VISION TRANSFORMER ADAPTER FOR DENSE PREDICTIONS

ICLR 2023

汇报人：丁伯瑞

2023.06.20

Author

Zhe Chen[1,2*], Yuchen Duan[2,3*], Wenhai Wang[2✉], Junjun He[2], Tong Lu[1✉], Jifeng Dai[2,3], Yu Qiao[2]
[1]Nanjing University, [2]Shanghai AI Laboratory, [3]Tsinghua University
czcz94cz@gmail.com, {duanyuchen, wangwenhai, hejunjun}@pjlab.org.cn
lutong@nju.edu.cn, {daijifeng, qiaoyu}@pjlab.org.cn

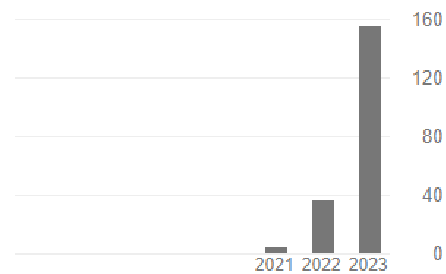
Zhe Chen

PhD candidate, Nanjing University
在 smail.nju.edu.cn 的电子邮件经过验证 - 首页
Computer Vision   Foundation Model

标题	引用次数	年份
Vision Transformer Adapter for Dense Predictions Z Chen, Y Duan, W Wang, J He, T Lu, J Dai, Y Qiao International Conference on Learning Representation (ICLR)	101	2022
Internimage: Exploring Large-scale Vision Foundation Models with Deformable Convolutions W Wang, J Dai, Z Chen, Z Huang, Z Li, X Zhu, X Hu, T Lu, L Lu, H Li, ... IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	53	2022
Towards Ultra-Resolution Neural Style Transfer via Thumbnail Instance Normalization Z Chen, W Wang, E Xie, T Lu, P Luo Proceedings of the AAAI Conference on Artificial Intelligence 36 (1), 393-400	10	2022
InternVideo-Ego4D: A Pack of Champion Solutions to Ego4D Challenges G Chen, S Xing, Z Chen, Y Wang, K Li, Y Li, Y Liu, J Wang, YD Zheng, ... Ego4D Challenge 2022 @ ECCV	7	2022
InternGPT: Solving Vision-Centric Tasks by Interacting with Chatbots Beyond Language Z Liu, Y He, W Wang, W Wang, Y Wang, S Chen, Q Zhang, Y Yang, Q Li, ... arXiv preprint arXiv:2305.05662	5	2023
DDP: Diffusion Model for Dense Visual Prediction Y Ji, Z Chen, E Xie, L Hong, X Liu, Z Liu, T Lu, Z Li, P Luo arXiv preprint arXiv:2303.17559	5	2023
FAST: Faster Arbitrarily-Shaped Text Detector with Minimalist Kernel Representation Z Chen, J Wang, W Wang, G Chen, E Xie, P Luo, T Lu arXiv preprint arXiv:2111.02394	5	2021
VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks W Wang, Z Chen, X Chen, J Wu, X Zhu, G Zeng, P Luo, T Lu, J Zhou, ... arXiv preprint arXiv:2305.11175	4	2023
SiameseCCR: A Novel Method for One-Shot and Few-Shot Chinese CAPTCHA Recognition using Deep Siamese Network	3	2020

