Swin Transformer

Артем Исмагилов

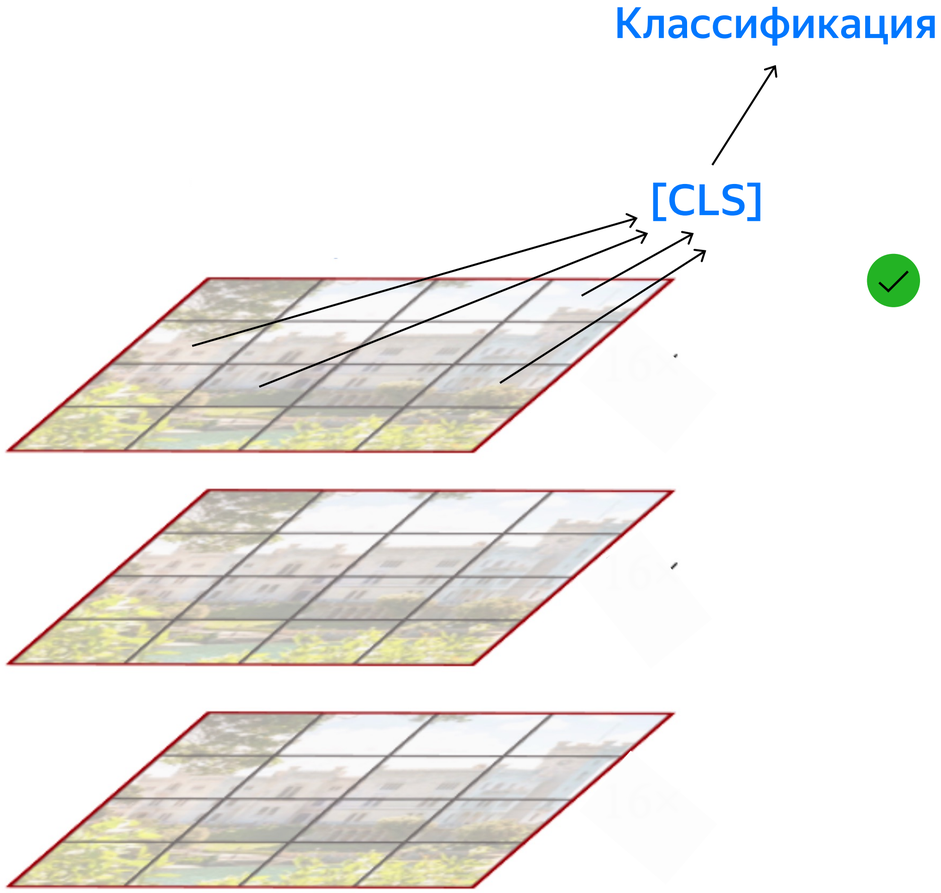
Intro

Swin Transformer

Как применять для разных задач

Zero Shot и One Shot детекция

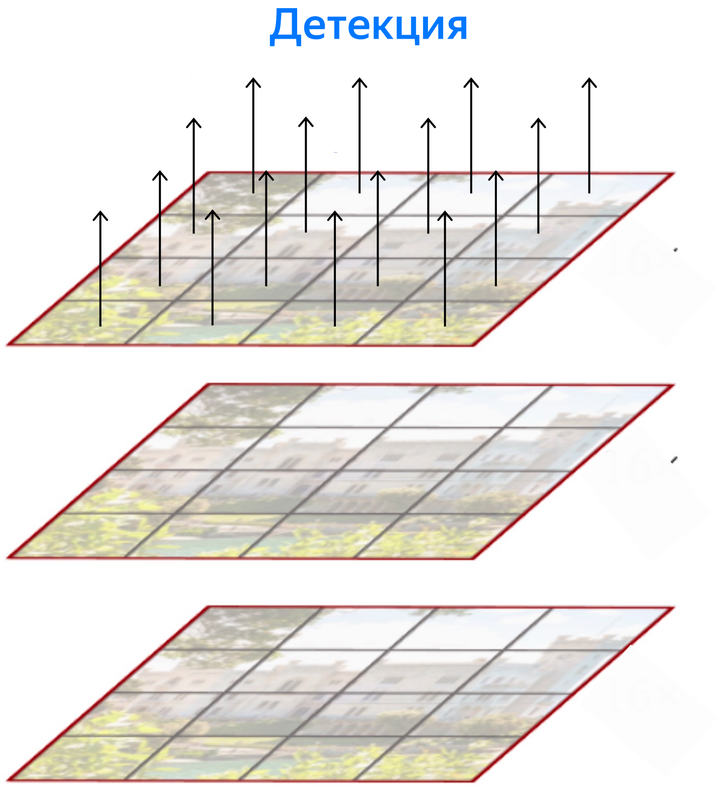
ViT



- Хорошо извлекает глобальные признаки

https://arxiv.org/abs/2103.14030

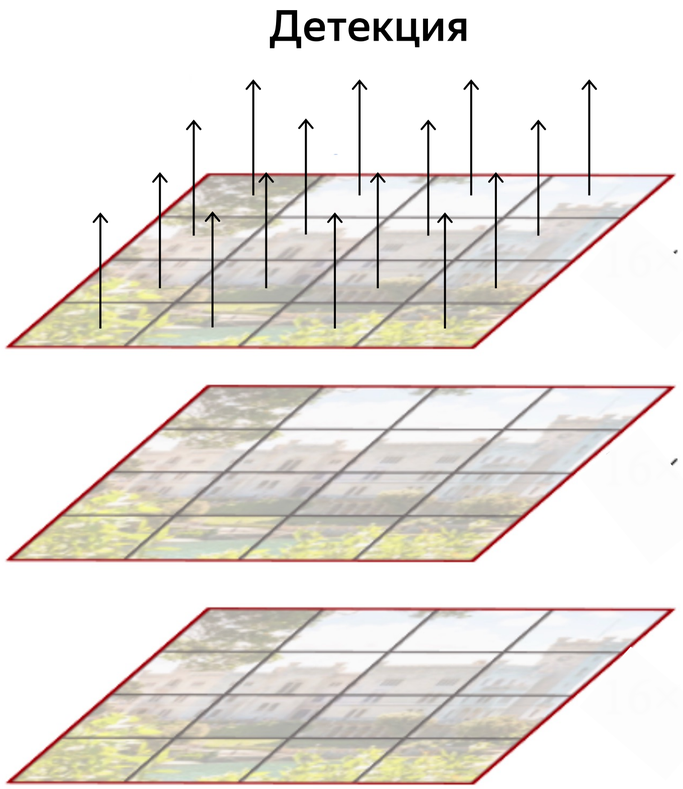
ViT



- Хорошо извлекает глобальные признаки
- Хорошо извлекает локальные признаки

https://arxiv.org/abs/2103.14030

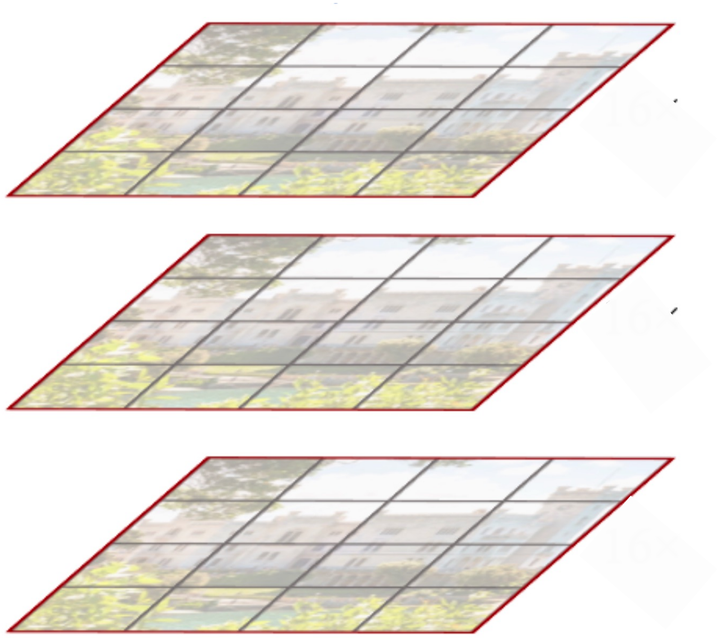
ViT



- Хорошо извлекает глобальные признаки
- Хорошо извлекает локальные признаки
- Локальные признаки содержат весь контекст

https://arxiv.org/abs/2103.14030

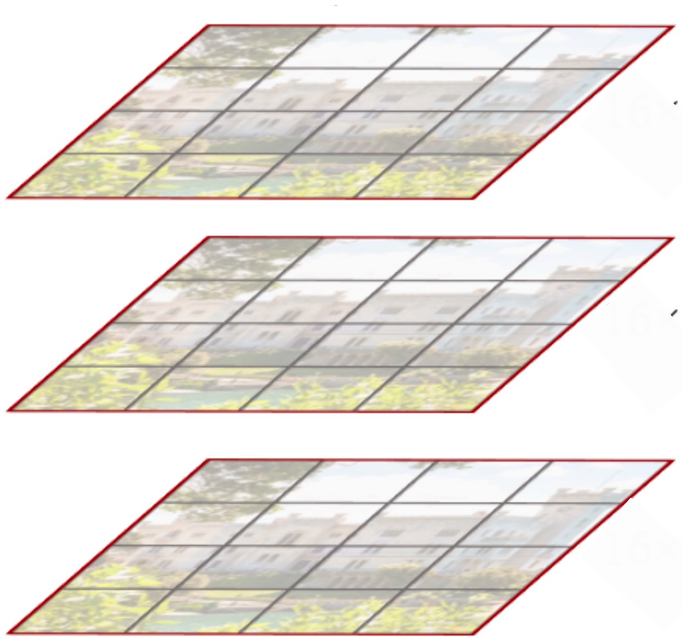
ViT



- ✓ Хорошо извлекает глобальные признаки
- ✓ Хорошо извлекает локальные признаки
- ✓ Локальные признаки содержат весь контекст
- ✗ В Self-Attention очень много патчей

https://arxiv.org/abs/2103.14030

ViT



- ✓ Хорошо извлекает глобальные признаки
- ✓ Хорошо извлекает локальные признаки
- ✓ Локальные признаки содержат весь контекст

