

Rethinking Vision Transformers for MobileNet Size and Speed

Yanyu Li
 Snap Inc., Northeastern University
 li.yanyu@northeastern.edu

Ju Hu
 Snap Inc.
 jhu3@snap.com

Yang Wen
 Snap Inc.
 yangwenca@gmail.com

Georgios Evangelidis
 Snap Inc.
 gevangelidis@snap.com

Kamyar Salahi
 UC Berkeley
 kam.salahi@berkeley.edu

Yanzhi Wang
 Northeastern University
 yanz.wang@northeastern.edu

Sergey Tulyakov
 Snap Inc.
 stulyakov@snap.com

Jian Ren
 Snap Inc.
 jren@snap.com

Abstract

With the success of Vision Transformers (ViTs) in computer vision tasks, recent arts try to optimize the performance and complexity of ViTs to enable efficient deployment on mobile devices. Multiple approaches are proposed to accelerate attention mechanism, improve inefficient designs, or incorporate mobile-friendly lightweight convolutions to form hybrid architectures. However, ViT and its variants still have higher latency or considerably more parameters than lightweight CNNs, even true for the years-old MobileNet. In practice, latency and size are both crucial for efficient deployment on resource-constraint hardware. In this work, we investigate a central question, can transformer models run as fast as MobileNet and maintain a similar size? We revisit the design choices of ViTs and propose a novel supernet with low latency and high parameter efficiency. We further introduce a novel fine-grained joint search strategy for transformer models that can find efficient architectures by optimizing latency and number of parameters simultaneously. The proposed models, EfficientFormerV2, achieve 3.5% higher top-1 accuracy than MobileNetV2 on ImageNet-1K with similar latency and parameters. This work demonstrate that properly designed and optimized vision transformers can achieve high performance even with MobileNet-level size and speed¹.

1. Introduction

The promising performance of Vision Transformers (ViTs) [20] has inspired many follow-up works to further

¹Code: <https://github.com/snap-research/EfficientFormer>.

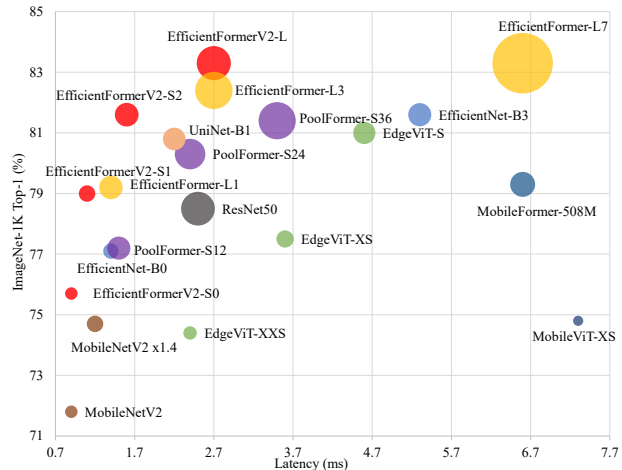


Figure 1. Comparison of model size, speed, and performance (top-1 accuracy on ImageNet-1K). Latency is profiled by iPhone 12 (iOS 16). The area of each circle is proportional to the number of parameters (model size). EfficientFormerV2 achieves high performance with small model sizes and fast inference speed.

refine the model architecture and improve training strategies, leading to superior results on most computer vision benchmarks, such as classification [49, 51, 8, 54], segmentation [80, 14, 6], detection [7, 44, 66], and image synthesis [21, 26]. As the essence of ViT, Multi Head Self Attention (MHSA) mechanism is proved to be effective in modeling spatial dependencies in 2D images, enabling a global receptive field. In addition, MHSA learns second-order information with the attention heatmap as dynamic weights, which is a missing property in Convolution Neural Networks (CNNs) [27]. However, the cost of MSHA is also obvious—quadratic computation complexity with respect to

the number of tokens (resolution). Consequently, ViTs tend to be more computation intensive and have higher latency compared to widely adopted lightweight CNNs [33, 32], especially on resource-constrained mobile devices, limiting their wide deployment in real-world applications.

Many research efforts [56, 57, 58, 45] are taken to alleviate this limitation. Among them, one direction is to reduce the quadratic computation complexity of the attention mechanism. Swin [50] and following works [19, 49] propose window-based attention such that the receptive field is constrained to a pre-defined window size, which also inspires subsequent work to refine attention patterns [11, 75, 78, 59]. With the pre-defined span of attention, the computation complexity becomes linear to resolution. However, sophisticated attention patterns are generally difficult to support or accelerate on mobile devices because of intensive shape and index operations. Another track is to combine lightweight CNN and attention mechanism to form a hybrid architecture [56, 13, 55]. The benefit comes two-fold. First, convolutions are shift invariant and are good at capturing local and detailed information, which can be considered as a good complement to ViTs [27]. Second, by placing convolutions in the early stages while placing MHSA in the last several stages to model global dependency, we can naturally avoid performing MHSA on high resolution and save computations [48]. Albeit achieving satisfactory performance, the latency and model size are still less competitive compared to lightweight CNNs. For instance, MobileViT [56] achieves better performance than MobileNetV2 while being at least $5\times$ slower on iPhone 12. As applicable to CNNs, architecture search, pruning, and quantization techniques are also thoroughly investigated [35, 36, 52, 37, 9, 45, 48]. Nevertheless, these models still emerge obvious weaknesses, *e.g.*, EfficientFormer-L1 [45] achieves comparable speed and better performance than MobileNetV2 $\times 1.4$, while being $2\times$ larger. Thus, a *practical yet challenging* question arises:

Can we design a transformer-based model that is both light and fast, and preserves high performance?

In this work, we address the above question and propose a new family of mobile vision backbones. We consider three vital factors: *number of parameters*, *latency*, and *model performance*, as they reflect disk storage and mobile applications. First, we introduce *novel architectural improvements* to form a strong design paradigm. Second, we propose a *fine-grained* architecture search algorithm that jointly optimizes model size and speed for transformer models. With our network design and search method, we obtain a series of models under various constraints of model size and speed while maintaining high performance, named EfficientFormerV2. Under the exact same size and latency (on iPhone 12), EfficientFormerV2-S0 outperforms MobileNetV2 by 3.5% higher top-1 accuracy on ImageNet-1K [18]. Compared to EfficientFormer-L1 [45], EfficientFormerV2-S1 has

similar performance while being $2\times$ smaller and $1.3\times$ faster (Tab. 2). We further demonstrate promising results in downstream tasks such as detection and segmentation (Tab. 3). Our contributions can be concluded as follows.