引用次数

	总计	2018 年至今
引用	196	196
h 指数	5	5
i10 指数	3	3



开放获取的出版物数量 查看全部
0 篇文章 2 篇文章
无法查看的文章 可查看的文章
根据资助方的强制性开放获取政策

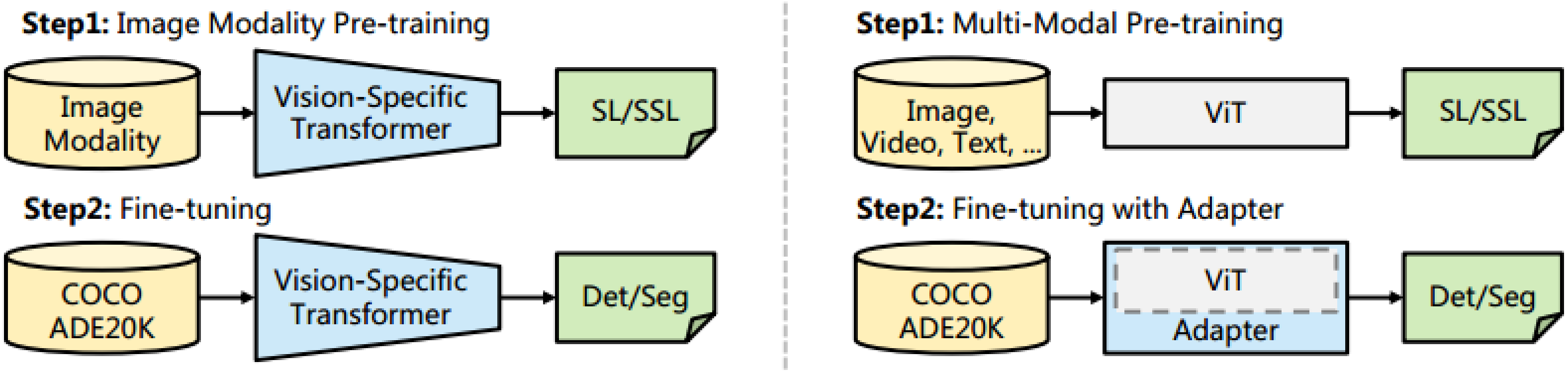
合著作者
Wenhai Wang (王文海)
Shanghai AI Laboratory
Yu Qiao
Professor of Shanghai AI Labora

Motivation

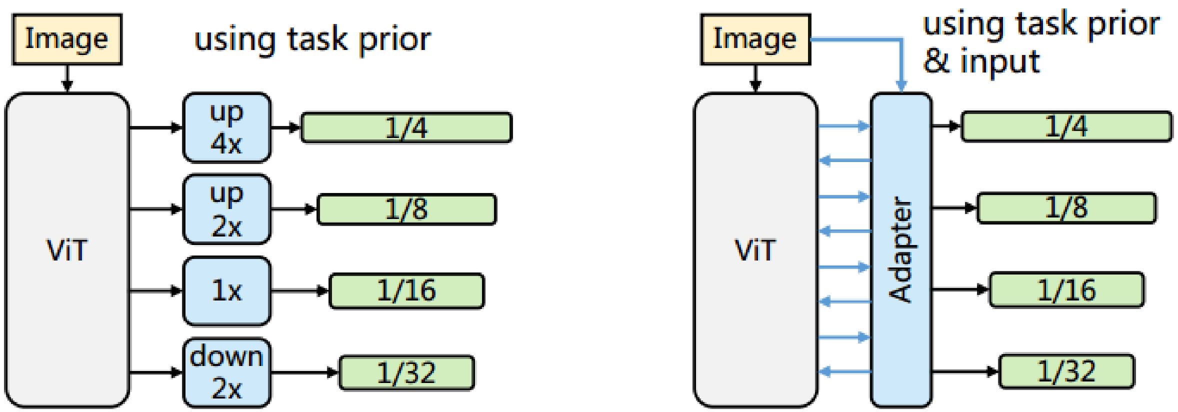
- 现在普遍的认知为ViT变体(如Swin-Transformer)会在目标检测，分割等领域可以产生更好的效果。因为其通过使用局部空间操作将视觉特定的归纳偏差引入其架构中。传统ViT缺乏图像相关的先验知识会导致收敛速度较慢，性能较低，因此在密集预测任务效果不好。

- 作者提出了一个Adapter优化方法：在不修改原始结构的情况下有效地将普通ViT适应下游密集预测任务。具体来说，为了将视觉特定的归纳偏差引入到普通ViT中，作者为ViT- adapter设计了三个定制模块。

