applied sciences

MDPI

Article

CGA-ViT: Channel-Guided Additive Attention for Efficient Vision Recognition

Yayue Zhao [1,2], Jingli Miao [1], Zhenping Li [2], Baiyang Li [2], Anqi Zhuo [2,3] and Yingxiao Zhao [2,*]

[1] School of Information and Electrical Engineering, Hebei University of Engineering, No. 19 Taiji Road, Economic and Technological Development Zone, Handan 056038, China; zyy763031609@gmail.com (Y.Z.)
[2] Information Research Center of Military Science, Academy of Military Sciences, Beijing 100142, China; lizhenping18@mails.ucas.ac.cn (Z.L.)
[3] Mathematics and Physics School, University of Science and Technology Beijing, Beijing 100083, China
* Correspondence: zhaoyingxiao12@nudt.edu.cn; Tel.: +86-1520-143-9261

Abstract

Vision transformers (ViTs) excel at global context modeling with self-attention. However, standard self-attention leads to quadratic computational complexity, which restricts its practical use in high-resolution or latency-sensitive tasks. Existing methods achieve linear complexity via local window constraints or additive approximations. However, they often compromise long-range dependency modeling. To address this issue, we propose the channel-guided additive attention vision transformer (CGA-ViT), which achieves synergistic optimization of multi-scale feature extraction and efficient global context modeling. First, we propose multi-scale dilated feature embedding (MDFE). By designing multi-scale sampling and spatial feature embedding, we can expand the receptive field and capture fine-grained features simply by adjusting the dilation rate in the early stages; second, we design channel-guided additive attention (CGA), dynamically modulating key vectors using query-derived descriptors, enabling long-range semantic interactions while maintaining linear complexity growth. We adopt a hierarchical structure, and in the shallow layers, we use CGA to carry out local-global interactions and use efficient additive attention in deep layers for global integration. Evaluations on ImageNet-1K show that CGA-ViT achieves 84.0% Top-1 accuracy with 4.7 GFLOPs, outperforming Swin-T (81.3%) and ConvNeXt-T (82.1%) by 2.7 and 1.9 percentage points under comparable computational costs. Ablation experiments verify MDFE and CGA, which together contribute to 65.0% of performance gains, with the rest from token-level supervision. Overall, CGA-ViT effectively balances the intrinsic tradeoff between efficiency and global modeling capability, significantly boosts visual recognition performance without extra computational overhead, and provides an efficient solution for lightweight ViT design.

Keywords: vision transformer; attention mechanism; lightweight network; image classification

Academic Editors: Thomas Lindner and Douglas O'Shaughnessy

Received: 30 December 2025
Revised: 2 February 2026
Accepted: 6 February 2026
Published: 10 February 2026

1. Introduction

Vision transformers (ViTs) rely on self-attention for global context modeling, achieving state-of-the-art performance on image classification [1–7], object detection [8–12], and semantic segmentation [13–15]. However, standard self-attention incurs quadratic complexity $O(N^2)$ with spatial resolution N, rendering ViTs impractical for high-resolution inputs or edge deployment [16–19]. Window-based methods [20] restrict attention to local neighborhoods, reducing complexity but limiting long-range dependencies. Linear attention approximations [16] improve throughput at the cost of contextual richness.

Both approaches trade global modeling for efficiency, motivating methods that preserve long-range dependencies without quadratic cost.

In this work, we present CGA-ViT, a hierarchical vision transformer with two core components. First, multi-scale dilated feature embedding applies variable dilation rates to key vectors, enlarging receptive fields without additional parameters. Second, channel-guided additive attention compresses multi-scale features via learned projections, generating low-dimensional representations that reweigh keys through element-wise multiplication. This preserves long-range dependencies while maintaining linear complexity. Because early layers retain multi-scale information, we introduce efficient additive attention for global modeling in later stages.

Our main contributions can be summarized as follows:

- Multi-scale Dilated Feature Embedding Mechanism (MDFE). A dilated convolution-analogous method achieves multi-scale context aggregation.
- Channel-Guided Additive Attention Mechanism (CGA). A channel-wise attention mechanism that guides the query vector to interact with the key vector processed by MDFE, preserving long-range dependencies with linear computational complexity.
- CGA-ViT Architecture. A hierarchical vision transformer that integrates our proposed CGA to achieve a balance between efficiency and global context modeling.

As evidenced in Figure 1, our experiments on the ImageNet-1K benchmark [21] show that our proposed architecture consistently outperforms state-of-the-art ViT variants, both in terms of accuracy metrics and computational efficiency parameters. These results validate our design for applications requiring efficient performance.

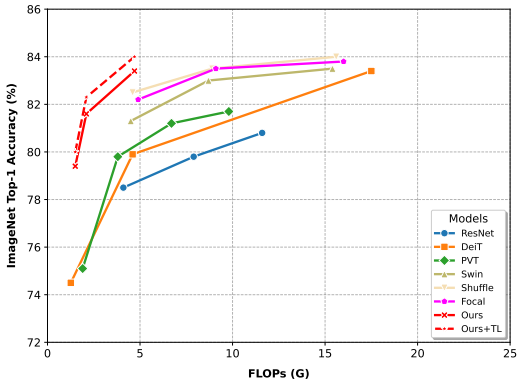

Figure 1. ImageNet-1K accuracy—curve of FLOPs at a 224^2 resolution. CGA-ViT achieves the best balance of accuracy and efficiency at tiny and small model scales.

2. Related Work

Recent work in vision transformers has developed along four directions that motivate our approach: (1) local attention and global attention in vision transformers, (2) multi-scale vision transformers, (3) efficient attention mechanisms, and (4) channel attention mechanisms.

