NeurIPS 2024 Spotlight

Parameter-Inverted Image Pyramid Networks

Xizhou Zhu[2,1]*, Xue Yang[1]*, Zhaokai Wang[3,1]*, Hao Li[4,1]

Wenhan Dou[2,5], Junqi Ge[2,5], Lewei Lu[5], Yu Qiao[1], Jifeng Dai[2,1]

[1]OpenGVLab, Shanghai AI Laboratory

[2]Tsinghua University

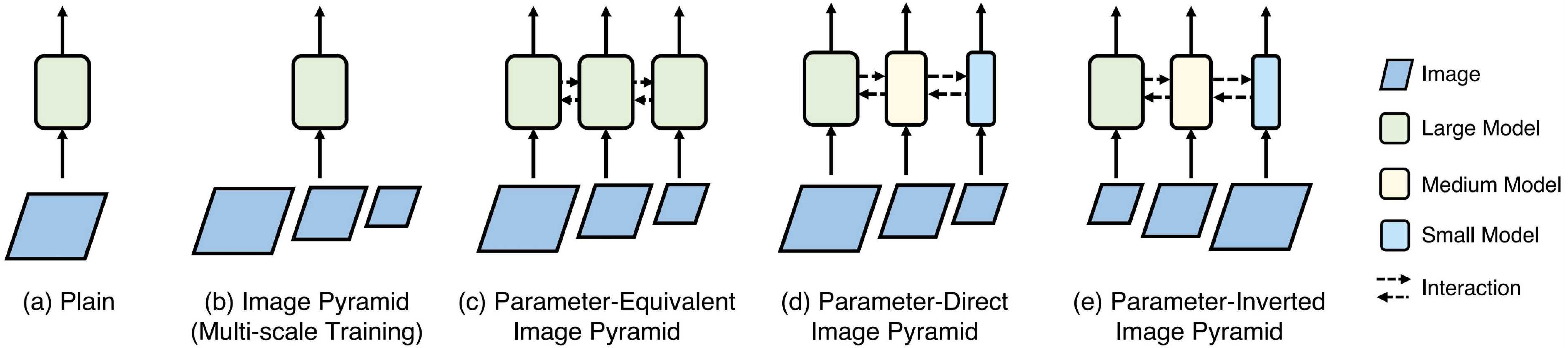
[3]Shanghai Jiao Tong University

[4]The Chinese University of Hong Kong

[5]SenseTime Research

Motivation

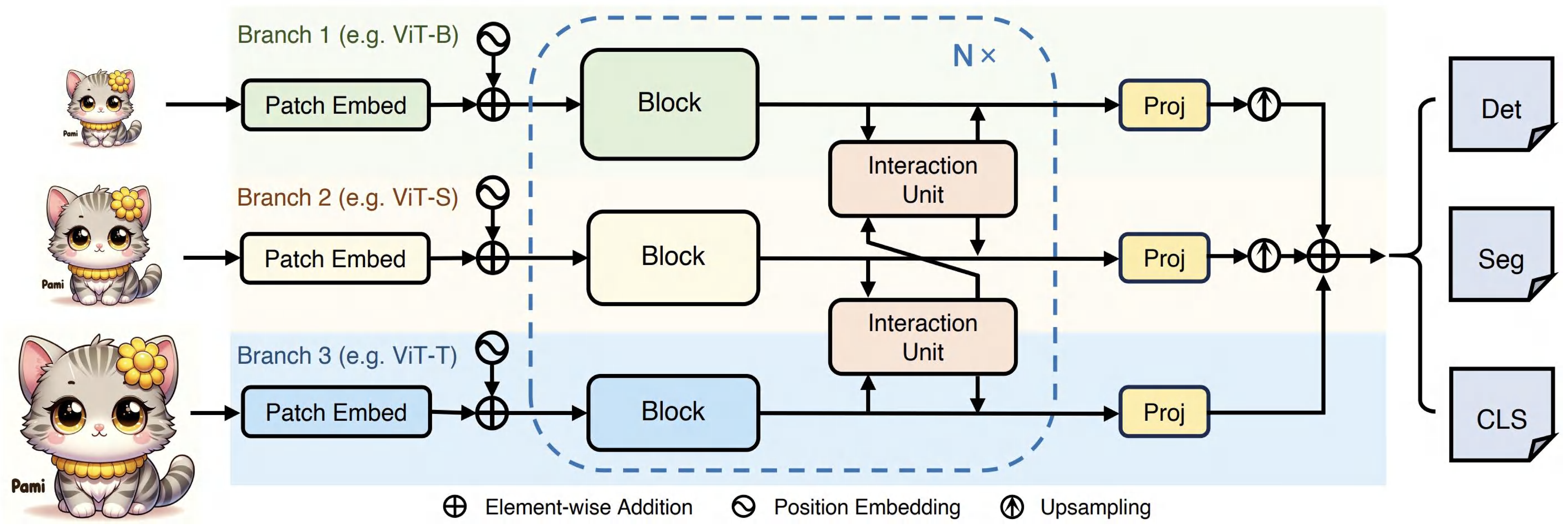
- Traditional image pyramids processing high-res images: significant computational overhead
- Parameter-inverted Design: large models for low-res images to extract rich context; small models for high-res images to focus on details.
- Cross-branch interactions improves efficiency and avoids redundant modeling.



