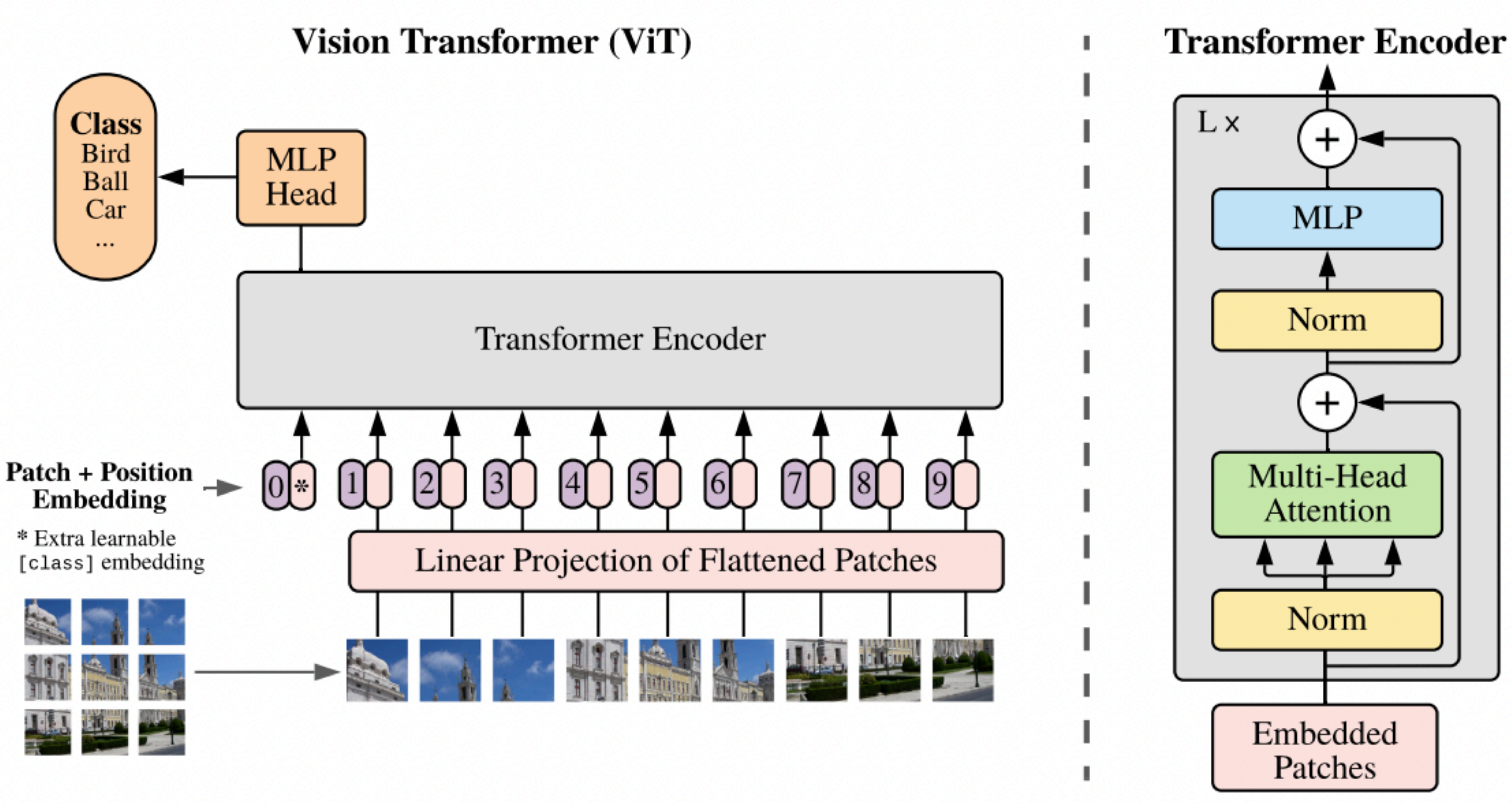
Exploring Plain Vision Transformer Backbones for Object Detection