2.1. LocalWindows Attention and Global Attention in Vision Transformers

Vision transformers rely on global self-attention, which incurs quadratic complexity $O(N^2)$ with input size, limiting scalability for high-resolution images [17,18]. Window-based methods reduce this cost by restricting attention to local neighborhoods. Swin Transformer [20] applies self-attention within non-overlapping windows with periodic shifts, reducing computation but introducing boundary discontinuities. CSWin [22] enhances inter-window communication via directional banded attention; however, its fixed-width

patterns constrain adaptive receptive field learning. Pyramid architectures such as PVT [23] employ hierarchical downsampling to reduce cost; however, this compression loses fine-grained details crucial for dense prediction. These limitations motivate our approach to preserve global context without quadratic cost or boundary artifacts.

These approaches typically isolate local windows or rely on aggressive downsampling. In contrast, our CGA module integrates multi-scale dilated embedding for local context modeling with channel-guided additive attention for global aggregation. This design preserves fine-grained details while enabling efficient global reasoning, mitigating the boundary discontinuities inherent to window-based methods.

2.2. Multi-Scale Vision Transformer

This design mismatches the multi-scale requirements of dense prediction tasks. Hierarchical approaches address this by progressively downsampling features. PVT [23] and Swin [20] reduce spatial resolution in deeper layers to capture coarse semantics. CrossFormer [24] extends this with multi-scale patch embedding at the input stage. Alternatively, MPViT [25] processes multiple scales via parallel paths, while Shunted Transformer [26] aggregates tokens of varying granularity. Hybrid models such as Conformer [27] and Mobile-Former [28] integrate convolutional branches for local feature extraction. ViTAE [29] embeds dilated convolutions within attention blocks, enabling multi-scale context without additional parameters.

Our multi-scale dilated feature embedding similarly avoids parameter overhead, but differs in applying dilation directly to key projections rather than attention blocks or patch embedding, enabling more flexible receptive field control.

2.3. Efficient Attention Mechanisms

As shown in Figure 2, researchers have developed various efficient variants to reduce the computational cost of global attention. Sparse attention, exemplified in BigBird [30], implements fixed global-local-random patterns but sacrifices spatial priors. Linearized attention [31,32] approximates the softmax operation through kernel functions or random projections, though it encounters approximation errors and adaptability limitations. Matrix-decomposition approaches like Linformer [33] reduce key/value matrices through low-rank projections, yet their static projections fail to capture dynamic spatial relationships.

Building upon the additive attention framework, we introduce a multi-scale dilated feature embedding mechanism that compresses spatial features through a learnable positional encoding matrix and cross-interacts them with channel-wise features, enhancing channel-wise representational capacity while maintaining low computational overhead.

2.4. Channel Attention Mechanisms

Channel attention originates with SENet [34], which applies global average pooling (GAP) to aggregate spatial information, followed by a bottleneck MLP to learn channel weights. However, SENet neglects spatial structure, limiting performance on localization tasks. CBAM [35] extends this with a spatial attention branch, combining GAP and global max pooling (GMP) for channel features while using channel pooling and 3×3 convolutions for spatial attention. This enables joint optimization of channel and spatial features, though the sequential design adds computational cost and restricts interaction between branches. ECA-Net [36] addresses SENet's dimensionality reduction with 1D convolutions using adaptive kernel sizes, capturing local channel dependencies without compression. This improves efficiency but maintains the limitation of missing spatial information.

We propose a multi-scale dilated feature embedding mechanism that applies dilated convolutions to key projections, expanding receptive fields without additional parame-

ters. A learnable positional encoding compresses spatial features, enabling an interaction between channel and spatial information while preserving computational efficiency.

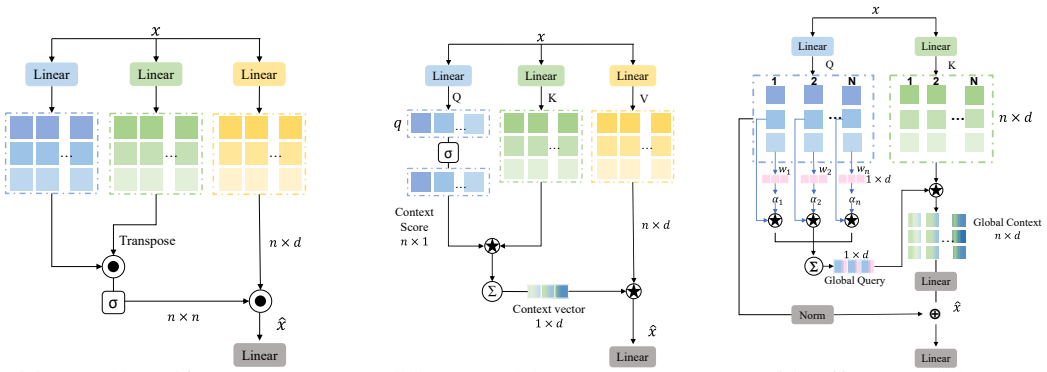

Figure 2. Comparison of self-attention modules. (**a**) Typical self-attention used in ViTs. (**b**) Separable attention in MobileViT-v2. It employs element-wise operations to compute the context vector based on the interactions between the Q and K matrices, and then multiplies the context vector by the V matrix to generate the final output. (**c**) Efficient additive attention. Here, the query matrix is multiplied by learnable weights and pooled to produce global queries. Then, the matrix K is element-wise multiplied by the broadcasted global queries, resulting the global context representation.

