FASTERVIT: FAST VISION TRANSFORMERS WITH HIERARCHICAL ATTENTION

Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov
NVIDIA
{ahatamizadeh, pmolchanov}@nvidia.com

ABSTRACT

We design a new family of hybrid CNN-ViT neural networks, named FasterViT, with a focus on high image throughput for computer vision (CV) applications. FasterViT combines the benefits of fast local representation learning in CNNs and global modeling properties in ViT. Our newly introduced Hierarchical Attention (HAT) approach decomposes global self-attention with quadratic complexity into a multi-level attention with reduced computational costs. We benefit from efficient window-based self-attention. Each window has access to dedicated carrier tokens that participate in local and global representation learning. At a high level, global self-attentions enable the efficient cross-window communication at lower costs. FasterViT achieves a SOTA Pareto-front in terms of accuracy and image throughput. We have extensively validated its effectiveness on various CV tasks including classification, object detection and segmentation. We also show that HAT can be used as a plug-and-play module for existing networks and enhance them. We further demonstrate significantly faster and more accurate performance than competitive counterparts for images with high resolution.

Code is available at https://github.com/NVlabs/FasterViT.

1 INTRODUCTION

Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have recently become popular in computer vision and achieved superior performance in various applications such as image classification (Liu et al., 2021; Dong et al., 2022; Lin et al., 2017), object detection (Zhang et al., 2021b; Fang et al., 2021) and semantic segmentation (Xie et al., 2021; Cheng et al., 2021). In addition to learning more uniform local and global representations across their architecture when compared to Convolutional Neural Networks (CNNs), ViTs scale properly to large-scale data and model sizes (Raghu et al., 2021; Paul & Chen, 2022). Recently, several efforts (He et al., 2022; Xie et al., 2022) have also shown the exceptional capability of ViTs in self-supervised learning of surrogate tasks such as masked image modeling which may significantly enhance the performance of downstream applications. Despite these advantages, lack of inductive bias in pure ViT models may require more training data and impede performance (Xu et al., 2021b). Hybrid architectures, which consist of both CNN and ViT-based components, could address this problem and achieve competitive performance without needing large-scale training datasets (Dosovitskiy et al., 2020) or other techniques such as knowledge distillation (Touvron

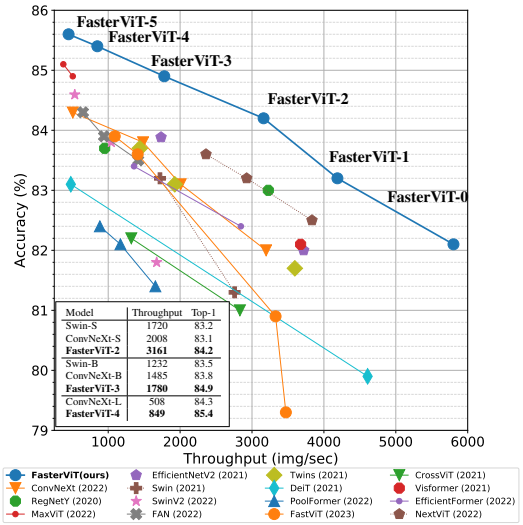
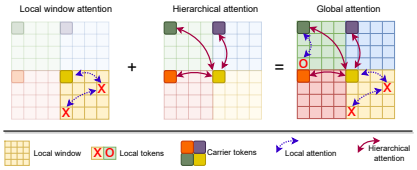


Figure 1: Comparison of image throughput and ImageNet-1K Top-1 accuracy. Throughput is measured on A100 GPU with batch size of 128.

et al., 2021a). An integral component of ViTs is the self-attention mechanism (Vaswani et al., 2017; Dosovitskiy et al., 2020) which enables modeling of both short and long-range spatial dependencies. However, the quadratic computational complexity of self-attention significantly impacts the efficiency and hinders its use for applications with high-resolution images. In addition, contrary to the isotropic architecture (*i.e.*, same feature resolution with no downsampling) of the original ViT model, learning feature representations in a multi-scale manner typically yields better performance (Fan et al., 2021; Wang et al., 2022), specifically for downstream applications (*e.g.*, detection, segmentation).

To address these issues, Swin Transformer (Liu et al., 2021) proposed a multi-scale architecture in which self-attention is computed in local windows, and window-shifting allows for interaction of different regions. However, due to the limited receptive field of these local regions and small area of coverage in window shifting (Liu et al., 2021; Lin et al., 2017), capturing cross-window interactions and modeling the long-range spatial dependencies become challenging for large-resolution input features. Furthermore, using self-attention blocks in early stages with larger resolution may impact the image throughput due to the increased number of local windows. Recently, the Swin Transformer V2 model (Liu et al., 2022a) was proposed to address training instabilities on high-resolution images by improving the self-attention mechanism. However, in addition to having a lower image throughput compared to the Swin Transformer (Liu et al., 2021), Swin Transformer V2 still relies on the original window-shifting mechanism for cross-interaction of different windows, which becomes less effective with large image sizes.

Figure 2: Visualization of the proposed Hierarchical Attention in the feature space. By performing local window attention and hierarchical attention we can achieve global information propagation at reduced costs.



