Vision Transformer

Macro Structure, Optimization, and High-level Applications

Xinggang Wang

Huazhong University of Science and Technology

https://xinggangw.info

Vision Transformer (ViT)

- **Macro Structure:**
  - Locality和Hierarchy的引入使得ViT在CV中独特的设计（不同于NLP）。在很多任务中都取得了极佳的性能。
  - 相比具体的算子(Attention or Conv or MLP), Macro Structure的对结果的影响更为重要。

- **Optimization:**
  - 目前Transformer作为视觉主干网络的成功,训练上和优化上的贡献也是不能忽视的。

- **High-level Applications:**
  - ViT以及Query Token的引入启发了新的框架设计。通过Query Token可以灵活高效的提取多尺度物体特征，实现高性能的物体检测和实例分割。

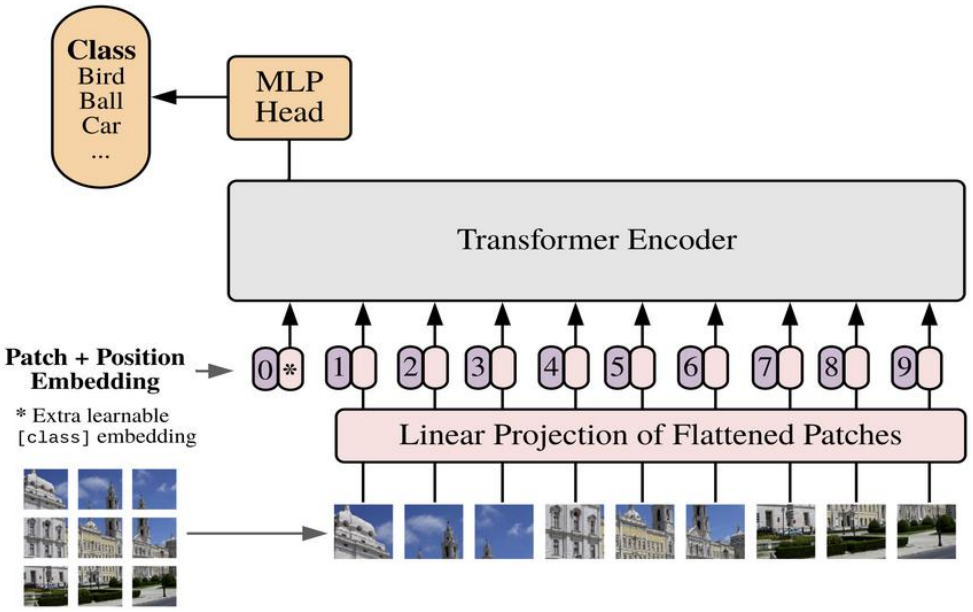
Vision Transformer (ViT)

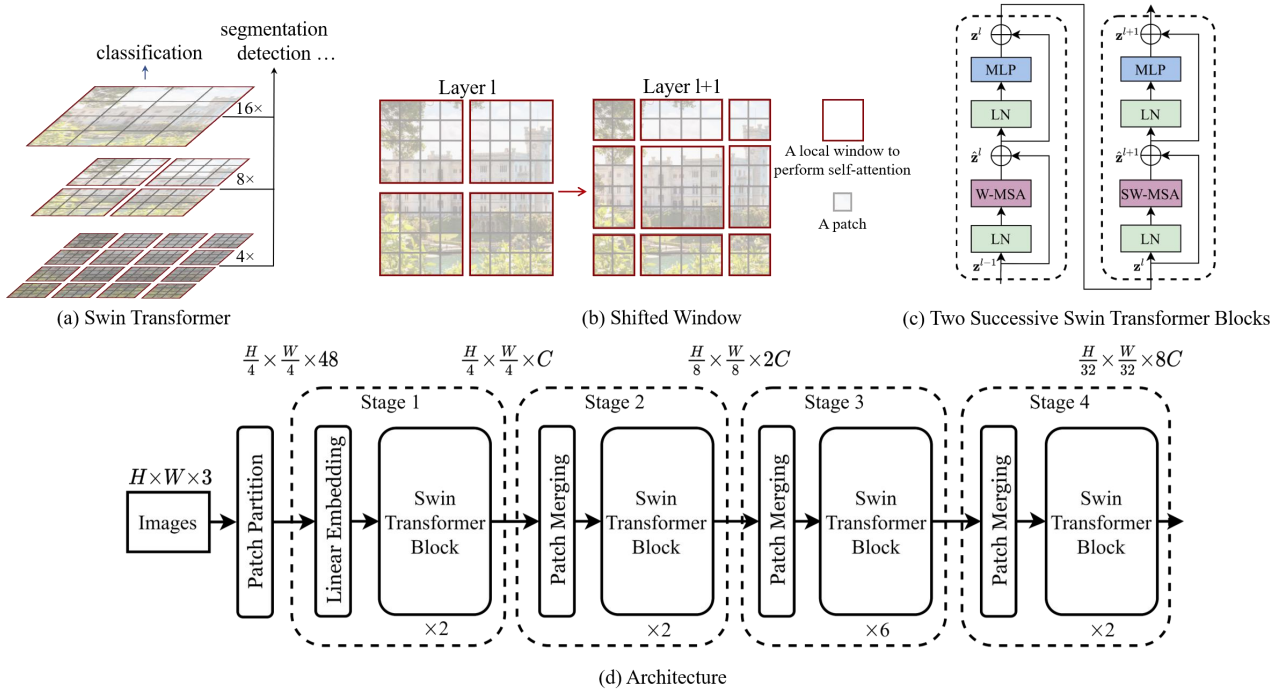
- **Macro Structure:**
  - Locality和Hierarchy的引入使得ViT在CV中独特的设计（不同于NLP）。在很多任务中都取得了极佳的性能。
  - 相比具体的算子(Attention or Conv or MLP), Macro Structure的对结果的影响更为重要。

- **Optimization:**
  - 目前Transformer作为视觉主干网络的成功,训练上和优化上的贡献也是不能忽视的。

- **High-level Applications:**
  - ViT以及Query Token的引入启发了新的框架设计。通过Query Token可以灵活高效的提取多尺度物体特征，实现高性能的物体检测和实例分割。