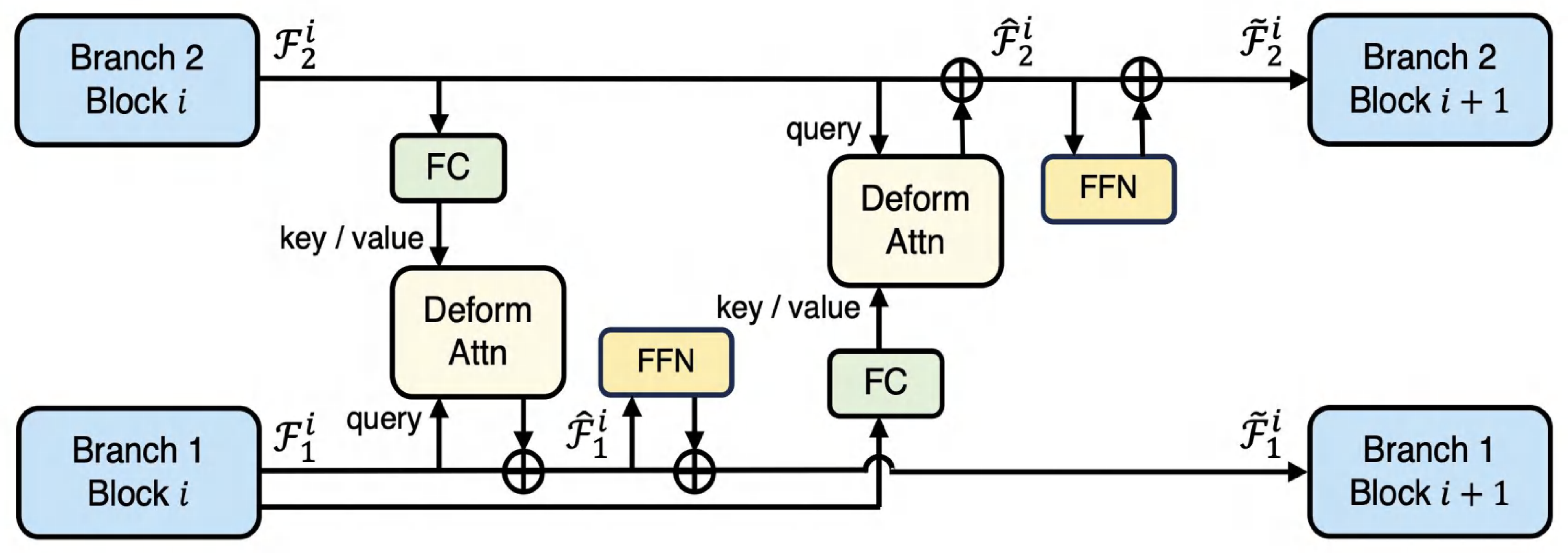
Different image pyramid network designs

Method

- Parameter-Inverted Image Pyramid Networks (PIIP)
- **Multi-resolution Branches:** Different-sized model for different resolutions with parameter-inverted design.
- **Cross-branch Interactions:** Added every few layers to integrate features of different scales.
- **Branch Merging:** combines outputs of all branches to form the final output.