- ✗ В Self-Attention очень много патчей
- ✗ Увеличить разрешение карты признаков очень дорого

https://arxiv.org/abs/2103.14030

Intro

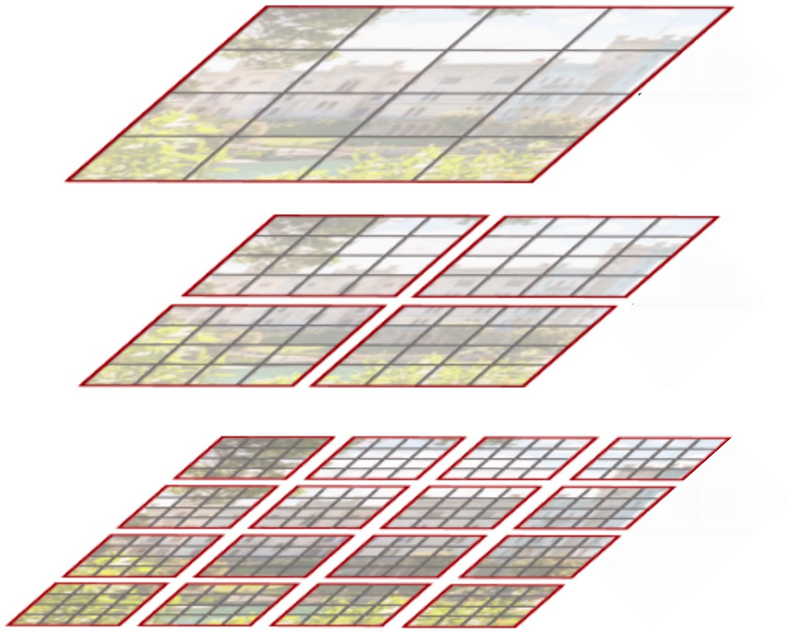
Swin Transformer

Как применять для разных задач

Zero Shot и One Shot детекция

Swin Transformer

Shifted Windows

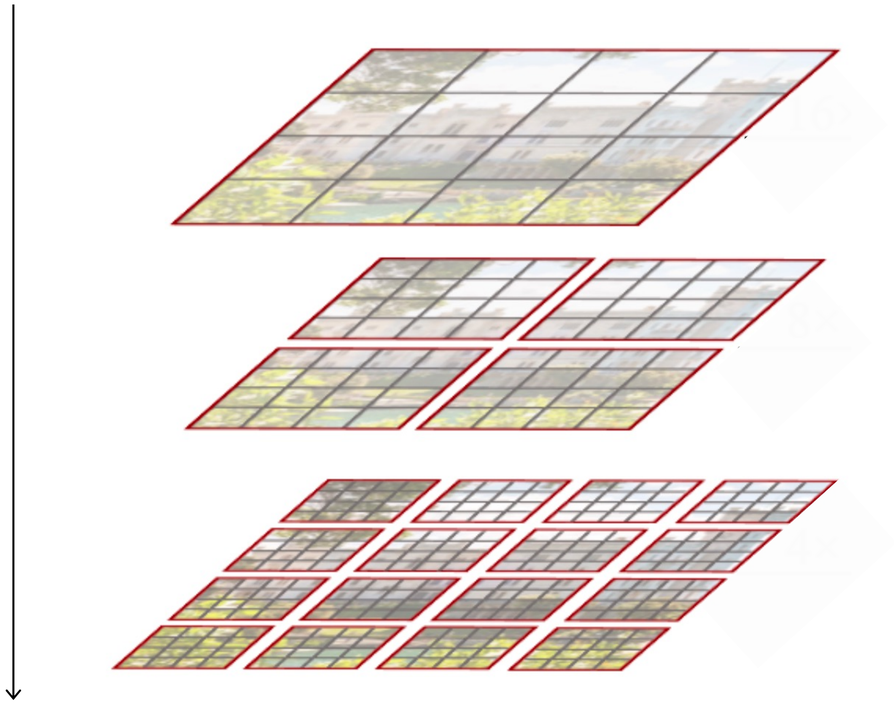


https://arxiv.org/abs/2103.14030

Swin Transformer

Shifted Windows

Больше разрешение

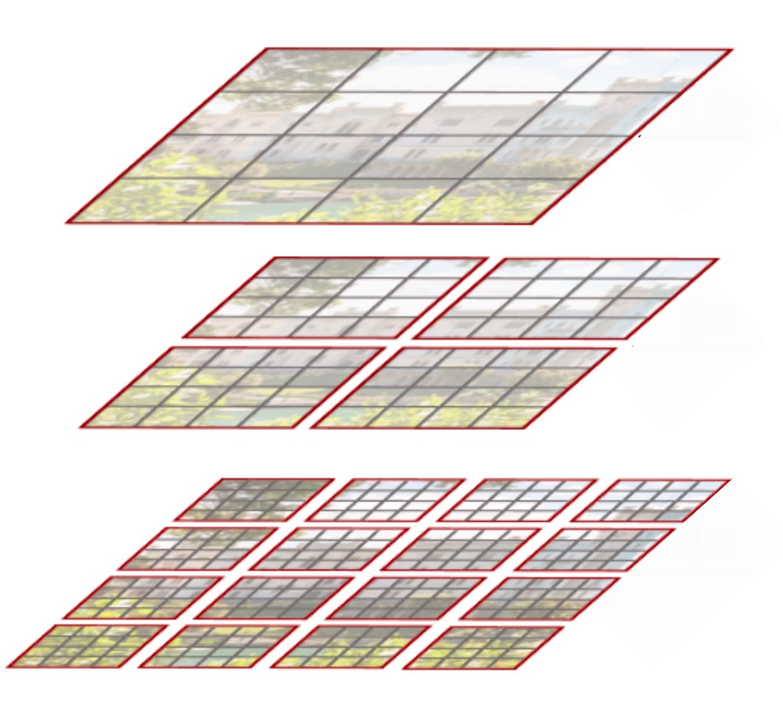


https://arxiv.org/abs/2103.14030

Swin Transformer

Shifted Windows

Больше разрешение

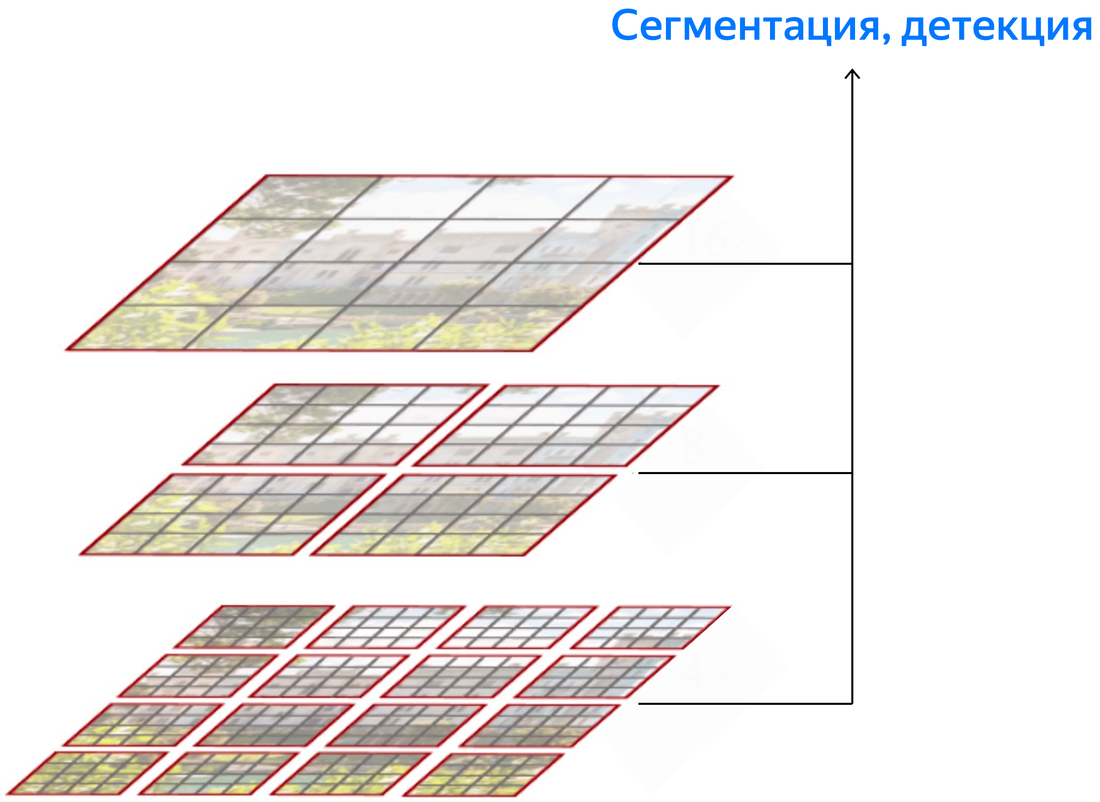