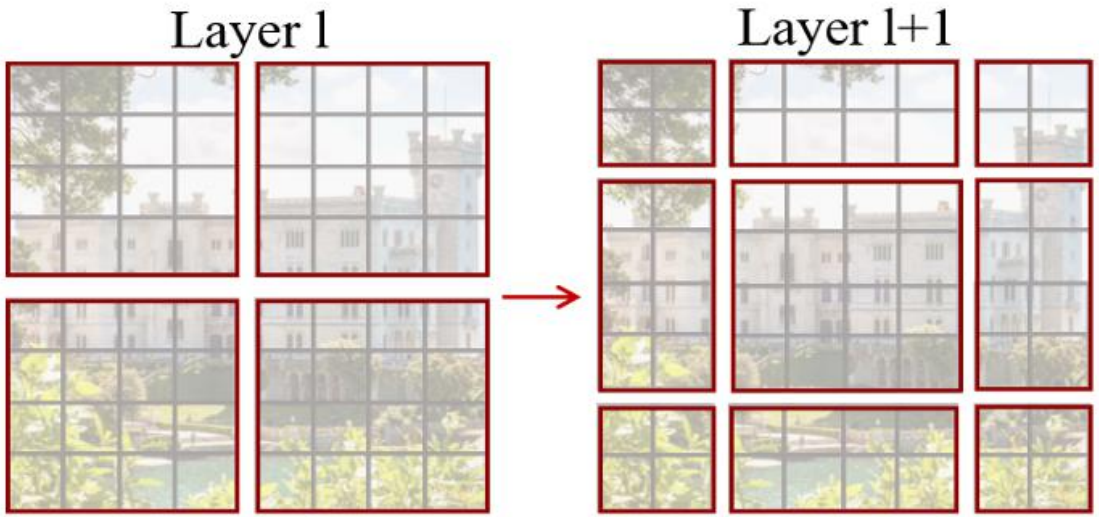
From ViT to Swin Transformer



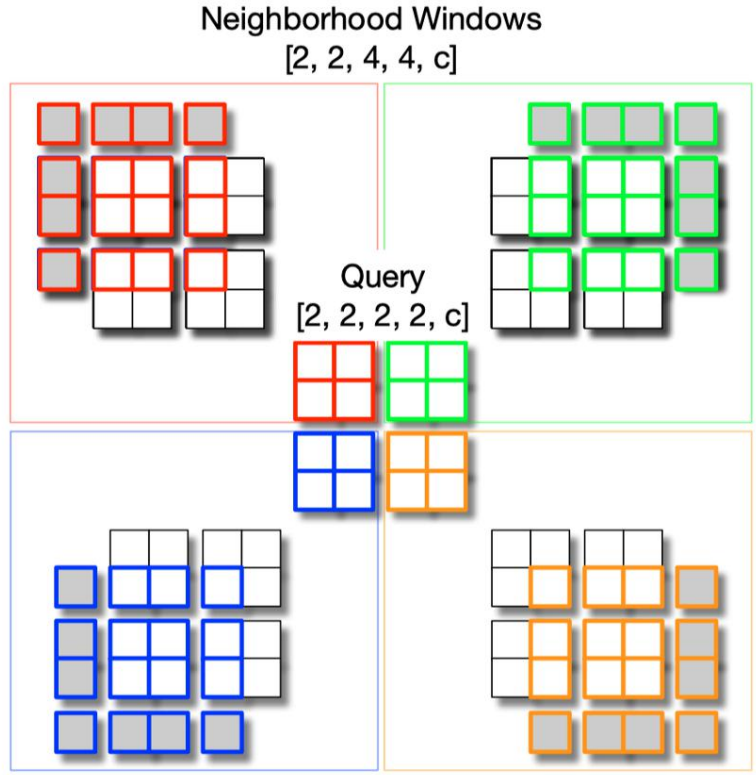


在计算机视觉中，从最开始的ViT到以Swin为代表的层次化VT，引入区域性（locality）和多层次特征表达（hierarchy）非常重要

Locality: local window and cross-window interaction

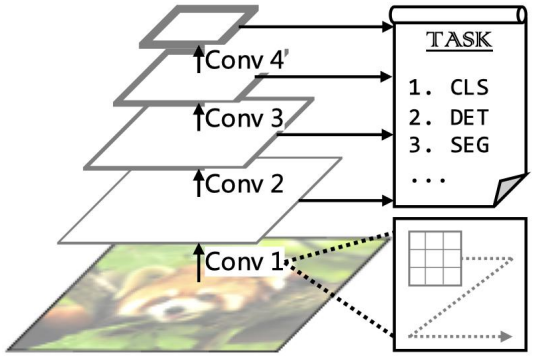


Shift window in Swin Trans

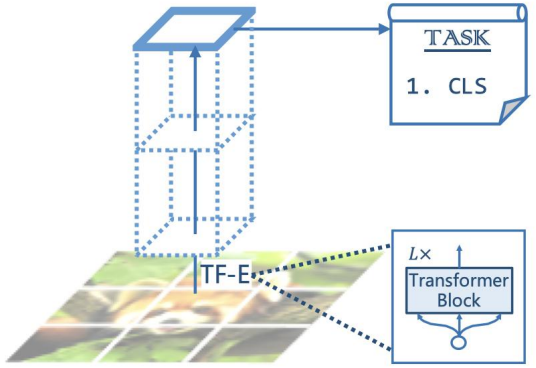


Overlapped window in HaloNet

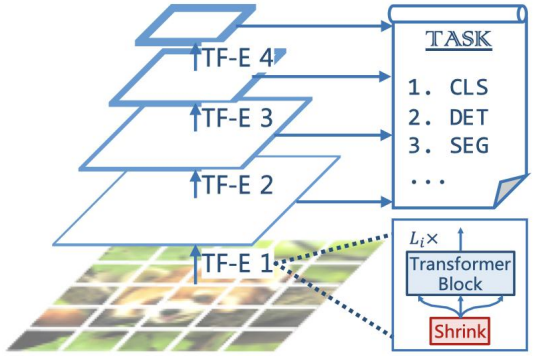
Hierarchy



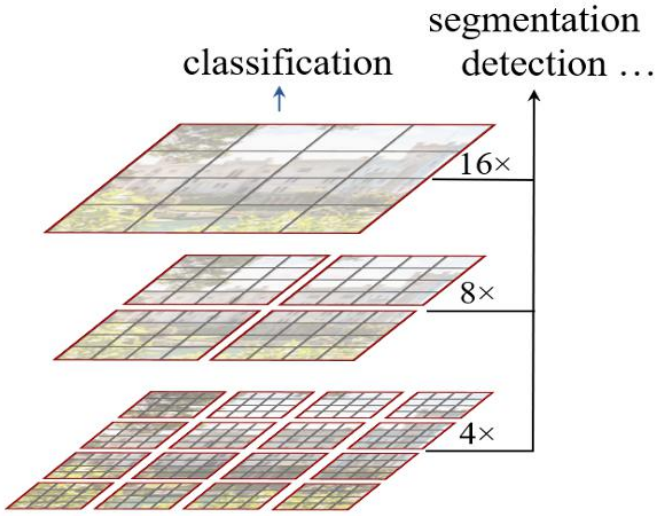
(a) CNNs: VGG [41], ResNet [15], *etc.*



(b) Vision Transformer [10]



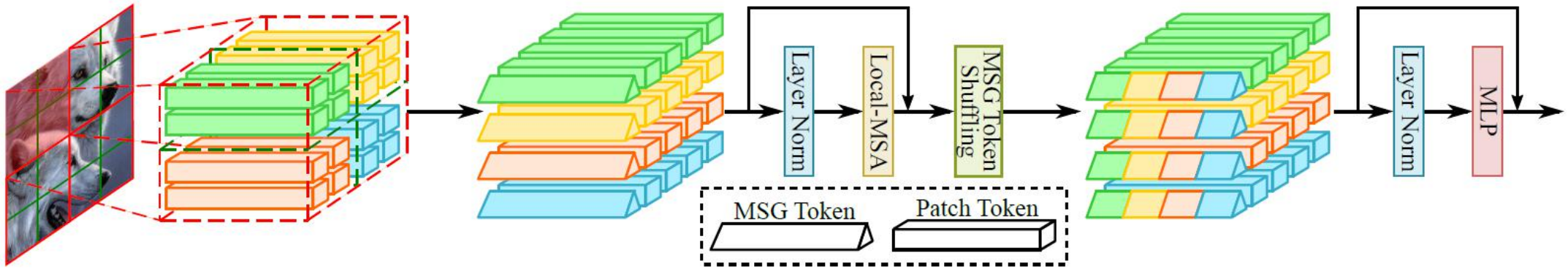
(c) Pyramid Vision Transformer (ours)