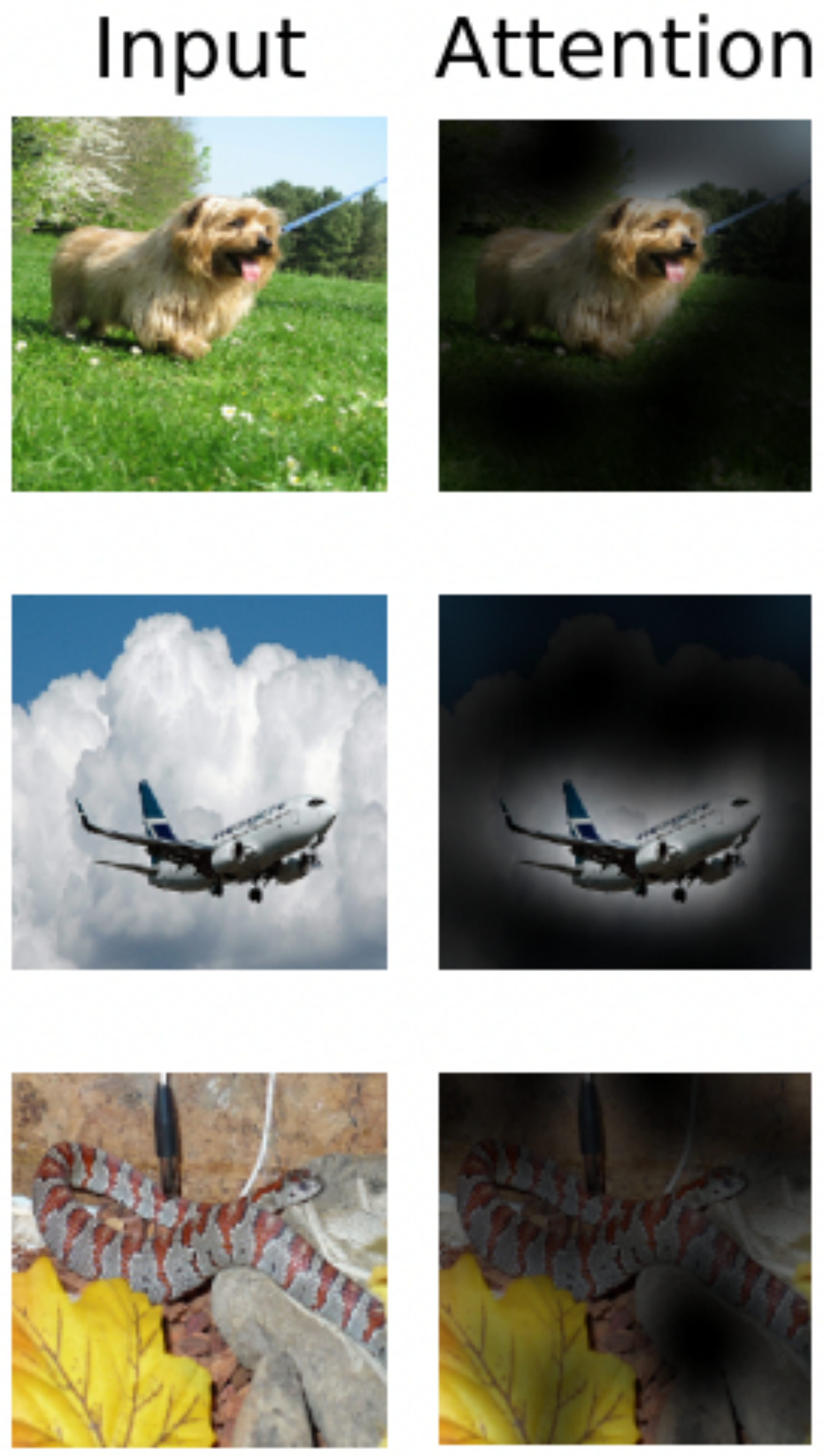
Facebook AI Research

Добряев Иван
Добросовестнов Иван
Поклонская Мария

Vision Transformer

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE





А теперь к сути статьи!

- Есть новая архитектура, которая лучше CNN подходов.
- Есть Задача Детекции в которой CNN подходы хороши.

Можем использовать ViT для детекции?

Цель

- Исследовать не иерархические тушки в задаче детекции объектов,

с минимальными изменениями

Зачем?

- Это позволит использовать ViT на прямую.

Ключевые части подхода

- Simple feature pyramid.
- Backbone adaptation.

Имплементация

- BackBones: vanilla pretrained ViT-B, ViT-L, ViT-H
- Map scale is 1/16
- Heads: Mask R-CNN / Cascade Mask R-CNN
- Input Image size: 1024×1024 + augmentation
- Fine-tune for up to 100 epochs in COCO.
- Optimizer: AdamW

А что делать с FPN?

