Lite vision transformer with enhanced self-attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 11998–12008. [\[CrossRef\]](#)

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.