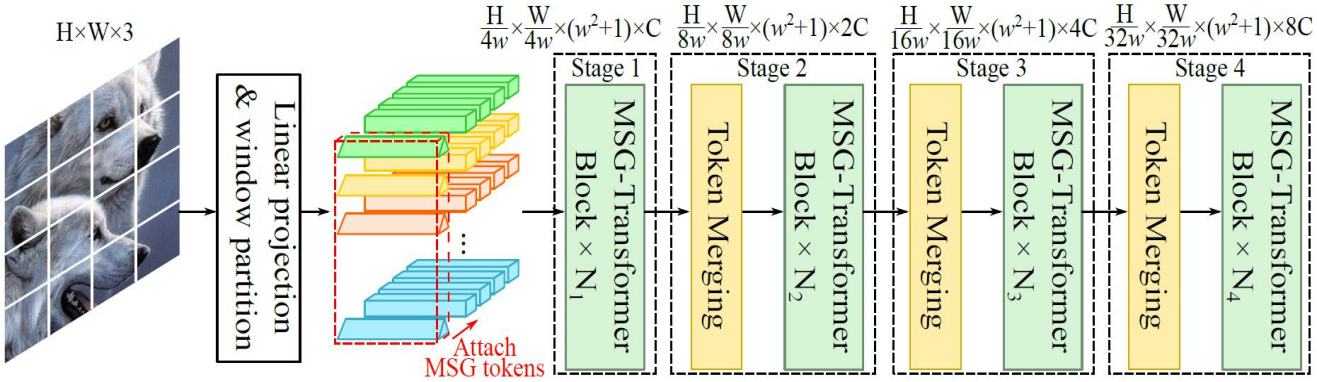


W. Wang, et al., Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. ICCV 2021

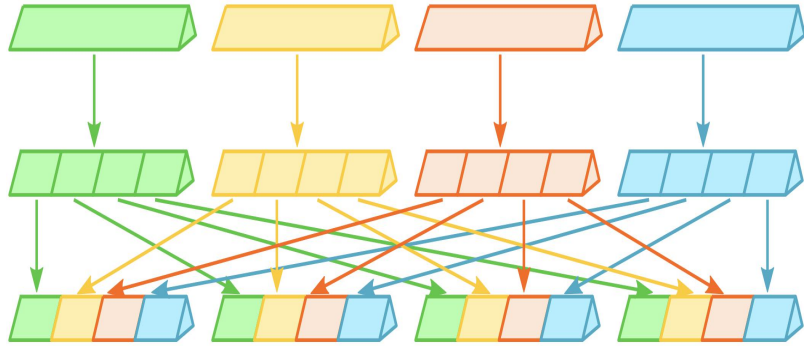
Macro Structure:

从Vanilla ViT到Hierarchical Local Window-based ViT: MSG Transformer





MSG Transformer



MSG Token Shuffle

Use MSG tokens to extract information from local windows and shuffle operation to exchange information between MSG tokens.

J. Fang, et al., MSG-Transformer: Exchanging Local Spatial Information by Manipulating Messenger Tokens, arXiv:2105.15168

Macro Structure: MSG Transformer

Method	Input size	Params	FLOPs	GPU throughput (images / s)	CPU latency	Top-1 Acc (%)
Transformer Networks						
DeiT-S [43]	224^2	22M	4.6G	898.3	118ms	79.8
T2T-ViT$_t$-14 [52]	224^2	22M	5.2G	559.3	225ms	80.7
PVT-Small [47]	224^2	25M	3.8G	749.0	146ms	79.8
TNT-S [13]	224^2	24M	5.2G	387.1	215ms	81.3
CoaT Mini [51]	224^2	10M	6.8G	-	-	80.8
Swin-T [28]	224^2	28M	4.5G	692.1	189ms	81.3
MSG-Transformer-T	224^2	24M	4.0G	711.8	151ms	**81.6**
DeiT-B [43]	224^2	87M	17.5G	278.9	393ms	81.8
T2T-ViT$_t$-19 [52]	224^2	39M	8.4G	377.3	314ms	81.4
T2T-ViT$_t$-24 [52]	224^2	64M	13.2G	268.2	436ms	82.2
PVT-Large [47]	224^2	61M	9.8G	337.1	338ms	81.7
TNT-B [13]	224^2	66M	14.1G	231.1	414ms	82.8
Swin-S [28]	224^2	50M	8.7G	396.6	346ms	83.0
MSG-Transformer-S	224^2	54M	8.9G	399.1	269ms	**83.0**
ViT-B/16 [11]	384^2	87M	55.4G	81.1	1218ms	77.9
ViT-L/16 [11]	384^2	307M	190.7G	26.3	4420ms	76.5
DeiT-B [43]	384^2	87M	55.4G	81.1	1213ms	83.1
Swin-B [28]	224^2	88M	15.4G	257.6	547ms	83.3
MSG-Transformer-B	224^2	95M	15.8G	258.1	437ms	**83.6**

Macro Structure: MSG Transformer

Method	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	Params	FLOPs	FPS
DeiT-S	48.0	67.2	51.7	41.4	64.2	44.3	80M	889G	-
ResNet-50	46.3	64.3	50.5	40.1	61.7	43.4	82M	739G	10.5
Swin-T	50.5	69.3	54.9	43.7	66.6	47.1	86M	745G	9.4
MSG-Transformer-T	**51.0**	**69.9**	**55.3**	**44.3**	**67.5**	**47.9**	82M	735G	9
ResNeXt101-32×4d	48.1	66.5	52.4	41.6	63.9	45.2	101M	819G	7.5
Swin-S	51.8	70.4	56.3	44.7	67.9	48.5	107M	838G	7.5
MSG-Transformer-S	**52.1**	**70.6**	**56.8**	**45.1**	**68.4**	**48.9**	111M	842G	7.5
ResNeXt101-64×4d	48.3	66.4	52.3	41.7	64.0	45.1	140M	972G	6.0
Swin-B	51.9	70.9	56.5	45.0	68.4	48.7	145M	982G	6.3
MSG-Transformer-B	**52.0**	70.8	56.2	**45.1**	68.3	**48.9**	152M	989G	6.2

Macro Structure: 采用和Swin相同Macro Structure的CNN(DW Conv)

Swin	DW Conv.
concat 4×4, linear 96-d, LN	concat 4×4, linear 96-d, LN
[LN, linear 96x3-d local sa. 7×7, head 3 linear 96-d LN, linear 384-d GELU, linear 96-d] × 2	[linear 96-d, BN, ReLU depthwise conv. 7×7, BN, ReLU linear 96-d, BN, ReLU BN, linear 384-d GELU, linear 96-d] × 2
concat 2×2, linear 192-d , LN	concat 2×2, linear 192-d , LN
[LN, linear 192x3-d local sa. 7×7, head 6 linear 192-d LN, linear 768-d GELU, linear 192-d] × 2	[linear 192-d, BN, ReLU depthwise conv. 7×7, BN, ReLU linear 192-d, BN, ReLU BN, linear 768-d GELU, linear 192-d] × 2

这里仅列出前两个stage的config

Q. Han, et al., Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight

Macro Structure: 采用和Swin相同Macro Structure的CNN(DW Conv)

Local attention: perform attention in local small windows					IN-1k Acc.	Real Acc.
Swin-T [35]	224^2	28M	4.5G	713.5	81.3	86.6
Swin-B [35]	224^2	88M	15.4G	263.0	83.3	87.9
Depth-wise convolution + point-wise 1×1 *convolution*						
DW-Conv.-T	224^2	24M	3.8G	928.7	81.3	86.8
DW-Conv.-B	224^2	74M	12.9G	327.6	83.2	87.9
D-DW-Conv.-T	224^2	51M	3.8G	897.0	81.9	87.3
D-DW-Conv.-B	224^2	162M	13.0G	322.4	83.2	87.9

	COCO Object Detection						ADE20K Semantic Segmentation		
	#param.	FLOPs	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	#param.	FLOPs	mIoU
Swin-T	86M	747G	50.5	69.3	54.9	43.7	60M	947G	44.5
DW Conv.-T	82M	730G	49.9	68.6	54.3	43.4	56M	928G	45.5
D-DW Conv.-T	108M	730G	50.5	69.5	54.6	43.7	83M	928G	45.7
Swin-B	145M	986G	51.9	70.9	56.5	45.0	121M	1192G	48.1
DW Conv.-B	132M	924G	51.1	69.6	55.4	44.2	108M	1129G	48.3
D-DW Conv.-B	219M	924G	51.2	70.0	55.4	44.4	195M	1129G	48.0