- We comprehensively study mobile-friendly design choices and introduce novel changes, which is a practical guide to obtaining ultra-efficient vision transformer backbones.
- We propose a novel fine-grained joint search algorithm that simultaneously optimizes model size and speed for transformer models, achieving superior Pareto optimality.
- For the first time, we show that vision transformer models can be as small and fast as MobileNetV2 while obtaining much better performance. EfficientFormerV2 can serve as a strong backbone in various downstream tasks.

2. Related Work

Vaswani *et al.* [73] propose attention mechanism to model sequences in NLP task, which forms transformer architecture. Transformers are later adopted to vision tasks by Dosovitskiy *et al.* [20] and Carion *et al.* [7]. DeiT [70] improves ViT by training with distillation and achieves competitive performance against CNNs. Later research further improves ViTs by incorporating hierarchical design [74, 71], injecting locality with the aid of convolutions [25, 17, 27, 66], or exploring different types of token mixing such as local attention [50, 19], spatial MLP mixer [69, 68], and non-parameterized pool mixer [83]. With appropriate changes, ViTs demonstrate strong performance in downstream vision tasks [80, 87, 86, 41, 40, 21, 84]. To benefit from the advantageous performance, efficient deployment of ViTs has become a research hotspot, especially for mobile devices [56, 13, 58, 55]. For reducing the computation complexity of ViTs, many works propose new modules and architecture design [39, 28, 22, 43, 63], while others eliminate redundancies in attention mechanism [75, 31, 10, 46, 15, 62, 72, 5]. Similar to CNNs, architecture search [12, 23, 9, 89, 48, 16, 77], pruning [85], and quantization [52] are also explored for ViTs.

We conclude two major drawbacks of the study in efficient ViT. First, many optimizations are not suitable for mobile deployment. For example, the quadratic computation complexity of the attention mechanism can be reduced to linear by regularizing the span or pattern of attention mechanism [50, 19, 11, 79]. Still, the sophisticated reshaping and indexing operations are not even supported on resource-constrained devices [45]. It is crucial to rethink the mobile-friendly designs. Second, though recent hybrid designs and network search methods reveal efficient ViTs with strong performance [56, 48, 45], they mainly optimize the Pareto curve for one metric while being less competitive in others. For example, MobileViT [56] is parameter efficient while being times slower than lightweight CNNs [64, 67]. Effi-

cientFormer [45] yields ultra-fast speed on mobile, but the model size is enormous. LeViT [24] and MobileFormer [13] achieve favorable FLOPs at the cost of redundant parameters.

3. Rethinking Hybrid Transformer Network

In this section, we study the design choices for efficient ViTs and introduce the changes that lead to the smaller size and faster speed without a performance drop. EfficientFormer-L1 [45] is used as a baseline model given its superior performance on mobile devices.

3.1. Token Mixers vs. Feed Forward Network

Incorporating local information can improve performance and make ViTs more robust to the absence of explicit positional embedding [6]. PoolFormer [83] and EfficientFormer [45] employ 3×3 average pooling layers (Fig. 2(a)) as local token mixer. Replacing these layers with depth-wise convolutions (DWCONV) of the same kernel size does not introduce latency overhead, while the performance is improved by 0.6% with negligible extra parameters (0.02M). Further, recent work [23, 6] suggest that it is also beneficial to inject local information modeling layers in the Feed Forward Network (FFN) in ViTs to boost performance with minor overhead. It is noteworthy that by placing extra depth wise 3×3 convolutions in FFNs to capture local information, the functionality of original local mixer (pooling or convolution) is duplicated. Based on these observations, we remove the explicit residual-connected local token mixer and move the dept-wise 3×3 CONV into the FFN, to get a unified FFN (Fig. 2(b)) with locality enabled. We apply the unified FFN to all stages of the network, as in Fig. 2(a,b). Such design modification simplifies the network architecture to only two types of blocks (local FFN and global attention), and boosts the accuracy to 80.3% at the same latency (see Tab. 1) with minor overhead in parameters (0.1M). More importantly, this modification allows us to *directly search the network depth* with the exact number of modules in order to extract local and global information, especially at the late stages of the network, as discussed in Sec. 4.2.

3.2. Search Space Refinement

With the unified FFN and the deletion of residual-connected token mixer, we examine whether the search space from EfficientFormer is still sufficient, especially in terms of depth. We vary the network depth (number of blocks in each stage) and width (number of channels), and find that deeper and narrower network leads to better accuracy (0.2% improvement), less parameters (0.13M reduction), and lower latency (0.1ms acceleration), as in Tab. 1. Therefore, we set this network as a new baseline (accuracy 80.5%) to validate subsequent design modifications, and enable a deeper supernet for architecture search in Sec. 4.2.

In addition, 5-stage models with further down-sized spatial resolution ($\frac{1}{64}$) have been widely employed in efficient ViT arts [24, 13, 48]. To justify whether we should search from a 5-stage supernet, we append an extra stage to current baseline network and verify the performance gain and overhead. It is noteworthy that though computation overhead is not a concern given the small feature resolution, the additional stage is parameter intensive. As a result, we need to shrink the network dimension (depth or width) to align parameters and latency to the baseline model for fair comparison. As seen in Tab. 1, the best performance of the 5-stage model surprisingly drops to 80.31% with more parameters (0.39M) and latency overhead (0.2ms), despite the saving in MACs (0.12G). This aligns with our intuition that the fifth stage is computation efficient but parameter intensive. Given that 5-stage network can not introduce more potentials in our size and speed scope, we stick to 4-stage design. This analysis also explains why some ViTs offer an excellent Pareto curve in MACs-Accuracy, but tend to be quite redundant in size [24, 13]. As the most important takeaway, optimizing single metric is easily trapped, and the proposed joint search in Sec. 4.2 provides a feasible solution to this issue.

3.3. MHSA Improvements

We then study the techniques to improve the performance of attention modules without raising extra overhead in model size and latency. As shown in Fig. 2(c), we investigate two approaches for MHSA. First, we inject local information into the Value matrix (V) by adding a depth-wise 3×3 CONV, which is also employed by [23, 66]. Second, we enable communications between attention heads by adding fully connected layers across head dimensions [65] that are shown as Talking Head in Fig. 2(c). With these modifications, we further boost the performance to 80.8% with similar parameters and latency compared to the baseline model.