In this work, we attempt to address these issues and propose a novel hybrid architecture, denoted as FasterViT, which is tailored for high-resolution input images, while maintaining a fast image throughput. FasterViT consists of four different stages in which the input image resolution is reduced by using a strided convolutional layer, while doubling the number of feature maps. We propose to leverage residual convolutional blocks in the high-resolution stages of the architecture (*i.e.*, stage 1, 2), while employing transformer-blocks in later stages (*i.e.*, stage 3, 4). This strategy allows for fast generation of high-level tokens which can be further processed with the transformer-based blocks. For each transformer block, we use an interleaved pattern of local and, newly proposed, Hierarchical Attention blocks to capture both short and long-range spatial dependencies and efficiently model the cross-window interactions. Specifically, our proposed Hierarchical Attention (see Fig. 2) learns carrier tokens as a summary of each local window and efficiently models the cross-interaction between these regions. The computational complexity of the Hierarchical Attention grows almost linearly with input image resolution, as the number of regions increases, due to the local windowed attention being the compute bottleneck. Hence, it is an efficient, yet effective way of capturing long-range information with large input features.

We have extensively validated the effectiveness of the proposed FasterViT model on various image tasks and datasets such as ImageNet-1k for image classification, MS COCO for object detection and instance segmentation and ADE20K dataset for semantic segmentation. FasterViT achieves state-of-the-art performance considering the trade-off between performance (*e.g.*, ImageNet-1K top-1 accuracy) and image throughput (see Fig. 1). To demonstrate the scalability of FasterViT for larger datasets, we have also pre-trained FasterViT on ImageNet-21K dataset and achieved state-of-the-art performance when fine-tuning and evaluating on larger-scale resolutions.

The summary of our contributions is as follows:

- We introduce FasterViT, which is a novel hybrid vision transformer architecture designed for an optimal trade-off between performance and image throughput. FasterViT scales effectively to higher resolution input images for different dataset and model sizes.
- We propose the Hierarchical Attention module which efficiently captures the cross-window interactions of local regions and models the long-range spatial dependencies.
- FasterViT achieves a new SOTA Pareto front in terms of image throughput and accuracy trade-off and is significantly faster than comparable ViT-based architectures yielding signifi-

cant speed-up compared to recent SOTA models. It also achieves competitive performance on downstream tasks of detection and instance segmentation on MS COCO dataset and semantic segmentation on ADE20K dataset.

2 Related Work

Vision Transformers. Oriented from the language processing domain, the first application of transformer architecture to vision task immediately offers an inspiring demonstration of the high efficacy of attention across image patches across varying scenarios (Dosovitskiy et al., 2020). The appealing strength of vision transformer and its architecture and logic simplicity has therefore triggered a quickly evolving literature in the past two years, where ViT performance is quickly boosted by an erupting new set of innovations: network-wise leveraging knowledge distillation for data-efficient training as in DeiT (Touvron et al., 2021a), hybriding convolution and self-attention for enhanced inductive biases as in LeViT (Graham et al., 2021), imposing CNN-inspired pyramid rules on ViTs (Wang et al., 2021; 2022), along with component-wise improvements such as improved token utilization as in T2T-ViT (Yuan et al., 2021), enhanced positional embedding (Chu et al., 2023), local window attention as shown in the inspiring work of the Swin family (Liu et al., 2021; 2022a) and CSwin (Dong et al., 2022), global attention in GCViT (Hatamizadeh et al., 2023), among many other architectural insights (Chu et al., 2021a; Zhang et al., 2021a; Yuan et al., 2022). Along with the increasing capacity comes the increasing computation burden. As similarly facing challenges in scaling up the models in language tasks (e.g., from BERT-Large 0.3B (Devlin et al., 2019), to Megatron-LM 8.3B (Shoeybi et al., 2019), and Switch-Transformer1.6T (Fedus et al., 2022)), scaling up vision transformers is also a highly challenging but highly important task (Dai et al., 2021; Liu et al., 2022a) due to the attention-extensive nature of transformers, urging efficiency for pervasive usage.

Towards Enhanced Efficiency. Boosting up ViT efficiency has therefore been a very vibrant area. One stream of approach roots in the efficient deep learning literature that cuts down on network complexity leveraging popular methods such as efficient attention (Bolya et al., 2022; Lu et al., 2021; Cai et al., 2022), network compression (Chen et al., 2021b;c; Liang et al., 2022; Yang et al., 2021a), dynamic inference (Yin et al., 2022; Rao et al., 2021), operator adaptation (Molchanov et al., 2022), token merging and manipulations (Marin et al., 2021; Xu et al., 2022), etc. These methods can yield off-the-shelf speedups on target ViT backbones, but are also limited to the original backbone's accuracy and capacity. Another stream of work, on the other hand, focuses on designing new ViT architectures with enhanced efficiency as an original design objective. For example, EfficientFormer (Li et al., 2022) entails mobile applications through dimension-consistent re-design of transformer block and removing redundant architectural components. VisFormer (Chen et al., 2021d) transits computation extensive transformer to a convolutional counterpart for enhanced vision efficiency. CrossViT (Chen et al., 2021a) learns multi-scale features and utilizes small/large-patch backed tokens that are channeled by efficient attention, offering linear time and memory complexity. Even with such a rapid progress in literature, enabling efficient ViTs remains a significant challenge, where we next further push the Pareto front of faster ViT on top of prior art by a large margin. Note that we focus on the second stream of architectural redesign for efficiency boost, and consider a joint exploration with the first acceleration stream of method like compression as orthogonal and fruitful future work.

