EMOv2: Pushing 5M Vision Model Frontier

Jiangning Zhang, Teng Hu, Haoyang He, Zhucun Xue, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—This work focuses on developing parameter-efficient and lightweight models for dense predictions while trading off parameters, FLOPs, and performance. Our goal is to set up the new frontier of the 5 M magnitude lightweight model on various downstream tasks. Inverted Residual Block (IRB) serves as the infrastructure for lightweight CNNs, but no counterparts have been recognized by attention-based design. Our work rethinks the lightweight infrastructure of efficient IRB and practical components in Transformer from a unified perspective, extending CNN-based IRB to attention-based models and abstracting a one-residual Meta Mobile Block (MMBlock) for lightweight model design. Following neat but effective design criterion, we deduce a modern Improved Inverted Residual Mobile Block (i^{2}RMB) and improve a hierarchical Efficient MOdel (EMOv2) with no elaborate complex structures. Considering the imperceptible latency for mobile users when downloading models under 4 G/5 G bandwidth and ensuring model performance, we investigate the performance upper limit of lightweight models with a magnitude of 5 M. Extensive experiments on various vision recognition, dense prediction, and image generation tasks demonstrate the superiority of our EMOv2 over state-of-the-art methods, e.g., EMOv2-1 M/2M/5 M achieve 72.3, 75.8, and 79.4 Top-1 that surpass equal-order CNN-/Attention-based models significantly. At the same time, EMOv2-5 M equipped RetinaNet achieves 41.5 mAP for object detection tasks that surpasses the previous EMO-5 M by +2.6↑ . When employing the more robust training recipe, our EMOv2-5M eventually achieves 82.9 Top-1 accuracy, which elevates the performance of 5M magnitude models to a new level.**

***Index Terms*—Computer vision, lightweight vision backbone, vision architecture design.**

I. Introduction

LIGHTWEIGHT models are particularly crucial in resource-constrained scenarios, drawing many research efforts [1], [2], [3], [4], [5], [6], [7] in various fields. Early work primarily can be divided into two categories: 1) models with fewer FLOPs and faster hardware-specific inference speeds [8], [9], [10], [11], [12], which do not emphasize parameter counts and perform poorly in high-resolution downstream tasks; 2) models that balance FLOPs and performance under limited parameter counts [2], [13], resulting in more compact models. With the development of computational devices, most current models achieve throughput of several thousand and latency within real-time 20 ms [1], [2], [14], where computational power is not the bottleneck for small model applications, even if we strive to reduce their computational requirements. Additionally, edge applications iterate models rapidly, as seen in short video platforms like TikTok, where effects frequently update lightweight real-time detection algorithms and small-scale generation models. Considering the imperceptible delay in downloading models under 4 G/5 G bandwidth and ensuring model performance, a lightweight model of 5 M magnitude is recommended as an appropriate size [15], [16]. Therefore, this paper explores the upper limits of lightweight model performance with a fixed parameter count, using a 5 M lightweight model as a typical representative.

MobileNetv2 [9] introduces an efficient *Inverted Residual Block* (IRB) based on *Depth-Wise Separable Convolution* (DW-Conv), which is widely regarded as the foundation of efficient models [10], [12], [17]. However, constrained by the natural induction bias of static convolution operations, the accuracy of CNN-based lightweight models is suboptimal due to the lack of global modeling capabilities. *This motivates us to explore the construction of a stronger fundamental block that surpasses the IRB by introducing global modeling capabilities.* On the other hand, benefiting from the dynamically global modeling capability of Multi-Head Self-Attention (MHSA), Vision Transformer (ViT) [18] and its derivatives [19], [20], [21], [22], [23], [24], [25], [26] have achieved significant improvements over CNNs. Some works attempt to address the quadratic computational complexity of MHSA by designing variants with linear complexity [27], [28], reducing the spatial resolution of features [19], [29], [30], rearranging channels [31], and employing local window attention [21], [22], among other strategies. Recently, researchers have introduced MHSA into certain layers of lightweight CNN models to improve complex blocks [2], [14], [17], [32], [33], [34] or have used multiple hybrid blocks. However, such designs lack uniformity, require meticulous design, and pose higher demands for adaptation to mobile device deployment. So far, no works explore MHSA-based counterparts as IRB, and this inspires us to think: *can we build a lightweight*

Received 10 December 2024; revised 29 May 2025; accepted 25 July 2025. Date of publication 7 August 2025; date of current version 3 October 2025. This work was supported in part by the "Leading Goose" Key R&D Program of Zhejiang Province under Grant 2025C01069, and in part by the National Research Foundation, Singapore, under its NRF Professorship under Grant NRF-P2024-001. Recommended for acceptance by R. K. Iyer. *(Corresponding authors: Zhucun Xue; Yong Liu.)*

Jiangning Zhang, Haoyang He, Zhucun Xue, and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China, and also with Youtu Lab, Tencent, Shenzhen 518057, China (e-mail: 12432038@zju.edu.cn; yongliu@iipc.zju.edu.cn).

Teng Hu is with Shanghai Jiao Tong University, Shanghai 200240, China.

Yabiao Wang and Chengjie Wang are with Youtu Lab, Tencent, Shenzhen 518057, China.

Xiangtai Li and Dacheng Tao are with Nanyang Technological University, Singapore 639798.

Code is available at https://github.com/zhangzjn/EMOv2.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3596776, provided by the authors.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3596776, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3596776

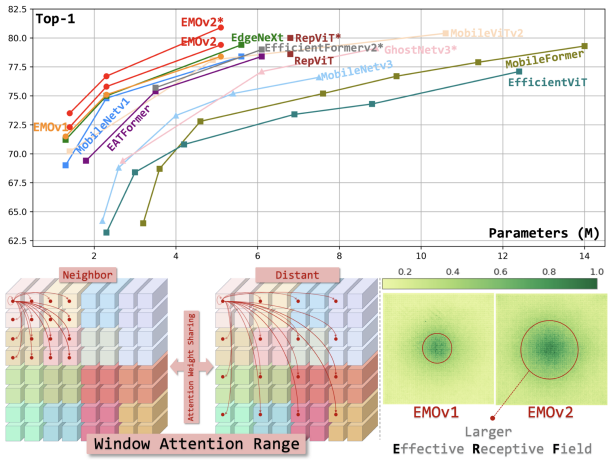

Fig. 1. *Top: Performance* versus *Parameters* with concurrent methods. Our EMOv2 achieves significant accuracy with fewer parameters. Superscript $*$: The comparison methods employ more robust training strategies described in their papers, while ours uses the strategy mentioned in Table XIX(e). *Bottom:* The range of token interactions varies with different window attention mechanisms. Our EMOv2, with parameter-shared spanning attention in Section III-C1, has a larger and correspondingly stronger Effective Receptive Field (ERF).

IRB-like infrastructure for attention-based models with only basic operators?

Based on the motivation above, we rethink the efficient IRB in MobileNetv2 [9] and the MHSA / FFN modules in Transformer [35] from a unified perspective, expecting to integrate their advantages at the infrastructure design level. As shown in Fig. 2-Left, while working to bring one-residual IRB with inductive bias into the attention model, we observe that MHSA/FFN submodules in two-residual Transformer share a similar meta-structure to IRB. Thus, we inductively abstract a one-residual Meta Mobile Block (MMBlock in Section III-B1) that takes parametric arguments' *expansion ratio* λ and *efficient operator* $\mathcal{F}$ to instantiate different modules, i.e., IRB, MHSA, and FFN. MMBlock reveals the consistent essence expression of the above three modules and can be regarded as an improved lightweight concentrated aggregate of Transformer. Furthermore, a neat yet effective *Inverted Residual Mobile Block* (iRMB) is deduced that only contains fundamental Depth-Wise Convolution and the improved EW-MHSA (*c.f.*, Section III-B2). And we build a ResNet-like 4-phase Efficient MOdel (EMOv1) with only iRMBs (*c.f.*, Section III-B).

Even though EMOv1 [13] achieves promising results, it is limited by window attention that can only model the interaction of neighbor information within a local window, as shown in Fig. 1-Bottom. This modeling approach leads to suboptimal performance in high-resolution downstream tasks due to the lack of distant information interaction. For instance, RetinaNet [36] using EMOv1-5 M only achieves 38.9 mAP that does not even reach 40. Recently, MobileViT [17] attempts to model long-range attention but performs moderately due to the loss of local dynamic modeling capability and a significant increase in FLOPs with higher resolutions. Thus, more balanced efforts between long-range modeling and lower GFlops are needed. To overcome these challenges, we explore the procedure of attention computation and discover that the neighbor window attention map can be reused to model the correlation between distant positions. Based on this, we design a novel spanning mechanism Section III-C2 (i.e., SEW-MHSA) that simultaneously models neighbor and distant features. As shown in Fig. 1, this mechanism does not increase the number of parameters and only adds a small number of FLOPs. It significantly enhances the model's effective receptive field, thereby improving performance in high-resolution downstream tasks (Section IV-B). Additionally, we improve the detailed structure of i^{2}RMB to enhance the performance further and explore different training strategies to maximize the model's potential in mainstream image classification tasks. Detailed comparison with state-of-the-art methods can be viewed in Fig. 1. Due to the neat structural design, i^{2}RMB can be easily extended to various downstream tasks, achieving significant and consistent performance improvements. Specifically, we apply EMOv2 to the temporal dimension for video recognition, and V-EMO-v2 obtains 65.2 Top-1 accuracy with 5.9 M parameters on Kinetics-400 for video classification that surpasses UniFormer-XXS's 63.2 with 9.8 M parameters. In addition, we enhance the recently popular UNet and DiT architectures for image segmentation and generation across multiple downstream tasks based on this module (Section III-C3). E.g., U-EMO-v2 obtains 88.3mAcc with 21.3 M parameters on HRF; D-EMO-v2 achieves 46.3/9.6 FID in generating 256$\times$256 ImageNet images with 400K training steps on S/XL scales, which significantly surpasses DiT's 68.4/19.5. In summary, we make the following significant extensions over the preliminary conference version (EMO [13] at ICCV'23):

1) Based on the abstracted one-residual *Meta Mobile Block* for lightweight model design, we extend the iRMB to a powerful i^{2}RMB block. Specifically, we design a parameter-sharing spanning attention mechanism, enabling interaction between neighborhood and distant spatial features within a single module without increasing the model's parameter count. This mechanism is also compatible with EW-MHSA, achieving efficient feature modeling for mobile applications. Additionally, we improve the post-attention and large local kernel structures to further enhance model performance.
2) We construct a 4-stage EMOv2 backbone solely based on the deduced i^{2}RMB block. This model significantly improves performance while maintaining the similar parameter count as EMOv1. For instance, EMOv2-5 M achieves a +1.0↑ improvement over EMOv1-5M in classification tasks. The performance gap widens further in high-resolution downstream tasks, with improvements of +1.7↑ and +2.6↑ mAP using SSDLite and RetinaNet, respectively. We also explore the impact of stronger training strategies on model performance, validating the model's scaling capability, with EMOv2-5M reaching up to 82.9 Top-1 accuracy.
3) Thanks to the general, neat, and powerful design of i^{2}RMB, we can easily extend it to a series of tasks, constructing various lightweight versions of different types of structures and achieving significant improvements. Finally, we provide detailed studies and experimental analysis to build our attention-based lightweight models in Section IV-C.

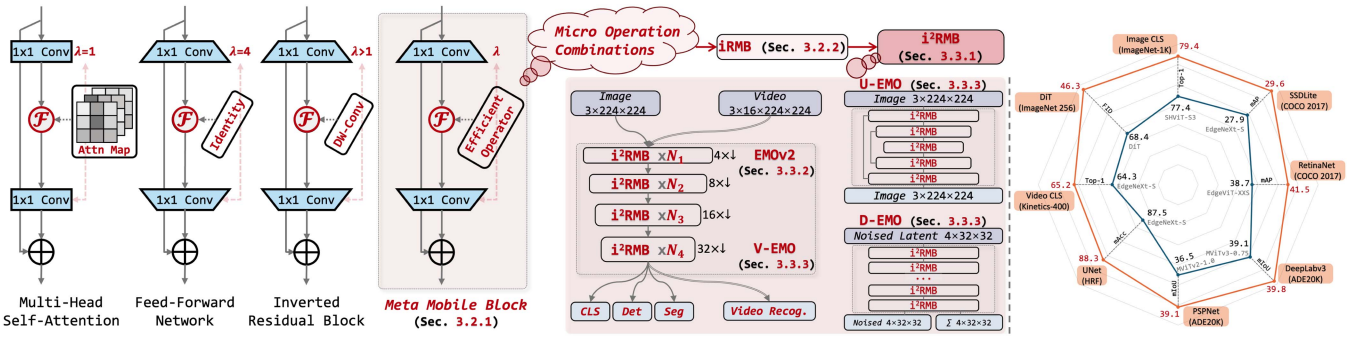

Fig. 2. *Left*: Abstracted unified $Meta-MobileBlock$ from *Multi-Head Self-Attention*, *Feed-Forward Network* [35], and *Inverted Residual Block* [9] (*c.f.* Section III-B1). The inductive block can be deduced into specific modules using different *expansion ratio* λ and *efficient operator* $\mathcal{F}$. *Middle*: We construct a family of vision models based on our i^{2}RMB module: 4-stage *EMOv2*, composed solely of the deduced *i^{2}RMB* (*c.f.* Section III-B2), for various perception tasks (image classification, detection, and segmentation in Section IV-B). Additionally, we introduce the temporally extended *V-EMO* for video classification, the U-EMO based on an encoder-decoder architecture, and D-EMO to replace the Transformer block in DiT [67]. These downstream models are typically built based on the i^{2}RMB. *Right*: Performance comparison with different SoTAs on various tasks.