3. Method

To address the challenge of capturing multi-scale dependencies in high-resolution medical images without excessive computational costs, we develop CGA-ViT, a hybrid architecture that bridges local convolutional inductive biases with global vision transformer capabilities. Rather than applying standard self-attention across dense feature maps, our approach first decomposes spatial information through a multi-scale dilated feature embedding (Section 3.1, then selectively enhances channel representations via a channel-guided additive attention that reweighs feature importance based on aggregated spatial contexts (Section 3.2, Figure 3). The complete integration of these components into an end-to-end trainable framework is detailed in Section 3.3.

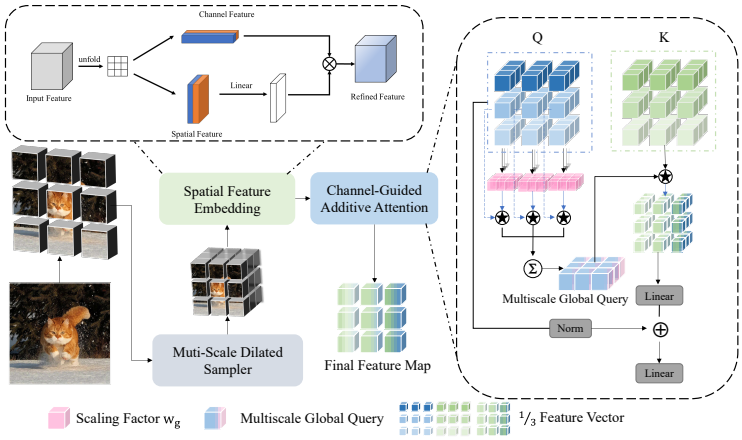

Figure 3. Channel-Guided Additive Attention Block Overview. (1) Multi-scale dilated sampler collects multi-scale context; we extracted features at three different scales. (2) Spatial feature embedding blends spatial and channel cues; we explicitly separated the spatial and channel features of the key vector, and then we performed spatial feature embedding. (3) Channel-guided additive attention combines channel information to guide attention operations.

3.1. Multi-Scale Dilate Feature Embedding

Conventional approaches rely heavily on single-scale features—a reliance that exposes representations to noise corruption and limits them within narrow receptive fields. Breaking free from these constraints, we design multi-scale dilated embedding modules that actively fuse features across the multi-window. The result is a model that shrugs off disturbances far more resiliently while markedly stretching its grasp of expansive contextual information.

3.1.1. Multi-Scale Dilated Sampler

To capture rich local context around each spatial position, we stack two complementary transformations on top of the feature map: we first flatten the 2D grid into continuous 1D patches via dimensional unfolding and then apply dilated sampling to these patches (shown in Figure 4), using a sliding window of size $k \times k$ with variable dilation rates $\delta \in \{1, 2, 3\}$. This sequential operation efficiently captures multi-scale spatial neighborhood information while controlling computational overhead:

$$Z = U_{k,\delta}(X), \tag{1}$$

where $U_{k,\delta}(\cdot)$ represents the unfolding operator. The operation consists of two steps: spatial unfolding followed by dilated sampling with rate. Given $X \in \mathbb{R}^{B \times d \times H \times W}$, this operation samples k^2 spatially distributed neighbors around each position, resulting in an unfolded tensor $Z \in \mathbb{R}^{B \times d \times k^2 \times N}$, where $N = H \times W$ is the total number of spatial locations.

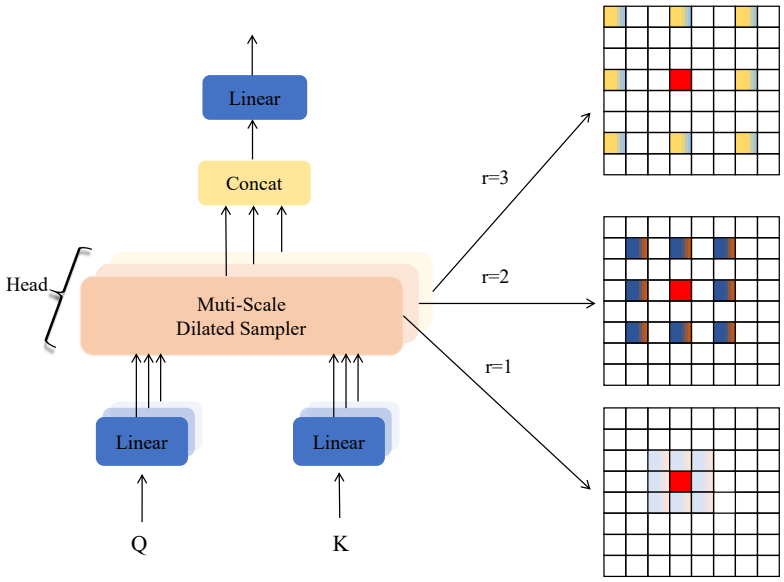

Figure 4. Multi-Scale Dilated Sampler. Channels are split into heads, each head sampling within a 3×3 window but with dilation rates $\delta \in \{1, 2, 3\}$. The red squares indicate the central position, and the mixed-color squares correspond to features extracted at different dilation rates. The outputs are passed through a linear projection layer that effectively fuses the information.

3.1.2. Spatial Feature Embedding

Direct fusion of channel-wise and spatial features escalates computational overhead and introducing redundancy that causes overfitting. We mitigate this via a learnable linear projection to adaptively compress spatial dimensions pre-fusion, enabling tight spatial-channel coupling without representational capacity loss:

$$\hat{X}_{b,n,c} = \underbrace{\sum_{i=1}^{k^2} w^{(i)} \cdot X_{b,n,c,i}}_{\text{Spatially weighted sum}} + b \tag{2}$$

Here, $W = [w^{(1)}, \ldots, w^{(k^2)}] \in \mathbb{R}^{k^2}$ is a shared projection matrix and b is a learnable bias term, gathering each channel's k^2 neighborhood features through weighted summation. Reshaping and aligning these spatial-channel features in latent space generates fused representations $K \in \mathbb{R}^{B \times d \times N}$ that act as keys for efficient attention computation:

$$K_{b,n,d} = f\left(\sum_{i=1}^{k^2} w_i \cdot K_{b,n,d,i}\right) \tag{3}$$

where $K \in \mathbb{R}^{B,N,d}$ represent the joint feature representations and f denotes the GELU activation function, which realizes the mapping of spatial information to the channel domain.