Больше контекст

https://arxiv.org/abs/2103.14030

Swin Transformer

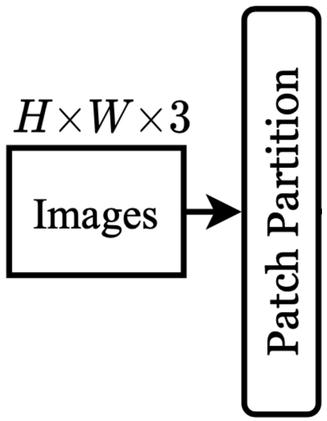
Shifted Windows



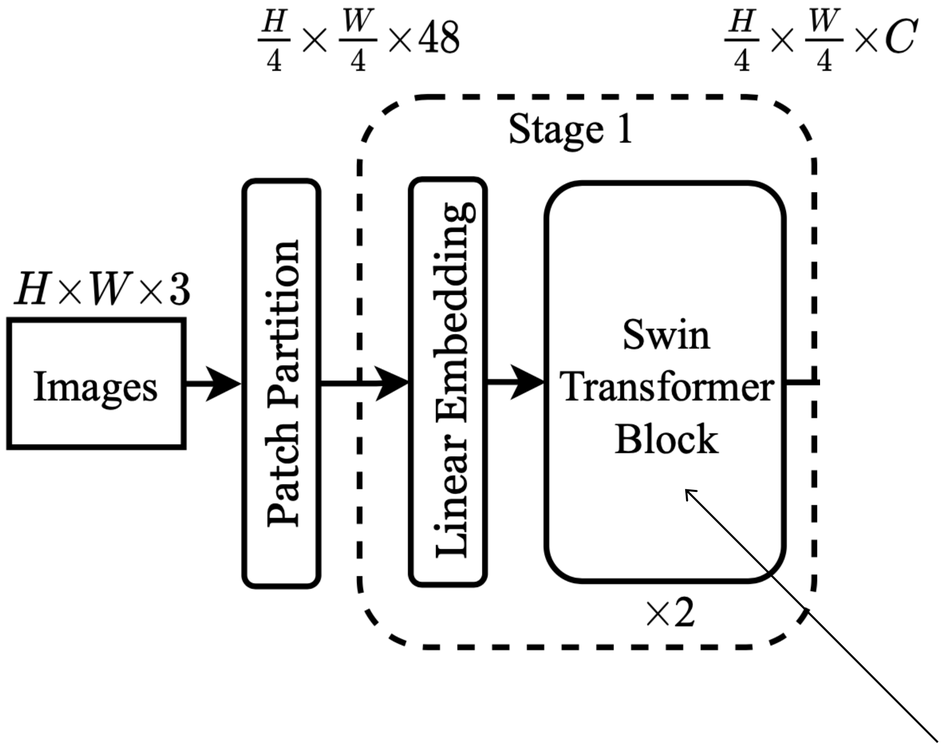
https://arxiv.org/abs/2103.14030

Swin Transformer

$\frac{H}{4} \times \frac{W}{4} \times 48$

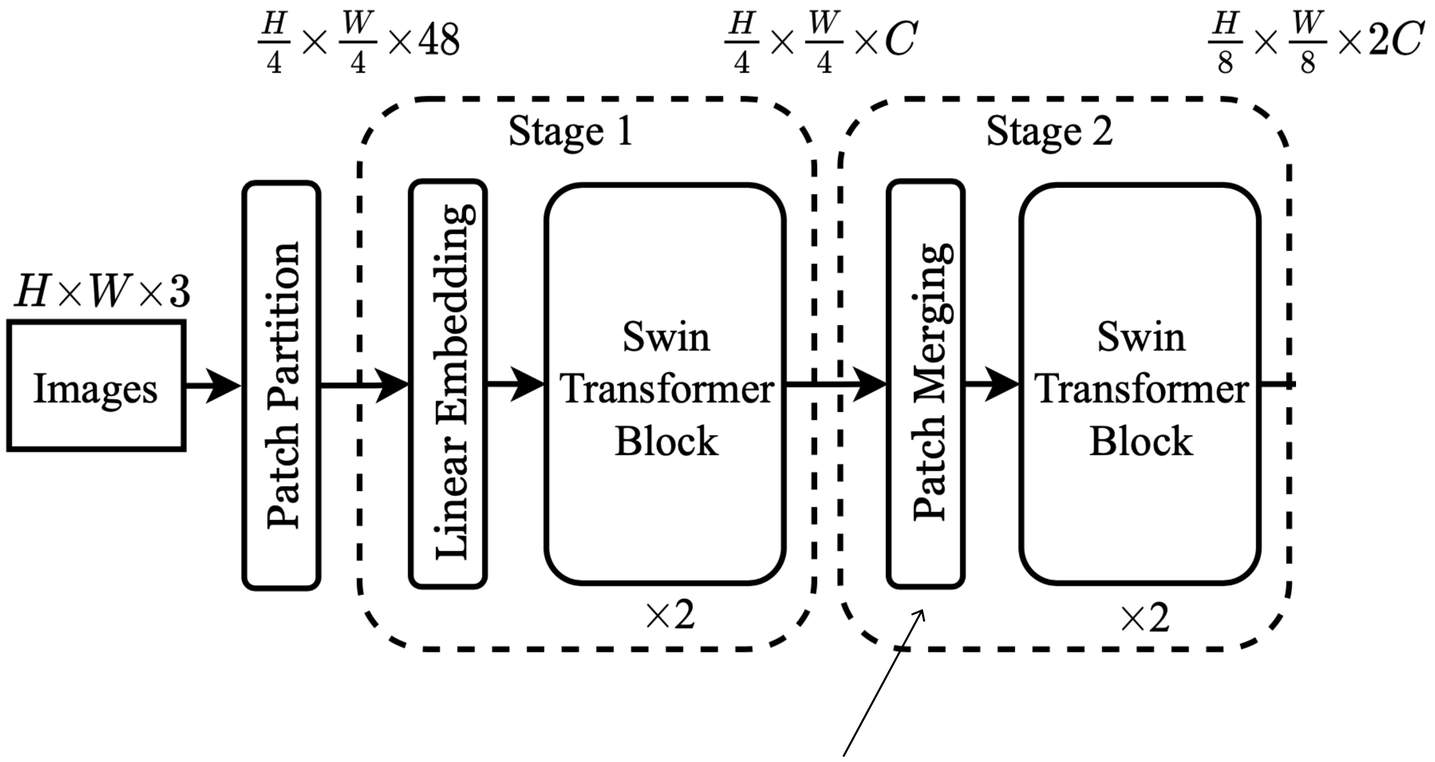


Swin Transformer



Обычный блок трансформера, но attention только внутри окон

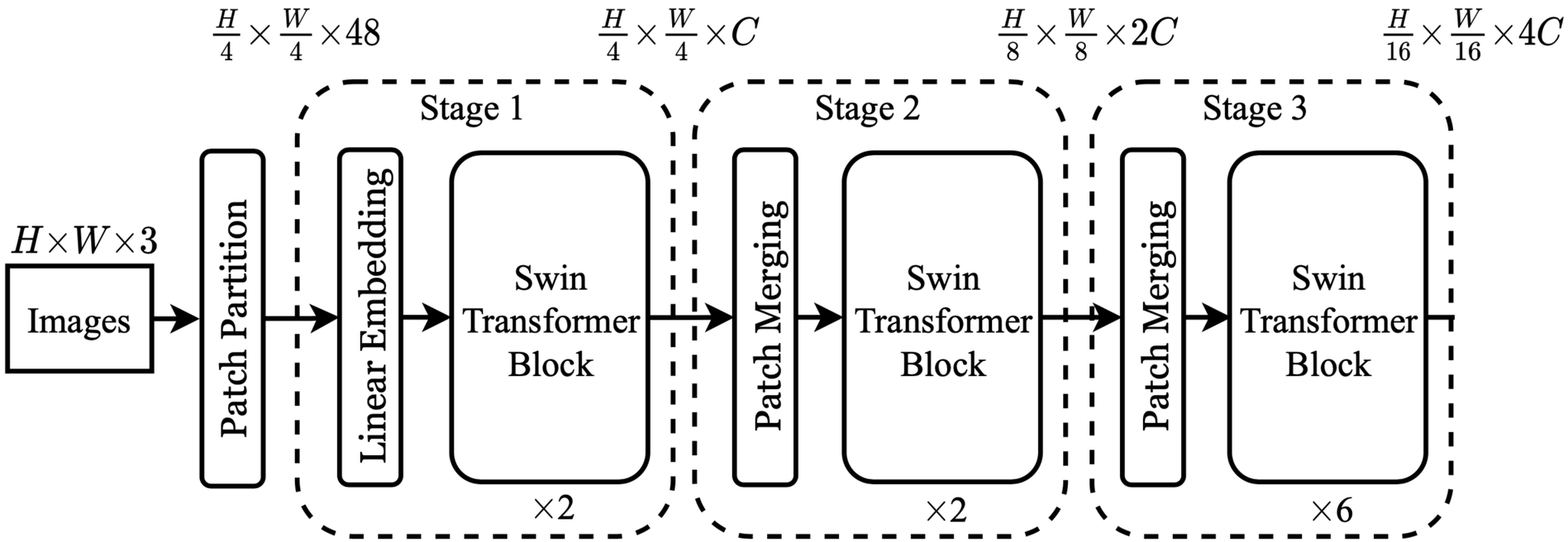
Swin Transformer



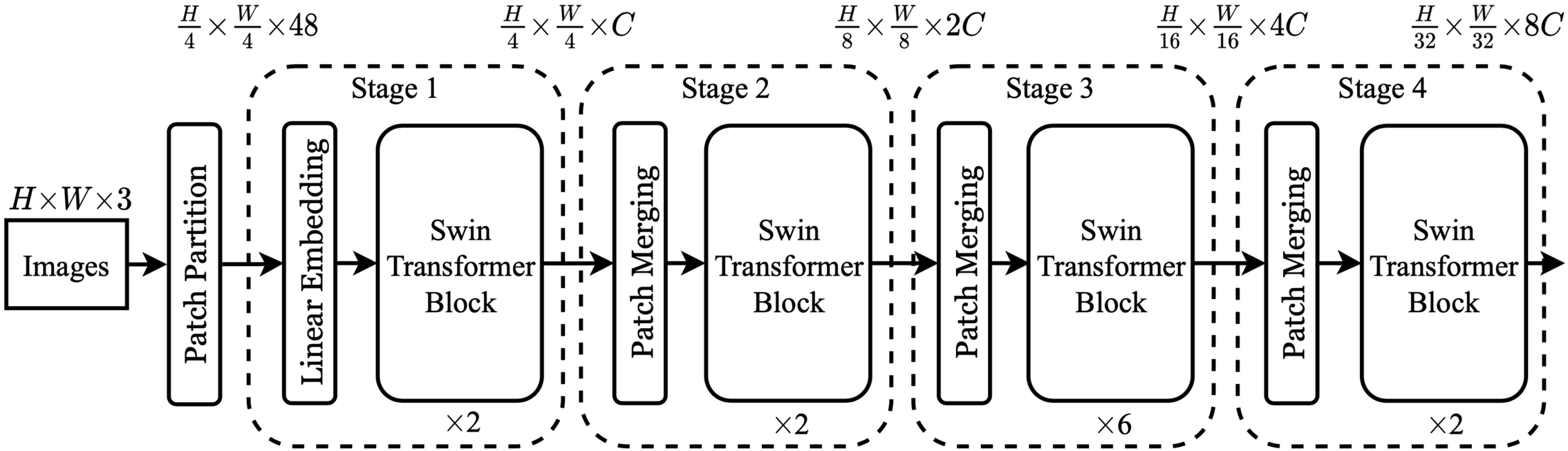
Конкатенируем эмбеддинги патчей в области 2 x 2
и уменьшаем размерность в 2 раза