Global Self-Attention. A number of efforts have introduced global self-attention to capture more contextual information. In NLP (*i.e., 1D*), BigBird (Zaheer et al., 2020) and LongFormer (Beltagy et al., 2020) proposed to select special tokens (*i.e. non-learnable*) as global tokens to attend to other tokens via a sliding-window dense self-attention. In computer vision, EdgeViT (Pan et al., 2022), Twins (Chu et al., 2021a) and Focal Transformer (Yang et al., 2021b) proposed hierarchical-like attention mechanisms which rely on heuristic token aggregation in the forms of pooling (Yang et al., 2021b) or linear projection (Pan et al., 2022; Chu et al., 2021a). There are three key differences between these efforts and our proposed hierarchical attention: (1) as opposed to using a pre-defined mechanism to select the global tokens (*e.g., random*), we propose to learn these tokens (*i.e., carrier token*) via summarizing the role of each region in the input feature space (2) we propose learnable token aggregation and propagation mechanisms by computing self-attention among carrier tokens (3) as opposed to using dense/dilated self-attention, our proposed HAT uses local window-based self-attention and has a smaller computational complexity.

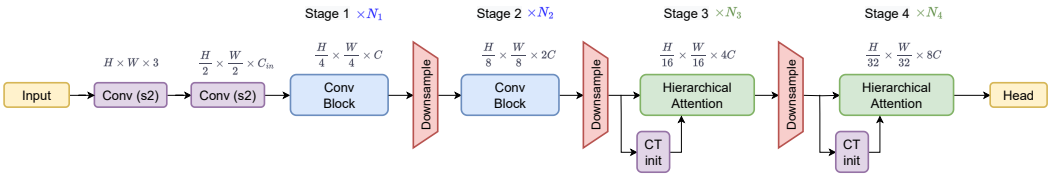


Figure 3: Overview of the FasterViT architecture. We use a multi-scale architecture with CNN and transformer-based blocks in stages 1, 2 and 3, 4, respectively. Best viewed in color.

3 FasterViT

3.1 Design Principals

We next detail our FasterViT architecture, offering Pareto accuracy-latency trade-off. We focus on highest throughput for computer vision tasks on mainstream off-the-shelf hardware such as GPUs that excel in parallel computing. Computation in this case involves a set of streaming multiprocessors (SMs) with CUDA and Tensor cores as computation units. It requires frequent data transfer for calculation and can be impacted by data movement bandwidth. As such, operations bounded by computation are math-limited, while those bounded by memory transfer are memory-limited. It requires a careful balance between the two to maximize throughput.

In hierarchical vision models, spatial dimension of intermediate representation shrinks as inference proceeds. Initial network layers mostly have larger spatial dimensions and fewer channel (*e.g.*, $112 \times 112 \times 64$), making them memory-bound. This makes a better fit for compute-intensive operations, such as dense convolution instead of depth-wise/sparse counterparts that impose extra transfer cost. Also operations not representable in matrix manipulation forms, *e.g.*, non-linearity, pooling, batch normalization, are also memory-bound and shall be minimized for usage. On the contrary, later layers tend to be math-limited with computationally expensive operations. For example, hierarchical CNNs have feature maps of size 14×14 with high dimensional kernels. This leaves room for more expressive operations such as Layer Normalization, squeeze-and-excitation, or attention, with fairly small effect on throughput. Guided by these insights we design a novel architecture that will benefit all stages from accelerated computing hardware.

3.2 Architecture

Our overall design is shown in Fig. 3. It exploits convolutional layers in the earlier stages that operate on higher resolution. The second half of the model relies on novel hierarchical attention layers to reason spatially across the entire feature maps. In this design, we optimize the architecture for compute and throughput. As a result, the first half of the network and downsampling blocks make use of dense convolutional kernels. We also avoid squeeze-and-excitation operators and minimize Layer Normalization for higher resolution stages (*i.e.*, 1, 2), as these layers tend to be math-limited. Later stages (*i.e.*, 3, 4) in the architecture tend to be math-limited as GPU hardware spends more time on compute compared to the memory transfer cost. As a result, applying multi-head attention will not be a bottleneck.

3.3 FasterViT Components

Stem An input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ is converted into overlapping patches by two consecutive 3×3 convolutional layers, each with a stride of 2, which project them into a D-dimensional embedding. The embedded tokens are further batch-normalized (Ioffe & Szegedy, 2015) and use the ReLU activation function after each convolution.

Downsampler Blocks FasterViT follows the hierarchical structure: the spatial resolution is reduced by 2 between stages by a downsampling block. We apply 2D layer normalization on spatial features, followed by a convolutional layer with a kernel of 3×3 and a stride of two.