TABLE I
OVERVIEW OF TECHNICAL TERMS AND ACRONYMS

Term	Full Name	Location
IRB	Inverted Residual Block	Sec. 3.2.1
MMBlock	Meta Mobile Block	Sec. 3.2.1
iRMB	Improved Inverted Residual Mobile Block	Sec. 3.2.2
i^{2}RMB	Improved Inverted Residual Mobile Block v2	Sec. 3.3.1
DW-Conv	Depth-Wise Separable Convolution	Sec. 3.2.2
EW-MHSA	Expanded Window Multi-Head Self-Attention	Sec. 3.2.2
SEW-MHSA	Spanning EW-MHSA	Sec. 3.3.1

4) We re-write the entire draft and add a more comprehensive discussion on close related works. We open-source our EMOv2 for the community.

We summarize the key technical terms and acronyms in Table I to facilitate paper reading.

II. Related Work

Lightweight CNN Models: With the increasing demands of neural networks for mobile vision applications, efficient model design has attracted extensive attention from researchers in recent years. SqueezeNet [37] replaces 3x3 filters with 1x1 filters and decreases channel numbers to reduce model parameters, while Inceptionv3 [38] factorizes the standard convolution into asymmetric convolutions. Later, MobileNet [8] introduces depth-wise separable convolution to alleviate a large amount of computation and parameters, followed in subsequent lightweight models [6], [9], [11], [39]. Besides the above handcraft methods, researchers exploit automatic architecture design in the pre-defined search space [1], [10], [12]. Specifically, RepViT [40] leverages the re-parameterization technique to enhance model performance, while recent GhostNetV3 [41] has further incorporated a Knowledge Distillation (KD) strategy. MobileNetv4 [42] employs both NAS algorithm and KD strategy to achieve impressive results, where a strong training recipe has already become a trend in lightweight model research. We draw on lightweight design principles from the CNN domain, such as depth-wise convolution and inverted residual designs, and integrate them with attention mechanisms to construct a stronger hybrid module.

Hugging Vision Transformer with CNN: Since ViT [18] first introduces Transformer structure [35] into visual tasks, massive improvements have successfully been developed. DeiT [43] provides a benchmark for efficient transformer training, subsequent works [19], [21] employ ResNet-like [44] pyramid structure to form pure Transformer-based models for dense prediction tasks. However, the absence of 2D convolution will potentially increase the optimization difficulty and damage the model accuracy for lacking local inductive bias, so researchers [45], [46] concentrate on how to better integrate convolution into Transformer for obtaining stronger hybrid models. E.g., work [47] incorporates convolution design into FFN, works [48], [49] regard convolution as the positional embedding for enhancing inductive bias of the model, and works [29] for attention and QKV calculations, respectively. Recently, MogaNet [50] encapsulates conceptual convolutions and gated aggregation into a compact module, and SHViT [51] uses a depthwise convolution layer for local feature aggregation or conditional position embedding. However, the above methods are still confined to the MetaFormer [52] architecture, where each block contains two residual connections. EMOv1 studies how to build a neat but effective lightweight model based on an improved one-residual attention block. In contrast, this paper further investigates the parameter-sharing mechanism for window attention, enabling it to simultaneously model neighbor and distant information interactions, thereby significantly enhancing the performance of downstream tasks.

Effective Transformer Improvements: Researchers [2], [53] have started to lighten Transformer-based models for low computational power. Tao et al. [53] introduces additional learnable tokens to capture global dependencies efficiently, and Chen et al. [53] design a parallel structure of MobileNet and Transformer with a two-way bridge in between. Works [54], [55] improve an efficient Transformer block by borrowing convolution operation, while EdgeNeXt [2] absorbs effective Res2Net [56] and transposed channel attention [57]. MobileVit series [14], [17], [32] fuse improved MobileViT blocks with Mobile blocks [9]. Recent EfficientFormerV2 [1] uses the NAS algorithm to search hardware-friendly modules, while ViG [58] introduces a gating

TABLE II
CRITERION COMPARISON FOR CURRENT EFFICIENT MODELS

Method *vs.* Criterion	①	②	③	④
MobileNet Series [8], [9], [32]	✔	✔	✚	✘
MobileViT Series [14], [17], [32]	✚	✚	✚	✘
EdgeNeXt [2]	✚	✘	✔	✘
EdgeViT [55]	✔	✚	✚	✘
RepViT [40]	✔	✘	✔	✘
EfficientFormerV2 [1]	✔	✚	✔	✘
EfficientVMamba [65]	✘	✘	✚	✘
MogaNet [50]	✔	✔	✔	✘
EMOv1	✔	✔	✔	✘
EMOv2	✔	✔	✔	✔

①: Usability; ②: Uniformity; ③: Efficiency and Effectiveness; ④: Generalization. ✔: Satisfied. ✚: Partially satisfied. ✘: Unsatisfied.

TABLE III
COMPLEXITY AND MAXIMUM PATH LENGTH (MPL) ANALYSIS OF MODULES

Module	#Params	FLOPs	MPL
MHSA	$4(C+1)C$	$8C^2L+4CL^2+3L^2$	$O(1)$
W-MHSA	$4(C+1)C$	$8C^2L+4CLl+3Ll$	$O(Inf)$
Conv	$(Ck^2/G+1)C$	$(2Ck^2/G)LC$	$O(2W/(k-1))$
DW-Conv	$(k^2+1)C$	$(2k^2)LC$	$O(2W/(k-1))$

Input/output feature maps are in $\mathbb{R}^{C\times W\times W}$, $L=W^2$, $l=w^2$, W and w are feature map size and window size, while k and G are kernel size and group number.

TABLE IV
TOY EXPERIMENTS FOR ASSESSING IRMB AND I^{2}RMB

Model	#Params ↓	FLOPs ↓	Top-1 ↑
DeiT-Tiny [43]	5.7M	1.3G	72.2
DeiT-Tiny w / iRMB	4.9M	1.1G	74.3 +2.1% ↑
DeiT-Tiny w / i^{2}RMB	5.0M	1.3G	75.0 +2.8% ↑
PVT-Tiny [19]	13.2M	1.9G	75.1
PVT-Tiny w / iRMB	11.7M	1.8G	75.4 +0.3% ↑
PVT-Tiny w / i^{2}RMB	11.9M	1.9G	76.1 +1.0% ↑

TABLE V
CORE CONFIGURATIONS OF EMOV2 VARIANTS

Items	EMOv2-1M	EMOv2-2M	EMOv2-5M
Depth	[2, 2, 8, 3]	[3, 3, 9, 3]	[3, 3, 9, 3]
Emb. Dim.	[32, 48, 80, 180]	[32, 48, 120, 200]	[48, 72, 160, 288]
Exp. Ratio	[2.0, 2.5, 3.0, 3.5]	[2.0, 2.5, 3.0, 3.5]	[2.0, 3.0, 4.0, 4.0]

TABLE VI
ABLATION STUDY ON COMPONENTS IN IRMB/I^{2}RMB

EMOv1 [13]			EMOv2		
EW-MHSA	DW-Conv	Top-1	SEW-MHSA	DW-Conv	Top-1
✘	✘	73.5	✘	✘	73.5
✔	✘	76.6 +3.1 ↑	✔	✘	77.7 +4.2 ↑
✘	✔	77.6 +4.1 ↑	✘	✔	78.1 +4.6 ↑
✔	✔	78.4 +4.9 ↑	✔	✔	79.4 +5.9 ↑

mechanism to facilitate the interaction of sequential and spatial information. However, most current approaches require *elaborate complex modules*, which limits the mobility and usability of the model. How to balance parameters, computation, and accuracy while designing easy-to-use lightweight models still needs further exploration.

RNN-reinvented Models: Due to the quadratic growth in computational complexity of Transformers with the number of tokens, some RNN-based models [59], [60], [61] have gradually gained attention, with Mamba [62] and RWKV [63] being the primary representatives. Zhu et al. [64] proposes vision Mamba, which applies SSM to visual tasks, while Duan et al. [60] also introduces a vision version based on RWKV. Recently, works [65], [66] explore the application of Mamba in lightweight visual tasks. These methods can seamlessly integrate into our proposed Meta Mobile Block, yielding favorable results. However, considering the verified stable performance of transformers across various fields, this paper explores improvements to the attention module based on a windowed operation.

III. METHODOLOGY

A. Criteria for General Lightweight Model

When designing light-weight visual models for mobile usages, we advocate the following criteria subjectively and empirically that an efficient model should satisfy as much as possible: ① *Usability:* Neat implementation that does not use complex operators and is easy to optimize for applications. ② *Uniformity.* As few core modules as possible to reduce model complexity and accelerate deployment. ③ *Efficiency and Effectiveness.* Balancing parameters and calculations with accuracy trade-off. ④ *Generalization.* Easily applied to perception tasks such as classification, detection, and segmentation, as well as to generative tasks, while compatible with architectures like ResNet and U-Net. We make a summary of current efficient models in Table II: *1)* Performance of MobileNet series [8], [9], [32] is now seen to be slightly lower, and its parameters are slightly higher than counterparts. *2)* Recent MobileViT series [14], [17], [32] achieve notable performances, but they suffer from higher FLOPs and slightly complex modules. *3)* EdgeNeXt [2] and EdgeViT [55] obtain pretty results, but their basic blocks also consist of elaborate modules. *4)* RepViT [40] employs multiple fundamental modules and introduces a re-parameterization strategy, while EfficientFormerV2 [1] utilizes NAS to search for hardware-friendly models, and EfficientVMamba [65] introduces a new SSM module. *5)* MogaNet [50] achieves a balance between performance and efficiency without introducing new complex operators. Comparably, the design principle of our EMO/v2 follows the above criteria without introducing complicated operations (*c.f.*, Section III-C2) while still obtaining impressive results on multiple vision tasks (*c.f.*, Section IV). Additionally, EMOv2 can be easily transferred to other models for various tasks, such as video classification, UNet-based image segmentation, and diffusion-based image generation (*c.f.*, Section III-C2).