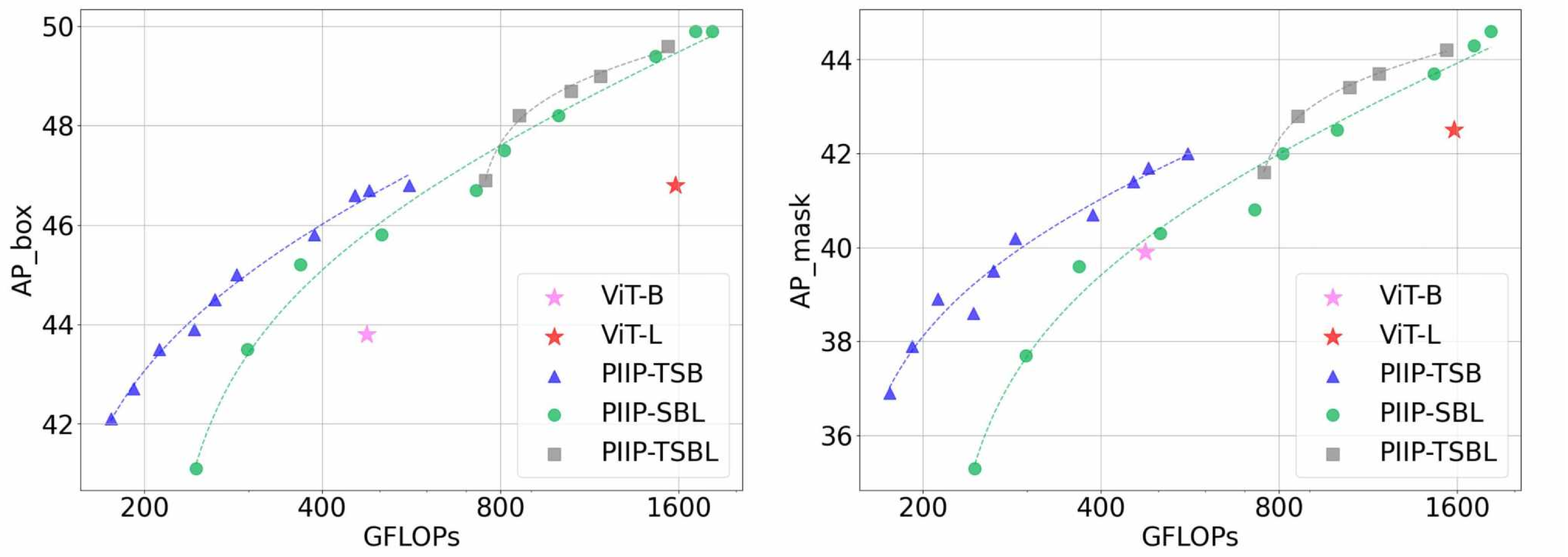
Overall architecture



Interaction unit

Experiments - Detection

Model	Resolution	#Param	#FLOPs	Mask R-CNN 1× schedule					
				AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ViTDet-B [23]	1024	90M	463G	43.8	67.6	47.7	39.9	63.6	42.2
PIIP-TSB (ours)	1120/896/448	146M	243G	43.9	65.7	47.5	38.6	61.8	40.6
	1568/896/448	147M	287G	45.0	67.0	48.7	40.2	63.8	42.6
	1568/1120/672	149M	453G	**46.6**	68.4	51.1	**41.4**	65.2	44.3
ViTDet-L [23]	1024	308M	1542G	46.8	70.8	51.4	42.5	67.3	45.3
PIIP-SBL (ours)	1120/672/448	493M	727G	46.7	69.0	50.6	40.8	65.2	42.8
	1344/896/448	495M	1002G	48.2	71.0	52.8	42.5	67.3	45.4
	1568/896/672	497M	1464G	49.4	71.9	53.9	43.7	68.4	46.6
PIIP-TSBL (ours)	1344/896/672/448	506M	755G	46.9	69.9	50.6	41.6	65.9	44.1
	1568/1120/672/448	507M	861G	48.2	70.5	52.7	42.8	66.9	45.6
	1792/1568/1120/448	512M	1535G	**49.6**	72.4	54.2	**44.2**	69.2	47.5