Method

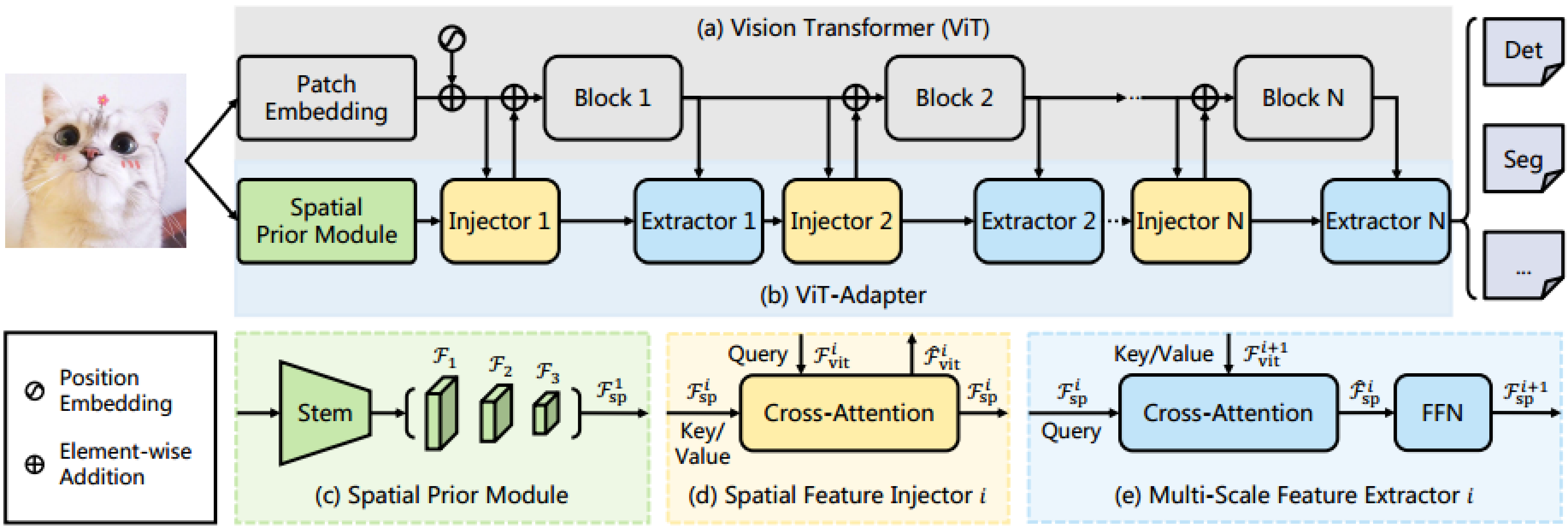


(a) Previous Paradigm
(b) Our Paradigm



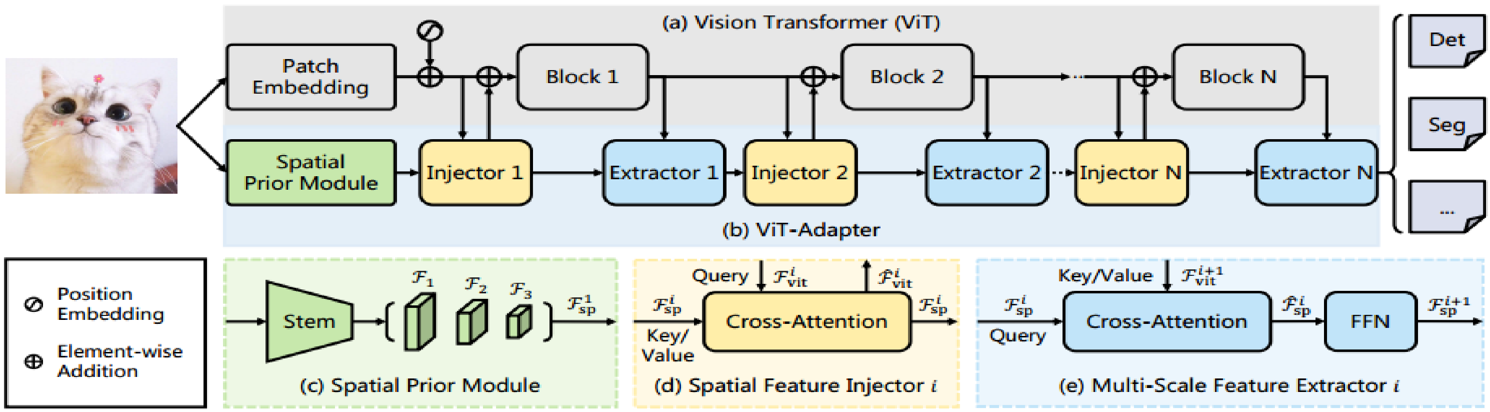
(a) Previous Method (Li et al., ViTDet)
(b) ViT-Adapter (ours)

