CoAtFormer: Vision Transformer with Composite Attention

Zhiyong Chang\textsuperscript{1}, Mingjun Yin\textsuperscript{2}, Yan Wang\textsuperscript{3}

\textsuperscript{1}Peking University \hspace{1cm} \textsuperscript{2}The University of Melbourne \hspace{1cm} \textsuperscript{3}Zuoyebang

changzy@pku.org.cn, mingjuny1@student.unimelb.edu.au wangyan11@zuoyebang.com

Abstract

Transformer has recently gained significant attention and achieved state-of-the-art performance in various computer vision applications, including image classification, instance segmentation, and object detection. However, the self-attention mechanism underlying the transformer leads to quadratic computational cost with respect to image size, limiting its widespread adoption in state-of-the-art vision backbones. In this paper we introduce an efficient and effective attention module we call Composite Attention. It features parallel branches, enabling the modeling of various global dependencies. In each composite attention module, one branch employs a dynamic channel attention module to capture global channel dependencies, while the other branch utilizes an efficient spatial attention module to extract long-range spatial interactions. In addition, we effectively blending composite attention module with convolutions, and accordingly develop a simple hierarchical vision backbone, dubbed CoAtFormer, by simply repeating the basic building block over multiple stages. Extensive experiments show our CoAtFormer achieves state-of-the-art results on various different tasks. Without any pre-training and extra data, CoAtFormer-Tiny, CoAtFormer-Small, and CoAtFormer-Base achieve 84.4\%, 85.3\%, and 85.9\% top-1 accuracy on ImageNet-1K with 24M, 37M, and 73M parameters, respectively. Furthermore, CoAtFormer also consistently outperform prior work in other vision tasks such as object detection, instance segmentation, and semantic segmentation. When further pretraining on the larger dataset ImageNet-22k, we achieve 88.7\% Top-1 accuracy on ImageNet-1K.

1 Introduction

In the past years, Convolution Neural Networks (CNNs) have become a defacto choice for a wide variety of computer vision tasks [Simonyan and Zisserman, 2014; He et al., 2015; Ren et al., 2015; He et al., 2017a] since AlexNet [Krizhevsky et al., 2012]. However, convolution operations merely capture local dependencies of pixels, which neglect the dependency modeling between distant pixels to some extent [Wang et al., 2017]. Recently, self-attention based Transformers [Vaswani et al., 2017] have achieved excellent performance in Natural Language Processing (NLP) benchmarks and became the dominant architecture for various applications. Meanwhile, numerous researchers attempt to introduce the Transformer-based architectures into vision domains, and attain promising performance in various tasks such as image classification [Dosovitskiy et al., 2020; Touvron et al., 2020], object detection [?; Zhu et al., 2020], and semantic segmentation [Zheng et al., 2020]. Vision Transformer (ViT) [Dosovitskiy et al., 2020], which utilises a sequence of embedded image patches as input to stacked standard Transformer encoders, is the first fully Transformer-based architecture that...
demonstrate comparable performance to CNNs. While ViTs present strong capabilities to model the long-range dependencies, their computational complexity grows quadratically with the image size, limiting its application to high resolution images [Yang et al., 2021] in which modeling multi-scale global context is essential for accurate representation modeling.

To reduce the computational cost, some works [Liu et al., 2021; Dong et al., 2021; Tu et al., 2022; Ding et al., 2022], most notably Swin Transformer [Liu et al., 2021], have limited attention region in a spatial local window. However, the limited receptive field of local attention challenges the capability of global self-attention to model global contextual information. Subsequent works such as MaxViT [Tu et al., 2022] attempted to mitigate this issue by developing highly intricate self-attention modules with increased model size. How to efficiently integrate various long-range interactions to balance the model complexity and generalizability under a computation budget still remains challenging.

In this work, we develop a novel Transformer block, called composite attention block, that capably serves as a fundamental architecture component which can model global channel and spatial interactions in a single module. In contrast to vanilla self-attention, composite attention exhibits greater efficiency and flexibility, specifically adapting seamlessly to varying input lengths with linear complexity. Compared to window/local attention, composite attention performs stronger perception capacity by simultaneously capturing global spatial and channel dependencies. Moreover, with only linear complexity, composite attention can serve as a basic stand-alone attention block in any layer of a neural network, even in high-resolution stages.