3.2. Channel-Guided Additive Attention

As illustrated in Figure 5, our channel-guided additive attention mechanism processes the token embeddings Q and $K \in \mathbb{R}^{B \times N \times d}$ from the MDFE module. The proposed architecture incorporates a learnable global channel-wise descriptor $G \in \mathbb{R}^{B \times d}$ that is adaptively generated through weighted aggregation of query representations across the sequence of tokens.

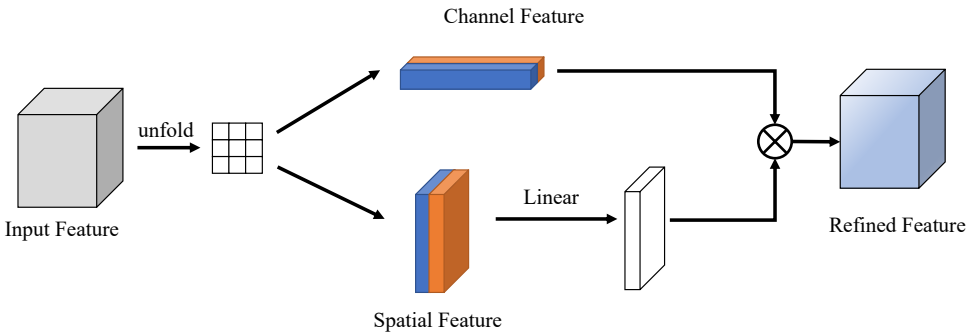

Figure 5. Spatial Feature Embedding. A dilated unfolding operator is a process that rearranges each neighborhood of the input feature map into a token sequence. The number of neighborhoods the operator rearranges is called the $k \times k$ neighborhood. The sequence is subdivided into two branches: a channel branch that preserves channel semantics and a spatial branch that compresses local geometry through a shared linear layer. The two latent representations are recombined through feature crossing, enabling the global channel context to refine every spatial location before the subsequent attention block.

We first derive per-token attention weights using a learnable vector $w_g \in \mathbb{R}^d$:

$$A_i = \frac{1}{\sqrt{d}}(Q_i \cdot w_g), \quad i = 1, \ldots, N \tag{4}$$

where the scaling factor $\frac{1}{\sqrt{d}}$ stabilizes gradient magnitudes. The global descriptor aggregates queries via the following weights:

$$G = \sum_{i=1}^{N} A_i * Q_i. \tag{5}$$

Based on this global descriptor, we propose a channel-guided additive attention mechanism. Specifically, we broadcast G to all spatial positions and perform element-wise multiplication with the key features $K_i \in \mathbb{R}^d$ at each location i. We then project the result back to the original feature space via learnable linear transformation:

$$\hat{x} = L(G * K) + \hat{Q}. \tag{6}$$

where $\hat{Q}$ denotes the normalized query matrix and L indicates the linear transformation.

3.3. Overall Architecture

As shown in Figure 6, CGA-ViT comprises four stages. Stages 1–2 use CGA modules for local detail modeling, while stages 3–4 adopt EAA [16] for computational efficiency. We employ overlapping 3×3 convolutions for patch embedding and inter-stage downsampling [37], preserving spatial continuity. Conditional position embedding (CPE) [38] handles variable input resolutions without explicit interpolation. This adaptive mechanism dynamically adjusts positional encodings to accommodate multi-scale feature resolutions. Specifically, our overall architecture is described as follows:

$$X = \mathrm{DwConv}(\hat{X}) + \hat{X}, \tag{7}$$

$$Y = \begin{cases} \mathrm{CGA}(\mathrm{Norm}(X)) + X, & \text{low-level stages,} \\ \mathrm{EAA}(\mathrm{Norm}(X)) + X, & \text{high-level stages,} \end{cases} \tag{8}$$

$$Z = \mathrm{MLP}(\mathrm{Norm}(Y)) + Y, \tag{9}$$

where $\hat{X}$ is the input of the current block, i.e., the image patches or the output from the last block. We implement CPE as a depth-wise convolution (DwConv) module with zero padding and 3×3 kernel size. We add MLP following prior works, which consists of two linear layers with a channel expansion ratio of 4 and one GELU activation.

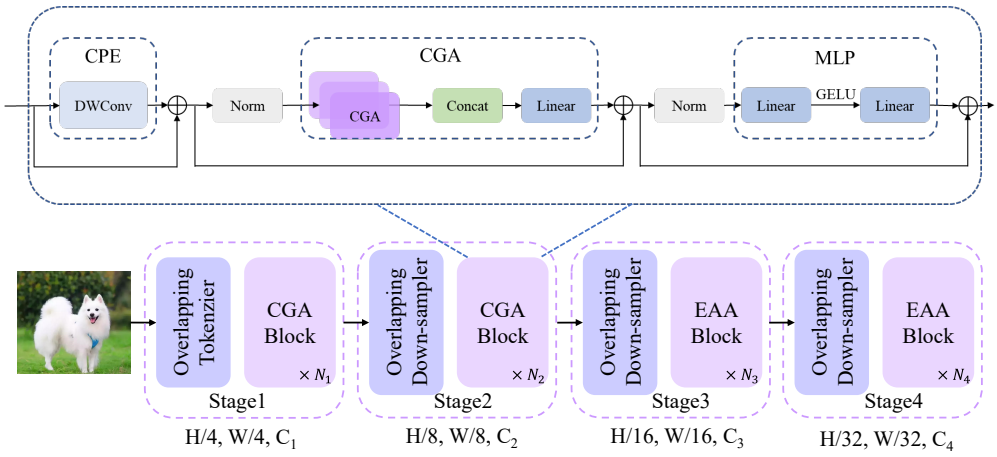

Figure 6. Overall architecture of CGA-ViT. The above figure shows the proposed channel-guided additive attention (CGA) module, which is composed of DwConv, a channel-guided additive attention block, and MLP. The lower figure shows CGA-ViT, which consists of overlapping tokenizers, overlapping downsamplers, a channel-guided additive attention block, and an efficient additive attention block.