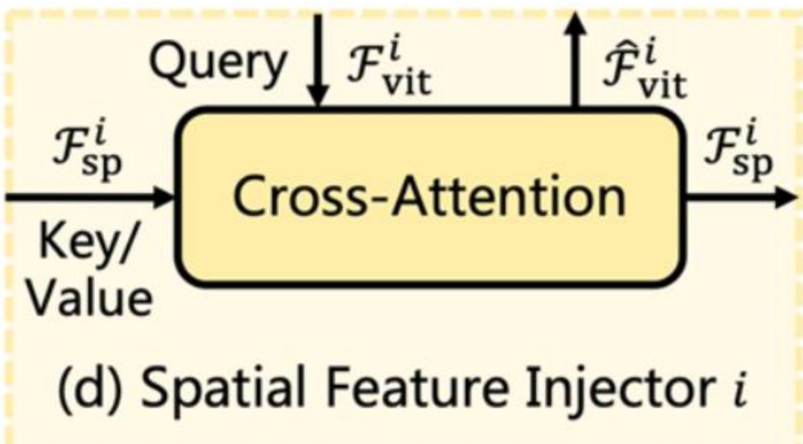
Method

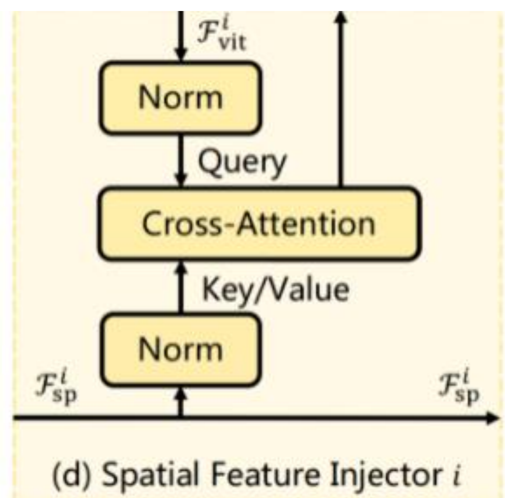


(a) Vision Transformer (ViT)
Patch Embedding
Block 1
Block 2
Block N
Spatial Prior Module
Injector 1
Extractor 1
Injector 2
Extractor 2
Injector N
Extractor N
(b) ViT-Adapter
Det
Seg
...
Position Embedding
Element-wise Addition
Stem
$\mathcal{F}_1$
$\mathcal{F}_2$
$\mathcal{F}_3$
$\mathcal{F}_{sp}^1$
(c) Spatial Prior Module
Query
$\mathcal{F}_{vit}^i$
$\hat{\mathcal{F}}_{vit}^i$
$\mathcal{F}_{sp}^i$
Key/ Value
Cross-Attention
$\mathcal{F}_{sp}^i$
(d) Spatial Feature Injector i
Key/Value
$\mathcal{F}_{vit}^{i+1}$
$\mathcal{F}_{sp}^i$
Query
Cross-Attention
$\hat{\mathcal{F}}_{sp}^i$
FFN
$\mathcal{F}_{sp}^{i+1}$
(e) Multi-Scale Feature Extractor i