Сторона окна Attention увеличивается в два раза

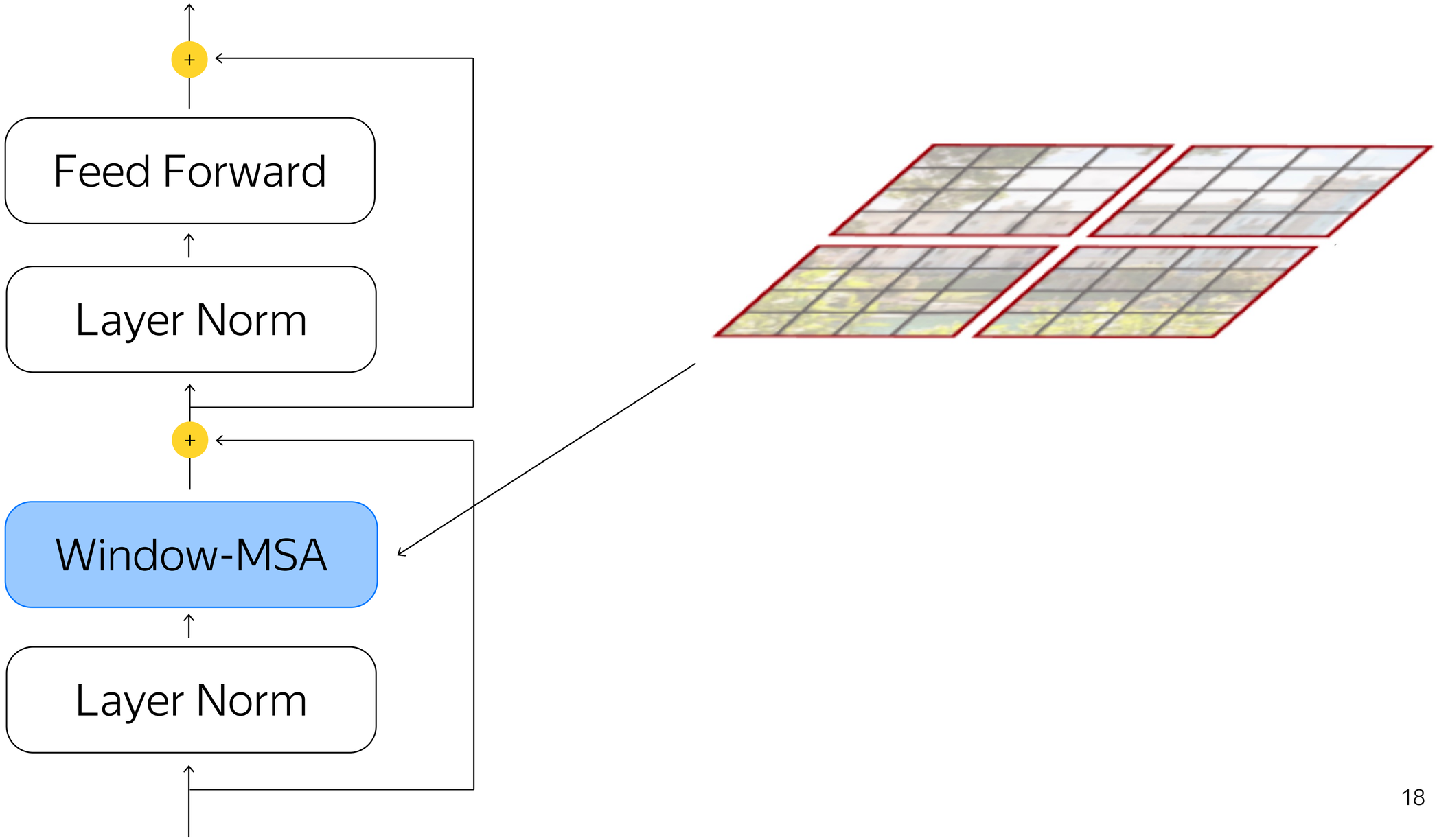
Swin Transformer



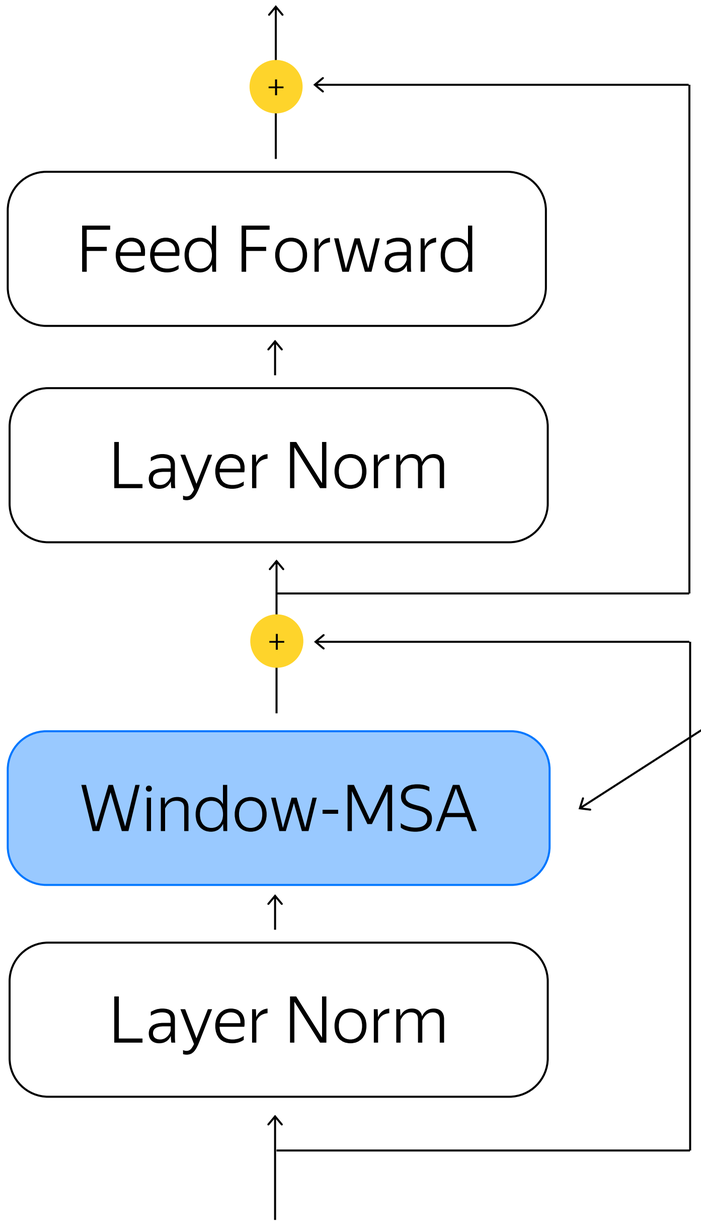
Swin Transformer

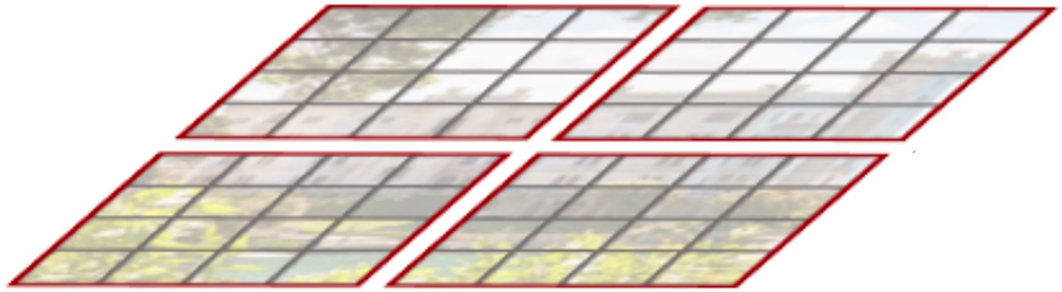


Swin Transformer Block



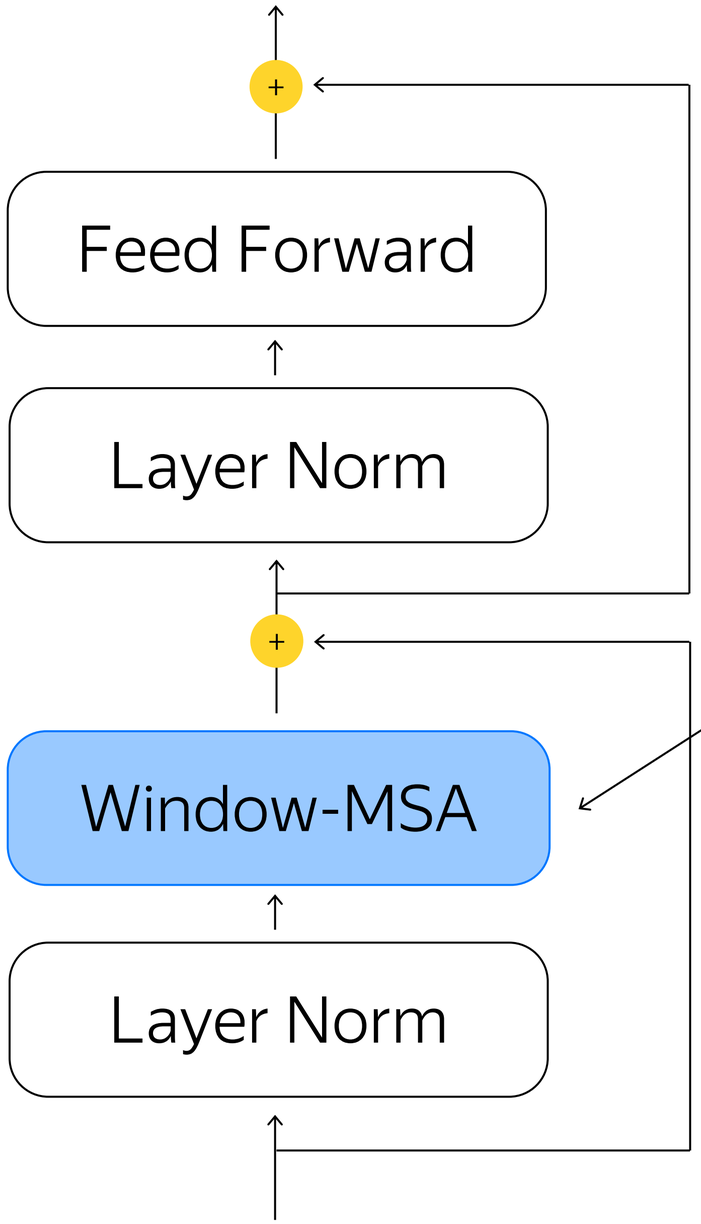
Swin Transformer Block

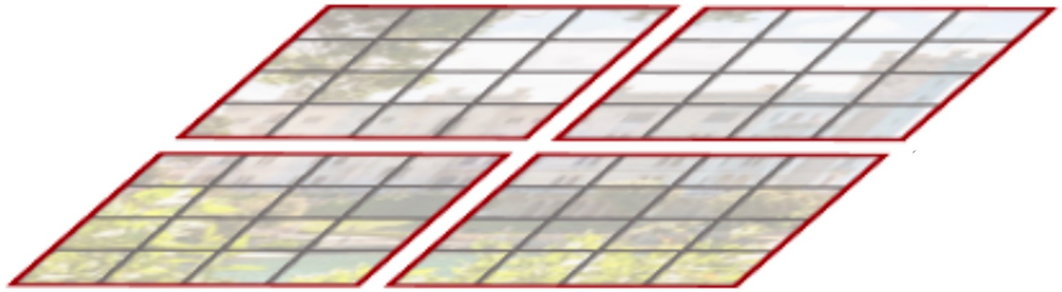