B. Efficient MOdel (EMOv1)

1) *Meta Mobile Block:*

a) Motivation: 1) Recent Transformer-based works [21], [68], [69], [70], [71], [72], [73] are dedicated to improving spatial token mixing under the MetaFormer [52] for high-performance network. CNN-based *Inverted Residual Block* [9] (IRB) is recognized as the infrastructure of efficient models [9], [12], but little work has been done to explore attention-based counterpart. This inspires us to build a lightweight IRB-like infrastructure for attention-based models. 2) While working to bring one-residual

TABLE VII
COMPARISON OF *TRAINING RECIPES AMONG POPULAR AND CONTEMPORARY METHODS* AND WE EMPLOY THE SAME SETTING IN ALL EXPERIMENTS

Hyper-Params.	MNetv3 [10] ICCV'19	ViT [18] ICLR'21	DeiT [43] ICML'21	MViTv1 [17] ICLR'22	MViTv2 [14] arXiv'22	EdgeNeXt [2] arXiv'22	EFormerv2 [1] ICCV'23	RepViT [40] CVPR'24	MogaNet [50] ICLR'24	Vim [64] ICLR'24	GNetv3 [41] arXiv'2404	MNetv4 [42] arXiv'2404	EMOv1/v2 Ours
Epochs	300	300	300	300	300	300	300	300	300	300	600	500	300
Batch size	512	4096	1024	1024	1024	4096	1024	2048	1024	1024	2048	4096	2048
Optimizer	RMSprop	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	LAMB	AdamW	AdamW
Learning rate	$6.4e^{-2}$	$3e^{-3}$	$1e^{-3}$	$2e^{-3}$	$2e^{-3}$	$6e^{-3}$	$1e^{-3}$	$4e^{-3}$	$1e^{-3}$	$1e^{-3}$	$5e^{-3}$	$4e^{-3}$	$6e^{-3}$
Learning rate decay	$1e^{-5}$	$3e^{-1}$	$5e^{-2}$	$1e^{-2}$	$5e^{-2}$	$5e^{-2}$	2.5^{-2}	2.5^{-2}	4^{-2}	1^{-1}	5^{-2}	1^{-1}	$5e^{-2}$
Warmup epochs	3	3.4	5	2.4	16	20	5	5	5	5	3	5	20
Label smoothing	0.1	✘	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Drop out rate	✘	✔	✘	✔	✘	✘	✘	✘	✘	✘	✔	✔	✘
Drop path rate	✔	✘	✔	✘	✘	✔	✔	✘	✔	✘	✘	✔	0.1
RandAugment	9/0.5/1	✘	9/0.5/1	✘	9/0.5/1	9/0.5/1	9/0.5/1	9/0.5/1	7/0.5/1	9/0.5/1	9/0.5/1	15/0.7/2	9/0.5/1
Mixup alpha	✘	✘	0.8	✘	0.8	✘	0.8	0.8	0.1	0.8	✘	✘	✘
Cutmix alpha	✘	✘	1.0	✘	1.0	✘	1.0	1.0	1.0	1.0	✘	✘	✘
Erasing probability	0.2	✘	0.25	✘	0.25	✘	0.25	0.25	0.25	0.25	✘	-	✘
Position embedding	✘	✔	✔	✘	✘	✔	✔	✔	✔	✔	✘	✘	✘
Multi-scale sampler	✘	✘	✘	✔	✘	✔	✘	✘	✘	✘	✘	✘	✘
NAS	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘	✔	✘
KD	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✔	✔	✘
#Repre.	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘	✔	✘	✘

Please zoom in for clearer comparisons. Abbreviations: MNet → MobileNet; MViT → MobileViT; EFormerv2 → EfficientFormerv2; GNet → GhostNet; NAS: Neural Architecture Search; KD: Knowledge Distillation; #Repre.: Re-parameterization strategy.

TABLE VIII
PERFORMANCE OF OUR EMOV1/V2 WITH DIFFERENT LIGHTWEIGHT MODEL TRAINING RECIPES

Recipe	MNetv3 [10]	DeiT [43]	EdgeNeXt [2]	Vim [64]	Ours
EMOv1 [13]	*NaN*	78.1	78.3	77.9	78.4
EMOv2	*NaN*	78.8	79.1	78.5	79.4

IRB with inductive bias into the attention model, we stumble upon two underlying sub-modules (i.e., FFN and MHSA) in two-residual Transformer that happen to share a similar structure to IRB. This inspires us to integrate these elements into a unified block representation, thereby constructing a more shallow foundational visual backbone. Compared to each ViT block, which contains two residual connections, our approach simplifies the architecture.

b) Induction: We rethink Inverted Residual Block in MobileNetv2 [9] with core MHSA and FFN modules in Transformer [35], and inductively abstract a general Meta Mobile Block (MMBlock) in Fig. 2, which takes parametric arguments *expansion ratio* λ and *efficient operator* $\mathcal{F}$ to instantiate different modules. We argue that *the MMBlock can reveal the consistent essence expression of the above three modules, and MMBlock can be regarded as an improved lightweight concentrated aggregate of Transformer*. Also, this is the basic motivation for our elegant and easy-to-use EMO/v2, which only contains one deduced iRMB/i^{2}RMB absorbing advantages of lightweight CNN and Transformer. Taking image input $\boldsymbol{X}(\in \mathbb{R}^{C\times H\times W})$ as an example, MMBlock first use an expansion MLP_e with output/input ratio equaling λ to expand channel dimension:

$$\boldsymbol{X}_e = \text{MLP}_e(\boldsymbol{X})(\in \mathbb{R}^{\lambda C\times H\times W}). \tag{1}$$

Then, intermediate operator $\mathcal{F}$ enhance image features further, e.g., identity operator, static convolution, dynamic MHSA, etc. Considering that MMBlock is suitable for efficient network design, we present $\mathcal{F}$ as the concept of *efficient operator*, formulated as:

$$\boldsymbol{X}_f = \mathcal{F}(\boldsymbol{X}_e)\left(\in \mathbb{R}^{\lambda C\times H\times W}\right). \tag{2}$$

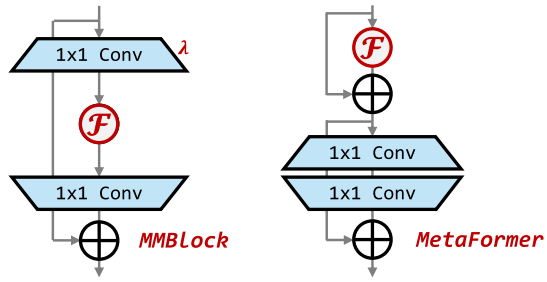

Fig. 3. Meta-paradigm comparison between our MMBlock and MetaFormer [52]. We integrate $\mathcal{F}$ into expended FFN to construct a more streamlined and shallower single-module block.

Finally, a shrinkage MLP_s with inverted input/output ratio equaling λ to shrink channel dimension:

$$\boldsymbol{X}_s = \text{MLP}_s(\boldsymbol{X}_f)\left(\in \mathbb{R}^{C\times H\times W}\right), \tag{3}$$

where a residual connection is used to get the final output $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{X}_s(\in \mathbb{R}^{C\times H\times W})$. For clarity, notice that normalization and activation functions are omitted. *Relation to MetaFormer:* We reveal the differences between our *Meta Mobile Block* and *MetaFormer* [52] in Fig. 3. ***1)*** From the structure, two-residual MetaFormer contains two sub-modules with two skip connections, while our Meta Mobile Block contains only one sub-module that covers one-residual IRB in the field of lightweight CNN. Also, shallower depths require less memory access and save costs [74] that is more general and hardware-friendly for optimization. ***2)*** From the motivation, MetaFormer is the induction of high-performance Transformer/MLP-like models, while our Meta Mobile Block is the induction of efficient IRB in MobileNetv2 [9] and effective MHSA/FFN in Transformer [18], [35] for designing lightweight infrastructure. ***3)*** Inductive one-residual Meta Mobile Block can be regarded as a conceptual extension of two-residual MetaFormer in the lightweight field. We hope our work inspires more future research dedicated to lightweight model design domain based on attention. ***4)*** From the result, our instantiated EMOv2-5 M (w/ 5.1 M #Params and 1.0 G FLOPs) exceeds instantiated PoolFormer-S12 (w/ 11.9 M #Params and 1.8 G FLOPs) by +2.1↑, illustrating that

TABLE IX
CLASSIFICATION PERFORMANCE COMPARISON AMONG DIFFERENT KINDS OF BACKBONES ON IMAGENET-1 K DATASET IN TERMS OF 5M-MAGNITUDE, AS WELL AS 1M-MAGNITUDE AND 2 M MODELS

	Model	#Params ↓	FLOPs ↓	Reso.	Top-1	Venue
1M mMagnitude	MNetv1-0.50 [8]	1.3	149	224^2	63.7	arXiv'1704
	MNetv3-L-0.50 [10]	2.6	69	224^2	68.8	ICCV'19
	MViTv1-XXS [17]	1.3	364	256^2	69.0	ICLR'22
	MViTv2-0.5 [14]	1.4	466	256^2	70.2	arXiv'22
	EdgeNeXt-XXS [2]	1.3	261	256^2	71.2	ECCVW'22
	EATFormer-Mobile [24]	1.8	360	224^2	69.4	IJCV'24
	☆ EMOv1-1M [13]	1.3	261	224^2	71.5	ICCV'23
	★ **EMOv2-1M**	1.4	285	224^2	**72.3**	-
	★ **EMOv2-1M**†	1.4	285	224^2	**73.5**	-
2M Magnitude	MNetv2-1.40 [9]	6.9	585	224^2	74.7	CVPR'18
	MNetv3-L-0.75 [10]	4.0	155	224^2	73.3	ICCV'19
	FasterNet-T0 [93]	3.9	340	224^2	71.9	CVPR'23
	GhostNetV3-0.5x [41]†, ‡	2.7	48	224^2	69.4	arXiv'2404
	MNetv4-Conv-S [42]*†	3.8	200	224^2	73.8	arXiv'2404
	MoCoViT-1.0 [94]	5.3	147	224^2	74.5	arXiv'22
	PVTv2-B0 [20]	3.7	572	224^2	70.5	CVM'22
	MViTv1-XS [17]	2.3	986	256^2	74.8	ICLR'22
	MFormer-96M [33]	4.6	96	224^2	72.8	CVPR'22
	EdgeNeXt-XS [2]	2.3	538	256^2	75.0	ECCVW'22
	EdgeViT-XXS [55]	4.1	557	256^2	74.4	ECCV'22
	tiny-MOAT-0 [75]	3.4	800	224^2	75.5	ICLR'23
	EfficientViT-M1 [95]	3.0	167	224^2	68.4	CVPR'23
	EfficientFormerV2-S0 [1]*†	3.5	400	224^2	75.7	ICCV'23
	EATFormer-Lite [24]	3.5	910	224^2	75.4	IJCV'24
	☆ EMOv1-2M [13]	2.3	439	224^2	75.1	ICCV'23
	★ **EMOv2-2M**	2.3	487	224^2	**75.8**	-
	★ **EMOv2-2M**†	2.3	487	224^2	**76.7**	-
5M Magnitude	MNetv3-L-1.25 [10]	7.5	356	224^2	76.6	ICCV'19
	EfficientNet-B0 [12]	5.3	399	224^2	77.1	ICML'19
	FasterNet-T2 [93]	15.0	1910	224^2	78.9	CVPR'23
	RepViT [40]‡	6.8	1100	224^2	78.6	CVPR'24
	RepViT [40]†, ‡	6.8	1100	224^2	80.0	CVPR'24
	GhostNetV3-1.3x [41]†, ‡	8.9	269	224^2	79.1	arXiv'2404
	MNetv4-Conv-M [42]*†	9.2	1000	224^2	79.9	arXiv'2404
	DeiT-Ti [43]	5.7	1258	224^2	72.2	ICML'21
	XCiT-T12 [57]	6.7	1254	224^2	77.1	NeurIPS'21
	LightViT-T [53]	9.4	700	224^2	78.7	arXiv'22
	MViTv1-S [17]	5.6	2009	256^2	78.4	ICLR'22
	MViTv2-1.0 [14]	4.9	1851	256^2	78.1	arXiv'22
	EdgeNeXt-S [2]	5.6	965	224^2	78.8	ECCVW'22
	PoolFormer-S12 [52]	11.9	1823	224^2	77.2	CVPR'22
	MFormer-294M [33]	11.4	294	224^2	77.9	CVPR'22
	MPViT-T [95]	5.8	1654	224^2	78.2	CVPR'22
	EdgeViT-XS [55]	6.7	1136	256^2	77.5	ECCV'22
	tiny-MOAT-1 [75]	5.1	1200	224^2	78.3	ICLR'23
	EfficientViT-M5 [94]	12.4	522	224^2	77.1	CVPR'23
	EfficientFormerV2-S1 [1]*†	6.1	650	224^2	79.0	ICCV'23
	ViG-T [58]	6.0	900	224^2	77.2	arXiv'2405
	SHViT-S3 [51]	14.2	601	224^2	77.4	CVPR'24
	EATFormer-Tiny [24]	6.1	1410	224^2	78.4	IJCV'24
	Vim-Ti [64]	7.0	1500	224^2	76.1	ICML'24
	EfficientVMamba-T [65]	6.0	800	224^2	76.5	arXiv'2403
	EfficientVMamba-S [65]	11.0	1300	224^2	78.7	arXiv'2403
	VRWKV-T [60]	6.2	1200	224^2	75.1	arXiv'2403
	MSVMamba-S [96]	7.0	900	224^2	77.3	arXiv'2405
	MambaOut-Femto [97]	7.0	1200	224^2	78.9	arXiv'2405
	☆ EMOv1-5M [13]	5.1	903	224^2	78.4	ICCV'23
	★ **EMOv2-5M**	5.1	1035	224^2	**79.4**	-
	★ **EMOv2-5M**†	5.1	1035	224^2	**80.9**	-
	★ **EMOv2-5M***	5.1	5627	512^2	**82.9**	-

White, grey, orange, and blue backgrounds indicate CNN-based, Transformerbased, RNN-based, and our EMO series, respectively. This kind of display continues for all subsequent experiments. Gray indicates the results obtained from the original paper. Comprehensive suggested models are marked in **bold**. Unit: #Params with (M) and FLOPs with (M). Abbreviations: MNet → MobileNet; MViT → Mobile- ViT; MFormer → MobileFormer. *: Neural Architecture Search (NAS) for elaborate structures. †: Using knowledge distillation. ‡: Re-parameterization strategy. *: Using stronger training strategy displayed in Tab. 19(e).