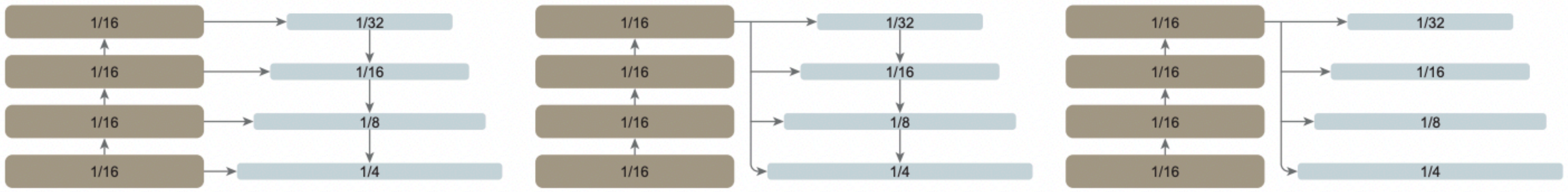
FPN

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

Swin

Попробовать сделать выходы ViT иерархическими (пирамидоидальный)

Simple feature pyramid

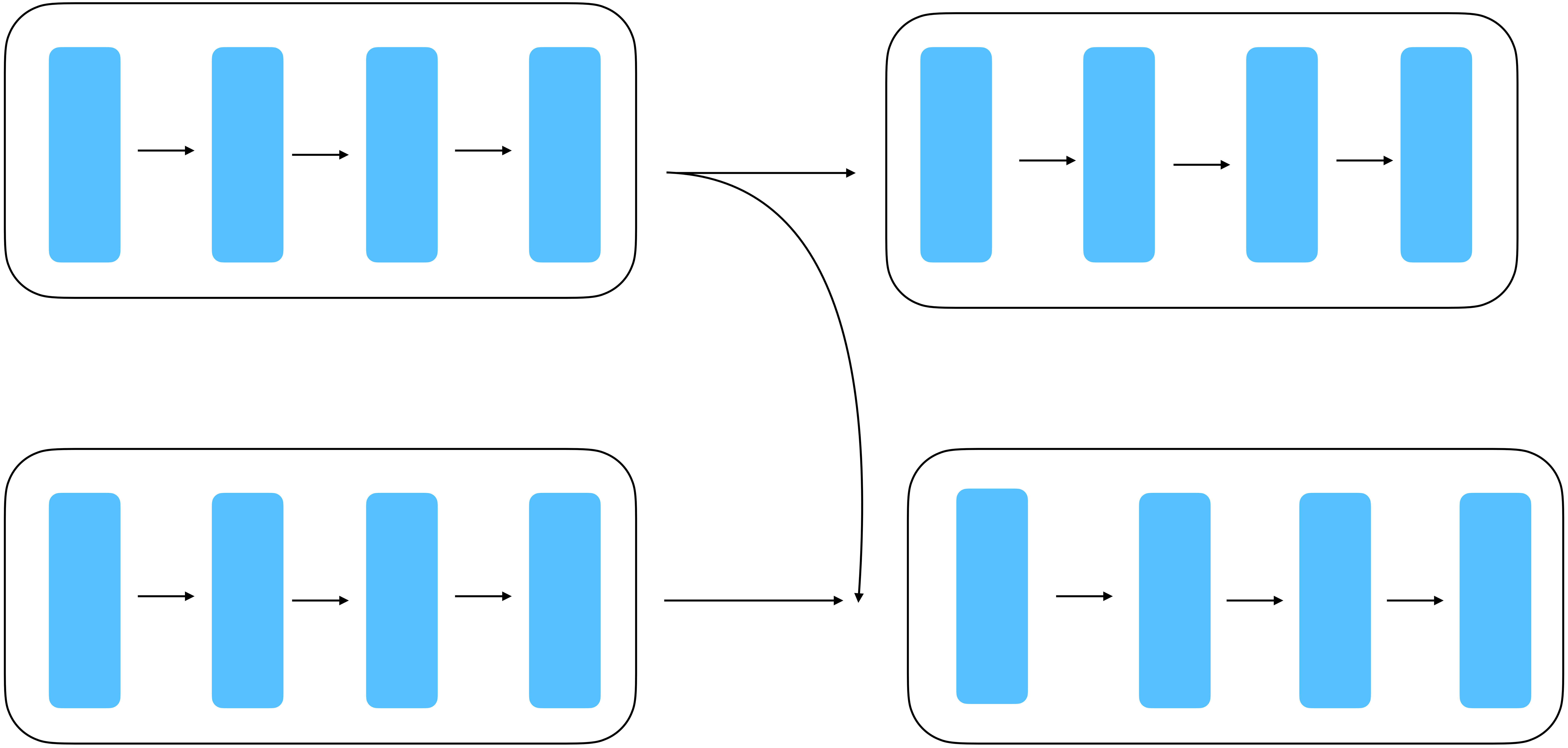


(a) FPN, 4-stages (b) FPN, last map (c) simple feature pyramid

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.