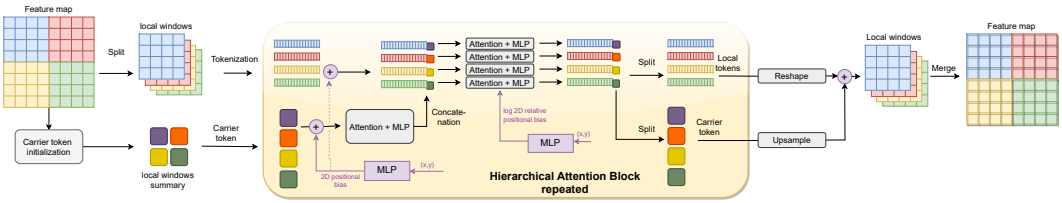


Figure 4: Proposed Hierarchical Attention block. Carrier tokens (CT) learn a summary of each local window and facilitate global information exchange between local windows. Local window tokens only have access to a dedicated subset of CT for efficient attention. CT undergo full self-attention to enable cross-window attention. "Attention" stands for MHSA (Vaswani et al., 2017), MLP for multi-layer perceptron. Best viewed in color.

Conv Blocks Stage 1 and 2 consist of residual convolutional blocks, which are defined as

$$\begin{aligned} \hat{\mathbf{x}} &= \text{GELU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{x}))), \\ \mathbf{x} &= \text{BN}(\text{Conv}_{3\times3}(\hat{\mathbf{x}})) + \mathbf{x}, \end{aligned} \tag{1}$$

where BN denotes batch normalization (Ioffe & Szegedy, 2015). Following the design principles, these convolutions are dense.

Hierarchical Attention In this work, we propose a novel formulation of windowed attention, summarized in Fig 2 and detailed presentation in Fig 4. We start with local windows introduced in Swin Transformer (Liu et al., 2021). Then, we introduce a notion of *carrier tokens* (CTs) that play the summarizing role of the entire local window. The first attention block is applied on CTs to summarize and pass global information. Then, local window tokens and CTs are concatenated, such that every local window has access only to its own set of CTs. By performing self attention on concatenated tokens we facilitate local and global information exchange at reduced cost. By alternating sub-global (CTs) and local (windowed) self-attention we formulate a concept of *hierarchical attention*. Conceptually, CTs can be further grouped into windows and have a higher order of carrier tokens.

Assume we are given an input feature map $\mathbf{x} \in \mathbb{R}^{H\times W\times d}$ in which H, W and d denote the height, width and number of feature maps, let us set $H = W$ for simplicity. We first partition the input feature map into $n \times n$ local windows with $n = \frac{H^2}{k^2}$, where k is the window size, as:

$$\hat{\mathbf{x}}_{\mathbf{l}} = \text{Split}_{k\times k}(\mathbf{x}). \tag{2}$$

The key idea of our approach is the formulation of *carrier tokens* (CTs) that help to have an attention footprint much larger than a local window at low cost. At first, we initialize CTs by pooling to $L = 2^c$ tokens per window:

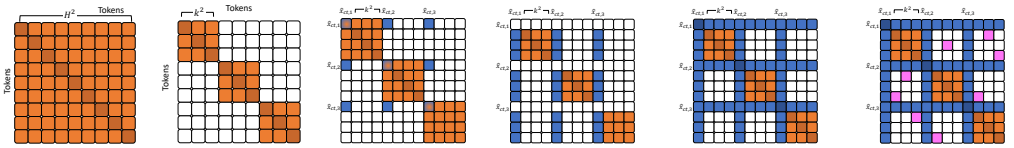
$$\begin{aligned} \hat{\mathbf{x}}_{\mathbf{c}} &= \text{Conv}_{3\times3}(\mathbf{x}), \\ \hat{\mathbf{x}}_{\mathbf{ct}} &= \text{AvgPool}_{H^2\rightarrow n^2L}(\hat{\mathbf{x}}_{\mathbf{c}}), \end{aligned} \tag{3}$$

where $\text{Conv}_{3\times3}$ represents efficient positional encoding inspired by (Chu et al., 2021b) and used in Twins (Chu et al., 2021a). $\hat{\mathbf{x}}_{\mathbf{ct}}$ and AvgPool denote the carrier tokens and feature pooling operation, respectively; c is set to 1, but can be changed to control latency. The current approach with conv+pooling gives flexibility with the image size. These pooled tokens represent a summary of their respective local windows, we set $L << k$. The procedure of CT initialization is performed only once for every resolution stage. Note that every local window $\hat{\mathbf{x}}_{\mathbf{l}}$ has unique set of carrier tokens, $\hat{\mathbf{x}}_{\mathbf{ct,l}}$, such that $\hat{\mathbf{x}}_{\mathbf{ct}} = \{\hat{\mathbf{x}}_{\mathbf{ct,l}}\}_{\mathbf{l}=0}^{n}$.

In every HAT block, CTs undergo the attention procedure:

$$\begin{aligned} \hat{\mathbf{x}}_{\mathbf{ct}} &= \hat{\mathbf{x}}_{\mathbf{ct}} + \gamma_1 \cdot \text{MHSA}(\text{LN}(\hat{\mathbf{x}}_{\mathbf{ct}})), \\ \hat{\mathbf{x}}_{\mathbf{ct}} &= \hat{\mathbf{x}}_{\mathbf{ct}} + \gamma_2 \cdot \text{MLP}_{d\rightarrow4d\rightarrow d}(\text{LN}(\hat{\mathbf{x}}_{\mathbf{ct}})), \end{aligned} \tag{4}$$

where LN represents layer normalization (Ba et al., 2016), MHSA represents multi-head self attention (Vaswani et al., 2017), γ is a learnable per-channel scale multiplier (Touvron et al., 2021b), $\text{MLP}_{d\rightarrow4d\rightarrow d}$ is a 2-layer MLP with GeLU (Hendrycks & Gimpel, 2016) activation function.



