TinyViM: Frequency Decoupling for Tiny Hybrid Vision Mamba

Supplementary Material

Table A. Ablation for the number of Mamba blocks. Throughput is tested on a Nvidia V100 GPU with maximum power-of-two batch size that fits in memory.

Variants	Params	GMACs	Throu.	Top-1
Vim-Ti [15]	7.0	1.5	617	76.1
EfficientVMamba-T [8]	6.1	1.0	1396	76.5
TinyViM-S*	**6.2**	**0.9**	**2235**	**78.5**
TinyViM-S	**5.6**	**0.9**	**2563**	**79.2**

1. Dataset and Implementation Details

ImageNet [1] is a large image categorization dataset for computer vision research, containing about 1.2 million labeled images of about 1000 categories. All of our models are trained on the Imagenet-1k dataset for 300 epochs using the AdamW optimizer and an initial learning rate of $2e^{-3}$. We use an image resolution of 224x224 for both training and testing. In addition, we use the same teacher model (i.e., the RegNetY-16GF model with a top-1 accuracy of 82.9%) as in [5, 11] for distillation. All experiments are conducted on 8 NVIDIA V100 GPUs and the throughput is tested with maximum power- of-two batch size that fits in memory.

MS-COCO 2017 [7] is a large-scale real-world image and annotation dataset that contains 118K training and 5K validation images with 80 object classes. We use TinyViM as the backbone for feature extraction and apply the Mask-RCNN [2] framework for object detection and instance segmentation. Similar to [5], we finetune our TinyViM for 12 epochs with an image resolution of 1333 × 800 and batch size of 32. The AdamW optimizer is used with a learning rate of $2e^{-4}$. We report the performance for object detection and instance segmentation in terms of mean average precision (mAP).

ADE20K [14] is a challenging segmentation dataset, which contains about 20,000 images and covers 150 categories. We apply the ImageNet pre-trained TinyVim as backbone to extract image features and Semantic FPN [4] as decode head for segmentation. Follwing [5, 11], we train the segmentation model with an image size of 512×512 and a batch size of 32 for 40K iterations. The AdamW optimizer with poly learning rate scheduling is applied and the initial learning rate is $2e^{-4}$. The semantic segmentation performance is reported with the mean intersection over union (mIoU) metric.

2. Model Configuration

We give the detailed architectural configuration of the TinyViM variant in Table A, which details each building block of the model variant and the corresponding hyperparameters. We mainly place the proposed TinyViM blocks at the end of each stage to capture the global context-enhanced features. For the third stage, we additionally place a TinyViM in the middle of the stage to obtain a better balance between accuracy and efficiency, and the effectiveness of this design of setting up more and larger blocks in the third stage has been confirmed by previous works [9, 13]. In addition, we set the downsampling ratios in TinyViM Blocks of the four stages from small to large to 8, 4, 2 and 1, respectively. In other words, the resolution of the feature maps input to Mamba block at each stage is 1/32 of the original images.

3. More Comparition Results

3.1. Qualitative results on ADE20K

We visualized the result of segmentation on ADE20K, 053 which is shown in Fig. A. It can be found that the segmen- 054 tation result of TinyViM is very close to the label. The cate- 055 gories and boundaries of objects are accurately segmented. 056 These visualizations demonstrate the state-of-the-art perfor- 057 mance of TinyViM in semantic segmentation

3.2. Qualitative results on MS-COCO

The detection and instance segmentation results on coco are 060 visualized in Fig. B. It can be observed that the people 061 and animals in the image are accurately segmented. The 062 location of the detection box is also very accurate. These 063 visualization results show that TinyViM achieves advanced 064 performance in both instance segmentation and detection 065 tasks.

4. Efficiency Analysis

We attribute the efficiency of TinyViM to two reasons: 1, the combination of mobile-friendly convolution and 2, the efficient laplace mixer design. For the former, we are not obsessed with designing a pure Mamba architecture such as ViM [15]. The inductive bias and mobile-friendliness of convolution make it naturally suitable for lightweight backbones. As a result, pure Convolution [3, 10, 13] or Convolution-ViT [5, 6, 11] models dominate the current lightweight backbone family. Inspired by these excellent Convolutional-ViT hybrid model [5, 6, 11, 12], we design the Convolution-Mamba hybrid architecture TinyViM. For the latter, we propose a variant TinyViM-S* to explore the impact of our reduced number of Mamba blocks on model

Table A. Our detailed model variants for ImageNet-1k, which describe the model configurations at each stage including the numbers of Local Block and TinyVim Block, the output channels and the resolution of the output features. In addition, the expansion factor for FFN in both Local Block and TinyViM Block defaults to 4.