Hybrid window attention



Backbone adaptation.

Convolutional propagation

Backbone adaptation.

Результаты

prop. locations	AP^{box}	AP^{mask}
none	52.9	47.2
first 4 blocks	52.9 (+0.0)	47.1 (–0.1)
last 4 blocks	54.3 (+1.4)	48.3 (+1.1)
evenly 4 blocks	**54.6** (+1.7)	**48.6** (+1.4)

(c) Locations of cross-window global propagation blocks.

prop. blks	AP^{box}	AP^{mask}
none	52.9	47.2
2	54.4 (+1.5)	48.5 (+1.3)
4	54.6 (+1.7)	48.6 (+1.4)
$24^{\dagger}$	**55.1** (+2.2)	**48.9** (+1.7)

(d) Number of global propagation blocks. †: Memory optimization required.

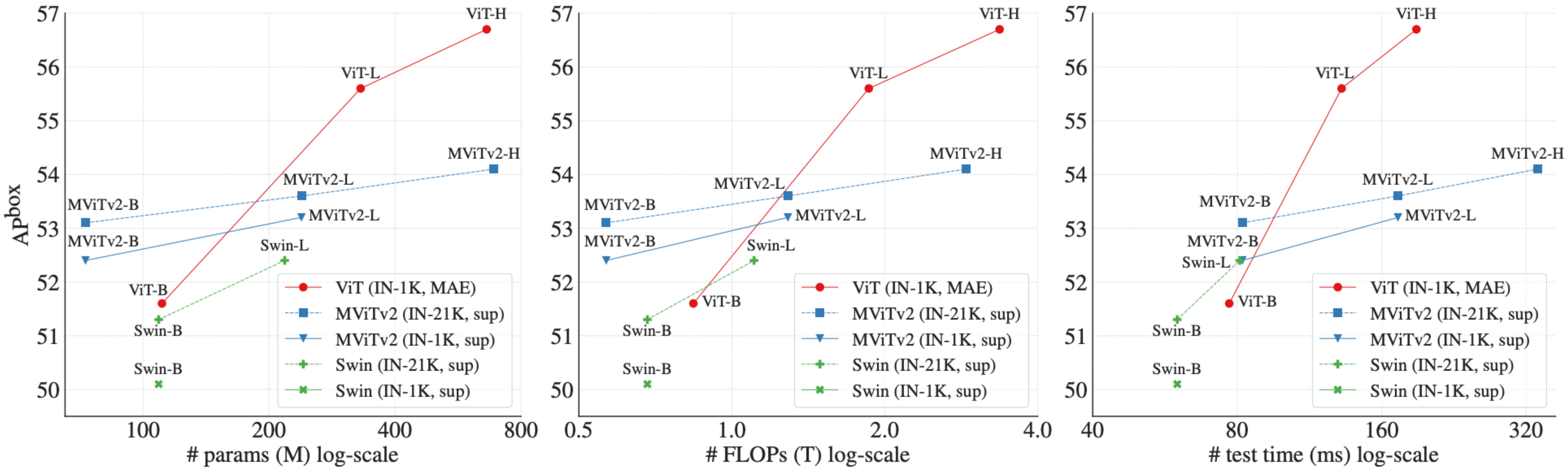
prop. strategy	AP^{box}	AP^{mask}
none	52.9	47.2
4 global blocks	54.6 (+1.7)	48.6 (+1.4)
4 conv blocks	**54.8** (+1.9)	**48.8** (+1.6)
shifted win.	54.0 (+1.1)	47.9 (+0.7)

(a) Window attention with various cross-window propagation strategies.

Сравнение резульатов

backbone	pre-train	Mask R-CNN AP^{box}	Mask R-CNN AP^{mask}	Cascade Mask R-CNN AP^{box}	Cascade Mask R-CNN AP^{mask}
hierarchical-backbone detectors:					
Swin-B	21K, sup	51.4	45.4	54.0	46.5
Swin-L	21K, sup	52.4	46.2	54.8	47.3
MViTv2-B	21K, sup	53.1	47.4	55.6	48.1
MViTv2-L	21K, sup	53.6	47.5	55.7	48.3
MViTv2-H	21K, sup	54.1	47.7	55.8	48.3
our plain-backbone detectors:					
ViT-B	1K, MAE	51.6	45.9	54.0	46.7
ViT-L	1K, MAE	55.6	49.2	57.6	49.8
ViT-H	1K, MAE	**56.7**	**50.1**	**58.7**	**50.9**