4. Experiments

To evaluate the performance of our CGA-ViT, we used the model as the visual backbone for the ImageNet-1K classification. In addition, we evaluated the effectiveness of key modules through ablation studies. All experiments were performed on servers equipped with 8 NVIDIA A100 GPUs.

4.1. Implementation Details

We train CGA-ViT on ImageNet-1K (1.28 M/50K train/val images) following DeiT [39] and PVT [23]: AdamW optimizer, 200 epochs (10 warm-up + 10 cool-down), cosine LR decay from 0.001, batch size 1024, weight decay 0.05. For enhanced supervision, we

adopt token labeling [40] with an auxiliary token-level head and MixToken augmentation (replacing CutMix/MixUp).

4.2. Image Classification

To comprehensively evaluate CGA-ViT, we conducted comparative analyses with three distinct model architectures: (i) lightweight ConvNets (MobileNet-V2/V3, EfficientNet-B0, ConvNeXt-T); (ii) pure ViTs (DeiT, Swin); and (iii) hybrid or MetaFormer variants (MobileViT-v2, PoolFormer, EfficientFormer, SwiftFormer). The comparison of our CGA-ViT models (Tiny, Small, Base) with these state-of-the-art methods is presented in Table 1.

Table 1. ImageNet-1K (224^2). * indicates models trained with token labeling. Lower is better: ↓. Higher is better: ↑. The latency is measured on an iPhone14 Neural Engine (iOS 16), and the throughput is measured on an NVIDIA A100 GPU. Our results are shown in bold for all model variants.

Model	Type	Params (M) ↓	FLOPs (G) ↓	Top-1 (%) ↑	Latency (ms) ↓	Throughput (A100) ↑
MobileViT-v2×1.0 [19]	Hybrid	4.9	1.8	78.1	1.7	3201
DeiT-T [39]	Transformer	5.7	1.3	72.2	3.8	5860
PVT-T [23]	Transformer	13.2	1.9	75.1	-	-
MobileNet-V3-Large [41]	CNN	5.4	0.6	75.1	0.9	10,351
ResNet-50 [1]	CNN	25.5	4.1	78.5	1.9	4835
EfficientNet-B0 [42]	CNN	5.3	0.7	77.1	1.3	8537
EdgeViT-XS [43]	Hybrid	6.7	1.9	77.5	2.7	4812
SwiftFormer-S [16]	Hybrid	6.1	1.6	78.5	0.8	5051
MobileNet-V2×1.4 [44]	CNN	6.1	0.8	74.7	0.9	7447
Ours-T	**Hybrid**	**6.2**	**1.9**	**Ours-T**	**1.8**	**6372**
Ours-T *	**Hybrid**	**6.5**	**1.9**	**Ours-T**	**1.8**	**6372**
MobileViT-v2×1.5 [19]	Hybrid	10.6	4.0	80.4	3.4	2356
PoolFormer-S12 [45]	Pool	12.0	2.1	77.2	1.2	3227
ResNet-101 [1]	CNN	44.5	7.9	79.8	3.3	2842
Swin-T [20]	Transformer	29.0	4.5	81.3	-	-
PVT-S [23]	Transformer	24.0	3.8	79.8	-	-
EfficientFormer-L1 [17]	Hybrid	12.3	2.4	79.2	1.1	5046
SwiftFormer-L1 [16]	Hybrid	12.1	2.5	80.9	1.1	4469
Ours-S	**Hybrid**	**9.6**	**2.2**	**81.6**	**2.4**	**5143**
Ours-S *	**Hybrid**	**10.0**	**2.2**	**82.3**	**2.4**	**5143**
DeiT-S [39]	Transformer	22.5	4.6	79.9	9.9	2990
Swin-S [20]	Transformer	50.0	8.7	83.0	-	-
PVT-L [23]	Transformer	61.4	9.8	81.7	-	-
ConvNeXt-T [3]	CNN	28.6	4.5	82.1	2.5	3235
ResNet-152 [1]	CNN	60.2	11.6	80.8	4.9	1856
PoolFormer-S36 [45]	Pool	31.0	5.2	81.4	2.8	1114
EfficientFormer-L3 [17]	Hybrid	31.3	5.4	82.4	2.0	2691
MobileViT-v2×2.0 [19]	Hybrid	18.5	7.5	81.2	7.5	1105
SwiftFormer-L3 [16]	Hybrid	28.5	5.8	83.0	1.9	2890
Ours-B	**Hybrid**	**23.2**	**4.7**	**83.4**	**3.8**	**3380**
Ours-B *	**Hybrid**	**23.2**	**4.7**	**84.0**	**3.8**	**3380**

Comparison with ConvNets. Our CGA-ViT demonstrates substantial performance advantages over conventional lightweight CNN counterparts. CGAViT-T achieves superior Top-1 accuracy by 2.3% and 4.3% margins compared to EfficientNet and MobileNet, respectively, while showing 0.5 ms and 0.9 ms higher inference latency. CGAViT-T exceeds ResNet-50 by 0.9% in Top-1 accuracy with 0.1 ms faster inference speed. The larger CGAViT-B variant surpasses ResNet-152 and ConvNeXt-T by 2.6% and 0.9% accuracy, respectively, while maintaining 1.1 ms lower latency than ResNet-50. These findings indicate superior performance metrics to lightweight CNNs despite slightly higher latency in specific configurations.