3.4. Attention on Higher Resolution

Attention mechanism is beneficial to performance. However, applying it to high-resolution features harms mobile efficiency since it has quadratic time complexity corresponding to spatial resolution. We investigate strategies to efficiently apply MHSA to higher resolution (early stages). Recall that in the current baseline network obtained in Sec. 3.3, MHSA is only employed in the last stage with $\frac{1}{32}$ spatial resolution of the input images. We apply extra MHSA to the second last stage with $\frac{1}{16}$ feature size, and observe 0.9% gain in accuracy. On the down side, the inference speed slows down by almost $2.7\times$. Thus, it is necessary to properly reduce complexity of the attention modules.

Although some work propose window-based attention [50, 19], or downsampled Keys and Values [42] to alleviate this problem, we find that they are not best-suited options for mobile deployment. Window-based attention is diffi-

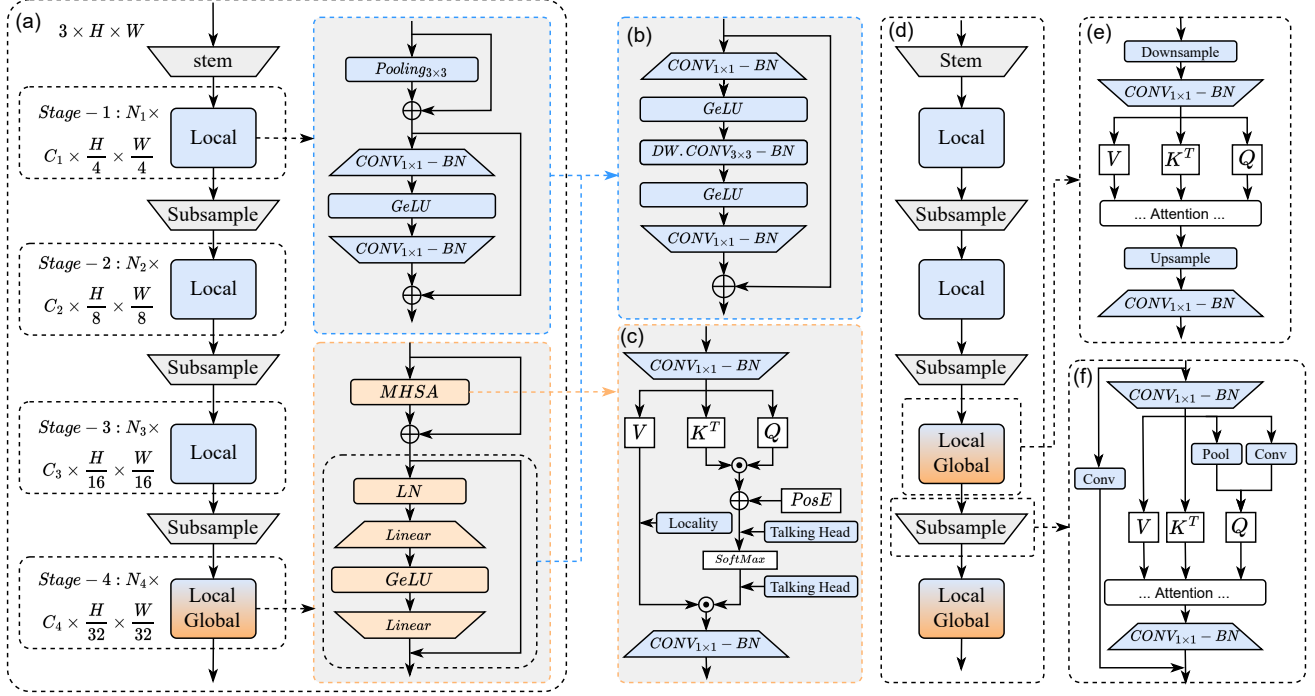


Figure 2. **Network architectures.** We consider three metrics, *i.e.*, model performance, size, and inference speed, and study the models that improve any metric without hurting others. (a) Network of EfficientFormer [45] that serves as a baseline model. (b) Unified FFN (Sec. 3.1). (c) MHSA improvements (Sec. 3.3). (d)&(e) Attention on higher resolution (Sec. 3.4). (f) Dual-Path Attention downsampling (Sec. 3.5).

Table 1. **Number of parameters, latency, and performance** for various design choices. The latency is tested on iPhone 12. Top-1 accuracy is obtained by validating models on ImageNet-1K for the classification task.

Section	Method	#Params (M)	MACs (G)	Latency (ms)	Top-1 (%)
(Baseline)	EfficientFormer-L1	12.25	1.30	1.4	79.2
Sec. 3.1	Pool Mixer $\rightarrow$ DWCONV $_{3 \times 3}$	12.27	1.30	1.4	79.8
	✓ Feed Forward Network	12.37	1.33	1.4	80.3
Sec. 3.2	✓ Vary Depth and Width	12.24	1.20	1.3	80.5
	5-Stage Network	12.63	1.08	1.5	80.3
Sec. 3.3	✓ Locality in V & Talking Head	12.25	1.21	1.3	80.8
Sec. 3.4	Attention at Higher Resolution	13.10	1.48	3.5	81.7
	✓ Stride Attention	13.10	1.31	1.5	81.5
Sec. 3.5	Attention Downsampling	13.18	1.33	1.6	81.4
	✓ Dual-Path Attention Downsampling	13.40	1.35	1.6	81.8

cult to accelerate on mobile devices due to the sophisticated window partitioning and reordering. As for downsampling Keys (K) and Values (V) in [42], full resolution Queries (Q) are required to preserve the output resolution (**Out**) after attention matrix multiplication:

$$\mathbf{Out}_{[B,H,N,C]} = (Q_{[B,H,N,C]} \cdot K_{[B,H,C,\frac{N}{2}]}^T) \cdot V_{[B,H,\frac{N}{2},C]}, \quad (1)$$

where B , H , N , C denotes batch size, number of heads, number of tokens, and channel dimension respectively. Based on our test, the latency of the model merely drops to 2.8ms, which is still $2\times$ slower than the baseline model.

Therefore, to perform MHSA at the earlier stages of the network, we downsample all Query, Key, and Value to a fixed spatial resolution ($\frac{1}{32}$) and interpolate the outputs from the attention back to the original resolution to feed into the next layer, as shown in Fig. 2((d)&(e)). We refer to this method as Stride Attention. As in Tab. 1, this simple approximation significantly reduces the latency from 3.5ms to 1.5ms and preserves a competitive accuracy (81.5% vs. 81.7%).