Tradeoffs



AP^box
57
56
55
54
53
52
51
50
100
200
400
800
params (M) log-scale
0.5
1.0
2.0
4.0
FLOPs (T) log-scale
40
80
160
320
test time (ms) log-scale
ViT-H
ViT-L
ViT-B
MViTv2-H
MViTv2-L
MViTv2-B
Swin-L
Swin-B
ViT (IN-1K, MAE)
MViTv2 (IN-21K, sup)
MViTv2 (IN-1K, sup)
Swin (IN-21K, sup)
Swin (IN-1K, sup)

Сравнение результатов

System-level comparisons with the leading results on COCO

method	framework	pre-train	single-scale test AP^{box}	AP^{mask}
hierarchical-backbone detectors:				
Swin-L [42]	HTC++	21K, sup	57.1	49.5
MViTv2-L [34]	Cascade	21K, sup	56.9	48.6
MViTv2-H [34]	Cascade	21K, sup	57.1	48.8
CBNetV2 [36]†	HTC	21K, sup	59.1	51.0
SwinV2-L [41]	HTC++	21K, sup	58.9	51.2
plain-backbone detectors:				
UViT-S [9]	Cascade	1K, sup	51.9	44.5
UViT-B [9]	Cascade	1K, sup	52.5	44.8
ViTDet, ViT-B	Cascade	1K, MAE	56.0	48.0
ViTDet, ViT-L	Cascade	1K, MAE	59.6	51.1
ViTDet, ViT-H	Cascade	1K, MAE	**60.4**	**52.0**

Ссылки на статьи

- Exploring Plain Vision Transformer Backbones for Object Detection https://arxiv.org/pdf/2203.16527.pdf
- AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE https://arxiv.org/pdf/2010.11929.pdf
- Swin Transformer: Hierarchical Vision Transformer using Shifted Windows https://arxiv.org/pdf/2103.14030.pdf
- MViTv2: Improved Multiscale Vision Transformers for Classification and Detection https://arxiv.org/pdf/2112.01526.pdf
- Mask R-CNN https://arxiv.org/pdf/1703.06870.pdf

Спасибо за внимание!

Exploring Plain Vision Transformer Backbones for Object Detection

Рецензия

Добросовестнов Иван

Авторы

Yanghao Li

Facebook AI Research
Verified email at fb.com - Homepage

Computer Vision

Hanzi Mao

Research Scientist, Facebook AI Research (FAIR)
Verified email at fb.com - Homepage

Machine Learning   Computer Vision

Ross Girshick

Research Scientist, Facebook AI Research (FAIR)
Verified email at eecs.berkeley.edu - Homepage

computer vision   machine learning

Kaiming He

Research Scientist, Facebook AI Research (FAIR)
Verified email at fb.com - Homepage

Computer Vision   Machine Learning

Facebook AI Research

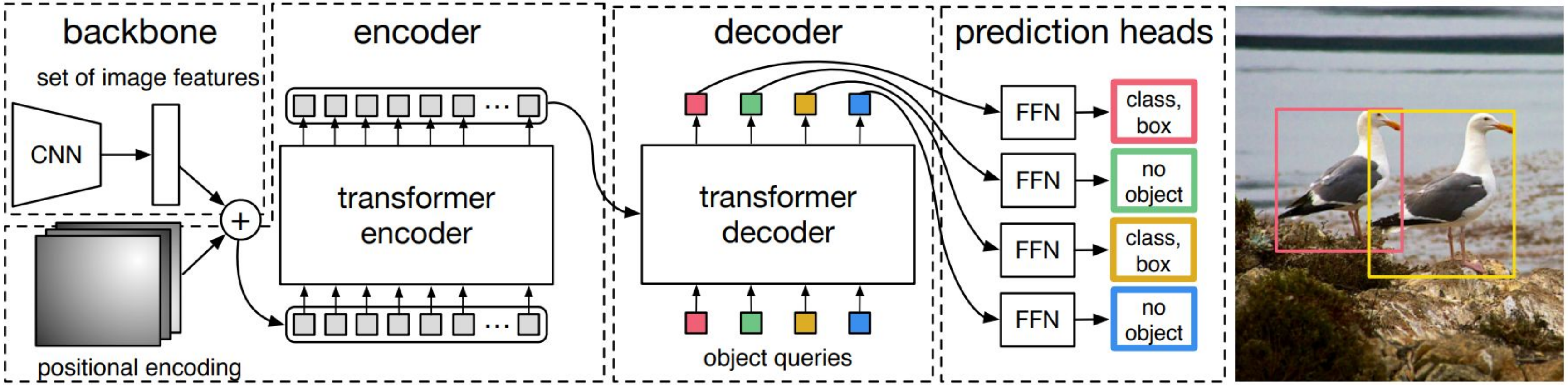
Влияние на работу

2020

28 мая

End-to-End Object Detection with Transformers

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (Facebook AI)



