ElasticViT: Conflict-aware Supernet Training for Deploying Fast Vision Transformer on Diverse Mobile Devices

Chen Tang*, Li Lyna Zhang*, Huiqiang Jiang, Jiahang Xu, Ting Cao, Quanlu Zhang, Yuqing Yang, Zhi Wang, Mao Yang

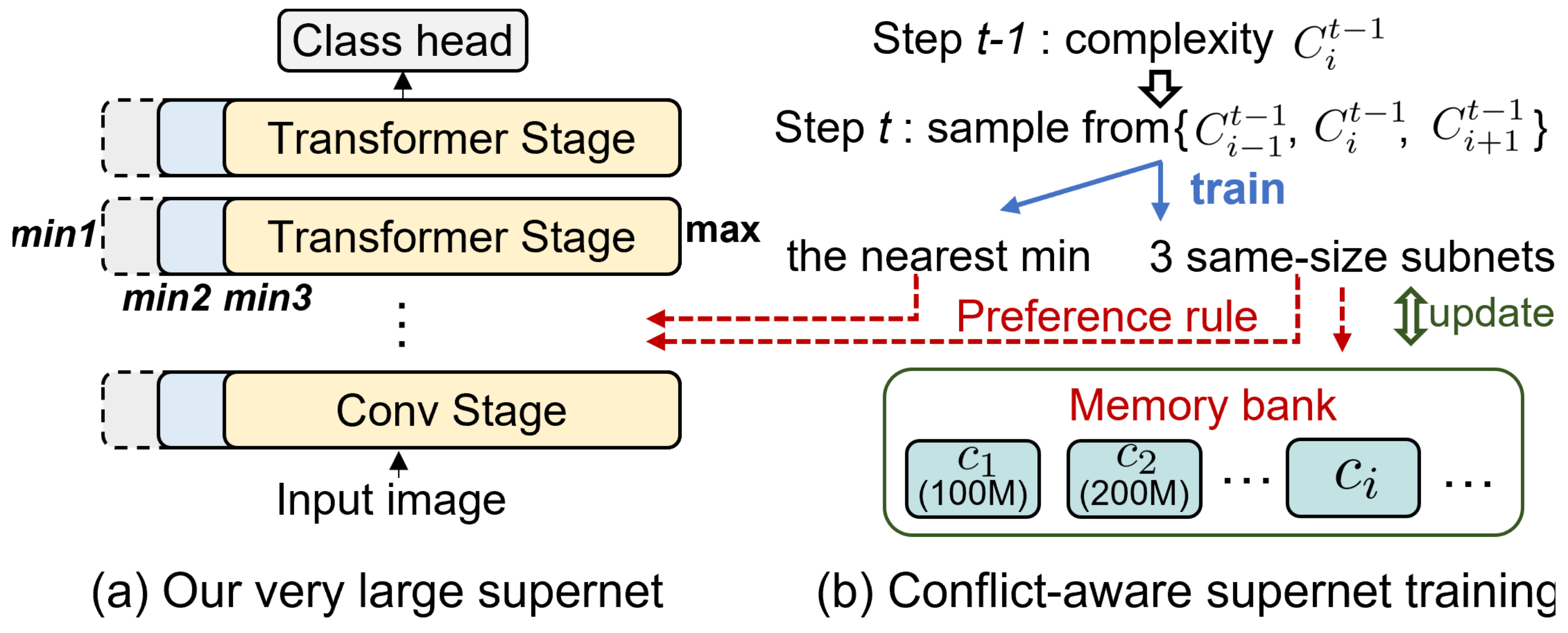
paper code

ICCV23 PARIS

Contributions of ElasticViT

- ❖ ElasticViT automates the design of accurate and low-latency ViTs for diverse mobile devices (i.e., from weak to strong devices) using a unified and retraining-free manner. For the first time we are able to train a single high-quality ViT supernet over a vast and mobile-regime search space.
- ❖ A thorough analysis of the poor-quality supernet trained by existing approaches, and reveals that uniform sampling results in subnets of vastly different sizes, leading to gradient conflicts.
- ❖ Conflict-aware training techniques (complexity-aware sampling and performance-aware sampling) are proposed to effectively address the gradient issues by sampling good subnets and limiting their FLOPs differences across adjacent training steps.
- ❖ Extensive experiments on multiple real mobile devices. This is the first time that ViT outperforms CNNs with a faster speed on mobile devices within the 200 MFLOPs range.