Method

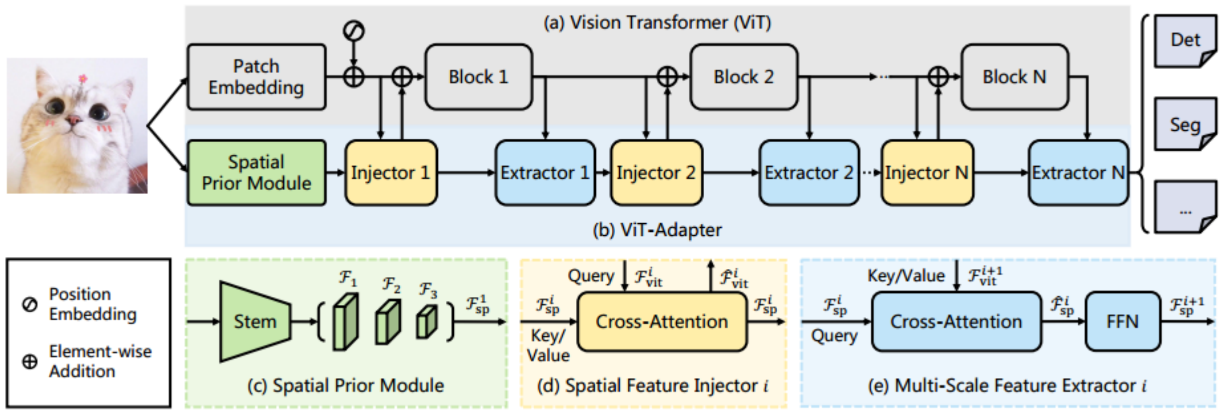


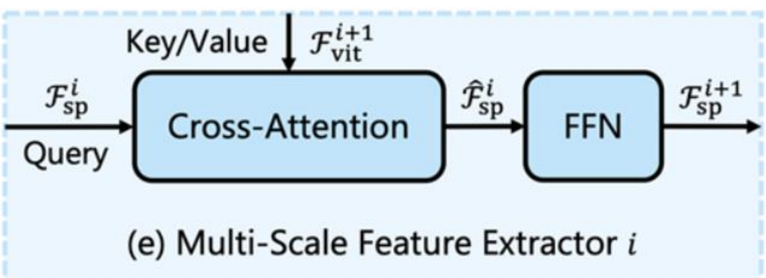


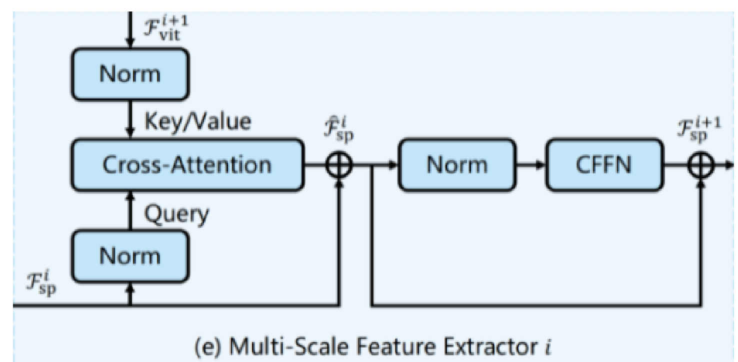


$$\hat{\mathcal{F}}_{\text{vit}}^{i} = \mathcal{F}_{\text{vit}}^{i} + \gamma^{i}\text{Attention}(\text{norm}(\mathcal{F}_{\text{vit}}^{i}), \text{norm}(\mathcal{F}_{\text{sp}}^{i}))$$

Method







$$\hat{\mathcal{F}}_{\mathrm{vit}}^{i} = \mathcal{F}_{\mathrm{vit}}^{i} + \gamma^{i}\,\mathrm{Attention}(\mathrm{norm}(\mathcal{F}_{\mathrm{vit}}^{i}), \mathrm{norm}(\mathcal{F}_{\mathrm{sp}}^{i}))$$