Q. Han, et al., Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight

Macro Structure: 采用和Swin相同Macro Structure的CNN

- CNN也可以极大的**从更现代的训练和优化策略**获益。在各个CV任务上取得极有竞争力的表现。
- 从某种意义上说， Vision Transformer的发展也促使人们反思卷积网络。**经典的卷积网络也得到了一轮复兴**[2,3,4]。
- CNN和Transformer之间的相互借鉴和融合，可以产生性能更好的Backbone网络[5]。

[1] H. Touvron, et al., Training data-efficient image transformers & distillation through attention, ICLR 2021
[2] I. Bello, et al., Revisiting ResNets: Improved Training and Scaling Strategies, NeurIPS 2021
[3] Q. Han, et al., Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight, arXiv 2106.04263
[4] R. Wightman, et al., ResNet strikes back: An improved training procedure in timm, arXiv 2110.00476
[5] W. Wu, et al. Co-scale conv-attentional image transformers, ICCV 2021

Macro Structure: 采用和Swin相同Macro Structure的MLP

	MHSA	Linear	DW Linear	MLP
Shift	80.5	79.7	79.8	79.9
Shuffle	80.6	79.6	79.6	79.8
MSG	80.4	79.5	79.7	79.7

Y. Fang, et al., What Makes for Hierarchical Vision Transformer? arXiv:2107.02174

- 具有Locality和Hierarchy的MLP-Like模型也可以取得非常有竞争力的结果。
- 在相同的Macro Structure下，不同的cross window communication方式得到的结果类似。

Macro Structure: 相同Macro Structure下的不同aggravation layer

Current "Most Interesting" **ConvMixer** Configurations *vs.* Other Simple Models							
Network	Patch Size	Kernel Size	# Params ($\times 10^6$)	Throughput (img/sec)	Act. Fn.	# Epochs	ImNet top-1 (%)
ConvMixer-1536/20	7	9	51.6	89	G	150	81.37
ConvMixer-768/32	7	7	21.1	203	R	300	80.16
ResNet-152	–	3	60.2	872	R	150	79.64
DeiT-B	16	–	86	703	G	300	81.8
ResMLP-B24/8	8	–	129	140	G	400	81.0

Patches Are All You Need ?

	Conv	MHSA	Linear & MLP
Columnar	"Missing Piece"	✓	✓
Hierarchical	✓	✓	**"Missing Piece"**

Y. Fang, et al., What Makes for Hierarchical Vision Transformer? (Ours)

	MHSA	Linear	DW Linear	MLP
Shift	80.5	79.7	79.8	79.9
Shuffle	80.6	79.6	79.6	79.8
MSG	80.4	79.5	79.7	79.7

Macro Structure: 相同Macro Structure下的不同aggravation layer

Patches Are All You Need ?

	Conv	MHSA	Linear & MLP
Columnar	"Missing Piece"	✓	✓
Hierarchical	✓	✓	**"Missing Piece"**

Y. Fang, et al., What Makes for Hierarchical Vision Transformer?

相同Macro Structure下，不同aggravation layer差异不大
主要矛盾在Macro Structure Design

Vision Transformer (ViT)

- **Macro Structure:**
  - Locality和Hierarchy的引入使得ViT在CV中独特的设计（不同于NLP）。在很多任务中都取得了极佳的性能。
  - 相比具体的算子(Attention or Conv or MLP), Macro Structure的对结果的影响更为重要。

- **Optimization:**
  - 目前Transformer作为视觉主干网络的成功,训练上和优化上的贡献也是不能忽视的。

- **High-level Applications:**
  - ViT以及Query Token的引入启发了新的框架设计。通过Query Token可以灵活高效的提取多尺度物体特征，实现高性能的物体检测和实例分割。

Optimization

- 相比CNN，Vision Transformer对训练和优化的要求较为苛刻。一些简单朴素的训练方式无法成功训练Vision Transformer [1]。

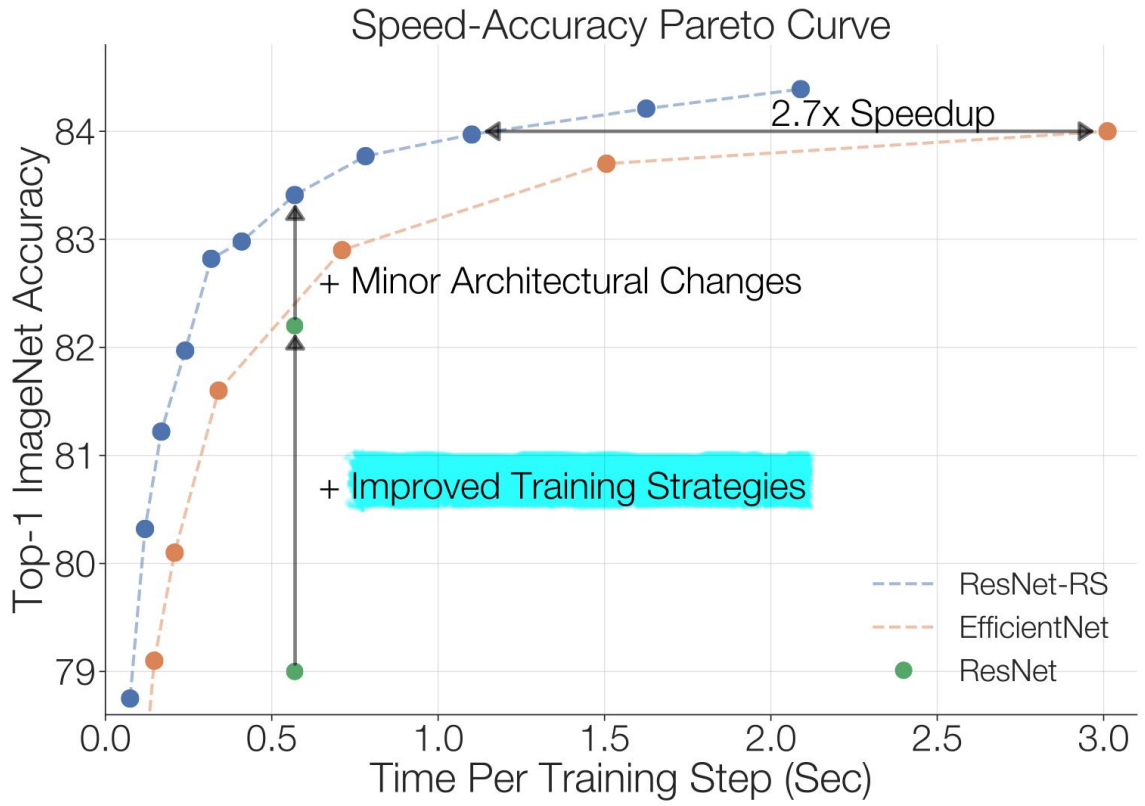
Ablation on ↓	Pre-training	Fine-tuning	Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	top-1 accuracy: pre-trained 224^2	top-1 accuracy: fine-tuned 384^2
none: DeiT-B	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8 ±0.2	83.1 ±0.1
optimizer	SGD	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
	adamw	SGD	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
data augmentation	adamw	adamw	✗	✗	✓	✓	✓	✓	✓	✗	✗	79.6	80.4
	adamw	adamw	✗	✓	✓	✓	✓	✓	✓	✗	✗	81.2	81.9
	adamw	adamw	✓	✗	✗	✓	✓	✓	✓	✗	✗	78.7	79.8
	adamw	adamw	✓	✗	✓	✗	✓	✓	✓	✗	✗	80.0	80.6
	adamw	adamw	✓	✗	✗	✗	✓	✓	✓	✗	✗	75.8	76.7
regularization	adamw	adamw	✓	✗	✓	✓	✗	✓	✓	✗	✗	4.3*	0.1
	adamw	adamw	✓	✗	✓	✓	✓	✗	✓	✗	✗	3.4*	0.1
	adamw	adamw	✓	✗	✓	✓	✓	✓	✗	✗	✗	76.5	77.4
	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✓	✗	81.3	83.1
	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✓	81.9	83.1

e.g., Erasing和Stoch. Depth对于DeiT的收敛来说是必要的

[1] H. Touvron, et al., Training data-efficient image transformers & distillation through attention