It is worthwhile to note that our proposed composite attention contains two parallel branches to capture various global context. In composite attention block, one branch exploits a dynamic channels module to model global channel dependencies, while the other branch employs an efficient spatial attention module to capture long-range spatial interactions. Concretely, We replace the self-attention, as originally introduced by Vaswani et al. [Dosovitskiy et al., 2020], with an efficient spatial attention. Instead of capturing the pairwise interactions between keys, queries, we model the interactions between a learnable global token embedding and queries only, computing the global context-aware spatial attention weights with linear complexity. Additionally, we use a simple graph neural network as the core component of dynamic channel attention to encourage the communication across channels.

Based on the composite attention mechanism, we further propose a simple but effective vision Transformer architecture named “CoAtFormer” by hierarchically stacking repeated blocks composed of composite attention and convolutions. This architecture exhibits significantly stronger modeling power while limiting computation cost. As a general-purpose vision backbone, the CoAtFormer demonstrates state-of-the-art (SOTA) performance for a broad range of visual tasks including image classification, object detection and segmentation. Specifically, for image classification using ImageNet-1K dataset, CoAtFormer with 24M, 37M, 73M parameters achieve new SOTA performance of 84.4%, 85.3%, 85.9% Top-1 accuracy and without using any extra data. As Figure 1 shows, CoAtFormer consistently outperforms both GCViT [Hatamizadeh et al., 2023], MaxViT [Tu et al., 2022] and ConvNeXt [Liu et al., 2022] models by a significant margin. Furthermore, for object detection and instance segmentation using MS COCO dataset, our model achieves a box mAP of 54.9 for object and a mask mAP of 47.5, surpassing recent state-of-the-art MaxViT counterpart by +1.5 and +1.8 respectively.

The main contributions of our work are summarized as follows:

- We propose a strong vision backbone, CoAtFormer, that can capture various global interactions throughout every stage of the network.
- We develop a simple but effective composite attention block composed of dynamic channel and efficient spatial attention, enjoining global context in linear complexity.
- Large amounts of performance analysis shows that CoAtFormer outperforms previous SOTA transformers on ImageNet dataset, with a significantly lesser number of parameters. In addition, CoAtFormer also has superior performance on other downstream tasks.

2 Related work

CNNs. Since AlexNet [Krizhevsky et al., 2012], convolutional neural networks (CNNs) have shown remarkable success in various vision applications [Chen et al., 2017; Tan and Le, 2019; Howard et al., 2017; Sandler et al., 2018; Simonyan and Zisserman, 2014; He et al., 2015]. VGGNet [Simonyan and Zisserman, 2014] and InceptionNets [Szegedy et al., 2014; Szegedy et al., 2015] show that a deep neural network consisted of convolutional layers and pooling layers can attain adequate performance in image recognition. ResNet [He et al., 2015] show stronger generalization ability by introducing skip connections every two layers to the base architecture. ConvNeXt [Liu et al., 2022] has re-introduced core designs of vision transformers and demonstrate a pure CNN can achieve performance comparable to vision transformers on various vision tasks.

Vision transformers. Since Transformers [Vaswani et al., 2017; Devlin et al., 2019] achieve tremendous successes in wide natural language processing (NLP) tasks, many efforts [Dosovitskiy et al., 2020; Liu et al., 2021; Dong et al., 2021; Tu et al., 2022; Ding et al., 2022; Hatamizadeh et al., 2023; Yang et al., 2023] have been devoted to developing stronger Transformer based architectures for various vision applications. The pioneering work ViT [Dosovitskiy et al., 2020] directly applies the transformer encoder architecture to a sequence of image patches. However, ViT requires large datasets such as JFT300M [Sun et al., 2017] for training. DeiT [Touvron et al., 2020] utilizes a new training paradigm to enable training of high-performance ViT architecture with fewer data. PVT [Wang et al., 2021] leverages the pyramid structure to generate multi-scale feature maps for general pixel-level dense prediction tasks. MaxViT [Tu et al., 2022] introduces multi-axis attention to capture both local and global context. DaViT [Ding et al., 2022] presents the dual attention mechanism, which contains spatial self-attention and
channel self-attention. DualViT [Yao et al., 2023] incorporates a critical semantic pathway to obtain global semantics with reduced order of complexity. SMT [Lin et al., 2023] proposes Multi-Head Mixed Convolution (MHMC) module and Scale-Aware Aggregation (SAA) module to enhance the convolutional modulation.