(a) Full attention (b) Windowed attention (c) **Hierarchical Attention (ours)** (d) Twins Chu et al. (2021a) (e) LongFormer Beltagy et al. (2020) (f) BigBird Zaheer et al. (2020)

Figure 5: Attention map comparison for a feature map of size $H \times H \times d$. □ - no attention, ■ - normal token attention, ■ - carrier token attention, ■ - random token attention. Full attention (a) has complexity of $O(H^4 d)$, windowed attention significantly reduces it to $O(k^2 H^2 d)$ but lacks global context.

Next, in order to model short-long-range spatial information, we compute the interaction between the local and carrier tokens, $\hat{\mathbf{x}}_{\mathbf{l}}$ and $\hat{\mathbf{x}}_{\mathbf{ct,l}}$, respectively. At first, local features and CTs are concatenated. Each local window only has access to its corresponding CTs:

$$\hat{\mathbf{x}}_{\mathbf{w}} = \text{Concat}(\hat{\mathbf{x}}_{\mathbf{l}}, \hat{\mathbf{x}}_{\mathbf{ct,l}}). \tag{5}$$

These tokens undergo another set of attention procedure:

$$\begin{aligned} \hat{\mathbf{x}}_{\mathbf{w}} &= \hat{\mathbf{x}}_{\mathbf{w}} + \gamma_1 \cdot \text{MHSA}(\text{LN}(\hat{\mathbf{x}}_{\mathbf{w}})), \\ \hat{\mathbf{x}}_{\mathbf{w}} &= \hat{\mathbf{x}}_{\mathbf{w}} + \gamma_2 \cdot \text{MLP}_{d \to 4d \to d}(\text{LN}(\hat{\mathbf{x}}_{\mathbf{w}})). \end{aligned} \tag{6}$$

Finally, tokens are further split back and used in the subsequent hierarchical attention layers:

$$\hat{\mathbf{x}}_{\mathbf{l}}, \hat{\mathbf{x}}_{\mathbf{ct,l}} = \text{Split}(\hat{\mathbf{x}}_{\mathbf{w}}), \tag{7}$$

Procedures described in Equations 4-7 are iteratively applied for a number of layers in the stage. To further facilitate long-shot-range interaction, we perform *global information propagation*, similar to the one in (Pan et al., 2022) in the end of the stage. Finally, the output of the stage is computed as:

$$\mathbf{x} = \text{Upsample}_{n^2 L \to H^2}(\hat{\mathbf{x}}_{\mathbf{ct,l}}) + \text{Merge}_{n^2 k^2 \to H^2}(\hat{\mathbf{x}}_{\mathbf{l}}) \tag{8}$$

MHSAs performed in Eq. 4 and 6 are token position invariant, however, the location of features in the spatial dimension are clearly informative. To address this, we first add absolute positional bias directly to CTs and local window tokens. We are inspired by SwinV2 (Liu et al., 2022a) and employ a 2-layer MLP to embed absolute 2D token location into feature dimension. Then, to facilitate image-like locality inductive bias we enhance the attention with log space relative positional bias from SwinV2 (Liu et al., 2022a) (2-layer MLP). It ensures that the relative position of tokens contribute to shared attention patterns. This approach yields flexibility regarding image size, as the positional encoding is interpolated by the MLP, and hence a trained model can be applied to any input resolution.

An attention map comparison between efficient global-local self attention is shown in Fig. 5. The proposed hierarchical attention splits full attention into local and sub-global, both compressible to 2 dense attentions. Carrier tokens participate in both attentions and facilitate information exchange.

Complexity Analysis of HAT The key features of the efficiency of our approach are (i) separation of attentions and (ii) local windows only have access to their CTs. The complexity of the most conventional and popular full attention is $O(H^4 d)$. Partitioning the feature size into windows of size k, and running the attention, simplifies the attention to $O(k^2 H^2 d)$ as proposed in (Liu et al., 2021). It is well known that such windowed attention is more efficient but lacks global feature interaction. Our approach takes this one step further and is based on carrier tokens that summarize and interact over the entire feature map, to remedy for missing global communication. Given L total carrier tokens per window, local window complexity is $O((k^2 + L)H^2 d)$. Local (windowed) attention is followed by attention on carrier tokens with complexity $O((\frac{H^2}{k^2} L)^2 d)$. The total cost of both attentions is $O(k^2 H^2 d + LH^2 d + \frac{H^4}{k^4} L^2 d)$.

An orthogonal approach for multilevel attention is to provide access to subsampled global information inside local attention. For example, Twins (Chu et al., 2021a) subsamples global feature map and

Table 1: Comparison of classification benchmarks on **ImageNet-1K** dataset (Deng et al., 2009). Image throughput is measured on A100 GPUs with batch size of 128.