Optimization

- 同样，CNN也可以极大的从更现代的训练和优化策略获益。在各个CV任务上取得极有竞争力的表现 [1,2,3]。



Improvements	Top-1	Δ
ResNet-200	79.0	—
+ Cosine LR Decay	79.3	**+0.3**
+ Increase training epochs	78.8 †	-0.5
+ EMA of weights	79.1	**+0.3**
+ Label Smoothing	80.4	**+1.3**
+ Stochastic Depth	80.6	**+0.2**
+ RandAugment	81.0	**+0.4**
+ Dropout on FC	80.7 ‡	-0.3
+ Decrease weight decay	82.2	**+1.5**
+ Squeeze-and-Excitation	82.9	**+0.7**
+ ResNet-D	83.4	**+0.5**

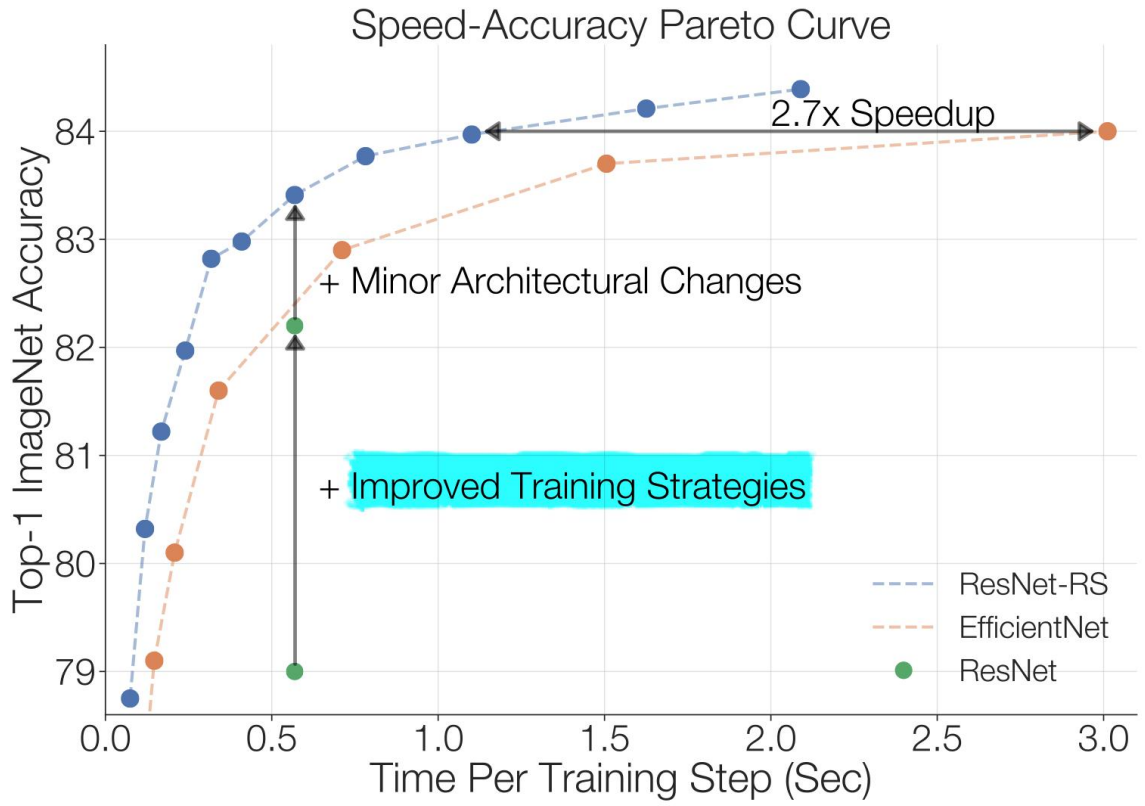
Training Methods , **Regularization Methods**

Architecture Improvements

[1] I. Bello, et al., Revisiting ResNets: Improved Training and Scaling Strategies
[2] Q. Han, et al., Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight
[3] R. Wightman, et al., ResNet strikes back: An improved training procedure in timm

Optimization

- 从某种意义上说， Vision Transformer的发展也促使人们反思卷积网络。经典的卷积网络也得到了一轮复兴[1,2,3]。



Improvements	Top-1	Δ
ResNet-200	79.0	—
+ Cosine LR Decay	79.3	**+0.3**
+ Increase training epochs	78.8 †	-0.5
+ EMA of weights	79.1	**+0.3**
+ Label Smoothing	80.4	**+1.3**
+ Stochastic Depth	80.6	**+0.2**
+ RandAugment	81.0	**+0.4**
+ Dropout on FC	80.7 ‡	-0.3
+ Decrease weight decay	82.2	**+1.5**
+ Squeeze-and-Excitation	82.9	**+0.7**
+ ResNet-D	83.4	**+0.5**

Training Methods , **Regularization Methods**

Architecture Improvements

[1] I. Bello, et al., Revisiting ResNets: Improved Training and Scaling Strategies
[2] Q. Han, et al., Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight
[3] R. Wightman, et al., ResNet strikes back: An improved training procedure in timm

Optimization

表中，A2和T2表示两种不同的训练&优化策略

R. Wightman, et al., ResNet strikes back: An improved training procedure in timm

	test set →	ImageNet-val	
↓ architecture	training →	A2	T2
ResNet-50		79.9	79.2
DeiT-S		79.6	80.4

使用A2训练，ResNet-50的表现优于DeiT-S

使用T2训练，DeiT-S的表现优于ResNet-50

结构的表现和训练策略深度耦合，不同的结构有其适配的优化方式。很难“公平”的去比较不同的结构。

Vision Transformer (ViT)

- **Macro Structure:**
  - Locality和Hierarchy的引入使得ViT在CV中独特的设计（不同于NLP）。在很多任务中都取得了极佳的性能。
  - 相比具体的算子(Attention or Conv or MLP), Macro Structure的对结果的影响更为重要。

- **Optimization:**
  - 目前Transformer作为视觉主干网络的成功,训练上和优化上的贡献也是不能忽视的。

- **High-level Applications:**
  - ViT以及Query Token的引入启发了新的框架设计。通过Query Token可以灵活高效的提取多尺度物体特征，实现高性能的物体检测和实例分割。