TABLE X
OBJECT DETECTION PERFORMANCE BY SSDLITE [10] ON MS-COCO 2017 [98] DATASET AT 320×320 RESOLUTION

Backbone	#Params ↓	FLOPs ↓	mAP
MNetv1 [8]	5.1	1.3G	22.2
MNetv2 [9]	4.3	0.8G	22.1
MNetv3 [10]	5.0	0.6G	22.0
MViTv1-XXS [17]	1.7	0.9G	19.9
MViTv2-0.5 [14]	2.0	0.9G	21.2
☆ EMOv1-1M [13]	2.3	0.6G	22.0
★ **EMOv2-1M**	2.4	0.7G	22.3
★ **EMOv2-1M**†	2.4	2.3G	26.6
MViTv2-0.75 [14]	3.6	1.8G	24.6
☆ EMOv1-2M [13]	3.3	0.9G	25.2
★ **EMOv2-2M**	3.3	1.2G	26.0
★ **EMOv2-2M**†	3.3	4.0G	30.7
ResNet50 [44]	26.6	8.8G	25.2
MViTv1-S [17]	5.7	3.4G	27.7
MViTv2-1.25 [14]	8.2	4.7G	27.8
EdgeNeXt-S [2]	6.2	2.1G	27.9
☆ EMOv1-5M [13]	6.0	1.8G	27.9
★ **EMOv2-5M**	6.0	2.4G	29.6
★ **EMOv2-5M**†	6.0	8.0G	34.8

Abbreviated MNet/MViT: MobileNet/MobileViT. †: 512 × 512 resolution.

TABLE XI
OBJECT DETECTION RESULTS BY RETINANET [36] ON MS-COCO 2017 [98] DATASET

Backbone	#Params	mAP^b	mAP^b_{50}	mAP^b_{75}	mAP^b_S	mAP^b_M	mAP^b_L
ResNet-50 [44]	37.7	36.3	55.3	38.6	19.3	40.0	48.8
PVTv1-Tiny [19]	23.0	36.7	56.9	38.9	22.6	38.8	50.0
PVTv2-B0 [20]	13.0	37.2	57.2	39.5	23.1	40.4	49.7
EdgeViT-XXS [55]	13.1	38.7	59.0	41.0	22.4	42.0	51.6
☆ EMOv1-5M	14.4	38.9	59.8	41.0	23.8	42.2	51.7
★ **EMOv2-5M**	14.4	41.5	62.7	44.1	25.7	45.5	55.5

TABLE XII
OBJECT DETECTION RESULTS BY MASK RCNN [99] ON MS-COCO 2017 [98] DATASET

Backbone	#Params ↓	mAP^b / mAP^m	mAP^b_{50} / mAP^m_{50}	mAP^b_{75} / mAP^m_{75}	mAP^b_S / mAP^m_S	mAP^b_M / mAP^m_M	mAP^b_L / mAP^m_L
PVT-Tiny [19]	33.0	36.7	59.2	39.3	-	-	-
		35.1	56.7	37.3	-	-	-
PVTv2-B0 [20]	23.0	38.2	60.5	40.7	-	-	-
		36.2	57.8	38.6	-	-	-
PoolFormer-S12 [52]	31.0	37.3	59.0	40.1	-	-	-
		34.6	55.8	36.9	-	-	-
MPViT-T [96]	28.0	42.2	64.2	45.8	-	-	-
		39.0	61.4	41.8	-	-	-
EATFormer-Tiny [24]	25.9	42.3	64.7	46.2	25.5	45.5	55.1
		39.0	61.5	42.0	22.4	42.0	52.7
☆ EMOv1-5M	24.8	39.3	61.7	42.4	23.5	42.3	51.1
		36.4	58.4	38.7	18.2	39.0	52.6
★ **EMOv2-5M**	24.8	42.3	64.3	46.3	25.8	45.6	56.3
		39.0	61.4	42.1	20.0	41.8	57.0

a stronger efficient operator makes a advantage. We further replace Token Mixer in MetaFormer with $\mathcal{F}$ in iRMB and build a 5.3M model. Compared with EMOv1-5M, it only achieves 77.5 Top-1 on ImageNet-1k that is -0.9↓ than our model, meaning that our proposed Meta Mobile Block has a better advantage for constructing lightweight models than two-residual MetaFormer.

c) Memory Access Cost (MAC) analysis: Within a single block, a two-residual Transformer-like block demands twice the number of MAC for intermediate activations compared to an single-residual MMBlock. Moreover, the unified expansion-shrinkage pathway in the MMBlock (as shown in (1), (2), and (3)) circumvents the gradient fragmentation issue that is commonly seen in MetaFormer [52] architectures. As a result, our models can achieve higher throughput. For example, EMOv2-2 M outperforms EdgeNeXt-XS [2] by 25% in GPU throughput

TABLE XIII
SEMANTIC SEGMENTATION RESULTS BY DEEPLABV3 [101], SEMANTIC FPN [102], SEGFORMER [103], AND PSPNET [104] ON ADE20K [105] DATASET AT 512×512 RESOLUTION

	Backbone	#Params ↓	FLOPs ↓	mIoU
DeepLabv3 [102]	MViTv2-0.5	6.3	26.1G	31.9
	MViTv3-0.5	6.3	-	33.5
	☆ EMOv1-1M	5.6	2.4G	33.5
	★ **EMOv2-1M**	5.6	3.3G	34.6
	MNetv2	18.7	75.4G	34.1
	MViTv2-0.75	9.6	40.0G	34.7
	MViTv3-0.75	9.7	-	36.4
	☆ EMOv1-2M	6.9	3.5G	35.3
	★ **EMOv2-2M**	6.6	5.0G	36.8
	MViTv2-1.0	13.4	56.4G	37.0
	MViTv3-1.0	13.6	-	39.1
	☆ EMOv1-5M	10.3	5.8G	37.8
	★ **EMOv2-5M**	9.9	9.1G	39.8
Semantic FPN [103]	ResNet-18	15.5	32.2G	32.9
	☆ EMOv1-1M	5.2	22.5G	34.2
	★ **EMOv2-1M**	5.3	23.4G	37.1
	ResNet-50	28.5	45.6G	36.7
	PVTv1-Tiny	17.0	33.2G	35.7
	PVTv2-B0	7.6	25.0G	37.2
	☆ EMOv1-2M	6.2	23.5G	37.3
	★ **EMOv2-2M**	6.2	25.1G	39.9
	ResNet-101	47.5	65.1G	38.8
	ResNeXt-101	47.1	64.7G	39.7
	PVTv1-Small	28.2	44.5G	39.8
	EdgeViT-XXS	7.9	24.4G	39.7
	EdgeViT-XS	10.6	27.7G	41.4
	PVTv2-B1	17.8	34.2G	42.5
	☆ EMOv1-5M	8.9	25.8G	40.4
	★ **EMOv2-5M**	8.9	29.1G	42.3
SegFormer [104]	MiT-B0	3.8	8.4G	37.4
	★ **EMOv2-2M**	2.6	10.3G	40.2
	MiT-B1	13.7	15.9G	42.2
	★ **EMOv2-5M**	5.3	14.4G	43.0
PSPNet [105]	MNetv2	13.7	53.1G	29.7
	MViTv2-0.5	3.6	15.4G	31.8
	☆ EMOv1-1M	4.3	2.1G	33.2
	★ **EMOv2-1M**	4.2	2.9G	33.6
	MViTv2-0.75	6.2	26.6G	35.2
	☆ EMOv1-2M	5.5	3.1G	34.5
	★ **EMOv2-2M**	5.2	4.6G	35.7
	MViTv2-1.0	9.4	40.3G	36.5
	☆ EMOv1-5M	8.5	5.3G	38.2
	★ **EMOv2-5M**	8.1	8.6G	39.1

TABLE XIV
SEMANTIC SEGMENTATION RESULTS BY UNET [107] ON HRF [108] DATASET AT 256×256 RESOLUTION

Backbone	#Params ↓	FLOPs ↓	mDice	aAcc	mAcc
UNet-S5-D16	29.0	204G	88.9	97.0	86.2
EdgeNeXt-S [2]	23.7	221G	89.1	97.1	87.5
★ **U-EMOv2-5M**	21.3	228G	89.5	97.1	88.3

TABLE XV
COMPARISON WITH THE STATE-OF-THE-ART ON KINETICS-400 [109] DATASET WITH FOUR INPUT FRAMES

Backbone	#Params ↓	FLOPs ↓	Top-1
UniFormer-XXS	9.8	1.0G	63.2
EdgeNeXt-S [2]	6.8	1.2G	64.3
★ **V-EMOv2-5M**	5.9	1.3G	65.2

and is nearly 4× faster on an iPhone 15. Simultaneously, it also shows a +0.8 increase in Top-1 accuracy (refer to Table XVIII).

2) Micro Designs for Deducted iRMB: Based on the inductive Meta Mobile Block, we instantiate an effective modern *Inverted Residual Mobile Block* (iRMB) for lightweight architecture design from a microscopic view in Fig. 4.

TABLE XVI
COMPARISON WITH DIT [67] FOR 400K TRAINING STEPS IN GENERATING 256×256 IMAGENET [79] IMAGES

Model	#Params ↓	FLOPs ↓	FID
DiT-S-2	33.0	5.5G	68.4
SiT-S-2	33.0	5.5G	57.6
D-EMOv2-S-2	24.6	5.4G	46.3
DiT-B-2	130.5	21.8G	43.5
SiT-B-2	130.5	21.8G	33.5
D-EMOv2-B-2	96.1	19.9G	24.8
DiT-L-2	458.1	77.5G	23.3
SiT-L-2	458.1	77.5G	18.8
D-EMOv2-L-2	334.8	69.3G	11.2
DiT-XL-2	675.1	114.5G	19.5
SiT-XL-2	675.1	114.5G	17.2
D-EMOv2-XL-2	492.7	101.5G	9.6

TABLE XVII
EFFICIENCY AND PERFORMANCE COMPARISON OF DIFFERENT DEPTH AND CHANNEL CONFIGURATIONS

Depth	Channels	#Params	FLOPs	Top-1
[2, 2, 10, 3]	[48, 72, 160, 288]	5.3M	1038M	79.1
[2, 2, 12, 2]	[48, 72, 160, 288]	5.0M	1127M	78.9
[4, 4, 8, 3]	[48, 72, 160, 288]	5.1M	1132M	79.4
[3, 3, 9, 3]	[48, 72, 160, 288]	5.1M	1035M	79.4
[2, 2, 12, 3]	[48, 72, 160, 288]	5.1M	1136M	79.1
[2, 2, 8, 2]	[48, 72, 224, 288]	5.1M	1117M	79.0

TABLE XVIII
COMPARISONS OF THROUGHPUT ON CPU/GPU AND RUNNING SPEED ON MOBILE IPHONE15 (MS)

Method	#Params ↓	FLOPs	CPU	GPU	iPhone15	Top-1
EdgeNeXt-XXS	1.3M	261M	73.1	2860.6	10.2	71.2
☆ **EMOv1-1M**	1.3M	261M	158.4	3414.6	3.0	71.5
★ **EMOv2-1M**	1.4M	285M	147.1	3182.2	3.6	72.3
EdgeNeXt-XS	2.3M	538M	69.1	1855.2	17.6	75.0
☆ **EMOv1-2M**	2.3M	439M	126.6	2509.8	3.7	75.1
★ **EMOv2-2M**	2.3M	487M	118.2	2312.4	4.3	75.8
EdgeNeXt-S	5.6M	965M	54.2	1622.5	22.5	78.8
☆ **EMOv1-5M**	5.1M	903M	106.5	1731.7	4.9	78.4
★ **EMOv2-5M**	5.1M	1035M	93.9	1607.8	5.9	79.4

a) Design principle: Following criteria in Section III-A, $\mathcal{F}$ in iRMB is modeled as cascaded *MHSA* and *Convolution* operations, formulated as $\mathcal{F}(\cdot) = \text{Conv}(\text{MHSA}(\cdot))$. This design absorbs CNN-like efficiency to model local features and Transformer-like dynamic modeling capability to learn long-distance interactions. However, naive implementation can lead to unaffordable expenses for two main reasons:

1) λ is generally greater than one that the intermediate dimension would be multiple to input dimension, causing quadratic λ increasing of parameters and computations. Therefore, components of $\mathcal{F}$ should be independent or linearly dependent on the number of channels.