Model	Image Size (Px)	#Param (M)	FLOPs (G)	Throughput (Img/Sec)	Top-1 (%)
Conv-Based					
ConvNeXt-T Liu et al. (2022b)	224	28.6	4.5	3196	82.0
ConvNeXt-S Liu et al. (2022b)	224	50.2	8.7	2008	83.1
ConvNeXt-B Liu et al. (2022b)	224	88.6	15.4	1485	83.8
RegNetY-040 Radosavovic et al. (2020)	288	20.6	6.6	3227	83.0
ResNetV2-101 Wightman et al. (2021)	224	44.5	7.8	4019	82.0
EfficientNetV2-S Tan & Le (2021)	384	21.5	8.0	1735	83.9
Transformer-Based					
Swin-T Liu et al. (2021)	224	28.3	4.4	2758	81.3
Swin-S Liu et al. (2021)	224	49.6	8.5	1720	83.2
SwinV2-T Liu et al. (2022a)	256	28.3	4.4	1674	81.8
SwinV2-S Liu et al. (2022a)	256	49.7	8.5	1043	83.8
SwinV2-B Liu et al. (2022a)	256	87.9	15.1	535	84.6
Twins-B Chu et al. (2021a)	224	56.1	8.3	1926	83.1
DeiT3-L	224	304.4	59.7	535	84.8
PoolFormer-M58 Yu et al. (2022)	224	73.5	11.6	884	82.4
Hybrid					
CoaT-Lite-S Xu et al. (2021a)	224	19.8	4.1	2269	82.3
CrossViT-B Chen et al. (2021a)	240	105.0	20.1	1321	82.2
Visformer-S Chen et al. (2021d)	224	40.2	4.8	3676	82.1
EdgeViT-S Pan et al. (2022)	224	13.1	1.9	4254	81.0
EfficientFormer-L7 Li et al. (2022)	224	82.2	10.2	1359	83.4
MaxViT-B Tu et al. (2022)	224	120.0	23.4	507	84.9
MaxViT-L Tu et al. (2022)	224	212.0	43.9	376	85.1
FasterViT					
FasterViT-0	224	31.4	3.3	**5802**	**82.1**
FasterViT-1	224	53.4	5.3	**4188**	**83.2**
FasterViT-2	224	75.9	8.7	**3161**	**84.2**
FasterViT-3	224	159.5	18.2	**1780**	**84.9**
FasterViT-4	224	424.6	36.6	**849**	**85.4**
FasterViT-5	224	957.5	113.0	**449**	**85.6**
FasterViT-6	224	1360.0	142.0	**352**	**85.8**

uses it as key and value for local window attention. It has a complexity of $O(k^2H^2d + \frac{H^4}{k^2}d)$ (from the paper). Under the same size of the local window (k), and H, we can get the difference of $O(L + \frac{H^2L^2}{k^4})$ for HAT and $O(\frac{H^2}{k^2})$ for Twins. HAT gets more efficient with higher resolution, for example, for $H = 32$, $k = 8$, with $L = 4$ we get $O(8)$ for HAT, whereas $O(16)$ for Twins.

4 Results

4.1 Image Classification

In Table 1, we demonstrate a quantitative comparison between the performance of FasterViT models and a variety of different hybrid, conv and Transformer-based networks on ImageNet-1K dataset. Comparing to Conv-based architectures, we achieve higher accuracy under the same throughput, for example, we outperform ConvNeXt-T by 2.2%. Considering the accuracy and throughput trade-off, FasterViT models are significantly faster than Transformer-based models such as the family of Swin

Table 2: Object detection and instance segmentation benchmarks using Cascade Mask R-CNN (He et al., 2017) on **MS COCO** dataset (Lin et al., 2014). All models employ 3× schedule. All model statistics are reported using a input test resolution of 1280 × 800.

Backbone	Throu. im/sec	AP^{box} Box	50	75	AP^{mask} Mask	50	75
Swin-T Liu et al. (2021)	161	50.4	69.2	54.7	43.7	66.6	47.3
ConvNeXt-T Liu et al. (2022b)	166	50.4	69.1	54.8	43.7	66.5	47.3
DeiT-Small/16 Touvron et al. (2021a)	269	48.0	67.2	51.7	41.4	64.2	44.3
FasterViT-2	**287**	**52.1**	**71.0**	**56.6**	**45.2**	**68.4**	**49.0**
Swin-S Liu et al. (2021)	119	51.9	70.7	56.3	45.0	68.2	48.8
X101-32 Xie et al. (2017)	124	48.1	66.5	52.4	41.6	63.9	45.2
ConvNeXt-S Liu et al. (2022b)	128	51.9	70.8	56.5	45.0	68.4	49.1
FasterViT-3	**159**	**52.4**	**71.1**	**56.7**	**45.4**	**68.7**	**49.3**
X101-64 Xie et al. (2017)	86	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B Liu et al. (2021)	90	51.9	70.5	56.4	45.0	68.1	48.9
ConvNeXt-B Liu et al. (2022b)	101	52.7	71.3	57.2	45.6	68.9	49.5
FasterViT-4	**117**	**52.9**	**71.6**	**57.7**	**45.8**	**69.1**	**49.8**

Transformers (Liu et al., 2021; 2022a). Furthermore, compared to hybrid models, such as the recent EfficientFormer (Li et al., 2022) and MaxViT (Tu et al., 2022) models, FasterViT on average has a higher throughput while achieving a better ImageNet top-1 performance. To validate the scalability of the proposed model, we pre-trained FasterViT-4 on ImageNet-21K dataset and fine-tuned it on various image resolutions on ImageNet-1K dataset. As shown in Table 3, FasterViT-4 has a better accuracy-throughput trade-off compared to other counterparts.