3.5. Dual-Path Attention Downsampling

Most vision backbones utilize strided convolutions or pooling layers to perform a static and local downsampling and form a hierarchical structure. Some recent works start to explore attention downsampling. For instance, LeViT [24] and UniNet [48] propose to halve feature resolution via attention mechanism to enable context-aware downsampling with the global receptive field. Specifically, the number of tokens in Query is reduced by half so that the output from the attention module is downsampled:

$$\text{Out}_{[B,H,\frac{N}{2},C]} = (Q_{[B,H,\frac{N}{2},C]} \cdot K_{[B,H,C,N]}^T) \cdot V_{[B,H,N,C]}. \quad (2)$$

However, it is nontrivial to decide how to reduce the number of tokens in Query. Graham *et al.* empirically use pooling to downsample Query [24], while Liu *et al.* propose to search for local or global approaches [48]. To achieve acceptable inference speed on mobile devices, applying attention downsampling to early stages with high resolution is not favorable, restricting the values of existing works that search different downsampling approaches at higher-resolution.

Instead, we propose a combined strategy, which is dual-path attention downsampling, that yields both locality and global dependency, as in Fig. 2(f). To get downsampled Queries, we use pooling as static local downsampling, 3×3 DWCONV as learnable local downsampling, and combine and project the results into Query dimension. In addition, the attention downsampling module is residual connected to a regular strided CONV to form a local-global manner, similar to the downsampling bottlenecks [30] or inverted bottlenecks [64]. As shown in Tab. 1, with slightly more parameters and latency overhead, we further improve the accuracy to 81.8% with dual-path attention downsampling, which also has better performance than only using attention module for subsampling, *i.e.*, attention downsampling.

4. EfficientFormerV2

As discussed, current arts merely focus on optimizing one metric, thus are either redundant in size [45] or slow in inference [56]. To find the most suitable vision backbones for mobile deployment, we propose to jointly optimize model size and speed. Furthermore, the network designs in Sec. 3 favor a deeper network architecture (Sec. 3.2) and more attentions (Sec. 3.4), calling for an improved search space and algorithm. In what follows, we present the supernet design of EfficientFormerV2 and its search algorithm.

4.1. Design of EfficientFormerV2

As discussed in Sec. 3.2, we employ a 4-stage hierarchical design which obtains feature sizes in $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of the input resolution. Similar to its predecessor [45], EfficientFormerV2 starts with a small kernel convolution stem

to embed input image instead of using inefficient embedding of non-overlapping patches,

$$\mathbb{X}_{i|_{i=1:j|_{j=1}}}^{B,C|_{j=1},\frac{H}{4},\frac{W}{4}} = \text{stem}(\mathbb{X}_0^{B,3,H,W}), \quad (3)$$

where B denotes the batch size, C refers to channel dimension (also represents the width of the network), H and W are the height and width of the feature, $\mathbb{X}_j$ is the feature in stage j , $j \in \{1, 2, 3, 4\}$, and i indicates the i -th layer. The first two stages capture local information on high resolutions; thus we only employ the unified FFN (FFN, Fig. 2(b)),

$$\mathbb{X}_{i+1,j}^{B,C_j,\frac{H}{2^{j+1}},\frac{W}{2^{j+1}}} = S_{i,j} \cdot \text{FFN}^{C_j,E_{i,j}}(\mathbb{X}_{i,j}) + \mathbb{X}_{i,j}, \quad (4)$$

where $S_{i,j}$ is a learnable layer scale [83] and the FFN is constructed by two properties: stage width C_j and a per-block expansion ratio $E_{i,j}$. Note that each FFN is residual connected. In the last two stages, both local FFN and global MHSA blocks are used. Therefore, on top of Eqn. 4, global blocks are defined as:

$$\mathbb{X}_{i+1,j}^{B,C_j,\frac{H}{2^{j+1}},\frac{W}{2^{j+1}}} = S_{i,j} \cdot \text{MHSA}(\text{Proj}(\mathbb{X}_{i,j})) + \mathbb{X}_{i,j}, \quad (5)$$

where Queries (Q), Keys (K), and Values (V) are projected from input features through linear layers $Q, K, V \leftarrow \text{Proj}(\mathbb{X}_{i,j})$, and

$$\text{MHSA}(Q, K, V) = \text{Softmax}(Q \cdot K^T + \text{ab}) \cdot V, \quad (6)$$

with ab as a learnable attention bias for position encoding.

4.2. Jointly Optimizing Model Size and Speed

Though the baseline network EfficientFormer [45] is found by latency-driven search and yields fast inference speed on mobile, there are two major drawbacks for the search algorithm. First, the search process is merely constrained by speed, resulting in the final models being parameter redundant, as in Fig. 1. Second, it only searches for depth (number of blocks N_j per stage) and stage width C_j , which is in a *coarse-grained* manner. In fact, the majority of computations and parameters of the network are in FFNs, and the parameter and computation complexity are linearly related to its expansion ratio $E_{i,j}$. $E_{i,j}$ can be specified independently for each FFN without the necessity to be identical. Thus, searching $E_{i,j}$ enables a more *fine-grained* search space where the computations and parameters can distribute *flexibly* and *non-uniformly* within each stage. This is a missing property in most recent ViT NAS arts [23, 48, 45], where $E_{i,j}$ remains identical per stage. We propose a search algorithm that enables a flexible per-block configuration, with joint constraints on size and speed, and finds vision backbones best suited for mobile devices.

4.2.1 Search Objective

First, we introduce the metric guiding our joint search algorithm. Given the fact that the size and latency of a network all matter when evaluating mobile-friendly models, we consider a generic and fair metric that better understands the performance of a network on mobile devices. Without loss of generality, we define a Mobile Efficiency Score (MES):

$$\text{MES} = \text{Score} \cdot \prod_i \left(\frac{M_i}{U_i} \right)^{-\alpha_i}, \quad (7)$$

where $i \in \{\text{size}, \text{latency}, \dots\}$ and $\alpha_i \in (0, 1]$ indicating the corresponding importance. M_i , and U_i represent the metric and its unit. *Score* is a pre-defined base score set as 100 for simplicity. Model size is calculated by the number of parameters, and latency is measured as running time when deploying models on devices. Since we focus on mobile deployment, the size and speed of MobileNetV2 are used as the unit. Specifically, we define $U_{\text{size}} = 3\text{M}$, and U_{latency} as 1ms latency on iPhone 12 (iOS 16) deployed with CoreML-Tools [1]. To emphasize speed, we set $\alpha_{\text{latency}} = 1.0$ and $\alpha_{\text{size}} = 0.5$. Decreasing size and latency can lead to a higher MES, and we search for Pareto optimality on MES-Accuracy. The form of MES is general and can be extended to other metrics of interest, such as inference-time memory footprint and energy consumption. Furthermore, the importance of each metric is easily adjustable by appropriately defining α_i .