Overview of ElasticViT



(a) Our very large supernet (b) Conflict-aware supernet training

Fig. 1

- ❖ Very Large Search Space: a single very large search space optimized for diverse mobile devices, containing a wide range of ViTs with sizes ranging from 37M FLOPs to 3G FLOPs.
- ❖ Conflict-aware Supernet Training: to circumvent the poor-quality supernet challenge posed by the very large search space, we address the gradient conflicts issue by constraining FLOPs differences among sampled subnets and sampling potentially good subnets, greatly improving supernet training quality.

Gradient Conflicts Problem Degrades the Overall Performance

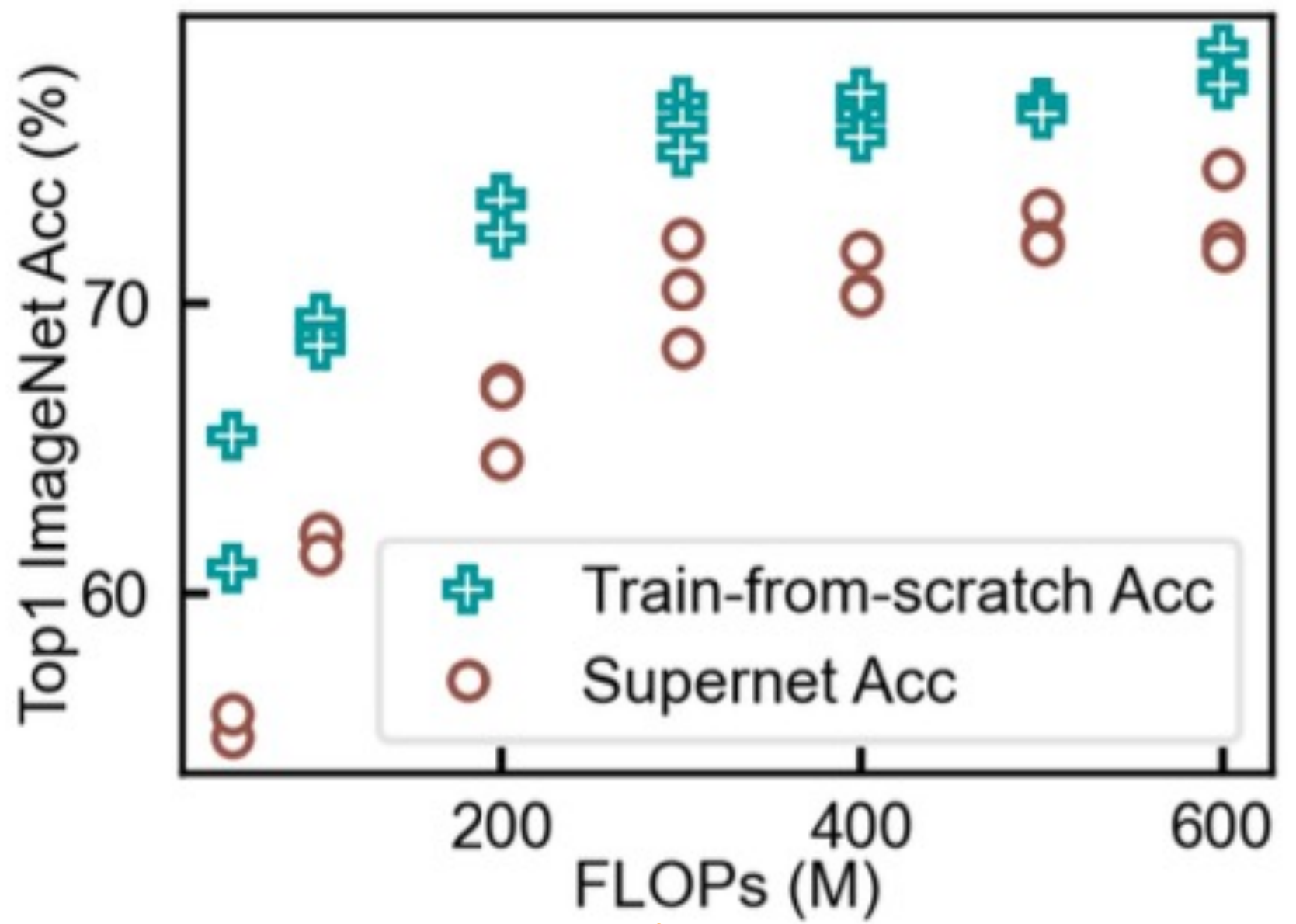


Fig. 2

- ❖ Optimization Difficulty: to support diverse mobile devices, the resulting very large search space is 10^7 times larger than the existing search spaces. Direct using existing two-stage NAS methods causes ~10% Top-1 accuracy degradation (Fig. 2).