Model	Image Size (Px)	#Param (M)	FLOPs (G)	Throughput (Img/Sec)	Top-1 (%)
ViT-L/16‡ Liu et al. (2021)	384	307.0	190.7	149	85.2
Swin-L‡ Liu et al. (2021)	224	197.0	34.5	787	86.3
Swin-L‡ Liu et al. (2021)	384	197.0	103.9	206	87.3
ConvNeXt-L‡ Liu et al. (2022b)	224	198.0	34.4	508	86.6
ConvNeXt-L‡ Liu et al. (2022b)	384	198.0	101.0	172	87.5
FasterViT-4‡	224	424.6	36.6	**849**	**86.6**
FasterViT-4‡	384	424.6	119.2	**281**	**87.5**

Table 3: **ImageNet-21K** pretrained classification benchmarks on **ImageNet-1K** dataset (Deng et al., 2009). Image throughput is measured on A100 GPUs with batch size of 128. ‡ denotes models that are pre-trained on ImageNet-21K dataset.

4.2 Dense Prediction Tasks

In Table 2, we present object detection and instance segmentation benchmarks on MS COCO dataset (Lin et al., 2014) with Cascade Mask R-CNN (He et al., 2017) network. We observe that FasterViT models have better accuracy-throughput trade-off when compared to other counterparts. Specifically, FasterViT-4 outperforms ConvNeXt-B and Swin-B by +0.2 and +1.0 in terms of box AP and +0.3 and +1.0 in terms of mask AP, while being 15% and 30% faster in terms of throughput, respectively. We also conduct additional object detection experiments with FasterViT-4 ImageNet-21K pre-trained backbone and the state-of-the-art DINO (Zhang et al., 2022) model and achieve a high detection accuracy of 58.7 box AP. In Table 4, we present the semantic segmentation benchmarks with UPerNet (Xiao et al., 2018) network for experiments conducted on ADE20K dataset (Zhou et al., 2017). Similar to previous tasks, FasterViT models benefit from a better performance-throughput trade-off.

Model	Throughput	FLOPs (G)	IoU(ss/ms)
Swin-T Liu et al. (2021)	350	945	44.5/45.8
ConvNeXt-T Liu et al. (2022b)	363	939	- /46.7
FasterViT-2	**377**	974	**47.2/48.4**
Twins-SVT-B Chu et al. (2021a)	204	-	47.7/48.9
Swin-S Liu et al. (2021)	219	1038	47.6/49.5
ConvNeXt-S Liu et al. (2022b)	234	1027	- /49.6
FasterViT-3	**254**	1076	**48.7/49.7**
Twins-SVT-L Chu et al. (2021a)	164	-	48.8/50.2
Swin-B Liu et al. (2021)	172	1188	48.1/49.7
ConvNeXt-B Liu et al. (2022b)	189	1170	- /49.9
FasterViT-4	**202**	1290	**49.1/50.3**

Table 4: Semantic segmentation on **ADE20K** (Zhou et al., 2017) with UPerNet (Xiao et al., 2018).

5 ABLATION

EdgeViT and Twins As shown in Table 5, we performed a comprehensive ablation study to validate the effectiveness of HAT by replacing all attention layers with attention mechanisms in EdgeViT (Pan et al., 2022) and Twins (Chu et al., 2021a) in the 3rd and 4th stages. For all model variants, FasterViT models with HAT achieve a better accuracy, sometimes by a significant margin. Twins achieves a higher throughput due to its small kernel size (*i.e.* $k = 2$), however, this significantly limits its accuracy. The better performance of HAT is attributed to its learnable information aggregation/propagation via CTs, and direct access to dedicated CTs in windowed attention.

Model	Attention	FLOPs (G)	Thr(Img/Sec)	Top-1 (%)
FasterViT-0	Twins (Chu et al., 2021a)	3.0	6896	80.8
FasterViT-0	EdgeViT (Pan et al., 2022)	3.2	5928	81.0
FasterViT-0	**HAT**	3.3	5802	**82.1**
FasterViT-1	Twins (Chu et al., 2021a)	4.7	4949	82.1
FasterViT-1	EdgeViT (Pan et al., 2022)	4.8	4188	82.5
FasterViT-1	**HAT**	5.3	4344	**83.2**
FasterViT-2	Twins (Chu et al., 2021a)	8.0	3668	82.9
FasterViT-2	EdgeViT (Pan et al., 2022)	8.5	3127	83.4
FasterViT-2	**HAT**	8.7	3161	**84.2**

Table 5: Ablation study on the effectiveness of HAT compared to EdgeViT (Pan et al., 2022) and Twins (Chu et al., 2021a) self-attention mechanisms. All attention blocks are replaced with the indicated attention type.

Carrier Token Size We investigated the effect of carrier token size and window size on the accuracy and image throughput of the model. We observed that increasing the carrier token size can improve the performance at the cost of decreased throughput, sometimes by a significant margin. In addition, increasing the window size slightly improves the Top-1 accuracy while also decreasing the throughput. In fact, increasing the window size does not scale properly to higher resolution images due to its significant impact on efficiency. As a result, HAT is a more effective and efficient mechanism that can be employed to model long-range spatial dependencies without sacrificing the throughput. Please refer to supplementary materials for more details.