**Channel-wise Attention.** Channel attention mechanisms have been widely adopted in CNNs [Hu et al., 2017; Woo et al., 2018; Qin et al., 2020], one of the most successful methods is SENet [Hu et al., 2017], which learns channel attention. It first squeezes the feature maps with global average pooling and captures the cross-channel relationships using two fully connected layers. FcaNet [Qin et al., 2020] learns many valuable frequency components by compressing channels using the discrete cosine transform (DCT). Some vision transformers architectures apply channel-wise attention to reduce the computational costs. XCiT [El-Noubry et al., 2021] proposes cross-covariance attention (XCA) to compute channel attention maps. DaViT [Ding et al., 2022] introduces channel group attention (CGA) to perform image-level interactions within each group.

Our work presents dynamic channel attention to capture global channel interactions in Transformers. We demonstrate its power when combined with efficient spatial attention, forming our composite attention mechanism.

### 3 Method

#### 3.1 Overall Architecture

We propose Composite Attention Vision Transformers (CoAtFormer), a general, efficient, yet effective Transformer backbone capturing both local and global dependencies. The overall architecture of CoAtFormer is illustrated in Figure 2. For an input image with size of $H \times W \times 3$, we leverage the Convolutional Token Embedding layer (two $3 \times 3$ convolution layer with a stride 2) to obtain $H' \times W' \times C_1$ feature maps. Following the design in modern CNNs (e.g., ResNet [He et al., 2015]), the whole network has four stages to generate feature maps of different scales which are important for dense prediction tasks. To produce the hierarchical representation, a downsampling layer ($3 \times 3$ convolution, stride 2) is used between two consecutive stages to reduce the number of tokens and increase the channel dimension. In each stage, several CoAtFormer blocks are stacked sequentially for feature transformation while maintaining the number of tokens. The CoAtFormer block is able to capture local-global representations.

#### 3.2 CoAtFormer Block

The proposed CoAtFormer block contains a conv encoder and a composite attention block, as illustrated in Figure 3. We will detail these parts in the following.

**Conv Encoder.** The baseline MaxViT [Tu et al., 2022] employs MBCov block [Sandler et al., 2018] as a local token mixer. Although MBCov block has been widely used in efficient models [Tu et al., 2022; Yang et al., 2022], replacing them with "modernize" convolution block [Liu et al., 2022] does not increase the computational cost. Further, it improves the performance and generalization without increasing the parameters. Using conv encoder before attention provides an additional benefit, as depth-wise convolutions can be considered as conditional position encoding (CPE) [Chu et al., 2021], eliminating the need for explicit positional encoding layers in our model. Specifically, a $3 \times 3$ depth-wise convolution (DWConv) is first applied to capture the local spatial interactions between pixels. Then, the derived features are fed into two point-wise convolutions with GELU activation. Finally, we introduce a residual connection [He et al., 2015] to enable information to flow across the network. Formally, given an input tensor $X \in \mathbb{R}^{C \times H \times W}$ ($H$, $W$ are its height, width, and channels), the conv encoder is represented as follows:

\[
\hat{X} = \text{BN}(\text{DWConv}_{3 \times 3}(X)), \quad (1) \\
\hat{X} = \text{Conv}_{1 \times 1}^\text{Conv}(\hat{X}), \quad (2) \\
\hat{X} = \text{GELU}(\hat{X}), \quad (3) \\
\hat{X} = \text{Conv}_{1 \times 1}^\text{Conv}(\hat{X}) + X, \quad (4)
\]

where BN, GELU, $\text{Conv}_{1 \times 1}^\text{Conv}$, and $\text{DWConv}_{3 \times 3}$ denote Batch Normalization [Ioffe and Szegedy, 2015], Gaussian error Linear Unit [Hendrycks and Gimpel, 2016], $1 \times 1$ point-wise convolution, and $3 \times 3$ depth-wise convolution, respectively.