(a) Object detection (b) Instance segmentation

Comparison with baseline on COCO val2017

Experiments - Detection

OpenGVLab

Method	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
	Mask R-CNN 1× schedule					
PVTv2-B5 [51]	47.4	68.6	51.9	42.5	65.7	46.0
ViT-B [24]	42.9	65.7	46.8	39.4	62.6	42.0
ViTDet-B [23]	43.2	65.8	46.9	39.2	62.7	41.4
Swin-B [30]	46.9	-	-	42.3	-	-
ViT-Adapter-B [7]	47.0	68.2	51.4	41.8	65.1	44.9
PIIP-TSB (ours)	47.9	70.2	52.5	42.6	67.2	45.5
ViT-L [24]	45.7	68.9	49.4	41.5	65.6	44.6
ViTDet-L [23]	46.2	69.2	50.3	41.4	65.8	44.1
ViT-Adapter-L [7]	48.7	70.1	53.2	43.3	67.0	46.9
PIIP-SBL (ours)	49.9	72.8	54.7	44.6	69.3	47.9
	DINO + MS schedule					
PIIP-SBL-3× (ours)	57.9	76.9	63.3	-	-	-
PIIP-H6B-1× (ours)	60.0	79.0	65.4	-	-	-

Method	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
	Cascade R-CNN 1× schedule					
Swin-L [30]	51.8	71.0	56.2	44.9	68.4	48.9
ConvNeXt-L [31]	53.5	72.8	58.3	46.4	70.2	50.2
PIIP-SBL (ours)	53.6	73.3	57.9	46.3	70.3	50.0
	Cascade R-CNN 3× + MS schedule					
Swin-B [30]	51.9	70.9	57.0	-	-	-
Shuffle-B [22]	52.2	71.3	57.0	-	-	-
ViT-B [24]	50.1	69.3	54.3	-	-	-
ViT-Adapter-B [7]	52.1	70.6	56.5	-	-	-
PIIP-TSB (ours)	53.1	72.3	57.4	46.5	70.1	51.1
Swin-L [30]	53.9	72.4	58.8	46.7	70.1	50.8
RepLKNet-31L [12]	53.9	72.5	58.6	46.5	70.0	50.6
ConvNeXt-L [31]	54.8	73.8	59.8	47.6	71.3	51.7
PIIP-SBL (ours)	54.5	73.8	59.1	47.7	71.6	52.1

Comparison with SoTA

Experiments - Segmentation & Classification

Table 5: **Comparison with baseline on ADE20K using UperNet.**

Method	Crop Size	#FLOPS	mIoU
ViT-B	640^2	159G	51.0
PIIP-TSB (ours)	896/448^2/336	118G	51.6
ViT-L	640^2	545G	53.6
PIIP-SBL (ours)	1120/448^2/336	456G	54.3

Table 6: **Semantic segmentation performance on ADE20K using UperNet.**

Method	Crop Size	mIoU
Swin-B [28]	512^2	48.1
ConvNeXt-B [29]	512^2	49.1
RepLKNet-31B [11]	512^2	49.9
SLaK-B [27]	512^2	50.2
InternImage-B [46]	512^2	50.2
PIIP-TSB (ours)	896/448^2/336	**51.6**
Swin-L [28]	640^2	52.1
RepLKNet-31L [11]	640^2	52.4
ConvNeXt-L [29]	640^2	53.2
ConvNeXt-XL [29]	640^2	53.6
InternImage-L [46]	640^2	53.9
PIIP-SBL (ours)	1120/448^2/336	**54.3**

Table 7: **Image classification performance on ImageNet.** Underline indicates FLOPs or metrics on par with the baseline.