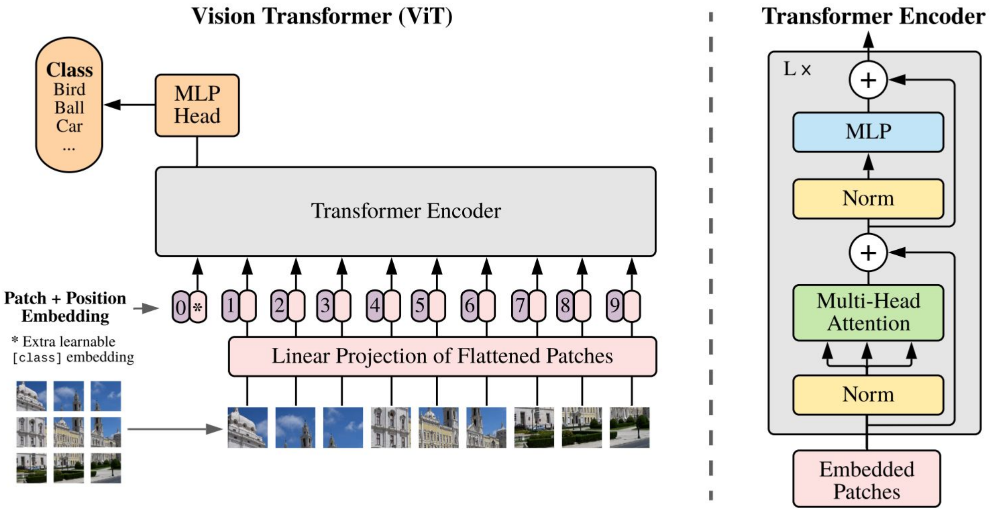
Влияние на работу

2021

3 июня

An image is worth 16x16 words: Transformers for image recognition at scale

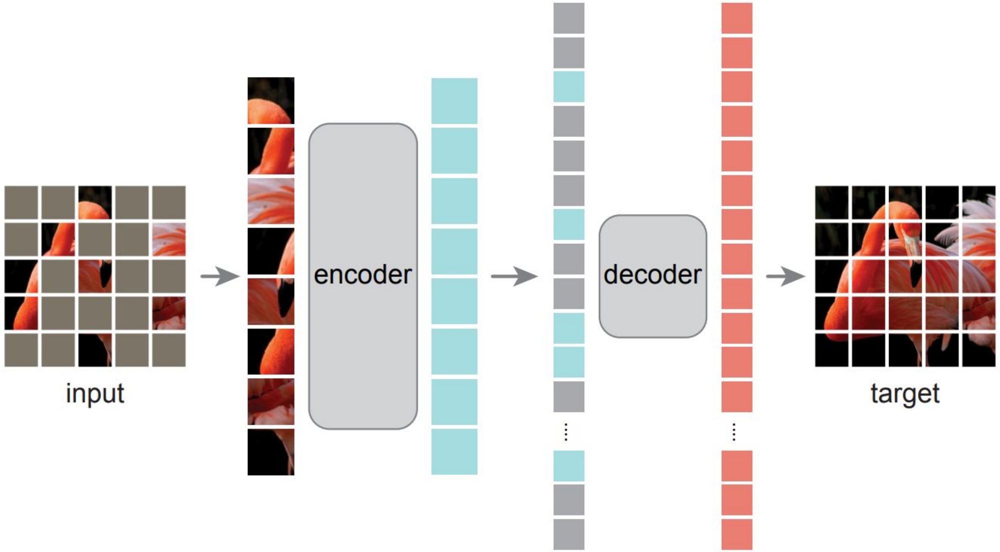
Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby



Влияние на работу

2022 Masked Autoencoders Are Scalable Vision Learners

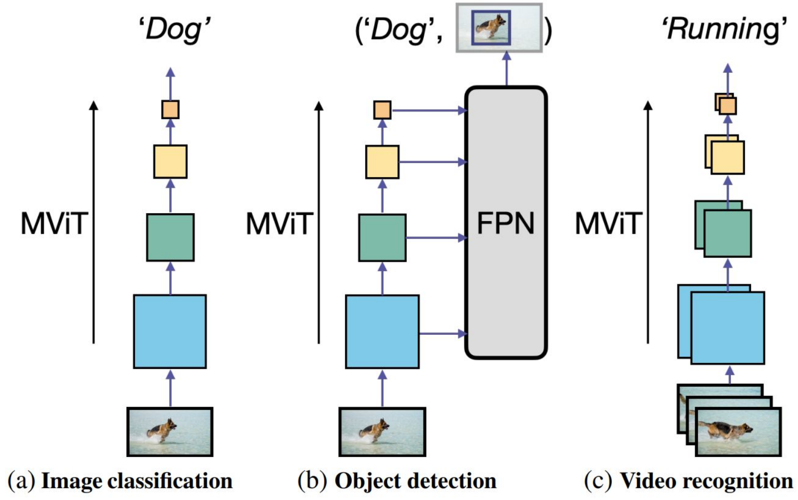
Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick

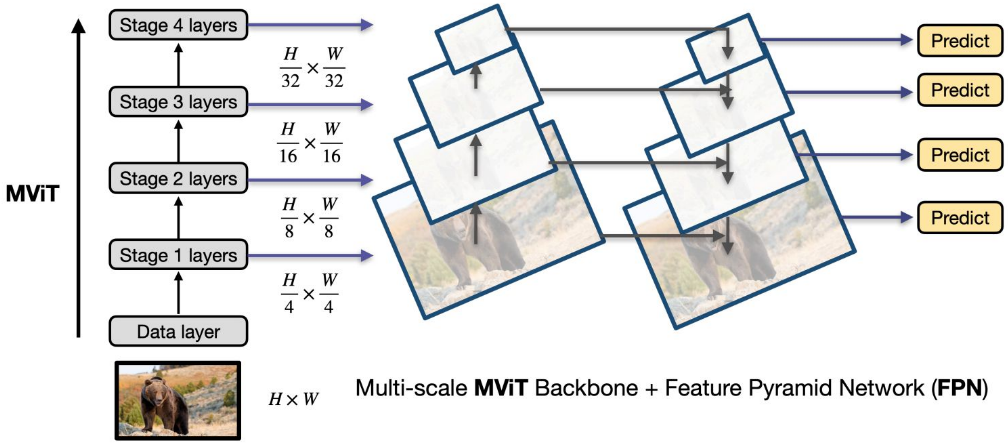


Влияние на работу

2022 MViTv2: Improved Multiscale Vision Transformers for Classification and Detection

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra, Malik, Christoph Feichtenhofer