**Composite Attention Block.** The detailed architecture of the composite attention block is shown in Figure 3. The composite attention block aims to learn enriched local-global features. It begins with local convolutional layers to extract local representations, followed by the composite attention mechanism. As the core element in a composite attention block, the composite attention contains two parallel branches. The two branches share the same input, but focus on relationships of different global context, which can be complementary to each other. In each branch, we first extract global channel or spatial dependencies using Dynamic Channel Attention (DCA) Module or Efficient Spatial Attention (ESA) Module, respectively, then the outputs of two branches can be merged using concatenation. We then describe our dynamic channel attention module and efficient spatial attention module.
Dynamic Channel Attention Module. In Figure 3, the top branch in composite attention aims to model cross-channel interactions. For an input feature tensor $X \in \mathbb{R}^{C \times H \times W}$, we first squeeze global spatial information into a channel descriptor. This is achieved by using T-Pool layer to reduce the spatial dimension of the input tensor to three by concatenating the average pooled, max pooled and std pooled features across that dimension. Mathematically, the channel responses $T \in \mathbb{R}^{C \times 3}$ are calculated by:

$$\mu_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{c,i,j}, \quad (5)$$

$$\sigma_c = \sqrt{\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} (x_{c,i,j} - \mu_c)^2}, \quad (6)$$

$$\nu_c = \max_{i,j} x_{c,i,j} \quad \text{and} \quad (7)$$

$$t_c = [\mu_c, \sigma_c, \nu_c]. \quad (8)$$

These channel features can be viewed as a set of unordered nodes which are denoted as $V = \{v_1, v_2, \ldots, v_C\}$. We can then build a graph $G = (V, E)$, where $E$ stands for all the edges. Each edge reflects the relation weight between two nodes. Based on this graph, we employ a double-layer graph convolution network [Kipf and Welling, 2016] to propagate information between nodes, the information diffusion process can be expressed as:

$$F = \text{h}(\text{A}h(\text{AT}W_1)W_2), \quad (9)$$

where $W_1 \in \mathbb{R}^{d_1 \times d_1}$ and $W_2 \in \mathbb{R}^{d_2 \times d_2}$ are state update matrix to be learned, $\text{h}$ denotes a non-linear activation, which is GELU [Hendrycks and Gimpel, 2016] in our work. $A \in \mathbb{R}^{C \times C}$ is a correlation matrix for propagating information, which contains the relation weight between nodes. In our experiments, A is randomly initialized and trained in an end-to-end manner along with the whole model. After graph-level processing, we aggregate node features $F \in \mathbb{R}^{C \times d_2}$ by taking average of node feature map on the feature dimension then exploit a sigmoid activation to generate channel-wise weights:

$$S = \text{sigmoid}\left(\frac{1}{d_2} \sum_{i=1}^{d_2} F_{c,i}\right). \quad (10)$$

Consequently, the output feature maps of DCA module can be formulated as:

$$Y_{\text{dca}} = X \ast S, \quad (11)$$

where $\ast$ denotes element-wise multiplication.

Compared to the recent SOTA channel group attention (CGA) [Ding et al., 2022], our proposed DCA enables a more comprehensive communication through a double-layer graph convolution network by treating each channel as a node in the graph. We will verify the effectiveness of DCA in subsequent sections.