	Pretrain		Finetune					
	W8, I256		W12, I384		W16, I512		W24, I768	
Model	acc	im/s	acc	im/s	acc	im/s	acc	im/s
SwinV2-T Liu et al. (2022a)	81.8	1674	83.2	573	83.8	168	84.2	72
SwinV2-S Liu et al. (2022a)	83.7	633	84.8	338	85.4	153	-	-
FasterViT-2	**84.3**	**2500**	**85.3**	**984**	**85.5**	**489**	**85.6**	**155**
SwinV2-B Liu et al. (2022a)	84.2	499	85.1	251	85.6	115	-	-
FasterViT-4 256	**85.3**	**653**	**86.0**	**254**	**86.1**	**133**	**86.0**	**44**

Table 6: Quantitative comparison between higher resolution fine-tuning of FasterViT and SwinV2. FasterViT is more accurate on average by 0.9%, and faster by 2x.

Plug-and-Play HAT We employed HAT as a plug-and-play module with Swin-T model Table 7. This change results in +0.9 and +0.4% improvement in terms of mIoU and Top-1 accuracy on ImageNet classification and ADE20K segmentation tasks. In addition, improvements on MS COCO by +0.5 box AP and +0.6 mask AP on object detection and instance segmentation tasks, respectively. In addition, we also provide throughput comparisons and show that HAT can be efficiently used with existing architectures with minimal overhead. Hence, it validates the effectiveness of HAT as a standalone self-attention.

	ImageNet		COCO			ADE20k	
	top-1	Thr	AP^{box}	AP^{mask}	Thr	mIoU	Thr
Swin-T	81.3	2758	50.4	43.7	161	44.5	350
Swin-T + HAT	**81.7**	2721	**50.9**	**44.3**	150	**45.4**	338

Table 7: Ablation study on the effectiveness of HAT as a plug-and-play module with Swin-T model for various CV tasks.Thr stands for throughput and is measure in image/sec.

6 CONCLUSION

In this work, we have presented a novel hybrid model, denoted as FasterViT, which achieves SOTA Pareto-front in terms of ImageNet Top-1 accuracy and throughput. We have extensively validated the effectiveness of FasterViT in downstream dense prediction tasks such as object detection, instance segmentation and semantic segmentation. Our benchmarks demonstrate better accuracy-throughput trade-off in comparison to counterpart models such as ConvNeXt and Swin Transformer.

7 ACKNOWLEDGEMENT

We thank Amanda Moran, Christopher Lamb, Sivakumar Thottakara, Sudeep Sabnis, Ranjitha Prasanna and other members of NVIDIA NGC team for providing highly-optimized GPU cloud infrastructures which were used for training and evaluation of FasterViT models.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 3, 6

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, and Judy Hoffman. Hydra attention: Efficient attention with many heads. In *European Conference on Computer Vision*, pp. 35–49. Springer, 2022. 3

Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. *arXiv preprint arXiv:2205.14756*, 2022. 3

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021a. 3, 7

Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. Autoformer: Searching transformers for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12270–12280, 2021b. 3

Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34:19974–19988, 2021c. 3

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021d. 3, 7

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021a. 3, 5, 6, 7, 8, 9

Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers? *CoRR*, abs/2102.10882, 2021b. 5

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *ICLR 2023*, 2023. URL https://openreview.net/forum?id=3KWnuT-R1bh. 3

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 7, 8

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019. 3

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134, 2022. 1, 3

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2, 3

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021. 2

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021. 1

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 3

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12259–12269, 2021. 3

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pp. 12633–12646. PMLR, 2023. 3

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017. 8

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. 1

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015. 4, 5

Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022. 3, 7, 8

Youwei Liang, Chongjian GE, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. EVit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn_. 3

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 8

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017. 1, 2

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. 1, 2, 3, 5, 6, 7, 8

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022a. 2, 3, 6, 7, 8, 9

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022b. 7, 8

Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: softmax-free transformer with linear complexity. *Advances in Neural Information Processing Systems*, 34:21297–21309, 2021. 3

Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860*, 2021. 3

Pavlo Molchanov, Jimmy Hall, Hongxu Yin, Jan Kautz, Nicolo Fusi, and Arash Vahdat. Lana: latency aware network acceleration. In *European Conference on Computer Vision*, pp. 137–156. Springer, 2022. 3

Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *ECCV*, 2022. 3, 6, 7, 9

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022. 1

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020. 7

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021. 1

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. 3

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. 3

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pp. 10096–10106. PMLR, 2021. 7

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021a. 1, 3, 8

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 32–42, 2021b. 5

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 459–479. Springer, 2022. 7, 8

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017. 2, 5

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021. 3

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2, 3

Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 7

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018. 8

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017. 8

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022. 1

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9981–9990, 2021a. 7

Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2964–2972, 2022. 3

Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34: 28522–28535, 2021b. 1

Huanrui Yang, Hongxu Yin, Pavlo Molchanov, Hai Li, and Jan Kautz. Nvit: Vision transformer compression and parameter redistribution. *arXiv preprint arXiv:2110.04869*, 2021a. 3

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021b. 3

Hongxu Yin, Arash Vahdat, Jose Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022. 7

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token ViT: Training vision transformers from scratch on imagenet. In *ICCV*, 2021. 3

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020. 3, 6

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 8

Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2998–3008, 2021a. 3

Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2799–2808, 2021b. 1

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017. 8