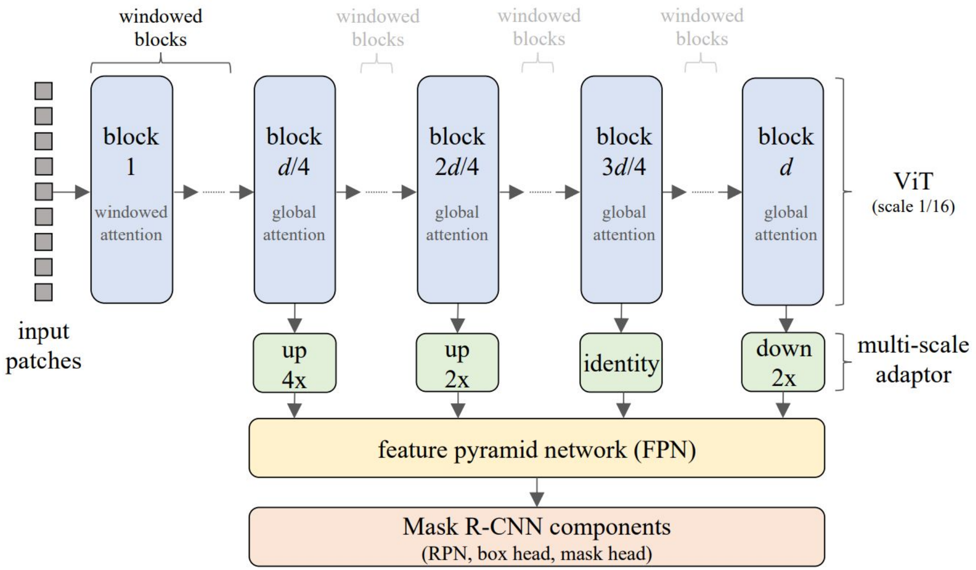
История статьи

2021

22 Ноября

Benchmarking Detection Transfer Learning with Vision Transformers

Yanghao Li Saining Xie Xinlei Chen Piotr Dollar Kaiming He Ross Girshick



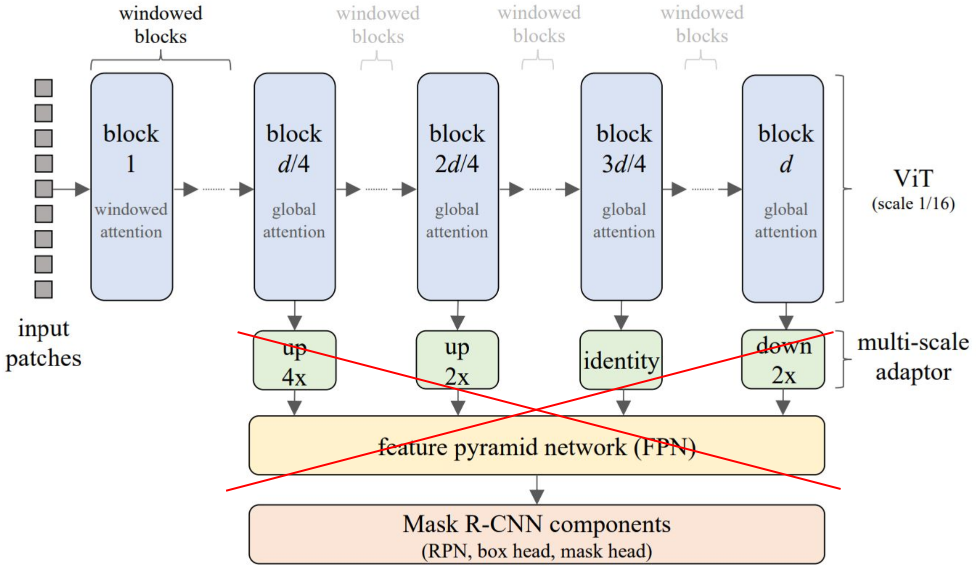
История статьи

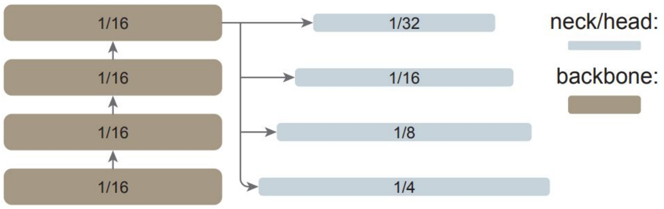
2022

30 Марта

Exploring Plain Vision Transformer Backbones for Object Detection

Yanghao Li Hanzi Mao Ross Girshick Kaiming He





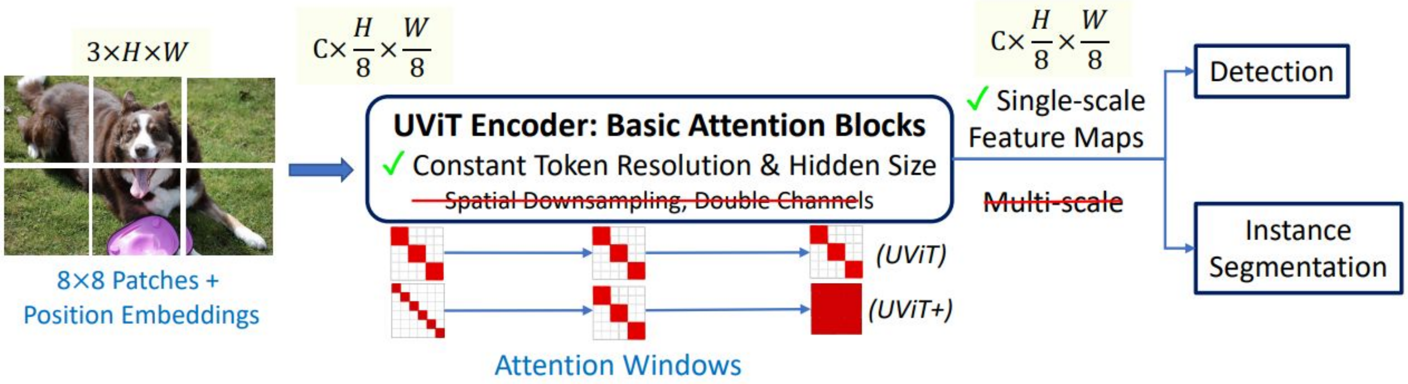
Конкуренты

2022

17 сентября

A Simple Single-Scale Vision Transformer for Object Detection and Instance Segmentation

Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, Zhangyang Wang, and Denny Zhou (Google Research & University of Texas)



Цитирующие работу

2022
14 июля

TransVG++: End-to-End Visual Grounding with Language Conditioned Vision Transformer

Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Senior Member, IEEE, Yanyong Zhang, Fellow, IEEE, Houqiang Li, Fellow, IEEE, Wanli Ouyang, Senior Member, IEEE