Efficient Spatial Attention Module. In Figure 3, the bottom branch of composite attention focuses on global spatial context. For an input feature tensor $X \in \mathbb{R}^{C \times H \times W}$, we reshape it to get the input token embedding $X \in \mathbb{R}^{n \times d}$, where $n = H \times W$ is the number of patches, $d = C$ is the dimensions of the token embedding. The input token embedding $X$ is linearly transformed into query $Q$, key $K$, and value $V$ using three matrices $W_q$, $W_k$, and $W_v$, where $Q, K, V \in \mathbb{R}^{n \times d}$, $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$. First, we compute the dot products of the query with a learnable global token vector $g \in \mathbb{R}^d$, and then apply a softmax function to produce the global context scores $c_{s} \in \mathbb{R}^{n}$ as:
\[ \alpha = \text{softmax}(Qg) \frac{\sqrt{d}}{\sqrt{d}}. \]  
(12)

The global context scores are able to capture the importance of each query element. Next, the global context scores are multiplied by the key and pooled, resulting in a single global context vector as follows:

\[ c = \sum_{i=1}^{n} \alpha_i \cdot K_i. \]  
(13)

Then, the global context vector is multiplied element-wise with the value to form the global spatial information and fed to a linear layer with weights \( W_o \in \mathbb{R}^{d \times d} \) to obtain the output tensor \( Y_{esa} \), it can be described as:

\[ Y_{esa} = (V \ast c) \cdot W_o. \]  
(14)

Finally, the output tensor \( Y_{esa} \) is reshaped to the original input feature dimensions \((C \times H \times W)\). Our proposed ESA module is comparatively cheap to compute compared to vanilla MHSA [Dosovitskiy et al., 2020] and has linear complexity with the token length.

After obtaining the outputs from these two branches, we concatenate them along the channel dimension and then project the result back to the original dimension:

\[ Y_{merged} = \text{concat}(Y_{deq}, Y_{esa}) \cdot W_m, \]  
(15)

where, \( W_m \in \mathbb{R}^{2C \times C} \) is a weight matrix. We employ the FFN of vanilla ViT [Dosovitskiy et al., 2020], which consists of two linear layers and a GELU activation.

With the aforementioned these components, the CoAtFormer block can be formulated as:

\[ Y_1 = \text{Conv-Encoder}(X_{t-1}), \]  
(16)

\[ M_t = \text{DWConv}_{3 \times 3}(\text{Conv}_{1 \times 1}(Y_t)), \]  
(17)

\[ Z_t = \text{Composite-Attention} (\text{Conv}_{1 \times 1}(Y_t)), \]  
(18)

\[ X_t = \text{FFN}(\text{BN}(Z_t)) + Z_t, \]  
(19)

where \( Y_t, M_t, \) and \( Z_t \) denote the intermediate output features in the l-th block, respectively. \( \text{BN} \) denotes the batch normalization [Ioffe and Szegedy, 2015].

### 3.3 Architecture Variants

For a fair comparison with the other vision Transformer, we consider four different models with various numbers of parameters and computational complexity. Specifically, we introduce CoAtFormer-T (tiny), CoAtFormer-S (small), CoAtFormer-B (base) and CoAtFormer-L (large) variants, which is corresponded to CoAtFormer-T, CoAtFormer-S, CoAtFormer-B and CoAtFormer-L, respectively. In all these variants, the expansion ratio of each FFN is set as 3.

The detail configurations of basic embedding channels \( C \) and number of blocks \( N_i \) are presented as following:

- CoAtFormer-T: \( C = 64, N_i = \{1, 3, 12, 1\} \)
- CoAtFormer-S: \( C = 80, N_i = \{1, 3, 12, 1\} \)
- CoAtFormer-B: \( C = 96, N_i = \{1, 6, 18, 1\} \)
- CoAtFormer-L: \( C = 128, N_i = \{1, 10, 22, 1\} \)