2) FLOPs of MHSA is proportional to the quadratic of total image pixels, so the cost of a naive Transformer is unaffordable for downstream application. The specific influences can be seen in Table III.

TABLE XIX
ABLATION STUDIES AND COMPARISON ANALYSIS ON IMAGENET [79]

(a) Attention mode analysis on classification and downstream RetinaNet [36] / DeepLabv3 [102].

Mode	#Params ↓	FLOPs ↓	Top-1	mAP	mIoU
None	4.3M	802M	77.9	39.3	37.2
None (Scaling to 5.1M)	5.1M	991M	78.4	39.6	37.7
Neighborhood Attention	5.1M	967M	78.8	40.4	39.0
Remote Attention	5.1M	967M	79.0	39.9	38.6
Spanning Attention	5.1M	1035M	79.4	41.5	39.8

(b) Applied stages of spanning attention.

Stage	#Params ↓	FLOPs ↓	Top-1
S-4	4.7M	832M	78.5
S-34	5.1M	1035M	79.4
S-234	5.1M	1096M	79.3
S-1234	5.2M	1213M	79.1

(c) Influence of DPR and BS hyperparameters.

DPR	Top-1	BS	Top-1
0.00	79.1	256	78.9
0.03	79.2	512	79.2
0.05	79.4	1024	79.4
0.10	79.3	2048	79.4
0.20	79.1	4096	79.4

(d) Convolution type. K: kernel size. D: Dilation.

Size	#Params ↓	FLOPs ↓	Top-1
K-1	4.8M	969M	78.6
K-3	4.9M	991M	79.0
K-5	5.1M	1035M	79.4
K-7	5.3M	1102M	79.2
K-9	5.5M	1184M	79.3
K-5 + D-2	5.1M	1035M	79.3
K-5 + D-3	5.1M	1035M	79.1
K-5 + DCNv2 [113]	6.7M	1625M	78.5

(e) Training strategies: image resolution, knowledge distillation, and 1000 training epochs.

Resolution	KD	Long Training	#Params.	FLOPs	Top-1
224	✘	✘	1.0G	5.1M	79.4
256	✘	✘	1.4G	5.1M	79.9
224	✔	✘	1.0G	5.1M	80.8
224	✘	✔	1.0G	5.1M	80.4
512	✘	✘	5.6G	5.1M	81.5
512	✔	✘	5.6G	5.1M	82.4
512	✔	✔	5.6G	5.1M	82.9

All the experiments use EMOv2-5M as default structure.

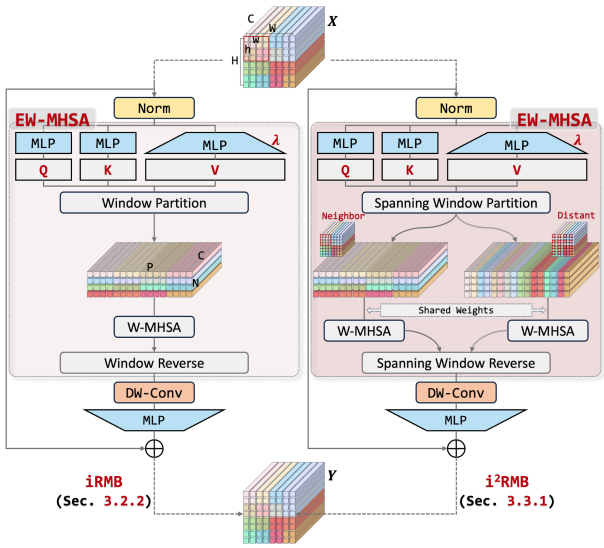

Fig. 4. Detailed implementation comparison of the Inverted Residual Mobile Block (*iRMB* in Section III-B2) and the improved version (*i²RMB* in Section III-C1). i²RMB designs a parameter-sharing spanning window attention mechanism that simultaneously models the interaction of distant and close window information.

b) Expanded Window MHSA: Parameters and FLOPs for obtaining Q, K in Window MHSA (W-MHSA) [21] is quadratic of the channel. Given the input $\boldsymbol{X}$ ($\in \mathbb{R}^{C\times H\times W}$), we obtain channel-unexpanded Q and K ($\in \mathbb{R}^{C\times H\times W}$) to compute the attention matrix M more efficiently, while the expanded V ($\in \mathbb{R}^{\lambda}C \times H \times W$) is used to capture finer-grained visual features. The essence of this expanding mechanism is that M models only the spatial positional relationships and is independent of the number of channels in V. This improvement is termed EW-MHSA, which is more applicable. Specifically, Window Partition operation flattens each feature map $F \in \{Q, K, V\}$ into N non-overlapping patches with each sequence length $P{=}w \times h$, where $N{=}H \times W/P$. The corresponding dimensional transformation can be described by the following formula: $[B, C, H, W] \rightarrow [BHW/P, C, P]$, and vice versa for the Window Reverse operation. To put it more directly, $w{=}4$, $h{=}4$, $P{=}16$, and $N{=}4$ for example in Fig. 4.

c) Structural deduction: Combining lightweight Depth-Wise Convolution (DW-Conv) and efficient EW-MHSA to trade-off model cost and accuracy, the process of the designed iRMB can be formulated as follows:

$$\mathcal{F}(\cdot) = \text{DW-Conv}(\text{EW-MHSA}(\cdot)). \tag{4}$$

This cascading manner can increase the expansion speed of the receptive field and reduce the maximum path length of the model to $O(2W/(k-1+2w))$, which has been experimentally verified with consistency in Section IV-C.

d) Flexibility: Empirically, current transformer-based methods [1], [2], [49], [50], [75] reach a consensus that inductive CNN in shallow layers while global Transformer in deep layers composition could benefit the performance. Unlike recent EdgeNeXt that employs different blocks for different depths, our iRMB satisfies the above design principle using only two switches to control whether two modules are used (Code level is also concise in #Supp). Therefore, we can easily implement the use of EW-MHSA for more semantic modeling only in the deeper layers, i.e., stage-3 and stage-4.

f) Efficient equivalent implementation: MHSA is typically employed in channel-consistent projection ($\lambda{=}1$), indicating that the FLOPs of multiplying the attention matrix by the expanded $\boldsymbol{X}_e$ ($\lambda{>}1$) will increase by a factor of λ - 1. Fortunately, the information flow from $\boldsymbol{X}$ to the expanded V ($\boldsymbol{X}_e$) involves only linear operations, allowing us to derive an equivalent proposition: "*When the number of groups in MLP$_e$ equals the number of heads in EW-MHSA, the result of the multiplication remains unchanged when the order is exchanged.*" To reduce FLOPs, matrix multiplication before MLP$_e$ is used by default, referred to as pre-attention.

g) Boosting naive transformer: To assess iRMB performance, we set λ to 4 and replace standard Transformer structure in columnar DeiT [43] and pyramidal PVT [19]. As shown in Table IV, we surprisingly found that iRMB can improve performance with fewer parameters and computations in the same training setting, especially for the columnar ViT. And the newly proposed i^{2}RMB further boosts the performance significantly. This proves that the one-residual iRMB/i^{2}RMB has obvious advantages over the two-residual Transformer in the lightweight model.

h) Parallel design of $\mathcal{F}$: We also implement the parallel structure of DW-Conv and EW-MHSA with half the number of channels in each component, and some configuration details are adaptively modified to ensure the same magnitude. Comparably, this parallel model gets 78.1 (-0.3↓) Top-1in ImageNet-1k dataset with 5.1M parameters and 964M FLOPs (+63M↑ than EMOv1-5M), but its throughput will slow down by about -7%↓.

Manner	#Params.	FLOPs	Top1	Throughput
Parallel	5.1M	964M	78.1	1618.4
Cascaded (Ours)	5.1M	903M	78.3	1731.7

This phenomenon is also discussed in the work [74] that: "Network fragmentation reduces the degree of parallelism".

C. Parameter-Efficient Extension (EMOv2)

Even though EMOv1 achieves satisfactory results, it only models the interaction of neighbor information within a local window. We further explore the performance frontier of lightweight models based on this module with a negligible increase in model parameters. Specifically, we leverage the principles of attention computation to reuse the neighbor window attention map for uniform sampling over a global window size, resulting in a novel spanning module termed SEW-MHSA. This mechanism simultaneously models both neighbor and distant features without increasing the number of parameters. Additionally, we elaborately improve structural details to further enhance the model's performance.

1) Improved Inverted Residual Mobile Block (i^{2}RMB): To avoid a significant increase in the number of parameters, we optimize the EW-MHSA and DW-Conv modules to construct a more powerful i^{2}RMB module in Fig. 4.

a) Spanning attention for EW-MHSA: This paper explores the potential of lightweight models under limited parameters, i.e., mainly 5 M for most mobile scenarios. We observe that in EW-MHSA, the attention map only computes feature interactions within windows. While this alleviates the computational explosion of global attention, it inevitably reduces the flow of the receptive field. Therefore, we extend the computation of the attention map to a parallel fusion of neighbor and distant window attention, introducing *Spanning Window Partition and Reverse* steps to achieve this goal. Compared to the naive Window Partition described in Section III-B2, this operation involves two parallel window partitions that separately segment the shared Q, K, and V into neighbor and distant partitions. In the former, each window contains only adjacent features. In the latter, feature selection within the window is performed based on a stride of $[H/h, W/w]$. This allows for feature interaction at different distances simultaneously, and its transformation can be described by the following formula: $[B, C, H, W] \rightarrow \{[BHW/P, C, P]_{neibor}, [BHW/P, C, P]_{distant}\}$. Followed by two parameter-shared MHSA, this powerful improvement is termed SEW-MHSA. The computation of Q and K remains in the non-extended dimension, following iRMB. This approach has two benefits: 1) A single module can accommodate global information in one forward pass, which is advantageous for downstream tasks requiring high resolution. 2) The parallel operation does not introduce additional parameters, reusing the parameters and computations of K, Q, and V, and only adds an extra attention map computation, thereby enhancing model accuracy with minimal computational cost.

b) Non-linearity for post-attention: We introduce a nonlinear activation function in the V computation of the attention mechanism, further filtering features before multiplying them with the attention map. This differs from the pre-attention described in Section III-B2, referred to as post-attention, which improves model performance without increasing the number of parameters.

c) Large kernel for local modeling: iRMB uses a kernel size of 3 for the DW-Conv in local modeling. Smaller values limit the model's receptive field. i^{2}RMB further investigates the impact of large kernels on accuracy. Considering the depth-wise modeling approach, this does not significantly increase the number of model parameters. Additionally, this structure provides the model with positional information, allowing it to achieve downstream structures without additional position embedding design.

d) Structural deduction: Combining lightweight Depth-Wise Convolution (DW-Conv) and efficient EW-MHSA to trade-off model cost and accuracy, the process of the designed iRMB can be formulated:

$$\mathcal{F}(\cdot) = \text{DW-Conv}(\text{SEW-MHSA}(\cdot)). \tag{5}$$

e) Accessibility analysis: Due to the fact that i^{2}RMB only includes convolution and multi-head self-attention operators, the constructed EMOv2 is built by stacking identical standard modules without employing hardware-aware search structures, and it uses a serial structure without multiple branches. This design is highly compatible with hardware acceleration, potentially offering strong generalizability for different hardware platforms and applications.

f) Parameter-shared spanning attention efficiency: As shown in Fig. 1, EW-MHSA only performs spatial information interaction in local regions within a single block cycle. Stacking multiple blocks gradually expands information to the global scope through DW-Conv, which is inefficient. Our improved SEW-MHSA achieves local and global information interaction within one block cycle using minimal FLOPs without increasing the number of parameters, with a complexity of

$O(NP^2) = O(HWP)$ that does not exhibit quadratic complexity $O((HW)^2)$ as the spatial resolution increases. $P{=}w \times h$ is sequence length and $N{=}H \times W/P$.