ARCHITECTURE CONFIGURATIONS

Variants	Settings of ViT					Settings of Adapter				Total
	Layers	Width	FFN	Heads	#Param	N	FFN	Heads	#Param	Param
Tiny (T)	12	192	768	3	5.5M	4	48	6	2.5M	8.0M
Small (S)	12	384	1536	6	21.7M	4	96	6	5.8M	27.5M
Base (B)	12	768	3072	12	85.8M	4	192	12	14.0M	99.8M
Large (L)	24	1024	4096	16	303.3M	4	256	16	23.7M	327.0M

Table 10: **Configurations of the ViT-Adapter.** We apply our adapters on four different settings of ViT, including ViT-T, ViT-S, ViT-B, and ViT-L, covering a wide range of different model sizes.

Experiment

Method	#Param (M)	Mask R-CNN 1× schedule						Mask R-CNN 3×+MS schedule					
		AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
PVT-Tiny (Wang et al., 2021)	32.9	36.7	59.2	39.3	35.1	56.7	37.3	39.8	62.2	43.0	37.4	59.3	39.9
PVTv2-B1 (Wang et al., 2022a)	33.7	41.8	64.3	45.9	38.8	61.2	41.6	44.9	67.3	49.4	40.8	64.0	43.8
ViT-T (Li et al., 2021b)	26.1	35.5	58.1	37.8	33.5	54.9	35.1	40.2	62.9	43.5	37.0	59.6	39.0
ViTDet-T (Li et al., 2022b)	26.6	35.7	57.7	38.4	33.5	54.7	35.2	40.4	63.3	43.9	37.1	60.1	39.3
ViT-Adapter-T (ours)	28.1	41.1	62.5	44.3	37.5	59.7	39.9	46.0	67.6	50.4	41.0	64.4	44.1
PVT-Small (Wang et al., 2021)	44.1	40.4	62.9	43.8	37.8	60.1	40.3	43.0	65.3	46.9	39.9	62.5	42.8
PVTv2-B2 (Wang et al., 2022a)	45.0	45.3	67.1	49.6	41.2	64.2	44.4	47.8	69.7	52.6	43.1	66.8	46.7
Swin-T (Liu et al., 2021b)	47.8	42.7	65.2	46.8	39.3	62.2	42.2	46.0	68.1	50.3	41.6	65.1	44.9
ConvNeXt-T (Liu et al., 2022)	48.1	44.2	66.6	48.3	40.1	63.3	42.8	46.2	67.9	50.8	41.7	65.0	44.9
Focal-T (Yang et al., 2021)	48.8	44.8	67.7	49.2	41.0	64.7	44.2	47.2	69.4	51.9	42.7	66.5	45.9
ViT-S (Li et al., 2021b)	43.8	40.2	63.1	43.4	37.1	59.9	39.3	44.0	66.9	47.8	39.9	63.4	42.2
ViTDet-S (Li et al., 2022b)	45.7	40.6	63.3	43.5	37.1	60.0	38.8	44.5	66.9	48.4	40.1	63.6	42.5
ViT-Adapter-S (ours)	47.8	44.7	65.8	48.3	39.9	62.5	42.8	48.2	69.7	52.5	42.8	66.4	45.9
PVTv2-B5 (Wang et al., 2022a)	101.6	47.4	68.6	51.9	42.5	65.7	46.0	48.4	69.2	52.9	42.9	66.6	46.2
Swin-B (Liu et al., 2021b)	107.1	46.9	-	-	42.3	-	-	48.6	70.0	53.4	43.3	67.1	46.7
ViT-B (Li et al., 2021b)	113.6	42.9	65.7	46.8	39.4	62.6	42.0	45.8	68.2	50.1	41.3	65.1	44.4
ViTDet-B (Li et al., 2022b)	121.3	43.2	65.8	46.9	39.2	62.7	41.4	46.3	68.6	50.5	41.6	65.3	44.5
ViT-Adapter-B (ours)	120.2	47.0	68.2	51.4	41.8	65.1	44.9	49.6	70.6	54.0	43.6	67.7	46.9
ViT-L† (Li et al., 2021b)	337.3	45.7	68.9	49.4	41.5	65.6	44.6	48.3	70.4	52.9	43.4	67.9	46.6
ViTDet-L† (Li et al., 2022b)	350.9	46.2	69.2	50.3	41.4	65.8	44.1	49.1	71.5	53.8	44.0	68.5	47.6
ViT-Adapter-L† (ours)	347.9	48.7	70.1	53.2	43.3	67.0	46.9	52.1	73.8	56.5	46.0	70.5	49.7