Model	Resolution	#FLOPs	Top-1 Acc
DeiT-B [42]	224	17.2G	81.8
PIIP-TSB (ours)	368/192/128	17.4G	82.1
ViT-L [40]	224	61.6G	84.0
ViT-L [40] (our impl.)	224	61.6G	85.2
PIIP-SBL (ours)	320/160/96	39.0G	85.2
PIIP-SBL (ours)	384/192/128	61.2G	85.9

Table 3: **Experiments on the large-scale vision foundation model InternViT-6B.**

Model	#Param	Mask R-CNN 1× schedule				UperNet 160k		
		#FLOPs	Resolution	AP^b	AP^m	Crop Size	#FLOPs	mIoU
InternViT-6B [8]	5919M	24418G	1024	53.8	48.1	512^2	6105G	58.36
PIIP-LH6B (ours)	7269M	5643G	1280/1024/256	53.5	47.5	640/512^2/192	1903G	57.82
	7271M	10368G	1280/1024/512	54.4	47.8	640/512^2/256	2592G	58.42
	7273M	13911G	1280/1024/640	**55.7**	**49.0**	640/512^2/384	4560G	**59.65**

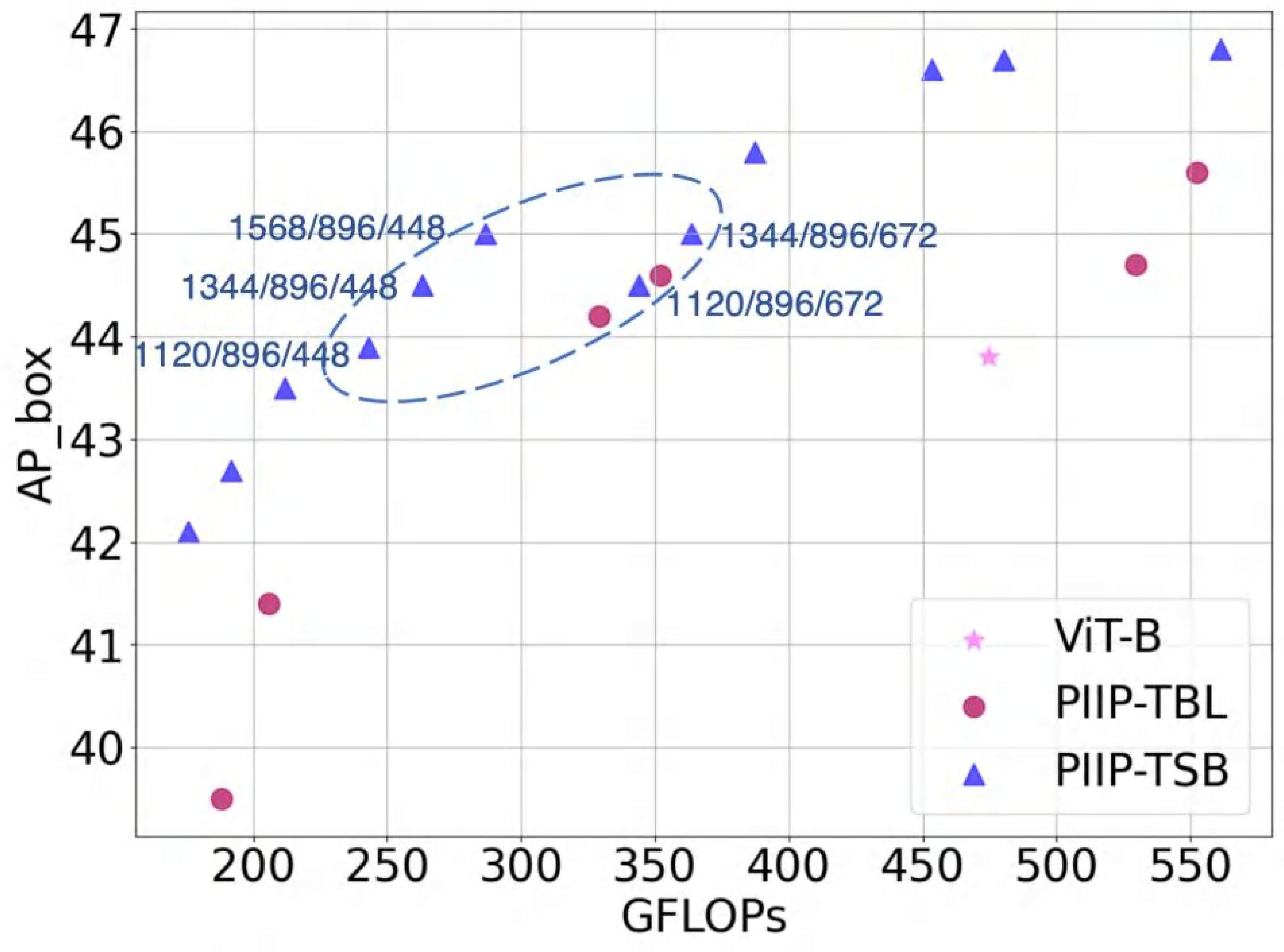
Experiments - Ablation

Table 9: **Ablation on image pyramid and parameter-inverted design.** 'PI', 'IP' and 'Inter.' represent parameter-inverted, image pyramid and interactions. 'MS' means multi-scale training, following [10].

Figure	Branches	PI	IP	Inter.	Resolution	#Param	#FLOPs	Mask R-CNN 1× schedule					
								AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Fig. 1(a)	B				1024	90M	463G	43.8	67.6	47.7	39.9	63.6	42.2