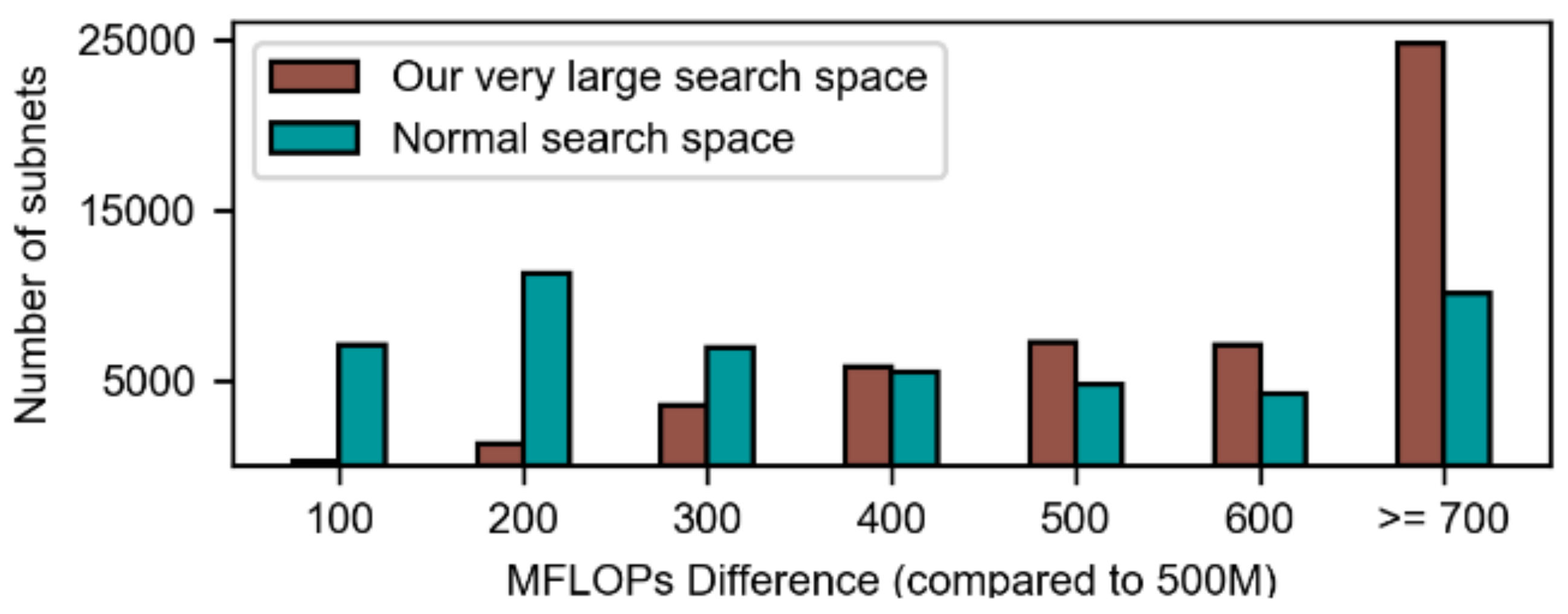