2) Macro Design of EMOv2 for Dense Prediction: Based on the above criteria, we design a ResNet-like 4-phase Efficient MOdel (EMO) based on a series of iRMBs for dense applications in our previous work [13]. In this extension work, we build a stronger vision backbone EMOv2 by the powerful i^{2}RMBs, as shown in Fig. 2-*Right*.

i) For the overall framework, EMOv2 consists of only i^{2}RMB without diversified modules②, which is a departure from recent efficient methods [2], [17] in terms of designing idea.

ii) For the specific module, i^{2}RMB consists of only convolution and multi-head self-attention without other complex operators ①. Also, benefitted by DW-Conv, i^2 RMB can adapt to down-sampling operation through the stride and does not require any position embeddings for introducing inductive bias to MHSA②. *The comparison of the requirements for position embedding across different methods is shown in Table VII.*

iii) For the configuration of different-scale models, we employ gradually increasing expansion rates and channel numbers, and detailed configurations are shown in Table V. Results for basic classification and downstream dense prediction tasks in Section IV demonstrate the superiority of our i^{2}RMB over SoTA lightweight methods on magnitudes of 1 M, 2 M, and core-focused 5 M③.

iv) i^{2}RMB can be easily extended to other foundational architectures and accomplish corresponding tasks④, such as temporal extension, UNet variant, and DiT-like model in Section III-C3.

a) Configuration details: Since MHSA is better suited for modeling semantic features for deeper layers, we only turn it on at stage-3/4 following previous works [2], [49], [75]. Note that this never violates the uniformity criterion, as the shutdown of MHSA was a special case of i^{2}RMB structure. To further increase the stability of EMO, BN [76]+SiLU [77] are bound to DW-Conv while LN [78]+GeLU [77] are bound to SEW-MHSA, and i^{2}RMB is competent for down-sampling operations.

b) Importance of instantiated efficient operator: Our defined *efficient operator* $\mathcal{F}$ contains two core modules, i.e., (S)EW-MHSA and DW-Conv. In Table VI, we conduct an ablation experiment to study the effect of both modules in iRMB/i^{2}RMB. The first row means that neither (S)EW-MHSA nor DW-Conv is used, i.e., the model is almost composed of MLP layers with several DW-Conv for down-sampling, and $\mathcal{F}$ degenerates to Identity operation. Surprisingly, this model still produces a respectable result, i.e., 73.5 Top-1. Comparatively, results of the second and third rows demonstrate that each component contributes to the performance, e.g., +3.1↑ and +4.1↑ when adding DW-Conv and EW-MHSA for EMO, respectively, while +4.2↑ and +4.6↑ for EMOv2. Our approach achieves the best result when both components are used. Besides, this experiment illustrates that the specific instantiation of iRMB/i^{2}RMB is very important to model performance.

c) Order of operators: Based on EMOv1-5 M, we switch the order of DW-Conv/EW-MHSA and find a slight -0.6↓, and a similar -0.7↓ drop is also observed in EMOv2 when switching DW-Conv/SEW-MHSA. Therefore, (S)EW-MHSA performs first by default.

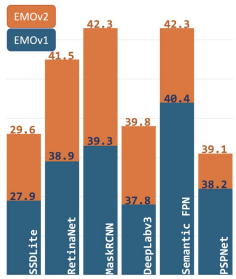

Fig. 5. Downstream gains of EMOv2-5 M over EMOv1-5 M.

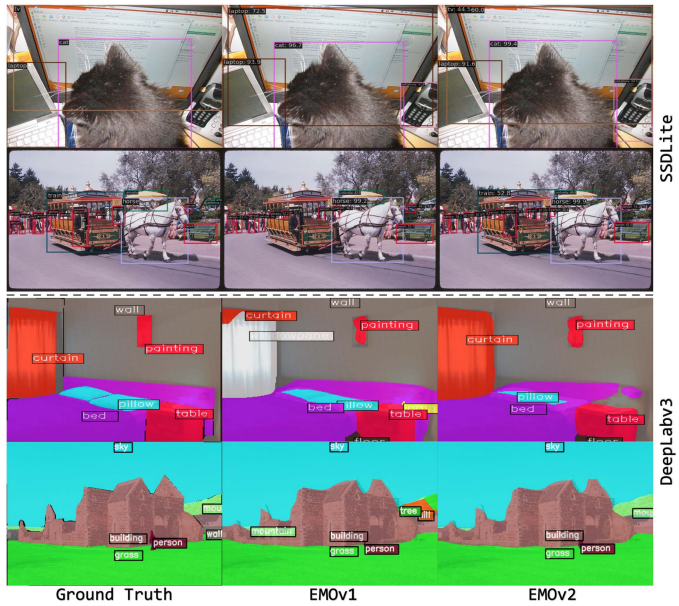

Fig. 6. Qualitative comparisons between EMOv1/v2 on downstream SSDLite [10] and DeepLabv3 [101]. EMOv2 demonstrates higher accuracy in class and boundary detection. Zoom in for more details.

d) Performance gains over EMOv1: The improved EMOv2-5 M achieves a Top-1 accuracy of 79.4, surpassing EMOv1-5 M by +1.0↑, without significantly increasing parameters and FLOPs.

Additionally, it demonstrates notable improvements across various high-resolution downstream tasks. For instance, in popular detection and segmentation tasks, as shown in Fig. 5, EMOv2 consistently achieves an enhancement of 1∼3 points across different frameworks.

e) Effective receptive field: Benefiting from parallel neighbor and distant modeling, our EMOv2 has a larger Effective Receptive Field (ERF) (see Fig. 1), which is further confirmed by the qualitative Grad-CAM subject attention (see Fig. 7).

3) i^{2}RMB-Centric Omni-Task Transformation: Thanks to the general, neat, and powerful i^{2}RMB design, we can easily extend it to various tasks in this extension work, as illustrated

Fig. 7. Visualizations by Grad-CAM. EMOv2 generates sharper and higher confidence attention maps than EMOv1.

in Fig. 2: *1)* video classification (V-EMO) extends the i^{2}RMB to the temporal dimension, *2)* UNet-based image segmentation (U-EMO) replaces the original convolutional blocks with i^{2}RMB, and *3)* diffusion-based image generation (D-EMO) replaces naive Transformer blocks with i^{2}RMB. We construct various lightweight versions of different types of structures and conduct extensive experiments to demonstrate the effectiveness and generalizability of i^{2}RMB in Section IV-B.

IV. Experimental Results

A. Image Classification

Setups: Different SoTA methods use various training recipes that could lead to potentially unfair comparisons, and we have summarized and compared these training strategies in Table VII. In contrast, our training strategy is weaker, yet it achieves impressive results without employing strong training tricks. All experiments are conducted on the ImageNet-1 K dataset [79] without using additional datasets or pre-trained models. Each model is trained for a standard 300 epochs from scratch at a resolution of 224×224 by default. AdamW [80] optimizer is employed with betas (0.9, 0.999), a weight decay of 5e^{-2}, a learning rate of 6e^{-3}, and a batch size of 2,048. We use a Cosine scheduler [81] with 20 warmup epochs, Label Smoothing 0.1 [38], stochastic depth [82], and RandAugment [83] during training. However, LayerScale [84], Dropout [85], MixUp [86], CutMix [87], Random Erasing [88], Position Embeddings [18], Token Labeling [89], and Multi-Scale training [17] are *disabled*. EMOv2 is implemented based on TIMM [90]. For comparative methods, we default to reporting the results from the original papers. For methods using strong training tricks, we apply symbolic markers (see Table IX) to avoid potential unfair comparisons. Additionally, to further conduct fair comparisons with these methods using strong tricks and simultaneously explore the upper performance limit of our EMOv2 model, we also report the results obtained with strong tricks.

Results analysis: We evaluate our method against SoTA models on three small magnitudes, and the quantitative results are presented in Table IX. Notably, our method achieves the best results without utilizing complex modules and strong training recipes employed by recent works, such as NAS in MobileNetv4 [42] and re-parameterization in RepViT [40]. For example, the smallest EMOv2-1 M achieves a SoTA Top-1 accuracy of 72.3, surpassing the CNN-based MobileNetv3-L-0.50 [10] by +3.5↑ with nearly half the parameters, and the Transformer-based MobileViTv2-0.5 [14] by +2.1↑ with only 61% of the FLOPs. The larger EMOv2-2 M achieves a SoTA Top-1 accuracy of 75.8 with only 487 M FLOPs, nearly half of MobileVit-XS [17] but with a +1.0↑ improvement. Comparatively, the latest EdgeViT-XXS [55] achieves a lower Top-1 accuracy of 74.4 while requiring +78%↑ more parameters and +14%↑ more FLOPs, whereas tiny-MOAT-0 [75] requires +48%↑ more parameters and +64%↑ more FLOPs to achieve a similar result. Consistently, EMOv2-5 M demonstrates a superior trade-off between #Params. (5.1 M), FLOPs (1.0 G), and accuracy (79.4), proving to be more efficient than contemporary counterparts. For example, it achieves +0.9↑ over EATFormer-Tiny [24] with better efficiency. When we further employ the KD training strategy (TResNet [91] with 83.9 accuracy as the teacher model), our three-magnitude EMOv2 models achieve 73.5, 76.7, and 80.9 Top-1 accuracy, respectively. This represents an increase of +2.0↑, +1.6↑, and +2.5↑ compared to our previous conference method [13]. Moreover, these results significantly exceed the latest models using strong training strategies, such as RepViT [40], EfficientFormerV2 [1], GhostNetV3 [41], and MobileNetv4 [42].

Training recipes matters: We evaluate EMO [13] and EMOv2 with different mainstream training recipes presented in Table VIII. We find that our simple training recipe is enough to get impressive results, while existing stronger recipes (especially used by EdgeNeXt [2]) will not improve performance further. *NaN* indicates that the model did not train well for the possibly unadapted hyper-parameters.

B. Downstream Applications

Thanks to the structural design of *spanning attention* in i^{2}RMB, our EMOv2 can simultaneously model global and local information interactions, which significantly enhances the performance of downstream tasks. It is noteworthy that current lightweight models have only reported limited results on downstream tasks, and different methods lack a unified experimental standard. Therefore, we have endeavored to find overlapping results from the original papers for a fair comparison. Additionally, we report the detailed results of our method with different magnitudes on multiple downstream tasks in the supplementary materials.

Object detection: We evaluate our EMOv2 (pre-trained on ImageNet-1 K) with other SoTA methods on MS-COCO 2017 [98] dataset, using the lightweight SSDLite [10] and heavy RetinaNet [36] / Mask RCNN [99]. Considering fairness and friendliness for the community, we employ standard MMDetection library [100] for experiments and replace the optimizer with AdamW [80] without tuning other parameters.

Comparison results on SSDLite are shown in Table X, and our EMOv1surpasses corresponding counterparts by apparent

advantages and the improved EMOv2 further boosts the performance. For example, SSDLite equipped with EMOv1-1 M achieves 22.0 mAP with only 0.6 G FLOPs and 2.3 M parameters, which boosts +2.1↑ compared with SoTA MobileViT [17] with only 66% FLOPs. Consistently, EMOv1-5M obtains the highest 27.9 mAP so far with much fewer FLOPs, e.g., 53% (1.8G) of MobileViT-S [17] (3.4G) and 0.3G less than EdgeNeXt-S (2.1G). EMOv2-5M further achieves 29.6 mAP with no significant increase in parameters, surpassing EMOv1-5M by +1.7↑. We also conduct experiments on heavy detection frameworks. Tables XI and XII present the results of different lightweight backbones on the RetinaNet [36] and Mask RCNN [99] methods, respectively. Our EMOv2 consistently achieves superior results compared to its counterparts, e.g., +5.2↑ mAP over the CNN-based ResNet-50, +2.8↑ mAP over the Transformer-based EdgeViT-XXS, and +2.6↑ mAP over our previous EMOv1under the RetinaNet framework. For the Mask RCNN framework, our EMOv2-5M obtains highly competitive results compared to the recently designed EATFormer for heavy architectures, with improvements of +3.0↑ mAP^b and +2.6↑ mAP^m over the previous generation EMOv1-5M model.

Semantic segmentation: ImageNet-1 K pre-trained EMOv2 is integrated with DeepLabv3 [101], Semantic FPN [102], SegFormer [103], and PSPNet [104] to adequately evaluate its performance on challenging ADE20K [105] dataset at 512×512 resolution. We employ the standard MMSegmentation library [106] with official configurations without tuning other parameters.

Due to the fact that different methods only report results on certain segmentation frameworks, we strive to find sufficient comparable models of similar magnitude under each method. Detailed results are presented in Table XIII. For lightweight models at the 1 M/2M/5 M magnitude, our method demonstrates significant advantages over comparative methods (including CNN, Transformer, and hybrid architectures), achieving a balance between parameters, computational cost, and performance. Notably, our conference version model (i.e., EMO [13]) achieves highly competitive results, and the improved EMOv2 model further significantly enhances the metrics. For instance, under the Deeplabv3 framework, our EMOv2-1 M/2M/5 M achieved 34.6/36.8/39.8 mIoU, respectively, representing improvements of +1.1↑/+1.5↑/+2.0↑ over EMOv1with fewer parameters. Similarly, under the Semantic FPN framework, our EMOv2-1M/2M/5M achieves 37.1/39.9/42.3 mIoU, respectively, representing improvements of +2.9↑/+2.6↑/+1.9↑ over EMOv1without increasing the number of parameters. More detailed results can be found in the supplementary materials.