Fig. 1(b)	B		✓		MS	90M	463G	44.8	69.2	49.1	41.0	65.8	43.9
-	BBB		✓		896/448/224	262M	369G	43.3	65.8	46.6	37.9	61.5	39.6
-	BBB		✓		896/672/224	263M	457G	43.8	66.3	47.3	38.2	62.2	39.7
Fig. 1(c)	BBB		✓	✓	896/448/224	341M	466G	44.5	66.5	48.2	38.7	62.6	40.6
-	TSB			✓	896/896/896	148M	468G	44.6	66.4	48.3	39.0	62.7	41.4
Fig. 1(d)	TSB		✓	✓	448/672/896	147M	452G	42.6	64.2	45.6	36.5	59.5	38.0
Fig. 1(e)	TSB	✓	✓	✓	1568/1120/672	149M	453G	**46.6**	68.4	51.1	**41.4**	65.2	44.3
Fig. 1(a)	L				1024	308M	1542G	46.8	70.8	51.4	42.5	67.3	45.3
Fig. 1(c)	LLL		✓	✓	896/448/224	1053M	1458G	46.9	69.7	51.2	40.8	65.3	43.3
-	SBL			✓	848/848/848	495M	1539G	47.2	69.4	51.0	41.1	65.4	43.7
Fig. 1(e)	SBL	✓	✓	✓	1568/896/672	497M	1464G	**49.4**	71.9	53.9	**43.7**	68.4	46.6

Experiments - Ablation

OpenGVLab



(a) Variants with different resolutions

Table 4: **Experiments of initializing with different pre-trained weights on COCO val2017** with PIIP-SBL 1568/1120/672.

ViT-S	ViT-B / ViT-L	AP^b	AP^m
AugReg [43]	AugReg [43]	48.3	42.6
DeiT III [46]	Uni-Perceiver [66]	48.8	42.9
DeiT III [46]	MAE [18]	49.1	43.0
DeiT III [46]	DeiT III [46]	50.0	44.4
DeiT III [46]	DINOv2 [38]	51.0	44.7
DeiT III [46]	BEiTv2 [39]	51.8	45.4