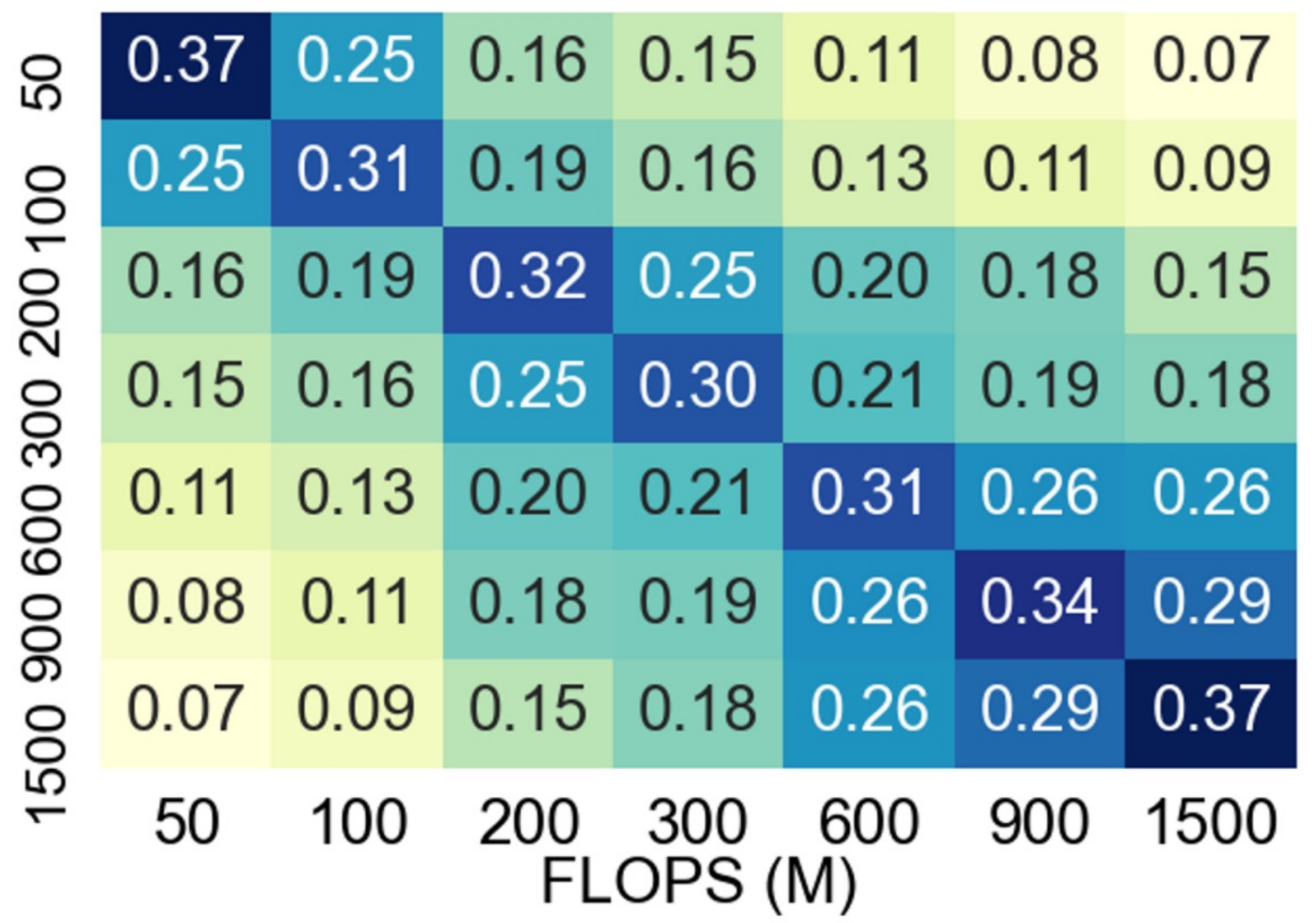