Previous studies have demonstrated the effectiveness of EMOv2 in classification and mainstream downstream detection/segmentation tasks. To further validate the superiority of EMOv2, we additionally extend it to UNet-like architectures, as well as video classification and DiT-based image generation.

UNet-based vision segmentation (U-EMO): Furthermore, we replace the basic convolutional block in UNet with the i^{2}RMB block to construct a more powerful U-EMO architecture, as described in Fig. 2, and we conduct experiments on the downstream segmentation task to demonstrate the generalizability of the proposed method across different architectures. Table XIV presents results of U-EMO, UNet [107], and the adapted EdgeNeXt [2] method on the HRF [108] dataset at 256×256 resolution. Our improved U-EMO achieves higher performance with fewer parameters without meticulous adjustments to the architecture and training recipes.

Video classification (V-EMO): By simply extending the temporal dimension of the convolution and spanning attention in the i^{2}RMB block, we obtain a basic i^{2}RMB-3D block for video processing. This allows us to replace modules while maintaining a structure similar to 2D EMOv2, resulting in the V-EMO model. We use ImageNet-1 K pretrained weights with temporal repetition to initialize the video classification model. Table XV presents a comparison of our method with UniFormer-XXS [49] and the adapted EdgeNeXt [2] method on the Kinetics-400 [109] dataset. Our V-EMO-5 M achieves a Top-1 accuracy of 65.2 with only 5.9 M parameters, outperforming UniFormer-XXS, which has 9.8 M parameters, by +2.0↑.

DiT-based image generation (D-EMO): The primary design goal of the i^{2}RMB is to simplify the Transformer block structure, making it suitable for mobile architecture design by reducing the depth of individual blocks while improving the modeling of both distant and neighboring features. Thanks to its plug-and-play characteristic, i^{2}RMB can easily replace the Transformer block in the DiT model for image generation tasks. Specifically, we fully adhere to the DiT [67] training framework, and the results on the 256×256 ImageNet generation task are shown in Table XVI. Compared to the baseline DiT [67] and the SiT [110] with improved training strategies, our D-EMO model, which replaces the basic Transformer block with i^{2}RMB, requires fewer parameters and computational resources while achieving significantly better FID scores. This demonstrates the advantage of spanning attention in downstream image generation task.

C. Structural Ablation and Analysis

This section uses EMOv2-5 M as the research backbone to ablate the proposed method modules and training hyperparameters, while also analyzing the model structure and results.

Depth and channel configurations: Using EMOv2-5 M as the baseline, we evaluate the impact of different depth configurations on model performance, as shown in the upper part of Table XVII. The selected depth configuration yields a relatively better performance. Furthermore, we assess the performance of slimmer and wider models with a similar number of parameters, as shown in the lower part of Table XVII. These models, despite having an increased computational load, do not result in further performance improvements, demonstrating the rationality of the current structural configuration.

Throughput comparison: Table XVIII presents throughput evaluation results compared with the state-of-the-art EdgeNeXt [2], which effectively balances parameters, computational load, and performance. The test platforms are an AMD EPYC 7K62 CPU and a V100 GPU, with a resolution of 224× 224 and a batch size of 256. Results indicate that EMOv1achieves faster speeds on both platforms with higher

Top-1 accuracy. For instance, EMOv1-1M achieves speed boosts of +20%↑ on the GPU and +116%↑ on the CPU compared to EdgeNeXt-XXS with the same FLOPs. The improved EMOv2 maintains nearly the same parameter count as EMOv1but significantly enhances performance with a slight increase in computational load. This performance gap is further widened on mobile devices (following the official classification project [111] on iPhone15), where our EMOv2 is 2.8× ↑, 4.1× ↑, and 3.9× ↑ faster than the state-of-the-art EdgeNeXt [2]. This improvement is attributed to our simple and device-friendly i^{2}RMB block, which does not rely on other complex structures such as the Res2Net module [56], transposed channel attention [57], etc.

Attention mode: The proposed i^{2}RMB in Section III-C1 includes two components: distant and neighbor window attention with shared parameters. Table XIX(a) evaluates the model's performance under different attention modes. When neighborhood and distant attention are added separately, the model shows significant improvement compared to the baseline model. It also outperforms models of similar magnitude without attention, especially in downstream task metrics, demonstrating the effectiveness of the proposed basic EW-MHSA (Section III-B2). Thanks to the shared parameter design, the model with integrated spanning attention achieves better Top-1 classification results without any additional parameters. This is particularly evident in detection and segmentation tasks, further proving the effectiveness of the spanning mechanism in i^{2}RMB.

Used stages of spanning attention: Table XIX(b) shows the changes in model accuracy when applying spanning attention to different stages based on EMOv2-5 M. As spanning attention is gradually added from the fourth stage (S-4) to all four stages (S-1234), the model's performance significantly increases (S-34) and then saturates and slightly decreases (S-234). Considering that more stages require additional parameters and computational resources, spanning attention is by default injected only in the last two stages. Interestingly, in the conference version of EMO [13], the accuracy of the model increases with the number of stages to which spanning attention is applied. This discrepancy may be due to the structure of i^{2}RMB, where EMOv2-5 M is closer to the performance upper limit for models with this parameter count.

Effect of training hyper-parameters: Table XIX(c) discusses the two most influential hyperparameters in model training. The proposed EMOv2-5 M exhibits strong robustness to the drop path rate (DPR) hyperparameter within the range of [0, 0.2], where the Top-1 accuracy fluctuates within 0.3, achieving the best result at a drop path rate of 0.05. Meanwhile, a smaller batch size (BS) of 256 slightly affects the model's performance, with the performance peaking at a batch size of 1024 and then stabilizing. Considering memory efficiency, a default batch size of 1024 is suggested. These ablation experiments demonstrate the robustness of EMOv2 to the above hyperparameter variations.

Neighborhood kernel size in i^{2}RMB: The size of the DW-Conv affects the local receptive field of i^{2}RMB, which significantly impacts the model's classification ability and perception capability in downstream tasks. As shown in Table XIX(d)-Top, when the kernel size gradually increases from 1 to 5, the model's performance improves from 78.6 to 79.4. However, further increases in kernel size do not yield noticeable gains and instead incur additional parameter and computational costs.

Convolution type in i^{2}RMB: Table XIX(d)-Bottom illustrates the impact of different convolution variants on EMOv2, which extend the receptive field. The use of dilated convolutions does not further improve the model's performance; in fact, when the dilation rate is set to 3, the model's performance slightly decreases. Deformable convolution significantly increases the model's parameter count and computational load. Therefore, we replace the DW-Conv in EMOv2-1 M with DCNv2 [112] with a group size of 1 to maintain a similar scale of the model. The results indicate that this substitution actually reduces the model's performance.

Stronger training strategy: Table XIX(e) presents three training strategies that enhance model performance without altering the model architecture or parameters. When employing higher resolutions (up to 512 in this paper), knowledge distillation (KD) with naive logit distribution (TResNet [91] in Section IV-A), and long training durations (up to 1000 epochs), the model's performance improves significantly. When all strategies are combined, the EMOv2-5 M achieves the best 82.9 Top-1 accuracy. This performance notably surpasses that of Swin-Transformer-T (28.2 M with 81.3 Top-1) and ResNet-152 (60.1 M with 82.0 Top-1).

Scale up assessment: We scale up EMOv2 to 20 M/50 M magnitudes to evaluate its scaling capability. The specific structure is presented in Table XX, and the comparison results with current backbones of similar magnitudes are shown in Table XXI. The results demonstrate that EMOv2 can be easily extended to large-scale models and achieve highly competitive results. This scaling capability is also reflected in Table XVI, proving the structural effectiveness and generalization of i^{2}RMB.

TABLE XX
CORE CONFIGURATIONS OF SCALED EMOV2 VARIANTS

Items	EMOv2-20M	EMOv2-50M
Depth	[3, 3, 13, 3]	[5, 8, 20, 7]
Emb. Dim.	[64, 128, 320, 448]	[64, 128, 384, 512]
Exp. Ratio	[2.0, 3.0, 4.0, 4.0]	[2.0, 3.0, 4.0, 4.0]

TABLE XXI
EVALUATION OF SCALING CAPABILITIES OF EMOV2 AT 20 M/50 M MAGNITUDES ON IMAGENET-1 K DATASET

	Model	#Params ↓	FLOPs ↓	Reso.	Top-1	Venue
20M Magnitude	ResNet-50 [44], [114]	25.5	4.1G	224^2	80.4	CVPR'16
	ConvNeXt-T [115]	28.5	4.5G	224^2	82.1	CVPR'22
	PVTv2-B2 [20]	25.3	4.0G	224^2	82.0	ICCV'21
	Swin-T [21]	28.2	4.5G	224^2	81.3	ICCV'21
	PoolFormer-S36 [52]	30.8	5.0G	224^2	81.4	CVPR'22
	ViTAEv2-S [116]	19.3	5.7G	224^2	82.6	IJCV'23
	EATFormer-Small [24]	24.3	4.3G	224^2	83.1	IJCV'24
	☆ EMOv1-20M [13]	20.5	3.8G	224^2	82.0	ICCV'23
	★ **EMOv2-20M**	20.1	4.0G	224^2	83.3	-
50M×80M Magnitude	ResNet-152 [44], [114]	60.1	11.5G	224^2	82.0	CVPR'16
	Swin-B [21]	87.7	15.5G	224^2	83.5	ICCV'21
	PoolFormer-M48 [52]	73.4	11.6G	224^2	82.5	CVPR'22
	ViTAEv2-48M [116]	48.6	13.4G	224^2	83.8	IJCV'23
	EATFormer-Base [24]	49.0	8.9G	224^2	83.9	IJCV'24
	★ **EMOv2-50M**	49.8	8.8G	224^2	84.1	-

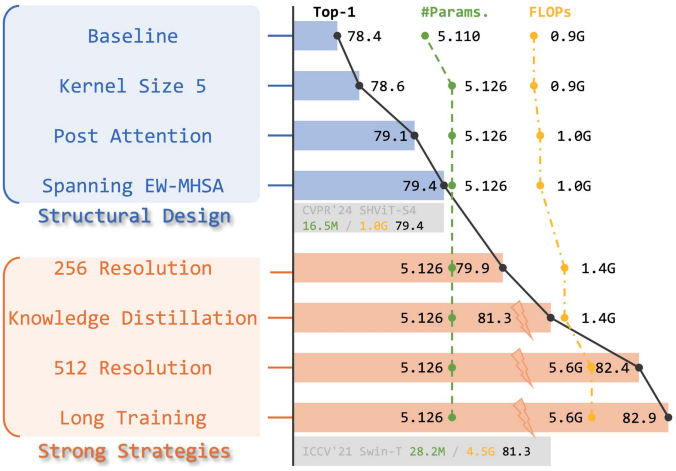

Fig. 8. *Overall incremental trajectory from baseline to modern EMOv2 at the 5 M magnitude:* Each line is based on a modification of the immediately preceding line. Detailed ablations in Section IV-C. Parameters and FLOPs are marked in green and yellow.

D. Visual Analysis Between EMOv1/v2

Quantitative downstream visualization: Fig. 6-Top presents the detection visualization results based on SSDLite. Compared to EMOv1, the improved EMOv2 demonstrates accurate classification and localization capabilities, even generalizing to objects that are missed in the ground truth. Thanks to the spanning attention mechanism, EMOv2 also achieves significant performance improvements in pixel-level dense prediction, as shown in Fig. 6-Bottom. *Class activation mapping comparison.* Fig. 7 presents the visualization results of Grad-CAM. The improved EMOv2 generates high-confidence class activations that are more closely aligned with the image subjects.

E. Summary

Starting from the EMOv1 baseline [13], we progressively explore factors influencing EMOv2 performance from the perspectives of *structural design* and *training strategy*. As shown in Fig. 8, the model parameters are controlled at 5.1 M, and each structural improvement incrementally enhances the model's performance without additional parameter increase: *1)* A larger kernel size improves the model's performance at the cost of only 0.016 M parameters. *2)* Post attention increases the Top-1 accuracy by 0.5 with an additional 0.1 G FLOPs. *3)* Spanning attention further enhances the model accuracy to 79.4, surpassing the baseline by +1.0↑. Additionally, this operation significantly improves the performance of EMOv2 on downstream tasks, as shown in Fig. 5. We use the structure at the end of the *structural design* phase as our default EMOv2-5 M, while higher resolution, extended training, and naive knowledge distillation strategies are employed to investigate the performance upper limits of our EMOv2 in the 5 M parameter magnitude. The detailed structure can be viewed in the attached source code.