4.2.2 Search Space and SuperNet

Search space consists of: (i) the depth of the network, measured by the number of blocks N_j per stage, (ii) the width of the network, *i.e.*, the channel dimension C_j per stage, and (iii) expansion ratio $E_{i,j}$ of each FFN. The amount of MHSA can be seamlessly determined during depth search, which controls the preservation or deletion of a block in the supernet. Thus, we set every block as MHSA followed by FFN in the last two stages of the supernet and obtain subnetworks with the desired number of global MHSA by depth search.

Supernet is constructed by using a slimmable network [82] that executes at elastic depth and width to enable a pure evaluation-based search algorithm. Elastic depth can be naturally implemented through stochastic drop path augmentation [34]. As for width and expansion ratio, we follow Yu *et al.* [81] to construct switchable layers with shared weights but independent normalization layers, such that the corresponding layer can execute at different channel numbers from a predefined set, *i.e.*, multiples of 16 or 32. Specifically, the expansion ratio $E_{i,j}$ is determined by the channels of the depth-wise 3×3 Conv in each FFN, and stage width C_j is determined by aligning the output channels of the last projection (1×1 Conv) of FFN and MHSA blocks. The

switchable execution can be expressed as:

$$\hat{\mathbb{X}}_i = \gamma_c \cdot \frac{w^{i,c} \cdot \mathbb{X}_i - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c, \quad (8)$$

where $w^{i,c}$ refers to slicing the first c filters of the weight matrix to obtain a subset of output, and γ_c, β_c, μ_c , and σ_c are the parameters and statistics of the normalization layer designated for width c . The supernet is pre-trained with Sandwich Rule [82] by training the largest, the smallest, and randomly sampled two subnets at each iteration (we denote these subnets as max, min, rand-1, and rand-2 in Alg. 1).

Algorithm 1 Evaluation-based search for size and speed

Require: Latency lookup table $T : \{\text{FFN}^{C,E}, \text{MHSA}^C\}$

Ensure: Subnet satisfying objectives: params, latency, or MES

→ **Super-net Pretraining:**

for epoch do

for each iter do

for subnet $\in \{\text{min}, \text{rand-1}, \text{rand-2}, \text{max}\}$ do

$\mathbb{Y} \leftarrow \prod_i \{\text{FFN}_i, \text{MHSA}_i\}(\mathbb{X}_i)$

$\mathcal{L} \leftarrow \text{criterion}(\mathbb{Y}, \text{label})$, backpropagation

end for ▷ Sandwich Rule

Update parameters (AdamW [53])

end for

end for ▷ finish supernet training

→ **Joint search for size and speed:**

Initialize state $S \leftarrow \{S_{N_{\text{max}}}, S_{C_{\text{max}}}, S_{E_{\text{max}}}\}$

while Objective not satisfied **do**

Execute action $\hat{A} \leftarrow \arg \min_A \frac{\Delta_{\text{Acc}}}{\Delta_{\text{MES}}}$

Update state frontier

end while

▷ get sub-net with target MES

→ **Train the searched architecture from scratch**

Discussion. The pruning of our supernet is partially inspired by the slimmable network [82]. However, the differences are also significant. First, the search objective is different. We apply the introduced joint search objective for optimizing model size and efficiency (Sec. 4.2.1). Second, the search actions are different. Depth is pruned through the reduction of each *block*, which is possible since we unify the design and only adopt two blocks: Feed Forward Network (Sec. 3.1) and attention block. The way of pruning the depth is different from the slimmable network. Unifying all the flexible search actions under one joint objective has not been studied for transformers before.

4.2.3 Search Algorithm

Now that search objective, search space, and supernet are formulated, we present the search algorithm. Since the supernet is executable at elastic depth and switchable width, we can search the subnetworks with the best Pareto curve by analyzing the efficiency gain and accuracy drop with respect to each slimming action. We define the action pool as:

$$A \in \{A_{N[i,j]}, A_{C[j]}, A_{E[i,j]}\}, \quad (9)$$

Table 2. **Classification results on ImgeNet-1K.** We report the number of parameters, *i.e.*, Params (M), GMACs, Training Epochs, and Top-1 accuracy for various methods. The latency results are obtained by running models on iPhone 12 (Neural Engine) compiled with CoreMLTools, Pixel 6 (CPU) compiled with XNNPACK, and Nvidia A100 (GPU) compiled with TensorRT. The batch size is 1 for models tested on iPhone 12 and Pixel 6, and 64 for A100. (-) denotes unrevealed or unsupported models. † denotes we re-train the previous models with the exact same training recipe as our work. Different training seeds result in about 0.1% fluctuation in accuracy. The latency is benchmarked with warmup and averaged over multiple runs, where the error ranges within 0.1 ms.

