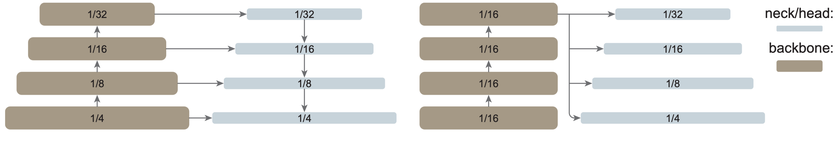
Exploring Plain Vision Transformer Backbones for Object Detection

Yanghao Li, Hanzi Mao, Ross Girshick*, and Kaiming He*
Meta AI

Code & models

Overview

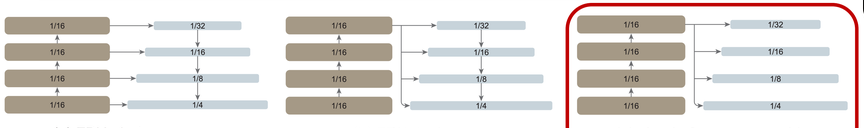
ViTDet: Detectors with ***plain***, ***non-hierarchical*** backbone
- ***Goal: Decouple*** *detection-specific designs from general backbone*
- ***How****: Two* ***Minimal*** *adaptations on ViT for detection*



hierarchical backbone, w/ FPN — **plain** backbone, w/ simple feature pyramid

Simple Feature Pyramid

Feature pyramid on a **plain** backbone
- FPN-like
- Simple feature pyramid (w/o FPN)
  - **Strongest** feature is from **last** feature map



(a) FPN, 4-stages (b) FPN, last map (c) simple feature pyramid

Simple feature pyramid is **sufficient**

pyramid design	ViT-B AP^{box}	ViT-B AP^{mask}	ViT-L AP^{box}	ViT-L AP^{mask}
no feature pyramid	47.8	42.5	51.2	45.4
(a) FPN, 4-stage	50.3 (+2.5)	44.9 (+2.4)	54.4 (+3.2)	48.4 (+3.0)
(b) FPN, last-map	50.9 (+3.1)	45.3 (+2.8)	**54.6** (+3.4)	48.5 (+3.1)
(c) simple feature pyramid	**51.2** (+3.4)	**45.5** (+3.0)	**54.6** (+3.4)	**48.6** (+3.2)

Backbone adaptation

Self-attention has high complexity for high-resolution
Ours: *Window attention + a few propagation blocks*
- Simple ***non-overlapping*** window attention
- Propagation blocks
  - **Global** propagation
  - **Convolution** propagation

Backbone adaptation schemes

prop. strategy	AP^{box}	AP^{mask}
none	52.9	47.2
4 global blocks	54.6 (+1.7)	48.6 (+1.4)
4 conv blocks	**54.8** (+1.9)	**48.8** (+1.6)
shifted win.	54.0 (+1.1)	47.9 (+0.7)

Global/Conv propagation *is better than "shifted win"*

Convolution block types

prop. conv	AP^{box}	AP^{mask}
none	52.9	47.2
naïve	54.3 (+1.4)	48.3 (+1.1)
basic	**54.8** (+1.9)	**48.8** (+1.6)
bottleneck	54.6 (+1.7)	48.6 (+1.4)

All propagation *strategies work*

Practical performance

prop. strategy	AP^{box}	# params	train mem	test time
none	52.9	1.00× (331M)	1.00× (14.6G)	1.00× (88ms)
4 conv (bottleneck)	54.6 (+1.7)	1.04×	1.05×	1.04×
4 global	54.6 (+1.7)	1.00×	1.39×	1.16×
24 global	55.1 (+2.2)	1.00×	3.34×†	1.86×

Conv prop. **:** ***<5%*** *increase for mem/time and* ***4%*** *more #param*

Global prop. **:** ***No increase*** *for model size*

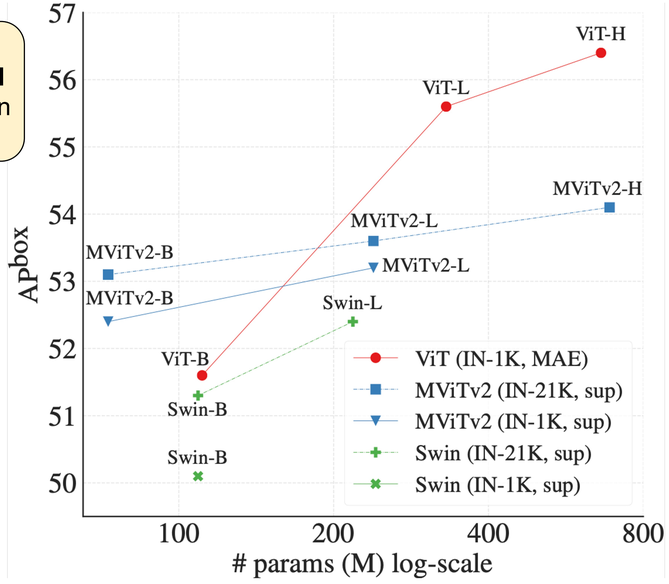
Results

Pre-training Strategies for ViTs

pre-train	ViT-B AP^{box}	ViT-B AP^{mask}	ViT-L AP^{box}	ViT-L AP^{mask}
none (random init.)	48.1	42.6	50.0	44.2
IN-1K, supervised	47.6 (–0.5)	42.4 (–0.2)	49.6 (–0.4)	43.8 (–0.4)
IN-21K, supervised	47.8 (–0.3)	42.6 (+0.0)	50.6 (+0.6)	44.8 (+0.6)
IN-1K, MAE	**51.2** (+3.1)	**45.5** (+2.9)	**54.6** (+4.6)	**48.6** (+4.4)

ViTDet enables ***easy use*** of powerful **MAE** pre-training

Plain *vs.* **Hierarchical** backbones on COCO



- ViTDet has ***better scale behavior***
  - ViT-H is **+2.6** better than MViTv2-H