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Forum</th>
<th>FLOPs</th>
<th>Top-1 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeiT-S [Touvron et al., 2020]</td>
<td>ICLR23</td>
<td>22.1M</td>
<td>4.3G</td>
<td>79.9</td>
</tr>
<tr>
<td>PVT-S [Wang et al., 2021]</td>
<td>ICCV23</td>
<td>21.8M</td>
<td>4.3G</td>
<td>81.9</td>
</tr>
<tr>
<td>Swin-T [Liu et al., 2021]</td>
<td>CVPR22</td>
<td>26.0M</td>
<td>4.3G</td>
<td>82.1</td>
</tr>
<tr>
<td>DaViT-T [Ding et al., 2022]</td>
<td>ICCV23</td>
<td>21.8M</td>
<td>4.3G</td>
<td>82.8</td>
</tr>
<tr>
<td>MOAT-0 [Yang et al., 2022a]</td>
<td>ICLR23</td>
<td>27.8M</td>
<td>5.7G</td>
<td>83.3</td>
</tr>
<tr>
<td>GPVLT-0 [Hatamizadeh et al., 2023]</td>
<td>ICLR23</td>
<td>23.8M</td>
<td>5.7G</td>
<td>83.3</td>
</tr>
<tr>
<td>Dual-ViT-S [Yao et al., 2020]</td>
<td>TPAMI23</td>
<td>24.9M</td>
<td>5.4G</td>
<td>83.3</td>
</tr>
<tr>
<td>GCViT-T [Hatamizadeh et al., 2023]</td>
<td>ICLR23</td>
<td>28.0M</td>
<td>5.7G</td>
<td>83.5</td>
</tr>
<tr>
<td>MaxViT-T [Tu et al., 2022]</td>
<td>ECCV22</td>
<td>31.0M</td>
<td>5.6G</td>
<td>83.6</td>
</tr>
<tr>
<td>SMT-S [Liu et al., 2023]</td>
<td>ICCV23</td>
<td>20.5M</td>
<td>4.7G</td>
<td>83.7</td>
</tr>
<tr>
<td>CoAtFormer-T (Ours)</td>
<td>ICCV23</td>
<td>24.3M</td>
<td>4.5G</td>
<td>84.4</td>
</tr>
<tr>
<td>Swin-X [Liu et al., 2021]</td>
<td>ICCV23</td>
<td>70.0M</td>
<td>15.4G</td>
<td>83.0</td>
</tr>
<tr>
<td>ConvNeXt-B [Liu et al., 2022]</td>
<td>CVPR22</td>
<td>50.0M</td>
<td>8.7G</td>
<td>83.1</td>
</tr>
<tr>
<td>Swin-B [Ding et al., 2021]</td>
<td>ICCV23</td>
<td>78.0M</td>
<td>15.4G</td>
<td>83.1</td>
</tr>
<tr>
<td>DaViT-B [Ding et al., 2022]</td>
<td>ICML23</td>
<td>51.0M</td>
<td>8.5G</td>
<td>84.3</td>
</tr>
<tr>
<td>CrossFormer++-L [Wang et al., 2022]</td>
<td>TPAMI23</td>
<td>52.0M</td>
<td>9.1G</td>
<td>84.2</td>
</tr>
<tr>
<td>GPVLT-L [Hatamizadeh et al., 2023]</td>
<td>ICLR23</td>
<td>36.2M</td>
<td>12.5G</td>
<td>84.1</td>
</tr>
<tr>
<td>GCViT-S [Hatamizadeh et al., 2023]</td>
<td>ICLR23</td>
<td>51.0M</td>
<td>8.5G</td>
<td>84.3</td>
</tr>
<tr>
<td>MaxViT-S [Tu et al., 2022]</td>
<td>ECCV22</td>
<td>69.0M</td>
<td>11.7G</td>
<td>84.4</td>
</tr>
<tr>
<td>Dilate-S [Ioffe et al., 2023]</td>
<td>TPAMI23</td>
<td>49.0M</td>
<td>10.0G</td>
<td>84.4</td>
</tr>
<tr>
<td>SMT-B [Lin et al., 2023]</td>
<td>ICCV23</td>
<td>32.0M</td>
<td>7.7G</td>
<td>84.3</td>
</tr>
</tbody>
</table>

Table 1: Comparison of image classification on ImageNet-1K for different models. All models are trained and evaluated with 224×224 resolution on ImageNet-1K by default, unless otherwise noted. † denotes the model is evaluated with resolution of 384×384.

### 4 Experiments

To validate the efficacy of CoAtFormer as a general vision backbone, we conduct experiments on ImageNet-1K [Deng et al., 2009] classification, COCO [Lin et al., 2014] object detection and instance segmentation, and ADE20K [Zhou et al., 2017] semantic segmentation. We also perform comprehensive ablation studies to evaluate the effectiveness of each component of CoAtFormer.