Свой относительный Posistional Encoding
в каждом окне отдельно

Swin Transformer Block

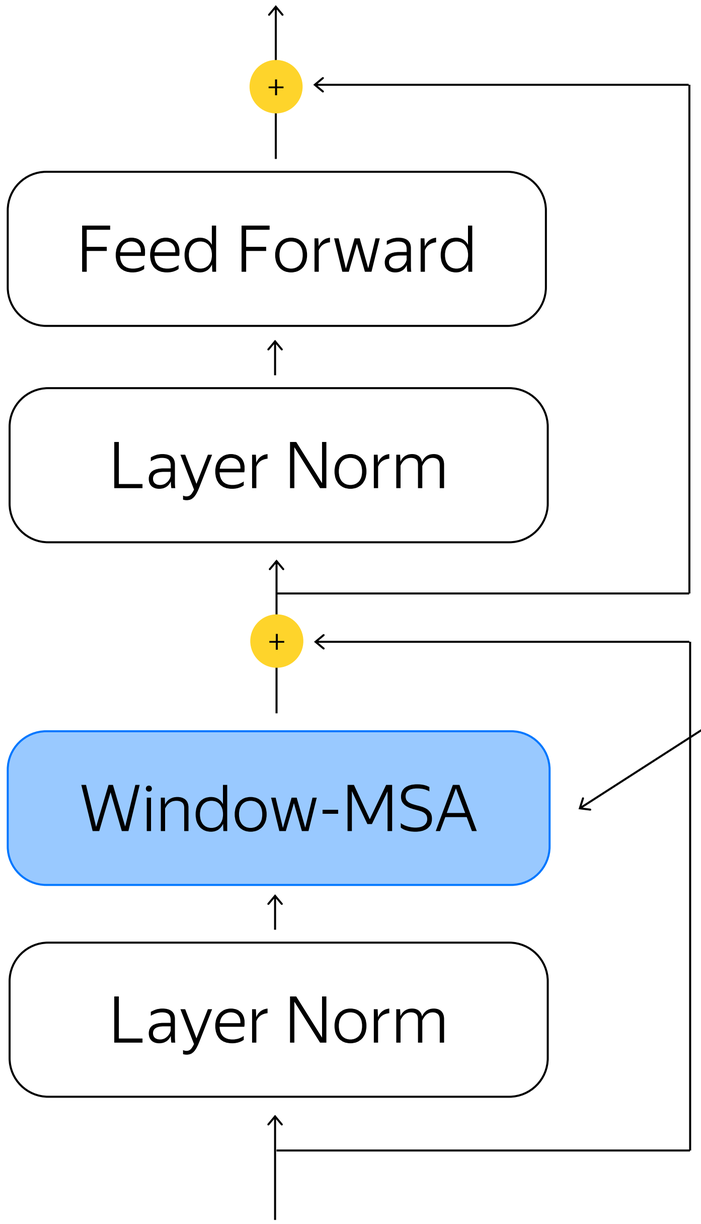


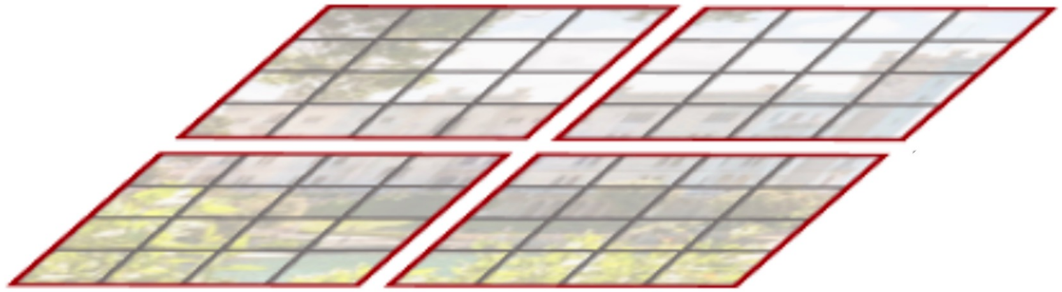


Обычный Self-Attention

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C,$$

Swin Transformer Block





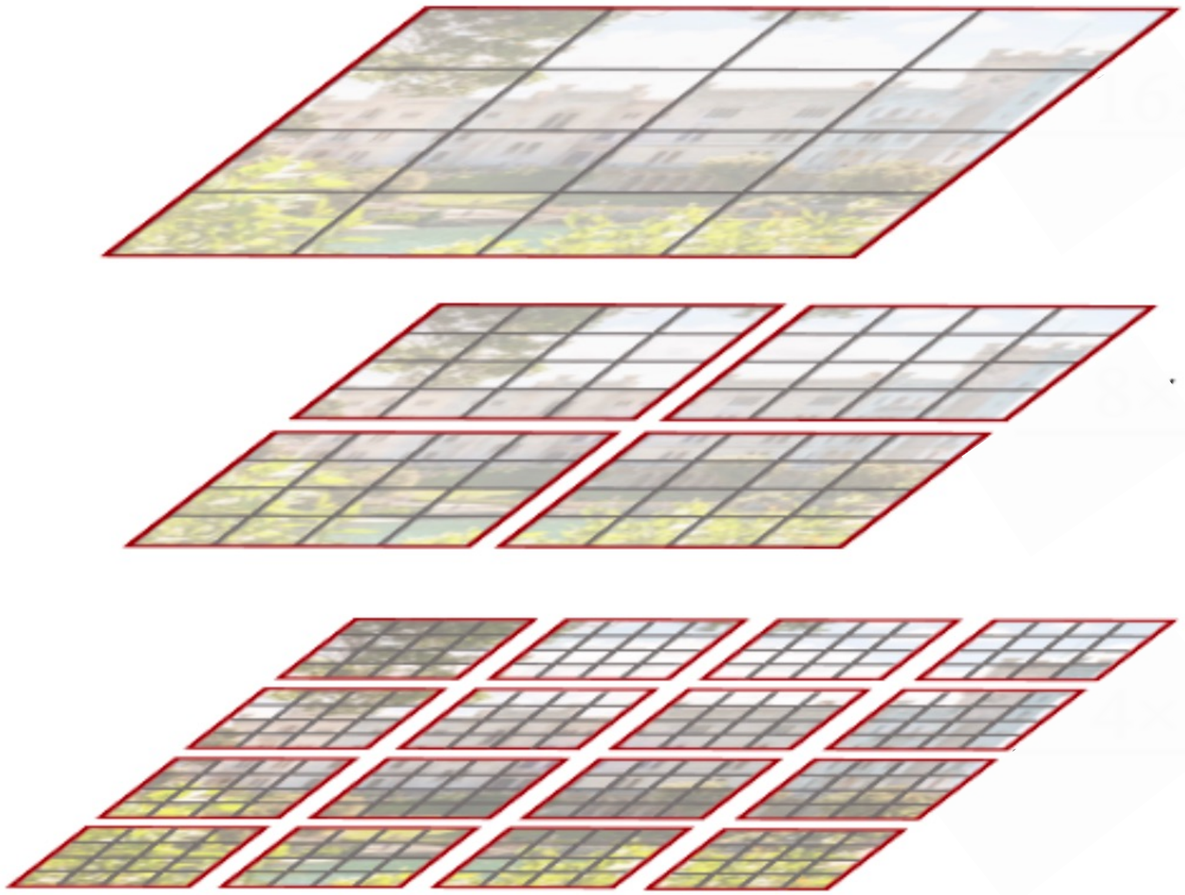