Model	Type	Params (M)	GMACs	Latency (ms)			MES ↑	Epochs	Top-1(%)
				iPhone 12	Pixel 6	A100			
MobileNetV2×1.0	CONV	3.5	0.3	0.9	25.3	5.0	102.9	300	71.8
MobileNetV2×1.0 †	CONV	3.5	0.3	0.9	25.3	5.0	102.9	300	72.2
MobileViT-XS	Hybrid	2.3	0.7	7.3	64.4	11.7	15.6	300	74.8
EdgeViT-XXS	Hybrid	4.1	0.6	2.4	30.9	11.3	35.6	300	74.4
Hydra Attention	Hybrid	3.5	0.38	4.1	-	9.7	22.6	300	75.6
EfficientFormerV2-S0	Hybrid	3.5	0.40	0.9	20.8	6.6	102.9	300 / 450	75.7 / 76.2
MobileNetV2×1.4	CONV	6.1	0.6	1.2	42.8	7.3	58.4	300	74.7
MobileNetV2×1.4 †	CONV	6.1	0.6	1.2	42.8	7.3	58.4	300	76.7
MobileNetV3-L	CONV	5.4	0.22	15.8	-	7.2	4.7	300	75.2
FBNet-V3	CONV	5.6	0.39	1.0	-	7.9	73.2	300	75.1
EfficientNet-B0	CONV	5.3	0.4	1.4	29.4	10.0	53.7	350	77.1
DeiT-T	Attention	5.9	1.2	9.2	66.6	7.1	7.8	300	74.5
EdgeViT-XS	Hybrid	6.7	1.1	3.6	55.5	14.3	18.6	300	77.5
LeViT-128S	Hybrid	7.8	0.31	19.9	15.5	3.4	3.1	1000	76.6
EfficientFormerV2-S1	Hybrid	6.1	0.65	1.1	33.3	8.8	63.8	300 / 450	79.0 / 79.7
EfficientNet-B3	CONV	12.0	1.8	5.3	123.8	35.0	9.4	350	81.6
PoolFormer-s12	Pool	12	2.0	1.5	82.4	14.5	33.3	300	77.2
LeViT-192	Hybrid	10.9	0.66	29.6	30.1	5.2	1.8	1000	80.0
MobileFormer-508M	Hybrid	14.0	0.51	6.6	55.2	14.6	7.0	450	79.3
UniNet-B1	Hybrid	11.5	1.1	2.2	57.7	16.9	23.2	300	80.8
EdgeViT-S	Hybrid	11.1	1.9	4.6	92.5	21.2	11.3	300	81.0
EfficientFormer-L1	Hybrid	12.3	1.3	1.4	50.7	8.4	35.3	300	79.2
EfficientFormerV2-S2	Hybrid	12.6	1.25	1.6	57.2	14.5	30.5	300 / 450	81.6 / 82.0
ResNet50	CONV	25.5	4.1	2.5	167.5	9.0	13.7	300	78.5
ResNet50 †	CONV	25.5	4.1	2.5	167.5	9.0	13.7	300	80.5
ConvNext-T	CONV	29.0	4.5	83.7	340.5	28.8	0.4	300	82.1
ResMLP-S24	SMLP	30	6.0	7.6	325.4	17.4	4.2	300	79.4
PoolFormer-s24	Pool	21	3.6	2.4	154.3	28.2	15.7	300	80.3
PoolFormer-s36	Pool	31	5.2	3.5	224.9	41.2	8.9	300	81.4
DeiT-S	Attention	22.5	4.5	11.8	218.2	15.5	3.1	300	81.2
PVT-Small	Attention	24.5	3.8	24.4	-	23.8	1.4	300	79.8
T2T-ViT-14	Attention	21.5	4.8	-	-	21.0	-	310	81.5
Swin-Tiny	Attention	29	4.5	-	-	22.0	-	300	81.3
CSwin-T	Attention	23	4.3	-	-	28.7	-	300	82.7
LeViT-256	Hybrid	18.9	1.12	31.4	50.7	6.7	1.3	1000	81.6
LeViT-384	Hybrid	39.1	2.35	48.8	102.2	10.2	0.6	1000	82.6
ConvMixer-768	Hybrid	21.1	20.7	11.6	-	-	3.3	300	80.2
NasViT-Supernet	Hybrid	-	1.9	-	-	-	-	360	82.9
EfficientFormer-L3	Hybrid	31.3	3.9	2.7	151.9	13.9	11.5	300	82.4
EfficientFormer-L7	Hybrid	82.1	10.2	6.6	392.9	30.7	2.9	300	83.3
EfficientFormerV2-L	Hybrid	26.1	2.56	2.7	117.7	22.5	12.6	300 / 450	83.3 / 83.5
Supernet	Hybrid	37.1	3.57	4.2	-	-	6.8	300	83.5

where $A_{N[i,j]}$ denotes slimming each block, $A_C[j]$ refers to shrinking the width of a stage, and $A_E[i,j]$ denotes slimming each FFN to a smaller expansion. Initializing the state with full depth and width (largest subnet), we evaluate the accuracy outcome (Δ_{Acc}) of each frontier action on a validation

partition of ImageNet-1K, which only takes about 4 GPU-minutes. Meanwhile, parameter reduction (Δ_{Params}) can be directly calculated from layer properties, *i.e.*, kernel size, in-channels, and out-channels. We obtain the latency reduction ($\Delta_{Latency}$) through a pre-built latency look-up

table measured on iPhone 12 with CoreMLTools. With the metrics in hand, we can compute Δ_{MES} through Δ_{Params} and $\Delta_{Latency}$, and choose the action with the minimum per-MES accuracy drop: $\hat{A} \leftarrow \arg \min_A \frac{\Delta_{Acc}}{\Delta_{MES}}$. It is noteworthy that though the action combination is enormous, we only need to evaluate the frontier one at each step, which is linear in complexity. Details can be found in Alg. 1.

5. Experiments

5.1. ImageNet-1K Classification

Implementation Details. We implement the model through PyTorch 1.12 [60] and Timm library [76], and use 16 NVIDIA A100 GPUs to train our models. We train the models from scratch by 300 and 450 epochs on ImageNet-1K [18], with AdamW [53] optimizer. Learning rate is set to 10^{-3} per 1, 024 batch size with cosine decay. We use a standard image resolution, *i.e.*, 224×224 , for both training and testing. Similar to DeiT [70], we use RegNetY-16GF [61] with 82.9% top-1 accuracy as the teacher model for hard distillation. We use three testbeds to benchmark the latency:

- **iPhone 12 - NPU.** We get the latency on iPhone 12 (iOS 16) by running the models on Neural Engine (NPU). The models (batch size of 1) are compiled with CoreML [1].
- **Pixel 6 - CPU.** We test model latency on Pixel 6 (Android) CPU. To obtain the latency for most works under comparison, we replace the activation from *all* models to ReLU to get fair comparisons. The models (batch size of 1) are compiled with XNNPACK [4].
- **Nvidia GPU.** We also provide the latency on a high-end GPU—Nvidia A100. The models (batch size of 64) are deployed in ONNX [2] and executed by TensorRT [3].