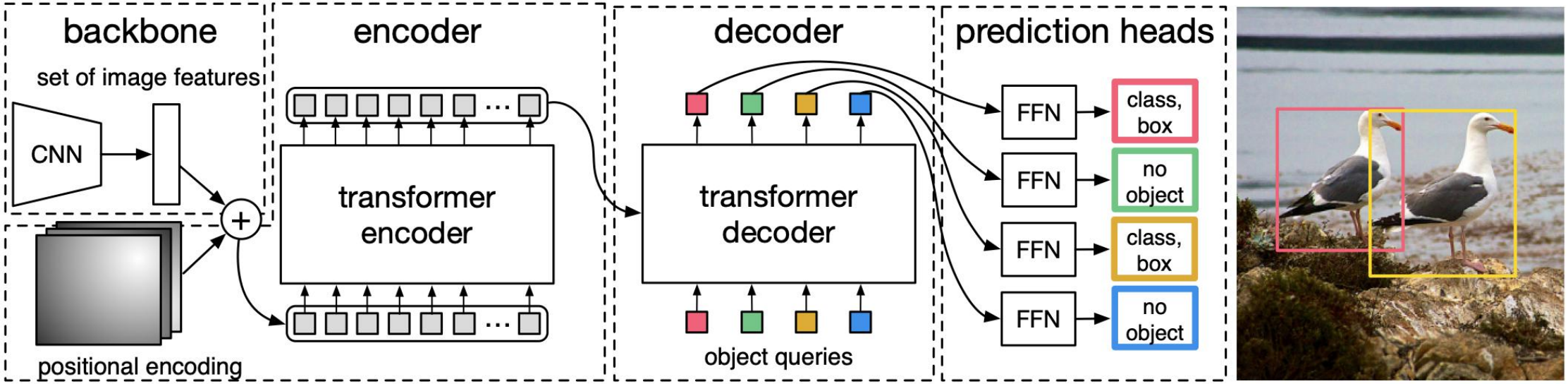
High-level Applications

- 在CV中，Transformer最先成功应用于2D目标检测(DETR)。
- 目前，Transformer已经被成功的应用在：
  - Tracking (e.g., Transformer Tracking[2])
  - MOT (e.g., TransTrack[3], TrackFormer[4], TransMOT[5])
  - Multi-modal Understanding (e.g., MDETR[6])
  - 2D Object Detection (e.g., DETR, Deformable DETR, Pix2Seq[7])
  - 3D Object Detection (e.g., 3DETR[8])
  - …

[1] H. Chen, et al., Pre-Trained Image Processing Transformer
[2] X. Chen , et al., Transformer Tracking
[3] P. Sun, et al., TransTrack: Multiple Object Tracking with Transformer
[4] T. Meinhardt, et al., TrackFormer: Multi-Object Tracking with Transformers
[5] P. Chu, et al., TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking
[6] A. Kamath, et al., MDETR: Modulated Detection for End-to-End Multi-Modal Understanding
[7] T. Chen, et al., Pix2seq: A Language Modeling Framework for Object Detection
[8] I. Misra, et al., An End-to-End Transformer Model for 3D Object Detection

Detection Transformer: DETR

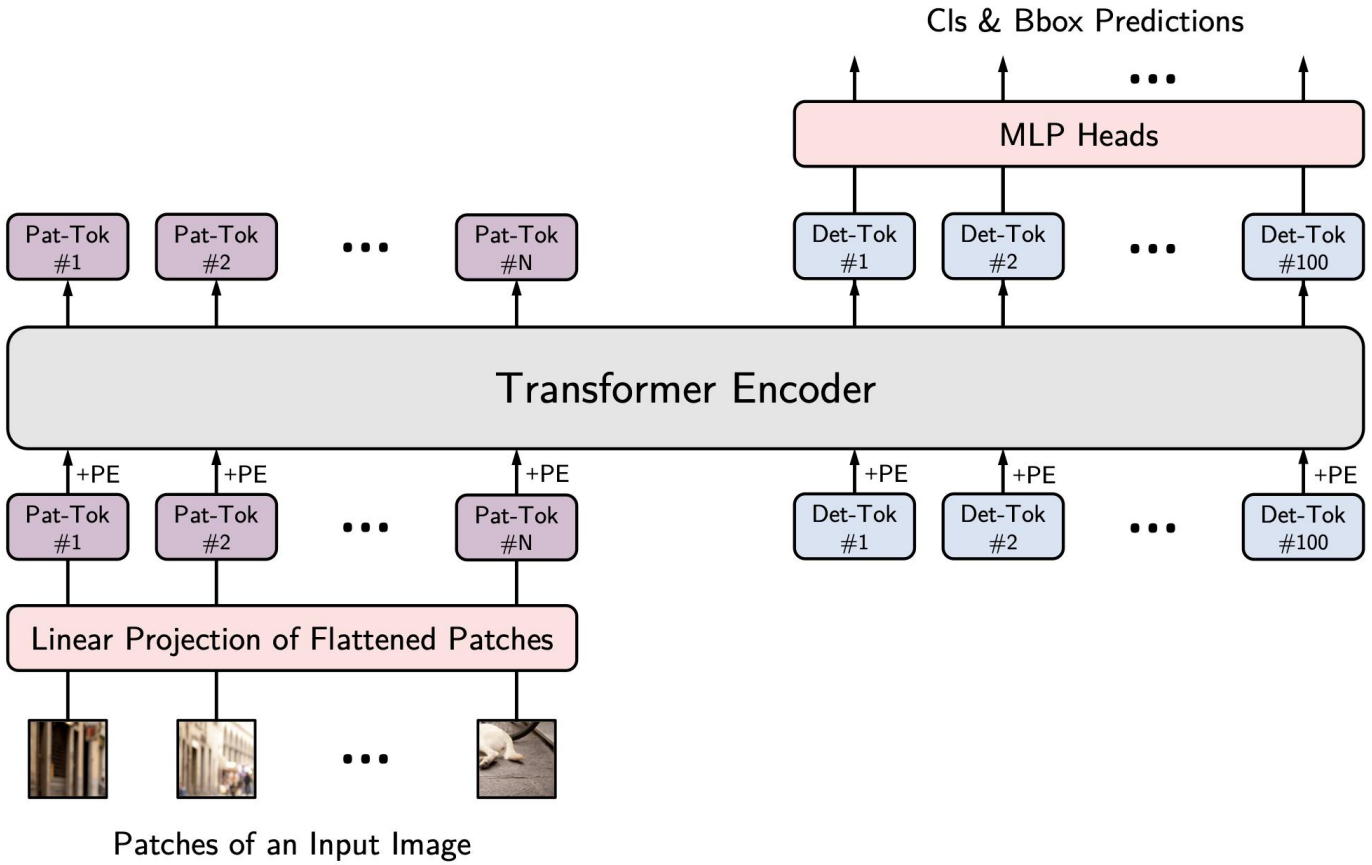
- 在CV中，Transformer最先成功应用于2D目标检测(DETR[1])。
- DETR采用Transformer Encoder & Decoder结构：
  - Encoder用于增强CNN Feature。
  - Decoder接收100个Object Query作为输入，通过Query之间的Self-attention以及Query和Encoder Feature之间的Cross-attention进行Parallel Decoding。
  - Decoder的输出的100个Object Query和GT进行最优匹配来完成label assignment。



[1] N. Carion, et al., End-to-End Object Detection with Transformers, ECCV 2020

Detection Transformer: YOLOS

- YOLOS[1]采用**纯粹Transformer Encoder结构**。将[PATCH] tokens和[DET] tokens拼接在一起作为输入。**没有做显式的Decoding**。



[1] Y. Fang, et al., You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection, NeurIPS 2021

Detection Transformer: YOLOS

- YOLOS表明2D Object Detection可以以纯粹Seq2Seq建模的方式实现。
- YOLOS是依赖2D归纳偏置和目标检测先验知识尽可能小的检测器之一。(Seq2Pix也在往这个方向研究)
- YOLOS可以将ViT预训练得到的表示迁移到2D Object Detection任务中。
- 为什么要做归纳偏置极小的Detection Transformer?
  - 有利于构建通用的Transformer大模型（类似NLP中的GPT）？也是亟待解决的难题。
  - 有利于将预训练的大模型尽可能少改动的用于下游识别任务（检测、分割等）。