Обычный Self-Attention

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C,$$

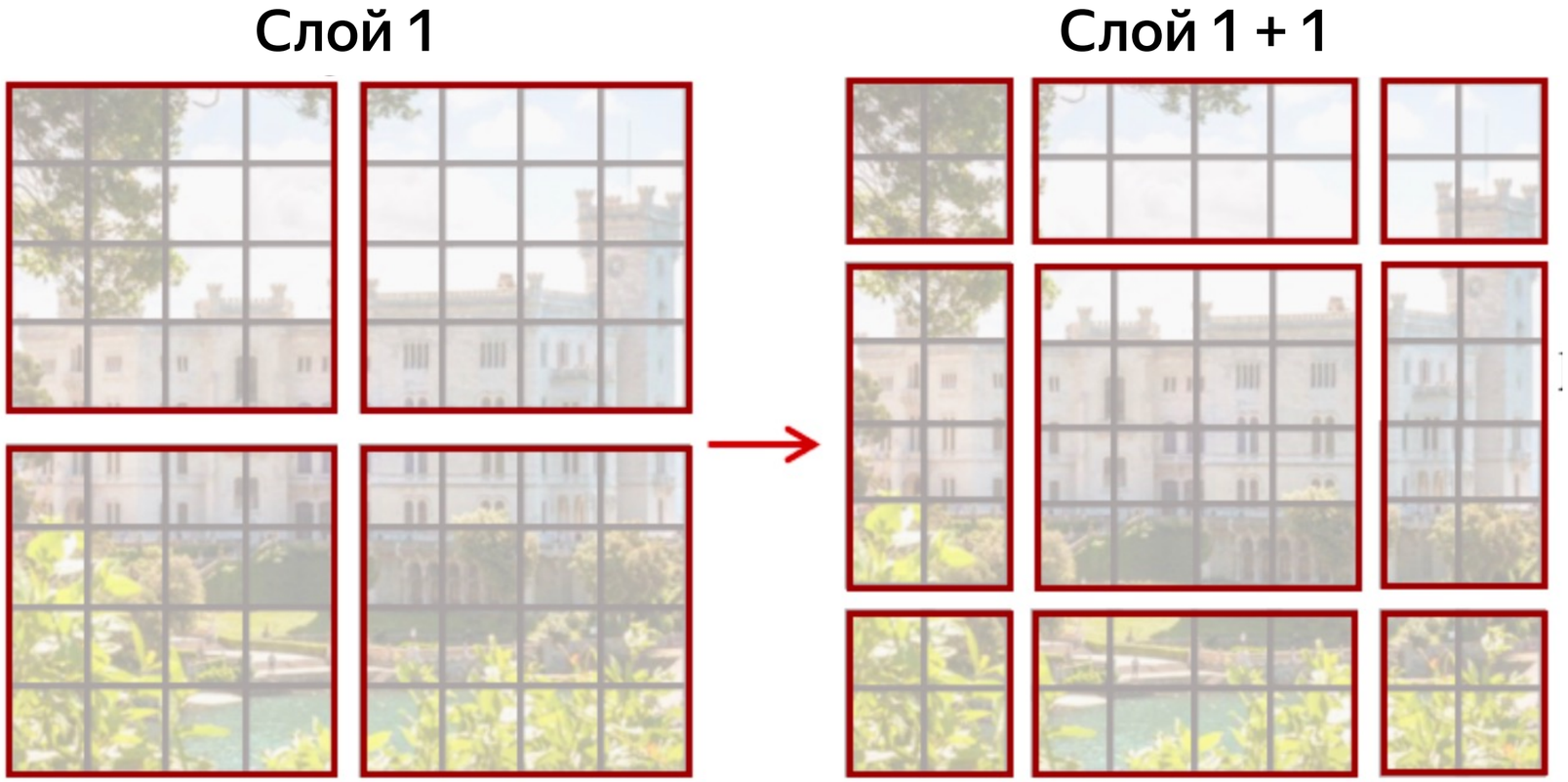
Window Self-Attention с окном M

$$\Omega(\mathrm{W\text{-}MSA}) = 4hwC^2 + 2M^2hwC,$$

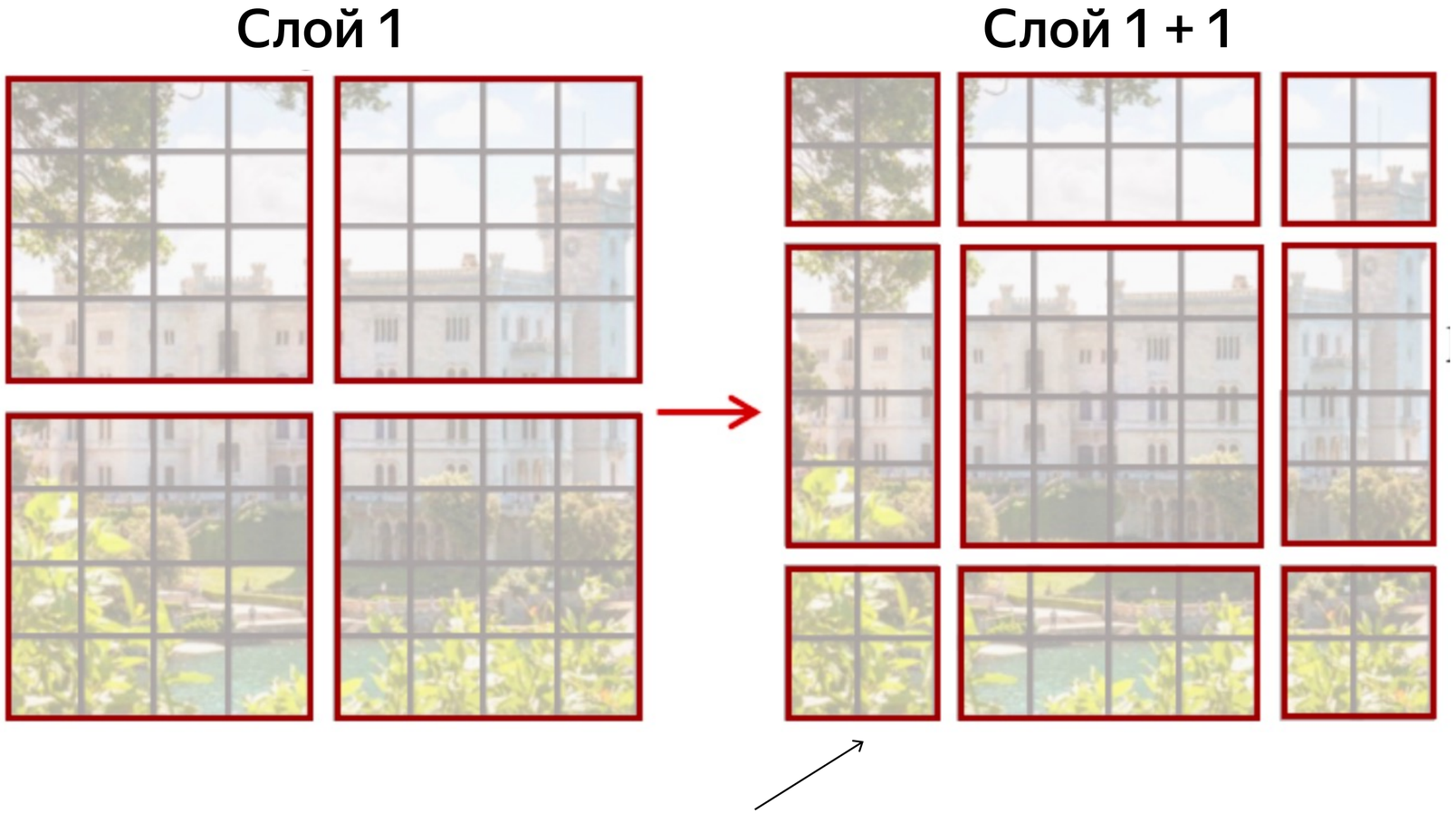
Причем тут shifted windows?



Причем тут shifted windows?



Причем тут shifted windows?