#### 4.1 Image Classification

**Setup.** For fair comparison, we follow the same training strategies as previous works [Touvron et al., 2020; Liu et al., 2021]. Specifically, we train all our models for 300 epochs with the input size of 224×224. We employ the AdamW optimizer with weight decay of 0.05. The default batch size and initial learning rate are set to 1024 and 0.001. Additionally, we explore the effectiveness of our models when pretrained on ImageNet-22K.

**Results.** In Table 1, we compare our proposed CoAtFormer with current state-of-the-art models. It shows that
Table 2 reports the results of our models on MS COCO dataset. Using a Mask-RCNN detector, our CoAtFormer-T (49.1/44.8) backbone outperforms counterparts with ConvNeXt-T (46.2/41.7) by +2.9 and +3.1 and DaViT-T (47.4/42.9) by +1.7 and +1.9 in terms of box AP and mask AP, respectively. Using a Cascade Mask-RCNN detector, we also observe substantial gains across all model configurations. Furthermore, we observe a almost saturated mAP in Swin Transformer [Liu et al., 2021] and GCViT [Hatemizadeh et al., 2023] from small to base model, while the mAP of our model consistently improves with larger model size, demonstrating enhanced scalability.

### 4.3 Semantic Segmentation on ADE20k

**Setup.** We further benchmark our method for semantic segmentation using the ADE20K dataset. We use UperNet [Xiao et al., 2018] as the segmentation method and our CoAtFormer as the backbone. We follow the same training recipe proposed