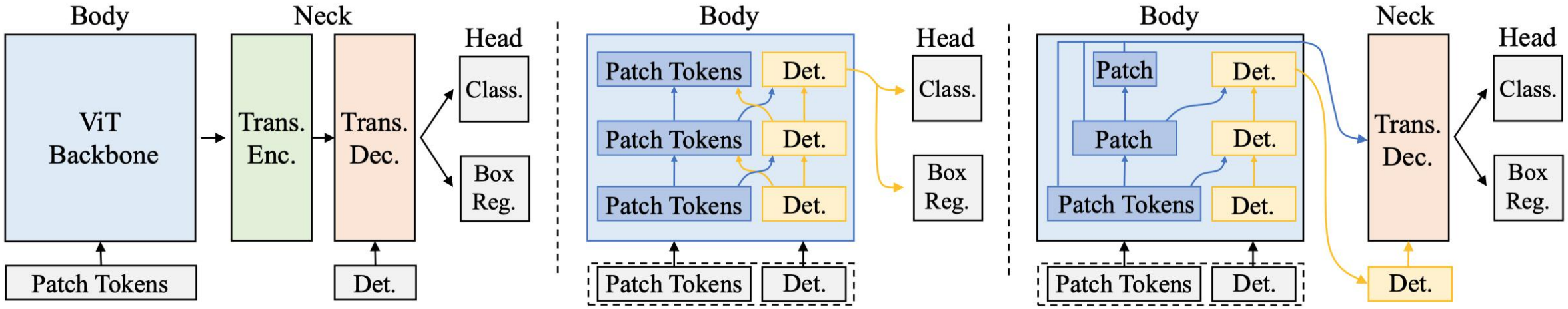
Detection Transformer: YOLOS

- YOLOS可以取得有竞争力的结果。
- YOLOS采用简单的设计，有极大的改进空间。

Method	Backbone	Epochs	Size	AP	Params. (M)	FLOPs (G)	FPS
Def. DETR [70]	FBNet-V3 [14]	150	800 × ∗	27.5	12.2	12.26	35
YOLOS-Ti	DeiT-Ti (⚗) [55]	300	432 × ∗	28.6	6.5	12.00	84
YOLOS-Ti	DeiT-Ti (⚗) [55]	300	528 × ∗	30.0	6.5	21.35	51
DETR [9]	ResNet-18-DC5 [25]	150	800 × ∗	36.9	28.7	128.9	7.4
YOLOS-S	DeiT-S [55]		800 × ∗	36.1	30.7	200.2	5.7
YOLOS-S (dwr)	DeiT-S [55] (dwr Scale [19])		704 × ∗	37.2	27.9	127.5	7.2
YOLOS-S (dwr)	DeiT-S [55] (dwr Scale [19])		784 × ∗	37.6	27.9	179.0	5.4
DETR [9]	ResNet-101-DC5 [25]	150	800 × ∗	42.5	60	253	–
YOLOS-B	DeiT-B (⚗) [55]			42.0	127	537	–

Detection Transformer: ViDT

- ViDT[1]受到DETR和YOLOS的启发，[PATCH] tokens和[DET] tokens均作为输入。
  - 采用Hierarchical ViT (e.g., Swin Transformer)处理[PATCH] tokens。
  - [DET] tokens与[PATCH] tokens之间进行单向交互([PATCH] tokens → [DET] tokens)。
  - 相对YOLOS增加了Decoder，加速收敛。



(a) DETR (b) YOLOS (c) ViDT

[1] H.Song, et al., ViDT: An Efficient and Effective Fully Transformer-based Object Detector

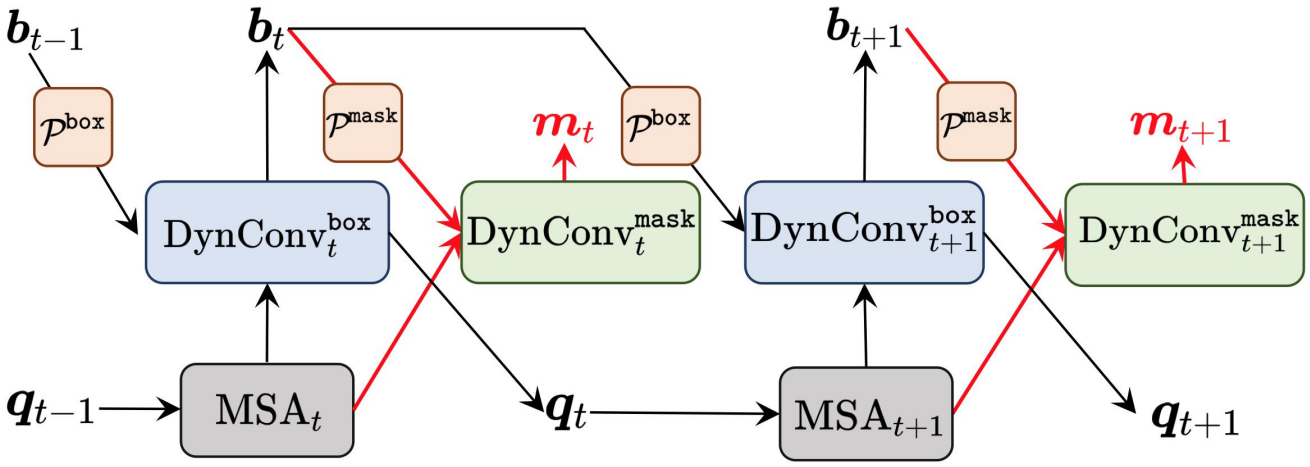
Detection Transformer: ViDT

Method	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Param.	FPS
DETR	DeiT-tiny	50	29.1	47.3	29.4	9.2	29.3	49.5	24M	10.9
	DeiT-small	50	30.8	50.9	31.0	10.5	31.0	52.1	39M	7.8
	Swin-nano	50	27.8	47.5	27.4	9.0	29.2	44.9	24M	24.7
	Swin-tiny	50	34.1	55.1	35.3	12.7	35.9	54.2	45M	19.3
	Swin-small	50	37.6	59.0	39.0	15.9	40.1	58.9	66M	13.5
Deformable DETR	DeiT-tiny	50	39.2	58.0	41.4	19.5	41.5	58.0	18M	12.4
	DeiT-small	50	40.9	60.3	43.3	20.4	43.8	59.6	35M	8.5
	Swin-nano	50	43.1	61.4	46.3	25.9	45.2	59.4	17M	7.0
	Swin-tiny	50	47.0	66.8	50.8	28.1	49.8	63.9	39M	6.3
	Swin-small	50	49.0	68.9	52.9	30.3	**52.8**	66.6	60M	5.5
YOLOS	DeiT-tiny	150	30.4	48.6	31.1	12.4	31.8	48.2	6M	28.1
	DeiT-small	150	36.1	55.7	37.6	15.6	38.4	55.3	30M	9.3
	DeiT-base	150	42.0	62.2	44.5	19.5	45.3	62.1	0.1B	3.9
ViDT (w.o. Neck)	Swin-nano	150	28.7	48.6	28.5	12.3	30.7	44.1	7M	36.5
	Swin-tiny	150	36.3	56.3	37.8	16.4	39.0	54.3	29M	28.6
	Swin-small	150	41.6	62.7	43.9	20.1	45.4	59.8	52M	16.8
	Swin-base	150	43.2	64.2	45.9	21.9	46.9	63.2	91M	11.5
ViDT	Swin-nano	50	40.4	59.6	43.3	23.2	42.5	55.8	16M	20.0
	Swin-tiny	50	44.8	64.5	48.7	25.9	47.6	62.1	38M	17.2
	Swin-small	50	47.5	67.7	51.4	29.2	50.7	64.8	61M	12.1
	Swin-base	50	**49.2**	**69.4**	**53.1**	**30.6**	52.6	**66.9**	0.1B	9.0

Table 2. Comparison of ViDT with other compared detectors on COCO2017 val set. Two neck-free detectors, YOLOS and ViDT (w.o. Neck) are trained for 150 epochs due to the slow convergence. FPS is measured with batch size 1 of 800×1333 resolution on a single Tesla V100 GPU.