Table 1: **Object detection and instance segmentation with Mask R-CNN on COCO val2017.** For fair comparison, we initialize all ViT-T/S/B models with the regular ImageNet-1K pre-training (Touvron et al., 2021), and ViT-L† with the ImageNet-22K weights from (Steiner et al., 2021).

Method	Pre-train	AP^b	AP^m
Swin-B (Mask R-CNN 3×+MS)	ImageNet-1K	48.6	43.3
	ImageNet-22K	49.6	44.3
	Multi-Modal	N/A	N/A
ViT-Adapter-B (Mask R-CNN 3×+MS)	ImageNet-1K	49.6	43.6
	ImageNet-22K	50.5	44.6
	Multi-Modal	**51.2**	**45.3**

Table 4: **Comparison of different pre-trained weights.** Our method retains the flexibility of ViT and thus could benefit from advanced multi-modal pre-training (Zhu et al., 2021).

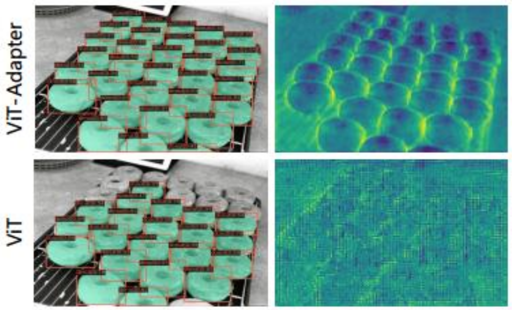
Experiment

- 消融实验：

Method	SPM	Injector	Extractor	Interaction Mode	Mask R-CNN 1× AP^b	AP^m	#Param
ViT-S (Li et al., 2021b)				-	40.2	37.1	43.8M
Variant 1	✓			Add	41.6	38.0	45.1M
Variant 2	✓	✓		Attention	42.6	38.8	46.6M
ViT-Adapter-S (ours)	✓	✓	✓	Attention	**44.7**	**39.9**	47.8M

N	AP^b	AP^m	#Param
0	40.2	37.1	43.8M
1	43.2	38.9	45.5M
2	43.9	39.4	46.2M
4	**44.7**	**39.9**	47.8M
6	44.7	39.8	49.4M

Attention Mechanism	Complexity	AP^b	AP^m	FLOPs	#Param	Train Time	Memory
Global Attention (Vaswani et al., 2017)	Quadratic	43.7	39.3	1080G	50.3M	1.61s	*19.0G
CSwin Attention (Dong et al., 2021)	Linear	43.5	39.2	456G	50.3M	0.56s	15.6G
Pale Attention (Wu et al., 2022a)	Linear	44.2	39.8	458G	50.3M	0.75s	17.4G
Deformable Attention (Zhu et al., 2020)	Linear	**44.7**	**39.9**	**403G**	**47.8M**	**0.36s**	**13.7G**



(b) Detection Results (c) Stride-8 Feature

Summary

- 本文提出了一种新的adapter方法，可以利用在普通的ViT上
- 因为特定的transformer无法利用多模态，该adapter可以利用多模态预训练模型
- 能够学习到CNN关注的纹理等信息，可以有效地加入到ViT中。