Evaluation on Single Metric. We show the comparison results in Tab. 2, which includes *the most recent and representative* works on vision transformers and CNNs. The works that *do not have public models or are not compatible with mobile devices* [6, 79, 49] are not contained in Tab. 2. EfficientFormerV2 series achieve the SOTA results on a single metric, *i.e.*, number of parameters or latency. For model size, EfficientFormerV2-S0 outperforms EdgeViT-XXS [58] by 1.3% top-1 accuracy with even 0.6M fewer parameters and MobileNetV2 $\times$ 1.0 [64] by 3.5% top-1 with similar number of parameters. For large models, EfficientFormerV2-L model achieves identical accuracy to recent EfficientFormer-L7 [45] while being $3.1\times$ smaller. As for speed, with comparable or lower latency, EfficientFormerV2-S2 outperforms UniNet-B1 [48], EdgeViT-S [58], and EfficientFormer-L1 [45] by 0.8%, 0.6% and 2.4% top-1 accuracy, respectively. We hope the results can provide practical insight to inspire future architecture design: *modern deep neural networks are robust to architecture permutation, optimizing*

the architecture with joint constraints, such as latency and model size, will not harm individual metrics.

Jointly Optimized Size and Speed. Further, we demonstrate the superior performance of EfficientFormerV2 when considering both model size and speed. Here we use MES as a more practical metric to assess mobile efficiency than using size or latency alone. EfficientFormerV2-S1 outperforms MobileViT-XS [56], EdgeViT-XXS [58], and EdgeViT-XS [58] by 4.2%, 4.6%, and 1.5% top-1, respectively, with far higher MES. With $1.8\times$ higher MES, EfficientFormerV2-L outperforms MobileFormer-508M [13] by 4.0% top-1 accuracy. The evaluation results answer the central question raised at the beginning: *with the proposed mobile efficiency benchmark (Sec. 4.2.1), we can avoid entering a pitfall achieving seemingly good performance on one metric while sacrificing too much for others. Instead, we can obtain efficient mobile ViT backbones that are both light and fast.*

5.2. Downstream Tasks

Object Detection and Instance Segmentation. We apply EfficientFormerV2 as backbone in Mask-RCNN [29] pipeline and experiment on MS COCO 2017 [47]. The model is initialized with ImageNet-1K pretrained weights. We use AdamW [53] optimizer with an initial learning rate as 2×10^{-4} and conduct training for 12 epochs with resolution as 1333×800 . Following Li *et al.* [42], we apply a weight decay as 0.05 and freeze the normalization layers in the backbone. As in Tab. 3, with similar model size, our EfficientFormerV2-S2 outperform PoolFormer-S12 [83] by 6.1 AP^{box} and 4.9 AP^{mask}. EfficientFormerV2-L outperforms EfficientFormer-L3 [45] by 3.3 AP^{box} and 2.3 AP^{mask}.

Semantic Segmentation. We perform experiments on ADE20K [88], a challenging scene segmentation dataset with 150 categories. Our model is integrated as a feature encoder in Semantic FPN [38] pipeline, with ImageNet-1K pretrained weights. We train our model on ADE20K for 40K iterations with batch size as 32 and learning rate as 2×10^{-4} with a poly decay by the power of 0.9. We apply weight decay as 10^{-4} and freeze the normalization layers. Training resolution is 512×512 , and we employ a single scale testing on the validation set. As in Tab. 3, EfficientFormerV2-S2 outperforms PoolFormer-S12 [83] and EfficientFormer-L1 [45] by 5.2 and 3.5 mIoU, respectively.

5.3. Ablation Analysis on Search Algorithm

We compare the proposed search algorithm with the random search and the one from EfficientFormer [45]. As seen in Tab. 4, our search algorithm obtains models with much better performance than random search, *i.e.*, Random 1 and Random 2. Compared with EfficientFormer [45], we achieve higher accuracy under similar parameters and latency, demonstrating the effectiveness of fine-grained search and joint optimization of latency and size.

Table 3. **Object detection & instance segmentation** on MS COCO 2017 with the Mask RCNN pipeline.

Backbone	Params (M)	Detection & Instance Segmentation						Semantic
		AP ^{box}	AP ^{box} ₅₀	AP ^{box} ₇₅	AP ^{mask}	AP ^{mask} ₅₀	AP ^{mask} ₇₅	mIoU
ResNet18	11.7	34.0	54.0	36.7	31.2	51.0	32.7	32.9
PoolFormer-S12	12.0	37.3	59.0	40.1	34.6	55.8	36.9	37.2
EfficientFormer-L1	12.3	37.9	60.3	41.0	35.4	57.3	37.3	38.9
EfficientFormerV2-S2	12.6	43.4	65.4	47.5	39.5	62.4	42.2	42.4
ResNet50	25.5	38.0	58.6	41.4	34.4	55.1	36.7	36.7
PoolFormer-S24	21.0	40.1	62.2	43.4	37.0	59.1	39.6	40.3
Swin-T	29.0	42.2	64.4	46.2	39.1	64.6	42.0	41.5
EfficientFormer-L3	31.3	41.4	63.9	44.7	38.1	61.0	40.4	43.5
EfficientFormerV2-L	26.1	44.7	66.3	48.8	40.4	63.5	43.2	45.2

Table 4. **Ablation analysis of search algorithms.** Our proposed fine-grained search with joint constraints on size and speed achieves better results than random search and the coarse-grained, single objective search from EfficientFormer [45]. Latency is measured on iPhone 12.

Search Algorithm	Params (M)	Latency (ms)	Top-1 (%)
Random 1	3.5	1.0	74.7
Random 2	3.5	1.0	75.0
EfficientFormerV2 (Ours)	3.5	0.9	75.7
EfficientFormer [45]	3.1	0.9	74.2
EfficientFormerV2 (Ours)	3.1	0.9	75.0

6. Discussion and Conclusion

In this work, we comprehensively study transformer backbones, identify inefficient designs, and introduce mobile-friendly novel architectural changes. We further propose a fine-grained joint search on size and speed and obtain the EfficientFormerV2 model family. We extensively benchmark and compare our work with existing studies on different hardware and demonstrate that EfficientFormerV2 is both lightweight, ultra-fast in inference speed and high performance. Since we focus on size and speed, one future direction is to apply the joint optimization methodology to subsequent research exploring other critical metrics, such as memory footprint and CO₂ emission.

7. Acknowledgements

This work is partly supported by the Army Research Office/Army Research Laboratory via grant W911-NF-20-1-0167 to Northeastern University, CNS1909172.