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Param (M)</th>
<th>FLOPs (G)</th>
<th>AP\textsuperscript{bbox}</th>
<th>AP\textsuperscript{mask}</th>
<th>AP\textsuperscript{bbox} \textsuperscript{50}</th>
<th>AP\textsuperscript{mask} \textsuperscript{50}</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVT-S [Wang et al., 2021]</td>
<td>44</td>
<td>245</td>
<td>43.0</td>
<td>65.5</td>
<td>39.9</td>
<td>62.5</td>
</tr>
<tr>
<td>Swin-T [Liu et al., 2021]</td>
<td>48</td>
<td>267</td>
<td>46.0</td>
<td>68.1</td>
<td>41.6</td>
<td>65.1</td>
</tr>
<tr>
<td>ConvNeXt-T [Liu et al., 2022]</td>
<td>48</td>
<td>262</td>
<td>46.2</td>
<td>67.9</td>
<td>41.7</td>
<td>65.0</td>
</tr>
<tr>
<td>DaViT-T [Ding et al., 2022]</td>
<td>48</td>
<td>263</td>
<td>47.4</td>
<td>69.5</td>
<td>42.9</td>
<td>66.8</td>
</tr>
<tr>
<td>GCViT-T [Hatemizadeh et al., 2023]</td>
<td>48</td>
<td>291</td>
<td>47.9</td>
<td>70.1</td>
<td>43.2</td>
<td>67.0</td>
</tr>
<tr>
<td>CoAtFormer-T</td>
<td>44</td>
<td>263</td>
<td>49.1</td>
<td>70.6</td>
<td>44.8</td>
<td>67.6</td>
</tr>
<tr>
<td>Cascade Mask-RCNN 3× schedule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeiT-S/16 [Touvron et al., 2020]</td>
<td>80</td>
<td>889</td>
<td>48.0</td>
<td>67.2</td>
<td>41.4</td>
<td>64.2</td>
</tr>
<tr>
<td>RccNet-50 [He et al., 2015]</td>
<td>82</td>
<td>739</td>
<td>46.3</td>
<td>64.3</td>
<td>40.1</td>
<td>61.7</td>
</tr>
<tr>
<td>Swin-T [Liu et al., 2021]</td>
<td>86</td>
<td>745</td>
<td>50.4</td>
<td>69.2</td>
<td>43.7</td>
<td>66.6</td>
</tr>
<tr>
<td>ConvNeXt-T [Liu et al., 2022]</td>
<td>86</td>
<td>741</td>
<td>50.4</td>
<td>69.1</td>
<td>43.7</td>
<td>66.5</td>
</tr>
<tr>
<td>GCViT-T [Hatemizadeh et al., 2023]</td>
<td>85</td>
<td>770</td>
<td>51.6</td>
<td>70.4</td>
<td>44.6</td>
<td>67.8</td>
</tr>
<tr>
<td>CoATFormer-T</td>
<td>81</td>
<td>742</td>
<td>52.8</td>
<td>71.4</td>
<td>46.4</td>
<td>69.2</td>
</tr>
<tr>
<td>Cascade Mask-RCNN 3× schedule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swin-S [Liu et al., 2021]</td>
<td>107</td>
<td>838</td>
<td>51.9</td>
<td>70.7</td>
<td>45.0</td>
<td>68.2</td>
</tr>
<tr>
<td>ConvNeXt-S [Liu et al., 2022]</td>
<td>108</td>
<td>827</td>
<td>51.9</td>
<td>70.8</td>
<td>45.0</td>
<td>68.4</td>
</tr>
<tr>
<td>GCViT-S [Hatemizadeh et al., 2023]</td>
<td>108</td>
<td>866</td>
<td>52.4</td>
<td>71.0</td>
<td>45.4</td>
<td>68.5</td>
</tr>
<tr>
<td>CoATFormer-S</td>
<td>86</td>
<td>825</td>
<td>53.5</td>
<td>72.5</td>
<td>47.1</td>
<td>70.7</td>
</tr>
<tr>
<td>Swin-B [Liu et al., 2021]</td>
<td>145</td>
<td>982</td>
<td>51.9</td>
<td>70.5</td>
<td>45.0</td>
<td>68.1</td>
</tr>
<tr>
<td>ConvNeXt-B [Liu et al., 2022]</td>
<td>146</td>
<td>964</td>
<td>52.7</td>
<td>71.3</td>
<td>45.6</td>
<td>68.9</td>
</tr>
<tr>
<td>GCViT-B [Hatemizadeh et al., 2023]</td>
<td>146</td>
<td>1018</td>
<td>52.9</td>
<td>71.7</td>
<td>45.8</td>
<td>69.2</td>
</tr>
<tr>
<td>MaxViT-B [Tu et al., 2022]</td>
<td>157</td>
<td>856</td>
<td>53.4</td>
<td>72.9</td>
<td>45.7</td>
<td>70.3</td>
</tr>
<tr>
<td>CoATFormer-B</td>
<td>138</td>
<td>960</td>
<td>54.9</td>
<td>74.1</td>
<td>47.3</td>
<td>71.2</td>
</tr>
</tbody>
</table>
Table 5 demonstrates the significance of composite attention block and conv encoder in our proposed architecture. Substituting composite attention block with convolution encoder leads to a decrease in accuracy by 2.9%, highlighting the effectiveness of composite attention block in our design. Moreover, removing conv encoder in all four stages of the network results in an additional accuracy reduction of 1.0%. In addition, we also measure the contributions of composite attention components (e.g., dynamic channel attention (DCA) and efficient spatial attention (ESA)) in Table 5. Removing DCA module and ESA module decrease the accuracy by 1.3 and 1.5, respectively.

Different fusion method. We try two general operations as the alternatives to the fusion method in composite attention. One is element-wise addition, the other is weighted sum. The results are shown in Table 6. It is clear that concatenation and an project layer as the fusion module achieves the best model size and accuracy trade-off among the different fusion methods we evaluated.

## 5 Interpretability

In Figure 4, we can find that learned spatial attention distributions align the region of image semantics, and hence demonstrate the effectiveness of efficient spatial attention. Additionally, corresponding Grad-CAM maps present accurate object localization with most intricate details.

## 6 Conclusion

In this work, we propose a novel architecture named CoAtFormer, which can efficiently capture global channel and spatial contexts by utilizing dynamic channel attention and efficient spatial attention. The proposed CoAtFormer architectures take advantages of both CNNs and transformers to capture local and global information, improving the representation ability of the model. Extensive experiments on ImageNet and other downstream vision applications demonstrate the effectiveness and superiority of the proposed architecture.
References