Name	Output	Small	Base	Large
stem	56×56	[RepDW-3 $\times$ 2, stride=2]		
stage1	56×56	[Local Block$\times$ 2, d=48 TinyViM Block$\times$ 1]	[Local Block$\times$ 3, d=48 TinyViM Block$\times$ 1]	[Local Block$\times$ 3, d=64 TinyViM Block$\times$ 1]
stage2	28×28	[RepDW-3 $\times$ 1, stride=2]		
		[Local Block$\times$ 2, d=64 TinyViM Block$\times$ 1]	[Local Block$\times$ 2, d=96 TinyViM Block$\times$ 1]	[Local Block$\times$ 3, d=128 TinyViM Block$\times$ 1]
stage3	14×14	[RepDW-3 $\times$ 1, stride=2]		
		[Local Block$\times$ 7, d=168 TinyViM Block$\times$ 2]	[Local Block$\times$ 8, d=192 TinyViM Block$\times$ 2]	[Local Block$\times$ 10, d=384 TinyViM Block$\times$ 2]
stage4	7×7	[RepDW-3 $\times$ 1, stride=2]		
		[Local Block$\times$ 5, d=224 TinyViM Block$\times$ 1]	[Local Block$\times$ 4, d=384 TinyViM Block$\times$ 1]	[Local Block$\times$ 5, d=512 TinyViM Block$\times$ 1]
Classifier	1×1	Average pool, 1000d fully-connected		
GFLOPs		0.9G	1.5G	4.7G
Params		5.6M	11.0M	31.7M

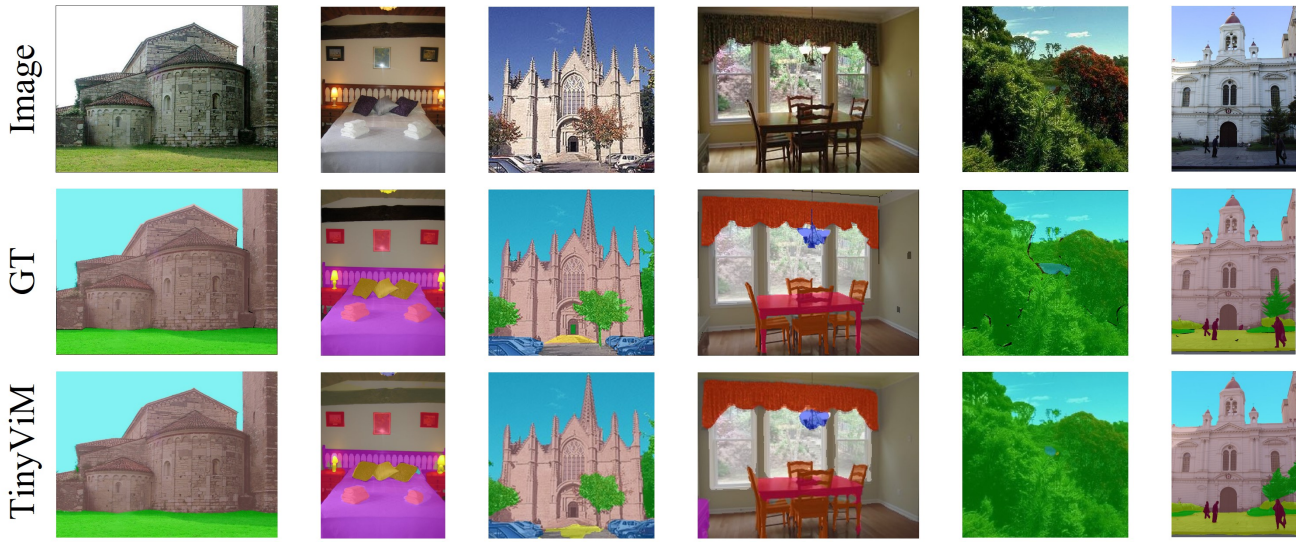

Figure A. Visualization of the detection and instance segmentation predictions on COCO. It shows that TinyViM can accurately localize and segment various objects in complex scenes.

Figure B. Visualization of segmentation predictions on ADE20K. Top: Input images. Mid: Ground truth masks. Bottom: The semantic segmentation results of TinyViM. It shows that TinyViM can accurately segment various objects in complex scenes.

efficiency, as shown in Table A. Specifically, TinyViM-S* adopts the same architecture as EfficientVMamba-T [8], i.e., for the four stages, the number of Mamba blocks is (2,2,4,2) and the feature dimensions are (48,96,192,384). It can be observed that TinyViM-S* shows significantly higher throughput compared to EfficientVMamba-T and ViM-Ti. Moreover, the Top-1 accuracy of TinyViM-S* is higher than that of EfficientVMamba-T and VIM-Ti by 2.4% and %2.0, respectively. These results validate the effectiveness of Laplace Mixer compared to other Mamba block designs. In addition, the performance and efficiency advantage is further increased when combining convolution at each stage of the network. Therefore, by combining Convolution and Laplace Mixer, TinyViM-S obtains a Top-1 accuracy of 79.2 and outperforms other Convolution, ViT and Mamba based models between accuracy and efficiency.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[3] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 1

[4] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 1

[5] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35: 12934–12949, 2022. 1

[6] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16889–16900, 2023. 1

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[8] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 1, 3

[9] Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, and Piotr Dollar. On network design spaces for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted

residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[11] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17425–17436, 2023. 1

[12] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5785–5795, 2023. 1

[13] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15909–15920, 2024. 1

[14] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[15] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*. 1