Limitation discussion: This study focuses on lightweight vision backbones and proposes EMOv2 model, extending them to the 20 M and 50 M parameter scales due to resource constraints. However, its Transformer-compatible architecture design potentially allows application to larger-scale vision backbones. Additionally, the spanning mechanism can be extended to the domain of large language models (LLMs), which warrants further exploration.

V. Conclusion

This work rethinks lightweight infrastructure from efficient IRB and effective components of Transformer from a unified perspective, proposing the abstracted concept of Meta Mobile Block for designing efficient models. Specifically, we deduce a modern infrastructural i^{2}RMB to build a parameter-efficient attention-shared EMOv2, while extending it to dense prediction and generation fields by adapting i^{2}RMB to different basic structures. Massive experiments on several downstream benchmarks demonstrate the superiority of our approach, and we also provide detailed studies and give some experimental findings on building an attention-based lightweight model.

Future work: Although EMOv2 exhibits excellent generalizability, this paper only trains the model on ImageNet-1 K through supervised learning. Larger-scale datasets and unsupervised training methods have the potential to further enhance the model's performance. Additionally, we plan to adapt EMOv2 to video generation models to improve the efficiency and quality of generative models. Moreover, the model itself can also benefit from improvements in basic attention mechanisms, such as linear complexity enhancements and better integration of position embeddings, which can further enhance the model's practical performance and efficiency.

References

[1] Y. Li et al., "Rethinking vision transformers for MobileNet size and speed," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 16843–16854.
[2] M. Maaz et al., "EdgeNext: Efficiently amalgamated CNN-transformer architecture for mobile vision applications," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2022, pp. 3–20.
[3] H. Shu et al., "TinySAM: Pushing the envelope for efficient segment anything model," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 20470–20478.
[4] C. Zhou, X. Li, C. C. Loy, and B. Dai, "EdgeSAM: Prompt-in-the-loop distillation for on-device deployment of SAM," 2023, arXiv: *2312.06660*.
[5] S. Xu et al., "RMP-SAM: Towards real-time multi-purpose segment anything," in *Proc. Int. Conf. Learn. Representations*, 2025.
[6] X. Li et al., "Semantic flow for fast and accurate scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 775–793.
[7] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "RTMO: Towards high-performance one-stage real-time multi-person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1491–1500.
[8] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, arXiv: *1704.04861*.
[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[10] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
[11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1577–1586.
[12] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[13] J. Zhang et al., "Rethinking mobile block for efficient attention-based models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 1389–1400.

[14] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," *Trans. Mach. Learn. Res.*, 2023.

[15] J. Nielsen, "The need for speed in AI," 2023. Accessed: Oct. 03, 2023. [Online]. Available: https://www.uxtigers.com/post/ai-response-time

[16] J. Nielsen, *Usability Engineering*. San Mateo, CA, USA: Morgan Kaufmann, 1994.

[17] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2022.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[19] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 548–558.

[20] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[21] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[22] Z. Liu et al., "Swin Transformer V2: Scaling up capacity and resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11999–12009.

[23] J. Zhang et al., "Analogous to evolutionary algorithm: Designing a unified sequence model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 26674–26688.

[24] J. Zhang et al., "EATFormer: Improving vision transformer inspired by evolutionary algorithm," *Int. J. Comput. Vis.*, vol. 132, pp. 3509–3536, 2024.

[25] X. Li et al., "Transformer-based visual segmentation: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10138–10163, Dec. 2024.

[26] D. Li et al., "Involution: Inverting the inherence of convolution for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12316–12325.

[27] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Representations*, 2020.

[28] K. M. Choromanski et al., "Rethinking attention with performers," in *Proc. Int. Conf. Learn. Representations*, 2021.

[29] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[30] J. Li et al., "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios," 2022, arXiv: *2207.05501*.

[31] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "DeLighT: Deep and light-weight transformer," in *Proc. Int. Conf. Learn. Representations*, 2021.

[32] S. N. Wadekar and A. Chaurasia, "MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," 2022, arXiv: *2209.15159*.

[33] Y. Chen et al., "Mobile-Former: Bridging MobileNet and transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5260–5269.

[34] Y. Li et al., "EfficientFormer: Vision transformers at MobileNet speed," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 940.

[35] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 733–743.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[37] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, arXiv: *1602.07360*.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[39] X. Li et al., "SFNet: Faster, accurate, and domain agnostic semantic segmentation via semantic flow," *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 466–489, 2024.

[40] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "RepViT: Revisiting mobile CNN from ViT perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 15909–15920.

[41] Z. Liu, Z. Hao, K. Han, Y. Tang, and Y. Wang, "GhostNetV3: Exploring the training strategies for compact models," 2024, arXiv: *2404.11202*.

[42] D. Qin et al., "MobileNetV4: Universal models for the mobile ecosystem," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 78–96.

[43] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," 2022, arXiv: *2204.07756*.

[46] K. Islam, "Recent advances in vision transformer: A survey and outlook of recent work," 2022, arXiv: *2203.01536*.

[47] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 559–568.

[48] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," in *Proc. Int. Conf. Learn. Representations*, 2023.

[49] K. Li et al., "UniFormer: Unified transformer for efficient spatial-temporal representation learning," in *Proc. Int. Conf. Learn. Representations*, 2022.

[50] S. Li et al., "MogaNet: Multi-order gated aggregation network," in *Proc. Int. Conf. Learn. Representations*, 2024.

[51] S. Yun and Y. Ro, "SHViT: Single-head vision transformer with memory efficient macro design," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 5756–5767.

[52] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10809–10819.

[53] T. Huang, L. Huang, S. You, F. Wang, C. Qian, and C. Xu, "LightViT: Towards light-weight convolution-free vision transformers," 2022, arXiv: *2207.05557*.

[54] Q. Zhang and Y.-B. Yang, "ResT: An efficient transformer for visual recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, Art. no. 1185.

[55] J. Pan et al., "EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 294–311.

[56] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[57] A. Ali et al., "XCiT: Cross-covariance image transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 20014–20027.

[58] B. Liao, X. Wang, L. Zhu, Q. Zhang, and C. Huang, "ViG: Linear-complexity visual sequence learning with gated linear attention," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 5182–5190.

[59] Q. He et al., "PointRWKV: Efficient RWKV-like model for hierarchical point cloud learning," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 3410–3418.

[60] Y. Duan et al., "Vision-RWKV: Efficient and scalable visual perception with RWKV-like architectures," in *Proc. Int. Conf. Learn. Representations*, 2025.

[61] H. Yuan et al., "Mamba or RWKV: Exploring high-quality and high-efficiency segment anything model," 2024, arXiv: *2406.19369*.

[62] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. 1st Conf. Lang. Model.*, 2024.

[63] B. Peng et al., "RWKV: Reinventing RNNs for the transformer era," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2023, pp. 14048–14077.

[64] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. Int. Conf. Mach. Learn.*, 2024, Art. no. 2584.

[65] X. Pei, T. Huang, and C. Xu, "EfficientVMamba: Atrous selective scan for light weight visual Mamba," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 6443–6451.

[66] H. He et al., "MobileMamba: Lightweight multi-receptive visual mamba network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 4497–4507.

[67] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 4172–4182.

[68] J. Yang et al., "Focal attention for long-range interactions in vision transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 30008–30022.

[69] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12114–12124.
[70] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 1707.
[71] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9204–9215.
[72] I. O. Tolstikhin et al., "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 24261–24272.
[73] H. Touvron et al., "ResMLP: Feedforward networks for image classification with data-efficient training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5314–5321, Apr. 2023.
[74] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 122–138.
[75] C. Yang et al., "MOAT: Alternating mobile convolution and attention brings strong vision models," in *Proc. Int. Conf. Learn. Representations*, 2023.
[76] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[77] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, arXiv: *1606.08415*.
[78] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv: *1607.06450*.
[79] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[80] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.
[81] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017.
[82] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 646–661.
[83] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3008–3017.
[84] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 32–42.
[85] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.
[86] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
[87] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6022–6031.
[88] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.
[89] Z.-H. Jiang et al., "All tokens matter: Token labeling for training better vision transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 18590–18602.
[90] R. Wightman, "Pytorch image models," 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models
[91] T. Ridnik, H. Lawen, A. Noy, E. Ben Baruch, G. Sharir, and I. Friedman, "TResNet: High performance GPU-dedicated architecture," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1399–1408.
[92] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12021–12031.
[93] H. Ma, X. Xia, X. Wang, X. Xiao, J. Li, and M. Zheng, "MoCoViT: Mobile convolutional vision transformer," 2022, arXiv: *2205.12635*.
[94] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14420–14430.
[95] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7277–7286.
[96] Y. Shi, M. Dong, and C. Xu, "Multi-scale VMamba: Hierarchy in hierarchy visual state space model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, Art. no. 808.
[97] W. Yu and X. Wang, "MambaOut: Do we really need mamba for vision?," 2024, arXiv: *2405.07992*.
[98] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[99] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[100] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, arXiv: *1906.07155*.
[101] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, arXiv: *1706.05587*.
[102] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6392–6401.
[103] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[104] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[105] B. Zhou et al., "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, pp. 302–321, 2019.
[106] M. Contributors, "MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark," 2020. [Online]. Available: https://github.com/open-mmlab/mmsegmentation
[107] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[108] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *Int. J. Biomed. Imag.*, vol. 2013, 2013, Art. no. 154860.
[109] W. Kay et al., "The Kinetics human action video dataset," 2017, arXiv: *1705.06950*.
[110] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie, "SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 23–40.
[111] A. Inc., "Optimize your core ML usage," 2022. [Online]. Available: https://developer.apple.com/documentation/vision/classifying_images_with_vision_and_core_ml
[112] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9308–9316.
[113] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in TIMM," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshops*, 2021.
[114] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.
[115] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *Int. J. Comput. Vis.*, vol. 131, pp. 1141–1162, 2023.

Jiangning Zhang received the BS degree from Electronic Information School, Wuhan University, Wuhan, China, in 2017, and the PhD degree from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2022. He is currently a research scientist with Youtu Lab, Tencent, Shanghai, China. His research interests include artificial intelligence generated content and deep learning.

Teng Hu received the BS degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2022. He is currently working toward the PhD degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include computer vision and computer graphics.

Haoyang He received the BS degree from Southwest Jiaotong University, Chengdu, China, in 2022. He is currently working toward the PhD degree in control science and engineering with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. His current research interests include anomaly detection, neural architecture design, and AIGC.

Zhucun Xue received the BS degree from Electronic Information School, Wuhan University, Wuhan, China, in 2017. She is currently working toward the PhD degree with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. Her research interests include artificial intelligence generated content and motion generation.

Yabiao Wang received the master's degree from Zhejiang University, in 2016. He is currently a research scientist with Tencent Youtu lab, China. He published more than 50 conference papers including CVPR, ICCV, ECCV, and AAAI etc. He won more than 20 challenge titles. His research interests include object detection, segmentation, few-shot learning, and AI generated content.

Chengjie Wang received the BS degree in computer science from Shanghai Jiao Tong University, China, in 2011, and the double MS degrees in computer science from Shanghai Jiao Tong University, China and Weseda University, Japan, in 2014. He is currently working toward the PhD degree with Shanghai Jiao Tong University, and the research director with Tencent YouTu Lab. His research interests include computer vision and machine learning. He has published more than 100 papers on major Computer Vision and Artificial Intelligence Conferences such as CVPR, ICCV, ECCV, AAAI, IJCAI, and NeurIPS, and holds more than 100 patents in these areas.

Yong Liu received the BS degree in computer science and engineering and the PhD degree in computer science from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively. He is currently a professor with the Institute of Cyber-Systems and Control, Zhejiang University. His main research interests include: robot perception and vision, deep learning, Big Data analysis, and multi-sensor fusion. His research interests on machine learning, computer vision, information fusion, and robotics.

Xiangtai Li received the PhD degree from Peking University, in 2022. He is working as a research scientist with Tiktok, Singapore. Previously, he worked as a research fellow with MMLab@NTU and a member of the Multimedia Laboratory, Nanyang Technological University. His research interests include computer vision and machine learning with a focus on scene understanding, segmentation, video understanding, and multi-modal learning. He regularly reviews top-tier conferences and journals, including CVPR, ICCV, ICLR, ECCV, ICML, NeurIPS, the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *International Journal of Computer Vision*.

Dacheng Tao (Fellow, IEEE) is with the Nanyang Technological University. He is also an advisor and chief scientist with the Digital Science Institute, University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and more than 200 publications in prestigious journals and proceedings at leading conferences. He received the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a fellow of the Australian Academy of Science, AAAS, and ACM.