Instance Segmentation Transformer: Instances as Queries

- 基于Sparse R-CNN，搭建基于Object Query的端到端实例分割模型QueryInst。
- 每一个Instance的属性(class, bbox, mask, identify, etc.)都和一个Object Query一一对应。
- Query通过动态卷积抽取Instance的信息来增强RoI特征，Query之间通过Self Attention来感知全图上下信息。



(c) QueryInst with dynamic mask head

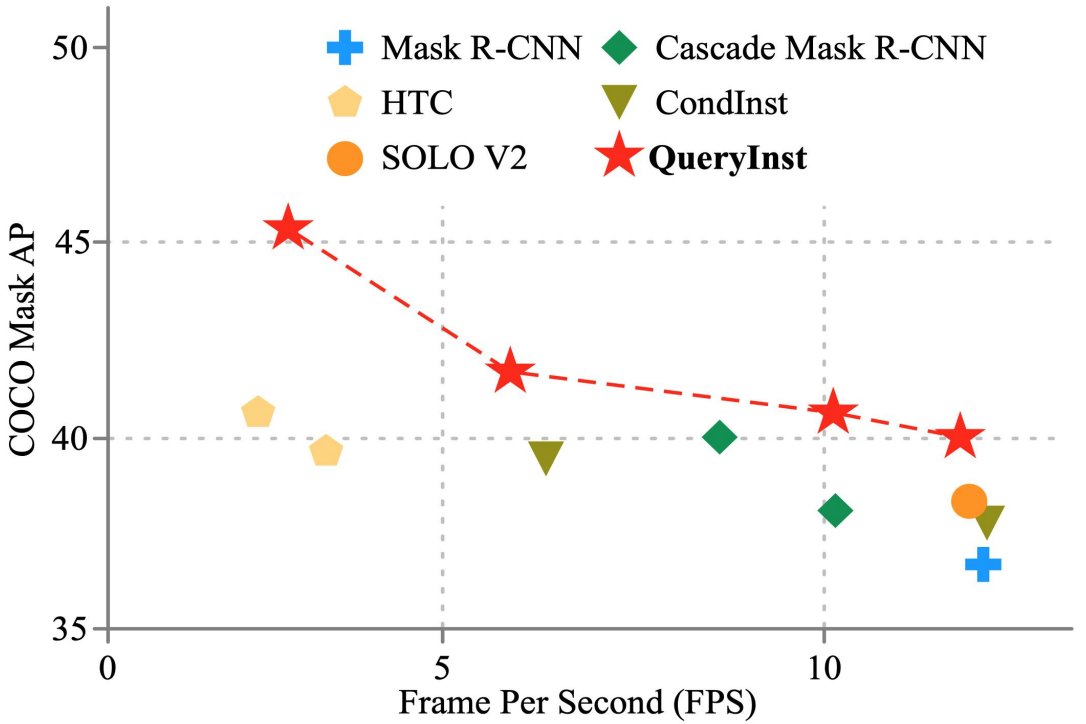
[1] Y. Fang et al., Instances as Queries, ICCV 2021

Instance Segmentation Transformer: Instances as Queries

Method	Mask AP	Box AP	FPS
Cascade Mask R-CNN	38.5	44.3	10.4
HTC	39.3	44.4	3.1
Sparse R-CNN	N / A	42.8	11.0
QueryInst	39.8	44.5	10.5

- Query can largely enhance the RoI feature
  → Improving both Mask AP & Box AP

- Training: Multi-stage parallel supervised mask head
- Inference: Only use the last stage predictions
  → High efficiency & FPS



Speed-accuracy Pareto curve on COCO `test-dev`

Instance Segmentation Transformer: Instances as Queries

Method	Backbone	Aug.	Epochs	AP^{box}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
Mask R-CNN [21]	ResNet-50-FPN	640 ~ 800	36	41.3	37.5	59.3	40.2	21.1	39.6	48.3	14.0
CondInst w/ sem. [46]				–	38.6	60.2	41.4	20.6	41.0	51.1	14.1
SOLOv2 [51]				40.4	38.8	59.9	41.7	16.5	**41.7**	**56.2**	13.8
QueryInst (5 Stage, 100 Queries)				**44.5**	**39.9**	**62.2**	**43.0**	**22.9**	**41.7**	51.9	13.5
Cascade Mask R-CNN [5]	ResNet-50-FPN	640 ~ 800	36	44.5	38.6	60.0	41.7	21.7	40.8	49.6	10.4
HTC [9]				44.9	39.7	61.4	43.1	22.6	42.2	50.6	3.1
QueryInst (100 Queries)				44.8	40.1	62.3	43.4	23.3	42.1	52.0	10.5
QueryInst (300 Queries)				**45.6**	**40.6**	**63.0**	**44.0**	**23.4**	**42.5**	**52.8**	7.0
Cascade Mask R-CNN	ResNet-101-FPN	640 ~ 800	36	45.7	39.8	61.6	43.0	22.4	42.2	50.8	8.7
HTC				46.2	40.7	62.7	44.2	23.1	43.4	52.7	2.5
QueryInst (300 Queries)				**47.0**	**41.7**	**64.4**	**45.3**	**24.2**	**43.9**	**53.9**	6.1
Cascade Mask R-CNN	ResNet-101-FPN	480 ~ 800 w/ crop	36	46.2	40.0	61.7	43.5	22.5	42.5	51.2	8.7
HTC				46.3	40.8	62.6	44.3	23.0	43.5	52.6	2.5
Sparse R-CNN (300 Queries)				46.3	–	–	–	–	–	–	6.9
QueryInst (300 Queries)				**48.1**	**42.8**	**65.6**	**46.7**	**24.6**	**45.0**	**55.5**	6.1
QueryInst (300 Queries)	ResNeXt-101-FPN w/ DCN	480 ~ 800 w/ crop	36	**50.4**	**44.6**	**68.1**	**48.7**	**26.6**	**46.9**	**57.7**	3.1
QueryInst (300 Queries) @ `val`	Swin-L	400 ~ 1200 w/ crop	50	**56.1**	**48.9**	**74.0**	**53.9**	**30.8**	**52.6**	**68.3**	$3.3^{\top}$
QueryInst (300 Queries)	Swin-L	400 ~ 1200 w/ crop	50	**56.1**	**49.1**	**74.2**	**53.8**	**31.5**	**51.8**	**63.2**	$3.3^{\top}$

Table 1: Main results on COCO `test-dev`. The numbers under "Aug." indicate the scale range of the shorter size of inputs with a stride of 32. AP^{box} denotes box AP. AP without superscript denotes mask AP. The best results are in **bold** for each configuration. Superscript "$\top$" indicates the FPS data are measured on a single RTX 2080Ti GPU with batch size 1.

Instance Segmentation Transformer: Instances as Queries

Summary

- **Macro Structure:**
  - Locality和Hierarchy的引入使得ViT在CV中得到广泛的应用。在很多任务中都取得了极佳的性能。
  - 相比具体的算子(Attention or Conv or MLP), Macro Structure的对结果的影响更为重要。
- **Optimization:**
  - 目前Transformer作为视觉主干网络的成功,训练上和优化上的贡献也是不能忽视的。
- **High-level Applications:**
  - ViT以及Query Token的引入启发了新的框架设计。通过Query Token可以灵活高效的提取多尺度物体特征，实现高性能的物体检测和实例分割。

敬请各位专家批评指正！

谢 谢！