References

- [1] Coremltools. <https://coremltools.readme.io/docs>.
- [2] Onnx. <https://onnx.ai>.
- [3] Tensorrt. <https://developer.nvidia.com/tensorrt>.
- [4] Xnnpack. <https://github.com/google/XNNPACK>.
- [5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, and Judy Hoffman. Hydra attention: Efficient attention with many heads. In *Computer Vision—ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 35–49. Springer, 2023.
- [6] Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. *arXiv preprint arXiv:2205.14756*, 2022.
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.
- [9] Arnav Chavan, Zhiqiang Shen, Zhuang Liu, Zechun Liu, Kwang-Ting Cheng, and Eric Xing. Vision transformer slimming: Multi-dimension searching in continuous optimization space. 2022.
- [10] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*, 2021.
- [11] 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*, pages 357–366, 2021.
- [12] 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*, pages 12270–12280, 2021.
- [13] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobileformer: Bridging mobilenet and transformer. *arXiv preprint arXiv:2108.05895*, 2021.
- [14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021.

- [15] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Re-visiting spatial attention design in vision transformers. *arXiv e-prints*, pages arXiv-2104, 2021.
- [16] Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, et al. Fbnetv3: Joint architecture-recipe search using predictor pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16276–16285, 2021.
- [17] 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.
- [18] 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*, pages 248–255. Ieee, 2009.
- [19] 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*, pages 12124–12134, 2022.
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.
- [22] Mohsen Fayyaz, Soroush Abbasi Kouhpayegani, Farnoush Rezaei Jafari, Eric Sommerlade, Hamid Reza Vaezi Joze, Hamed Pirsiavash, and Juergen Gall. Ats: Adaptive token sampling for efficient vision transformers. *arXiv preprint arXiv:2111.15667*, 2021.
- [23] Chengyue Gong, Dilin Wang, Meng Li, Xinlei Chen, Zhicheng Yan, Yuandong Tian, qiang liu, and Vikas Chandra. NASVit: Neural architecture search for efficient vision transformers with gradient conflict aware supernet training. In *International Conference on Learning Representations*, 2022.
- [24] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve Jegou, 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 (ICCV)*, pages 12259–12269, October 2021.
- [25] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. *arXiv preprint arXiv:2107.06263*, 2021.
- [26] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022.
- [27] Qi Han, ZeJia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *International Conference on Learning Representations*, 2021.
- [28] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.
- [29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [31] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021.
- [32] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.
- [33] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [34] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.
- [35] Qing Jin, Jian Ren, Oliver J Woodford, Jiazhao Wang, Geng Yuan, Yanzhi Wang, and Sergey Tulyakov. Teachers do more than teach: Compressing image-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13600–13611, 2021.
- [36] Qing Jin, Jian Ren, Richard Zhuang, Sumant Hanumante, Zhengang Li, Zhiyu Chen, Yanzhi Wang, Kaiyuan Yang, and Sergey Tulyakov. F8net: Fixed-point 8-bit only multiplication for network quantization. *arXiv preprint arXiv:2202.05239*, 2022.
- [37] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR, 2021.
- [38] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.
- [39] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR. OpenReview.net*, 2020.

- [40] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021.
- [41] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.
- [42] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022.
- [43] Wei Li, Xing Wang, Xin Xia, Jie Wu, Xuefeng Xiao, Min Zheng, and Shiping Wen. Sepvit: Separable vision transformer. *CoRR*, abs/2203.15380, 2022.
- [44] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021.
- [45] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. In *NeurIPS*, 2022.
- [46] Yawei Li, Kai Zhang, Jiezhong Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021.
- [47] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.
- [48] Jihao Liu, Xin Huang, Guanglu Song, Hongsheng Li, and Yu Liu. Uninet: Unified architecture search with convolution, transformer, and mlp. In *European Conference on Computer Vision*, pages 33–49. Springer, 2022.
- [49] 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. *arXiv preprint arXiv:2111.09883*, 2021.
- [50] 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*, pages 10012–10022, 2021.
- [51] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.
- [52] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021.
- [53] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [54] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.
- [55] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. *arXiv preprint arXiv:2206.10589*, 2022.
- [56] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.
- [57] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*, 2022.
- [58] Juntong 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 *European Conference on Computer Vision*, 2022.
- [59] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *arXiv preprint arXiv:2205.13213*, 2022.
- [60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [61] 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*, pages 10428–10436, 2020.
- [62] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [63] Cédric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to merge tokens in vision transformers. *CoRR*, abs/2202.12015, 2022.
- [64] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [65] Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. Talking-heads attention. *arXiv preprint arXiv:2003.02436*, 2020.
- [66] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *arXiv preprint arXiv:2205.12956*, 2022.
- [67] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [68] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

- [69] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.
- [70] 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*, pages 10347–10357. PMLR, 2021.
- [71] 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*, pages 32–42, 2021.
- [72] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *CoRR*, abs/2204.01697, 2022.
- [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [74] 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*, pages 568–578, 2021.
- [75] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. *arXiv preprint arXiv:2108.00154*, 2021.
- [76] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019.
- [77] Bichen Wu, Chaojian Li, Hang Zhang, Xiaoliang Dai, Peizhao Zhang, Matthew Yu, Jialiang Wang, Yingyan Lin, and Peter Vajda. Fbnetv5: Neural architecture search for multiple tasks in one run. *arXiv preprint arXiv:2111.10007*, 2021.
- [78] Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. Pale transformer: A general vision transformer backbone with pale-shaped attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2731–2739, 2022.
- [79] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.
- [80] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.
- [81] Jiahui Yu and Thomas S Huang. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1803–1811, 2019.
- [82] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks, 2018.
- [83] 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. *arXiv preprint arXiv:2111.11418*, 2021.
- [84] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. Improving visual quality of image synthesis by a token-based generator with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [85] Qingru Zhang, Simiao Zuo, Chen Liang, Alexander Bukharin, Pengcheng He, Weizhu Chen, and Tuo Zhao. Platon: Pruning large transformer models with upper confidence bound of weight importance. In *International Conference on Machine Learning*, pages 26809–26823. PMLR, 2022.
- [86] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation, 2022.
- [87] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. 2022.
- [88] 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*, 2017.
- [89] Qinqin Zhou, Kekai Sheng, Xiawu Zheng, Ke Li, Xing Sun, Yonghong Tian, Jie Chen, and Rongrong Ji. Training-free transformer architecture search. *arXiv preprint arXiv:2203.12217*, 2022.