Comparison with Transformer Models. Table 2 compares CGAViT with conventional transformers under similar parameter budgets. CGAViT-T improves DeiT-T by 7.2% Top-1 accuracy while reducing latency by 2 ms, despite identical model sizes. It also outperforms PVT-T by 4.4% with 50% fewer parameters. At larger scales, CGAViT-B reaches 84.0% accuracy with token labeling, exceeding Swin-S and DeiT-S by 1.0% and 4.1%, respectively. These gains suggest that the cross-grained attention design achieves better parameter efficiency than standard self-attention.

Comparison with Hybrid Architectures. Hybrid architectures reduce latency and parameters compared to pure transformers, but their fusion of local and global features often introduces semantic misalignment. For example, SwiftFormer achieves nearly twice the throughput of CGAViT, yet trails by 1.0–1.5% in accuracy. Conversely, CGAViT outperforms MobileViT-v2 on both accuracy and latency. These results suggest that tightly coupled multi-scale attention mitigates the local-global feature misalignment common in hybrid designs, improving the accuracy-efficiency tradeoff.

Table 2. Performance comparison of different backbones on detection, instance segmentation, and semantic segmentation tasks. Our results are shown in bold for all model variants.

Backbone	Detection & Instance Segmentation						Semantic mIoU%
	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	
ResNet18 [1]	34.0	54.0	36.7	31.2	51.0	32.7	32.9
PoolFormer-S12 [45]	37.3	59.0	40.1	34.6	55.8	36.9	37.2
EfficientFormer-L1 [17]	37.9	60.3	41.0	35.4	57.3	37.3	38.9
PVT-L [23]	40.4	62.9	43.8	37.8	60.1	40.3	39.8
SwiftFormer-L1 [16]	41.2	63.2	44.8	38.1	60.2	40.7	41.4
CGAViT-S (Ours)	**42.3**	**64.3**	**45.6**	**38.8**	**61.4**	**41.6**	**42.6**
ResNet50 [1]	38.0	58.6	41.4	34.4	55.1	36.7	36.7
PVT-S [23]	42.9	65.0	46.6	39.5	61.9	42.5	44.8
PoolFormer-S24 [45]	40.1	62.2	43.4	37.0	59.1	39.6	40.3
Swin-T [20]	42.2	64.4	46.2	39.1	64.6	42.0	41.5
EfficientFormer-L3 [17]	41.4	63.9	44.7	38.1	61.0	40.4	43.5
SwiftFormer-L3 [16]	42.7	64.4	46.7	39.1	61.7	41.8	43.9
CGAViT-B (Ours)	**43.8**	**65.6**	**47.7**	**40.3**	**62.8**	**43.0**	**45.1**

4.3. CGAViT as Backbone

Object Detection and Instance Segmentation.We adopt CGAViT as the backbone network for object detection and instance segmentation tasks, integrating it into the Mask-RCNN [8] framework to evaluate its performance. Experiments are conducted on the COCO 2017 [46] dataset, which comprises 118,000 training images and 5000 validation images. Boasting diverse scenarios and object categories, these datasets can fully validate the model's generalization capability. For model initialization, we use pre-trained weights from ImageNet-1K for the CGAViT backbone; this ensures efficient convergence and solid initial performance right from the start. On the optimization front, we employ AdamW, the mainstream optimizer in this field, with an initial learning rate set to 2×10^{-4}. The model is trained for 12 total epochs, and input images are fixed at a size of 1333×800. To guarantee fair comparisons, all experimental settings are aligned with those of existing related works.

We compare CGAViT against CNN and transformer backbones on object detection and instance segmentation (Table 2). At similar computational budgets, CGAViT shows improvements of 3.4 AP^{box} and 3.7 AP^{mask} over ResNet-50 and outperforms SwiftFormer-L3 by 1.1% and 1.2%, respectively. These consistent gains across different architectures suggest that CGAViT is an effective backbone for visual tasks.

Semantic Segmentation.We evaluate CGA-ViT on ADE20K [47,48], which contains 20K training and 2K validation images across 150 categories. Following standard practice,

we pair our backbone with Semantic FPN [49] and initialize from ImageNet-1K pre-trained weights. We train for 40K iterations with a batch size of 32 (8 GPUs), using AdamW with an initial learning rate of 2×10^{-4} and poly decay (power = 0.9). Input images are resized to 512×512 during training; at test time, the short side is resized to 512 with proportional scaling.

The segmentation results in Table 2 validate CGA-ViT's scalability to dense prediction tasks. At comparable complexity, CGA-ViT-B achieves 45.1% mIoU on ADE20K, exceeding SwiftFormer-L3 by 1.2 percentage points (43.9%) with fewer parameters. This gain stems from the channel-guided attention's explicit modeling of global context, which reduces ambiguity in boundary regions where CNNs and local attention struggle.

4.4. Ablation Study

To evaluate the contribution of each proposed component, the analysis begins with the SwiftFormer-XS backbone and progressively incorporates multi-scale dilated feature embedding (MDFE), channel-guided additive attention (CGA), and token labeling (TL). All experiments utilize the unified training recipe and identical random seeds, ensuring rigorous comparability. Table 3 presents Top-1 accuracy on ImageNet-1K.