Fig. 3



	50	100	200	300	600	900	1500
50	0.37	0.25	0.16	0.15	0.11	0.08	0.07
100	0.25	0.31	0.19	0.16	0.13	0.11	0.09
200	0.16	0.19	0.32	0.25	0.20	0.18	0.15
300	0.15	0.16	0.25	0.30	0.21	0.19	0.18
600	0.11	0.13	0.20	0.21	0.31	0.26	0.26
900	0.08	0.11	0.18	0.19	0.26	0.34	0.29
1500	0.07	0.09	0.15	0.18	0.26	0.29	0.37

Fig. 4

- ❖ Gradient Conflicts Issue: in this work, we reveal that the core problem is caused by the uniform sampling in very large search space preferred to sample subnets with significantly different FLOPs (Fig.3) and the gradient similarity of shared weights between two subnets is easily close to 0 when the FLOPs differences grow (Fig. 4).

Conflict-aware Training Techniques

- ❖ Adjacent Step Sampling: constrain the complexity level (e.g., FLOPs) of the sampled subnets to be close to the previous step (*complexity-aware sampling*):

$$g(s^{(t)}; C_j^{(t)}) = |C_j^{(t)} - C_i^{(t-1)}| \leq Z,$$

- ❖ Multiple Min Training: sample the nearest smallest subnet based on the current complexity level, ensuring performance bounds without introducing a large complexity difference with other subnets (*performance-aware sampling*):

$$\arg\min_{w} \left[\sum_{s_m^{(t)} \in \mathbf{U}}^{M} \mathcal{L}_D\left(f(w_{s_m^{(t)}})\right) + \sum_{s_n \in \hat{\mathbf{S}}} \sigma(s_n, C_j^{(t)}) \mathcal{L}_D\left(f\left(w_{s_n}\right)\right) \right] \quad (3)$$

and $\sigma(\cdot)$ selects the nearest smallest subnet from HSS:

$$\sigma(s_n, C_j^{(t)}) = \begin{cases} 1 & \text{if } s_n \text{ is the nearest min that is smaller than } C_j^{(t)} \\ 0 & \text{otherwise.} \end{cases}$$

- ❖ Performance-aware Sampling: sample subnets with higher potential accuracy from a prior distribution while building using a memory bank $\mathbf{B}_j$ and a ViT architecture preference rule $U(\hat{\mathcal{A}}_{C_j})$ (*performance-aware sampling*):

$$\mathbb{E}_{s_m^{(t)} \sim \Gamma(\mathcal{A})} = q \cdot U(\mathbf{B}_j) + (1-q) \cdot U(\hat{\mathcal{A}}_{C_j}),$$