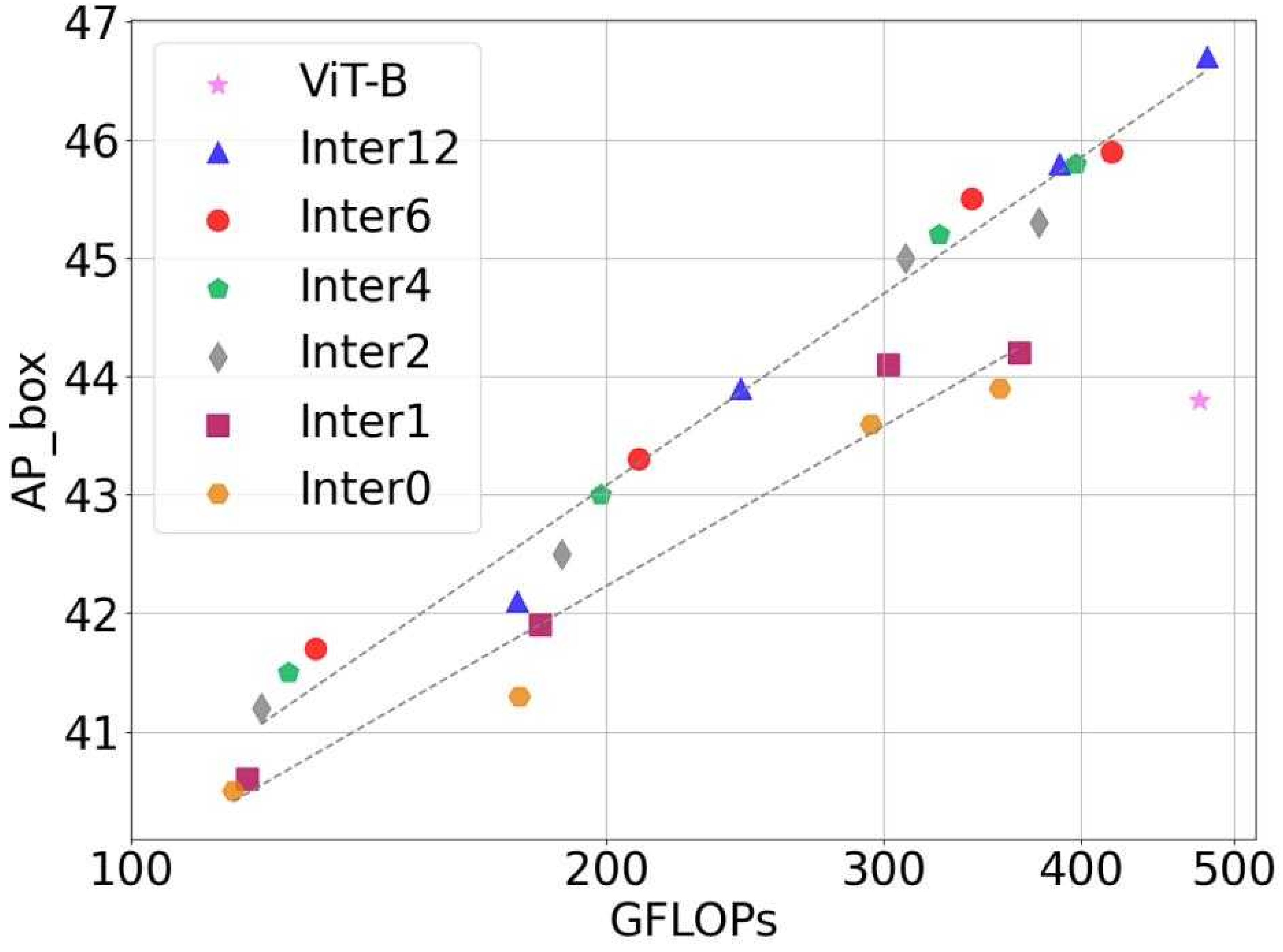
Table 8: **Ablation on Branch Merging on COCO val2017.** We use PIIP-TSB 1568/896/672.

Out Branch	AP^b	AP^m
B	43.1	37.0
S	44.7	39.1
T	45.6	40.6
B+S	45.4	39.8
B+T	46.3	41.1
S+T	46.2	40.9
B+S+T	**46.6**	**41.4**

Table 10: **Baseline with higher resolution.**

Model	Resolution	#Param	#FLOPs	AP^b
ViTDet-L	1024	308M	1542G	46.8
ViTDet-L	1792	308M	6458G	48.3
PIIP-TSBL	1792/1568/1120/448	512M	1535G	**49.6**

Experiments - Ablation

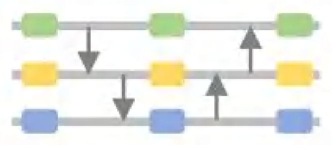
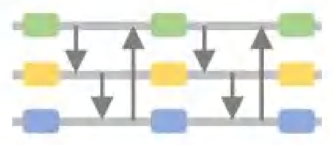
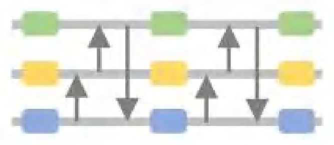
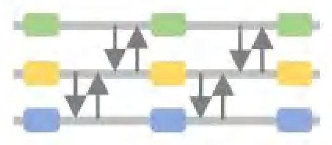
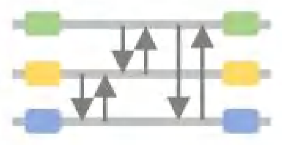