Table 3. Ablation on CGAViT-T with 224^2 inputs. Modules are added cumulatively; △Top-1 denotes the gain over the SwiftFormer-XS baseline.

Configuration	Top-1 (%)	△Top-1 (%)
SwiftFormer-XS (baseline)	75.7	–
+ MDFE †	78.5	+2.8
+ MDFE + CGA ‡	79.4	+3.7
+ MDFE + CGA + TL §	80.0	+4.3

† MDFE enlarges each attention head's receptive field through dilated token mixing, providing multi-scale contextual information with minimal computational overhead. ‡ CGA integrates channel attention with spatial feature processing; channel attention mechanisms guide Query and Key calculations. § TL provides pixel-level supervisory signals, enabling early network layers to capture position-sensitive visual cues without introducing any additional inference cost.

Discussion: (1) MDFE independently achieved a significant +2.8 pp improvement, demonstrating that rich feature representation provides significantly higher performance for small models compared to simply increasing depth. (2) The introduction of CGA contributed to an additional +0.9 pp improvement. Analysis indicates that more flexible attention distribution enables the model to focus on key areas, enhancing its ability to model long-range dependencies and local details. (3) TL yields a further +0.6 pp without affecting inference cost, making it a highly cost-effective add-on for lightweight networks. Roughly 65% of the total 4.3 pp gain arises from architectural changes (MDFE + CGA), with the remaining 35% credited to finer-grained supervision (TL).

Dilation Rate:When different dilation rates share the same attention head, features end up competing for channel space; this "channel competition" destabilizes training and sends gradients haywire. We sidestep this by partitioning heads so their count always matches dilation scales as integer multiples; specifically, we redistribute heads and per-head dimensions across dilation groups while keeping the total feature length intact. Based on the performance on the ImageNet-1K classification task, we analyze the impact of dilation scales. Table 4 presents the corresponding relationships among the number of heads, dilation scales, and Top-1 accuracy in Stage 1 or Stage 2, revealing that multi-scale dilated attention—with dilation rates [1, 2, 3]—hits 79.4% Top-1 accuracy, leaving single-scale setups ([1], [2], or [3] alone) trailing at 78.7–78.9%. The gap makes sense: blending scales harvests richer hierarchical semantics than any single receptive field can capture. In

addition, the dilation rates need to be kept moderate: excessively narrow scales will lose global dependencies, while overly wide scales tend to introduce redundant information. In contrast, the combination [1, 2, 3] can simultaneously balance the locality and sparsity of attention, thus avoiding the redundancy caused by indiscriminate modeling. Therefore, we set the default number of dilation scales of the model to 3 groups, corresponding to the dilation rate combination [1, 2, 3].

Table 4. Top-1 accuracy on ImageNet-1K of different dilation scales.

Scale	Head Num in Stage 1/2	Window Size	Dilation Rate	Top-1 (%)
Single	[3, 6]	3×3	[1]	78.7
		3×3	[2]	78.9
		3×3	[3]	78.8
Multi	[2, 4]	3×3	[1, 2]	79.0
	[3, 6]	3×3	[1, 2, 3]	79.4
		3×3	[2, 3, 4]	79.1
		3×3	[3, 4, 5]	79.0
	[4, 8]	3×3	[1, 2, 3, 4]	79.2

Scaling Factor: We ran targeted ablations on ImageNet-1K to test whether the scaling factor actually matters (as shown in Table 5).The experimental results demonstrate that removing the scaling factor entirely causes the model to implode—gradients explode within the first 20 epochs, training never stabilizes even after 200 epochs, and Top-1 accuracy collapses 2.8% below our best setup. Locking in a fixed scaling term prevents the explosion, yet convergence drags and performance still suffers. The real payoff comes from pairing learnable weights with the scaling factor: training snaps 40 epochs into place faster than the fixed-term baseline, while Top-1 accuracy climbs an additional 1.2%. These results fully prove that the w_g can adaptively calibrate the importance weights of tokens, thereby effectively enhancing the capability of global context modeling.

Table 5. Ablation study on scaling factor.

Scaling Factor	Epochs	Top-1 (%)	Gradient Explosion
$w_g / \sqrt{d}$	110	79.4	No
$1 / \sqrt{d}$	150	78.2	No
No Scaling factor	200 (Not converged)	76.6	Yes

4.5. Qualitative Results

The attention map visualizations are presented as illustrated in Figure 7. CGA-ViT exhibits high precision in delineating object regions, particularly in scenes of considerable complexity. Integrating multi-scale dilate feature embedding and channel-guided additive attention facilitates the comprehensive capture of detailed and contextual information, thereby enhancing prediction accuracy.

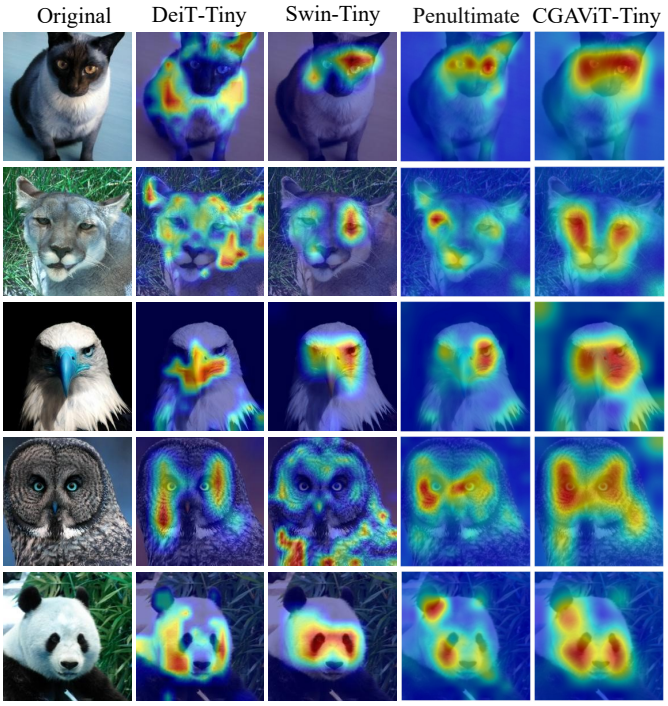

Figure 7. A qualitative comparison of attention maps. CGA-ViT exhibits superior spatial focus and semantic understanding.