$$\text{where } \mathbf{B}_j \cup \hat{\mathcal{A}}_{C_j} = \mathcal{A}_{C_j}$$

Results: ImageNet Classification Task

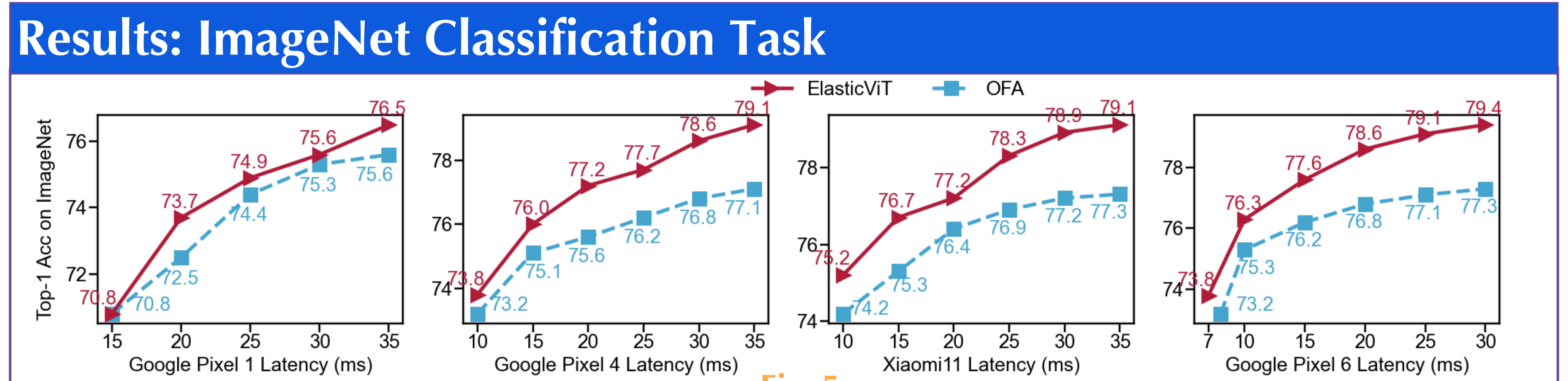


Fig. 5

- ❖ Hardware-aware NAS without Retraining: our discovered ViT models consistently outperform Once-For-All (OFA) under various latency constraints, on both weak (e.g., Google Pixel 1) and strong (e.g., Google Pixel 6) mobile devices.
- ❖ Joint Optimization: all models are trained and searched within a unified CNN-Transformer hybrid search space, this eliminates the need of costly search space design and supernet training.
- ❖ SOTA Performance: ElasticViT achieves top-1 accuracy from 67.2% to 80.0% on ImageNet from 60M to 800M FLOPs without extra retraining, outperforming all prior CNNs and ViTs in terms of accuracy and latency (please refer to our paper and GitHub repository for more high-performing models).

Results: Transfer Learning

Model	Latency	CIFAR10	CIFAR100	Food-101	Flowers	Pets
MobileNetV3	24.2 ms	97.1	83.3	86.8	94.3	87.7
ElasticViT-S1	**21.0 ms**	**97.5**	**86.1**	**87.2**	**94.3**	**92.1**
LeViT-128S	30.5 ms	96.8	85.0	73.6	86.2	90.1
ElasticViT-S2	**29.6 ms**	**97.5**	**86.9**	**88.3**	**95.2**	**92.9**
EfficientNet-B0	55.1 ms	97.9	86.9	89.1	92.4	92.2
LeViT-128	40.2 ms	97.8	86.6	80.8	86.2	92.2
ElasticViT-M	**37.5 ms**	**97.9**	**87.0**	88.8	**95.6**	**93.3**

- ❖ ElasticViT achieves up to 9.4% top-1 accuracy improvement on down-stream tasks.

Ablation Study: Proposed Conflict-aware Training Techniques

Method	MFLOPs							
	100	200	300	400	500	600	700	800
Baseline (Sandwich rule)	69.9	74.2	76.5	77.4	77.8	78.3	78.7	79.0
Adjacent step sampling	73.2	75.6	76.7	77.5	78.2	78.3	78.7	79.0
+Multiple min (HSS)	72.8	76.7	78.4	79.0	79.3	79.4	79.6	79.7
++Perf-aware sampling	**73.8**	**77.2**	**78.6**	**79.1**	**79.4**	**79.6**	**79.8**	**80.0**

- ❖ Compared to the baseline, conflict-aware training improve the overall performance.

Gradient Similarity Measurements

Method	MFLOPs					
	50		200		600	
	Random	Good	Random	Good	Random	Good
Sandwich rule	0.37	0.47	0.32	0.41	0.31	0.46
Ours	**0.50**	**0.56**	**0.51**	**0.67**	**0.51**	**0.67**

- ❖ ElasticViT significantly improves the gradient similarity and therefore greatly stabilizes the supernet training.