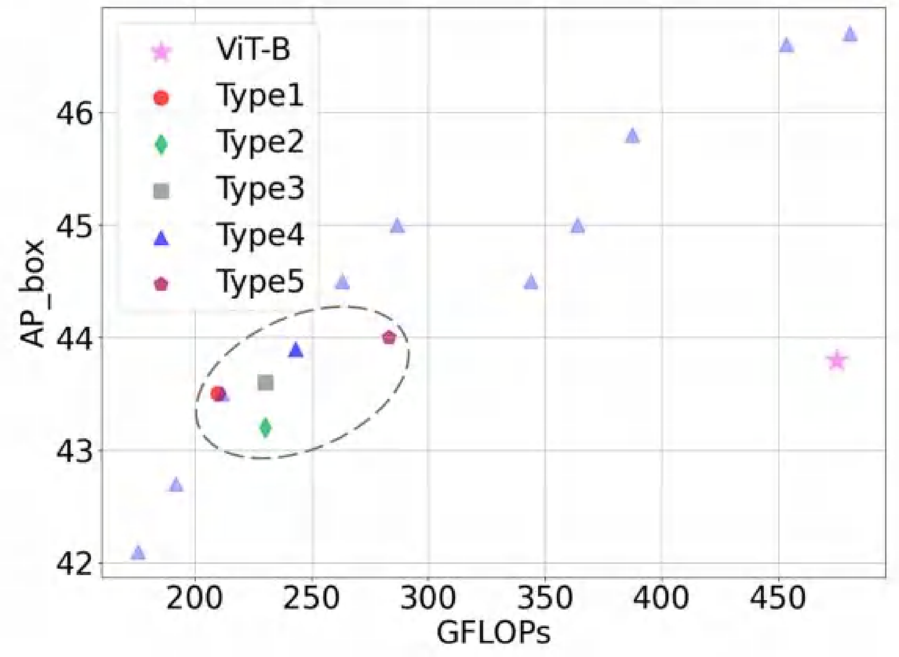
(b) Number of interactions

Table 11: **Ablation on attention type and number of interactions** with PIIP-TSB 1120/896/448.

#Interaction	Regular Attention					Deformable Attention				
	#FLOPs	AP^b	AP^b_l	AP^b_m	AP^b_s	#FLOPs	AP^b	AP^b_l	AP^b_m	AP^b_s
0	176G	41.3	59.0	44.6	22.5	176G	41.3	59.0	44.6	22.5
1	211G	41.1	59.1	44.9	22.6	182G	41.9	59.8	45.5	22.4
2	245G	41.7	59.5	45.2	22.7	187G	42.5	60.5	46.4	23.1
4	315G	41.6	59.2	45.3	22.8	198G	43.0	61.0	47.3	23.3
6	384G	42.1	59.7	45.8	23.2	210G	43.3	61.8	46.9	23.6
12	592G	42.0	60.0	45.9	23.1	243G	**43.9**	62.4	47.9	24.4

Table 12: **Ablation on interaction directions** with PIIP-TSB under resolution 1120/896/448.

Type					
#FLOPs	210G	230G	230G	243G	283G
AP^b	43.5	43.2	43.6	43.9	44.0
AP^m	38.7	38.3	38.6	38.6	38.7



Conclusion

OpenGVLab

- Introduces the Parameter-Inverted Image Pyramid Networks (PIIP) to address the computational challenges of traditional image pyramids.
- PIIP balances computational efficiency and performance with the parameter-inverted design and feature interaction mechanism.
- Experiments on detection, segmentation and classification tasks demonstrate that PIIP outperforms traditional single-branch networks while reducing computational costs.
- Provides an efficient and effective framework of multi-scale feature integration for future research.

Thanks for Listening !

Code Link: https://github.com/OpenGVLab/PIIP

Contact: wangzhaokai@sjtu.edu.cn