5. Conclusions

In this work, we present CGA-ViT, a hierarchical vision transformer that achieves efficient local-global modeling through multi-scale dilated feature embedding and channel-guided additive attention. Our key insight is replacing quadratic pairwise attention with element-wise global aggregation, reducing complexity from $O(N^2)$ to $O(N)$ while preserving channel-wise contextual information. Although we achieve a superior balance between accuracy and efficiency, our MDFE module uses fixed dilation rates [1, 2, 3], which limits adaptability to tasks requiring different receptive field combinations. In the future, we will explore adaptive dilation strategies for NPU deployment and extend CGA-ViT to spatiotemporal modeling through temporal convolutions.

Author Contributions: Conceptualization, Y.Z. (Yayue Zhao) and Y.Z. (Yingxiao Zhao); methodology, Y.Z. (Yayue Zhao) and Z.L.; validation, Y.Z. (Yayue Zhao), B.L. and A.Z.; data curation, A.Z.; writing—original draft preparation, Y.Z. (Yayue Zhao), Y.Z. (Yingxiao Zhao), J.M. and Z.L.; writing—review and editing, Y.Z. (Yayue Zhao) and Y.Z. (Yingxiao Zhao); visualization, Z.L.; supervision, Y.Z. (Yayue Zhao), J.M. and Y.Z. (Yingxiao Zhao); project administration, Y.Z. (Yingxiao Zhao); funding acquisition, Y.Z. (Yingxiao Zhao). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used Kimi K2.5 and Doubao to check and correct grammar, spelling, and punctuation, and to assist in translating initial drafts from Chinese to English. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25.
3. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
4. Peng, Y.-X.; Jiao, J.; Feng, X.; Zheng, W.-S. Consistent Discrepancy Learning for Intra-Camera Supervised Person Re-Identification. *IEEE Trans. Multimed.* **2022**, *25*, 2393–2403. [CrossRef]
5. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
6. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; 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.
7. Zhao, B.; Xiong, H.; Bian, J.; Guo, Z.; Xu, C.Z.; Dou, D. COMO: Efficient Deep Neural Networks Expansion with Convolutional Maxout. *IEEE Trans. Multimed.* **2021**, *23*, 1722–1730. [CrossRef]
8. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
9. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
11. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
13. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
14. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
16. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.-H.; Khan, F.S. SwiftFormer: Efficient Additive Attention for Transformer-Based Real-Time Mobile Vision Applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 17425–17436.
17. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 12934–12949.
18. Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S.W.; Anwer, R.M.; Shahbaz Khan, F. EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 3–20.
19. Mehta, S.; Rastegari, M. Separable Self-Attention for Mobile Vision Transformers. *arXiv* **2022**, arXiv:2206.02680. [CrossRef]
20. 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, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
21. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
22. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
23. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
24. Wang, W.; Yao, L.; Chen, L.; Cai, D.; He, X.; Liu, W. Crossformer: A versatile vision transformer based on cross-scale attention. *arXiv* **2021**, arXiv:2108.00154.

25. Lee, Y.; Kim, J.; Willette, J.; Hwang, S.J. Mpvit: Multi-path vision transformer for dense prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7287–7296.
26. Ren, S.; Zhou, D.; He, S.; Feng, J.; Wang, X. Shunted self-attention via multi-scale token aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10853–10862.
27. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 367–376.
28. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-former: Bridging mobilenet and transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5270–5279.
29. Xu, Y.; Zhang, Q.; Zhang, J.; Tao, D. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28522–28535.
30. Zaheer, M.; Guruganesh, G.; Dubey, K.A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. Big Bird: Transformers for Longer Sequences. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 17283–17297.
31. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 5156–5165.
32. Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. Rethinking attention with performers. *arXiv* **2020**, arXiv:2009.14794.
33. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768. [CrossRef]
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
37. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood attention transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6185–6194.
38. Chu, X.; Zhang, B.; Tian, Z.; Wei, X.; Xia, H. Do we really need explicit position encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
39. 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 (ICML), Virtual, 18–24 July 2021; pp. 10347–10357.
40. Jiang, Z.; Hou, Q.; Yuan, L.; Zhou, D.; Jin, X.; Wang, A.; Feng, J. All Tokens Matter: Token Labeling for Training Better Vision Transformers. *arXiv* **2021**, arXiv:2104.10858. [CrossRef]
41. 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; pp. 1314–1324.
42. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
43. Pan, J.; Bulat, A.; Tan, F.; Zhu, X.; Dudziak, L.; Li, H.; Tzimiropoulos, G.; Martinez, B. EdgeViTs: Competing Light-Weight CNNs on Mobile Devices with Vision Transformers. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 294–311.
44. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
45. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. MetaFormer Is Actually What You Need for Vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10819–10829.
46. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
47. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene Parsing through ADE20K Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 633–641.

48. Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic Understanding of Scenes through the ADE20K Dataset. *Int. J. Comput. Vis.* **2019**, *127*, 302–321. [CrossRef]
49. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic Feature Pyramid Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 6399–6408.