Для более эффективного вычисления обрезки окон объединяются в нормальные окна

К Attention в этом случае добавляется маска

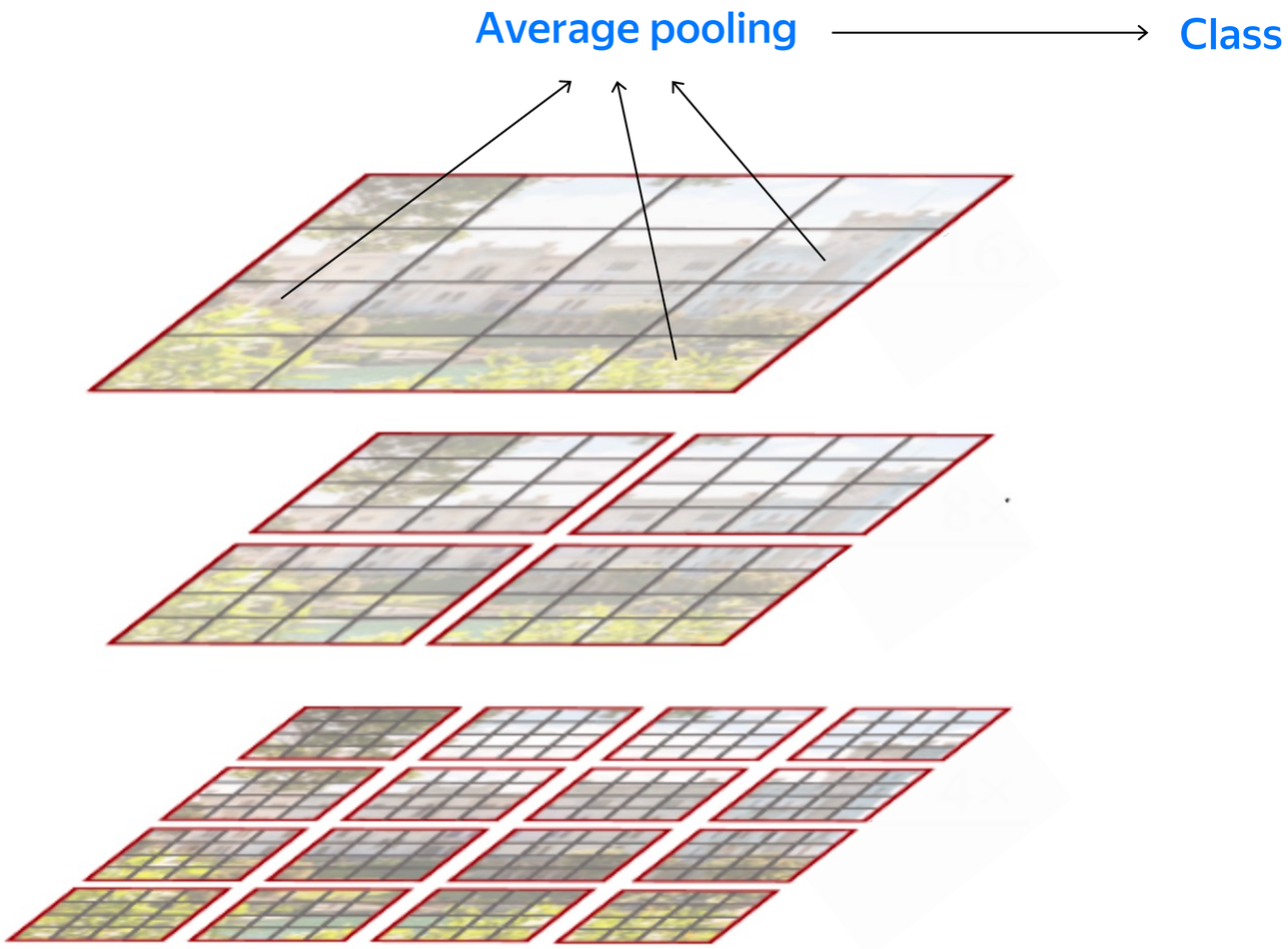
Intro

Swin Transformer

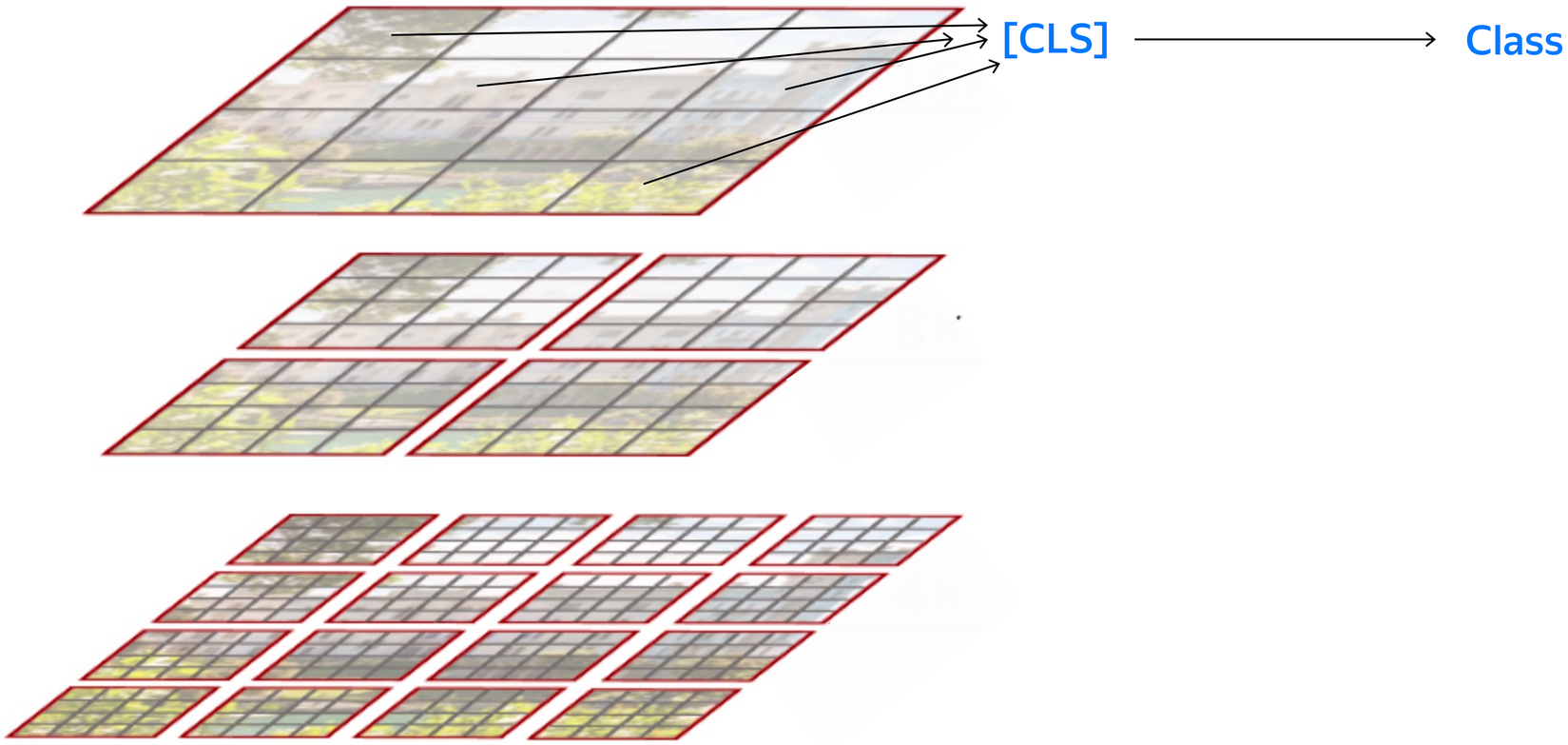
Как применять для разных задач

Zero Shot и One Shot детекция

Классификация для бедных



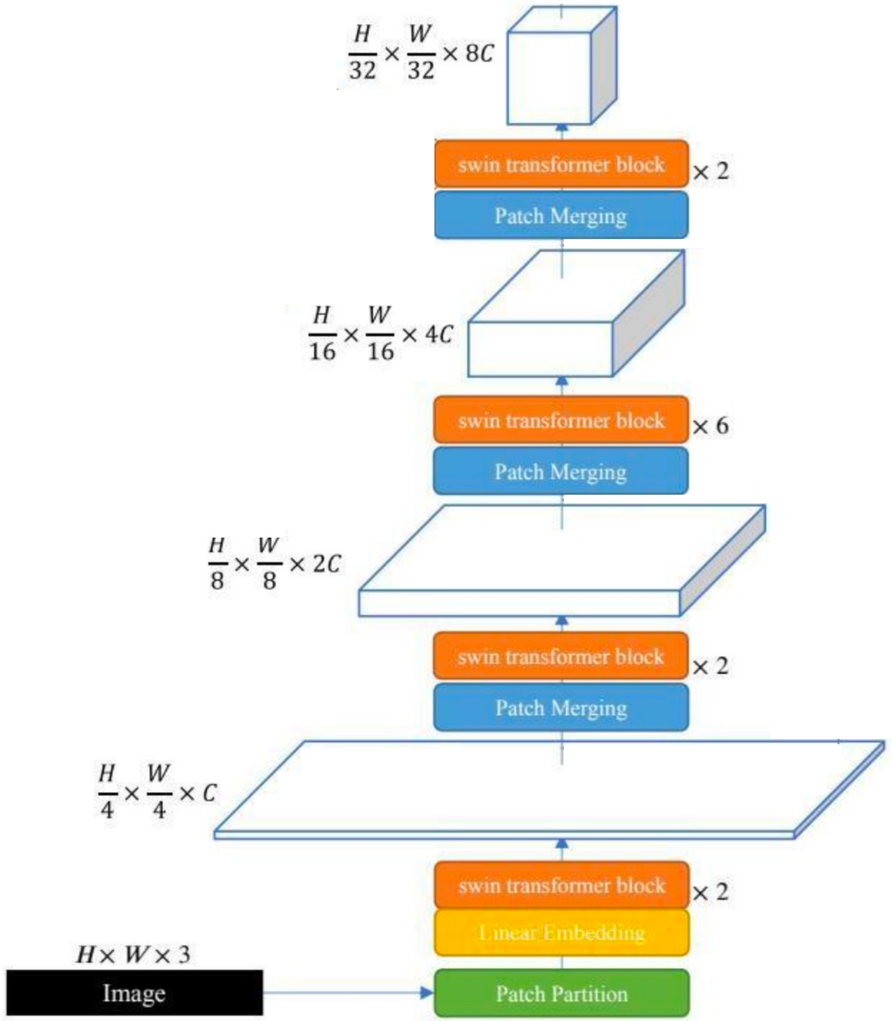
Классификация



Классификация

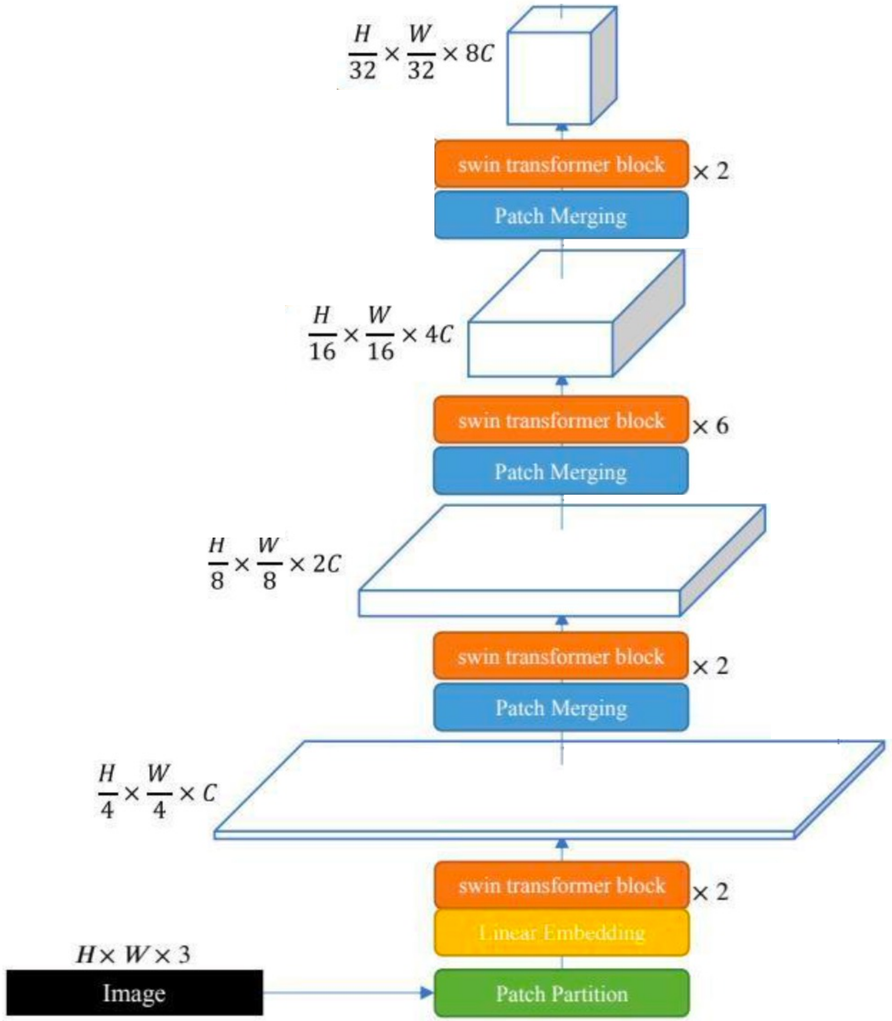
Regular ImageNet-1K trained models					
method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
RegNetY-4G	224^2	21M	4.0G	1156.7	80.0
RegNetY-8G	224^2	39M	8.0G	591.6	81.7
RegNetY-16G	224^2	84M	16.0G	334.7	82.9
EffNet-B3	300^2	12M	1.8G	732.1	81.6
EffNet-B4	380^2	19M	4.2G	349.4	82.9
EffNet-B5	456^2	30M	9.9G	169.1	83.6
EffNet-B6	528^2	43M	19.0G	96.9	84.0
EffNet-B7	600^2	66M	37.0G	55.1	84.3
ViT-B/16	384^2	86M	55.4G	85.9	77.9
ViT-L/16	384^2	307M	190.7G	27.3	76.5
DeiT-S	224^2	22M	4.6G	940.4	79.8
DeiT-B	224^2	86M	17.5G	292.3	81.8
DeiT-B	384^2	86M	55.4G	85.9	83.1
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5
Swin-B	384^2	88M	47.0G	84.7	84.5

Сегментация



https://arxiv.org/abs/2208.02034

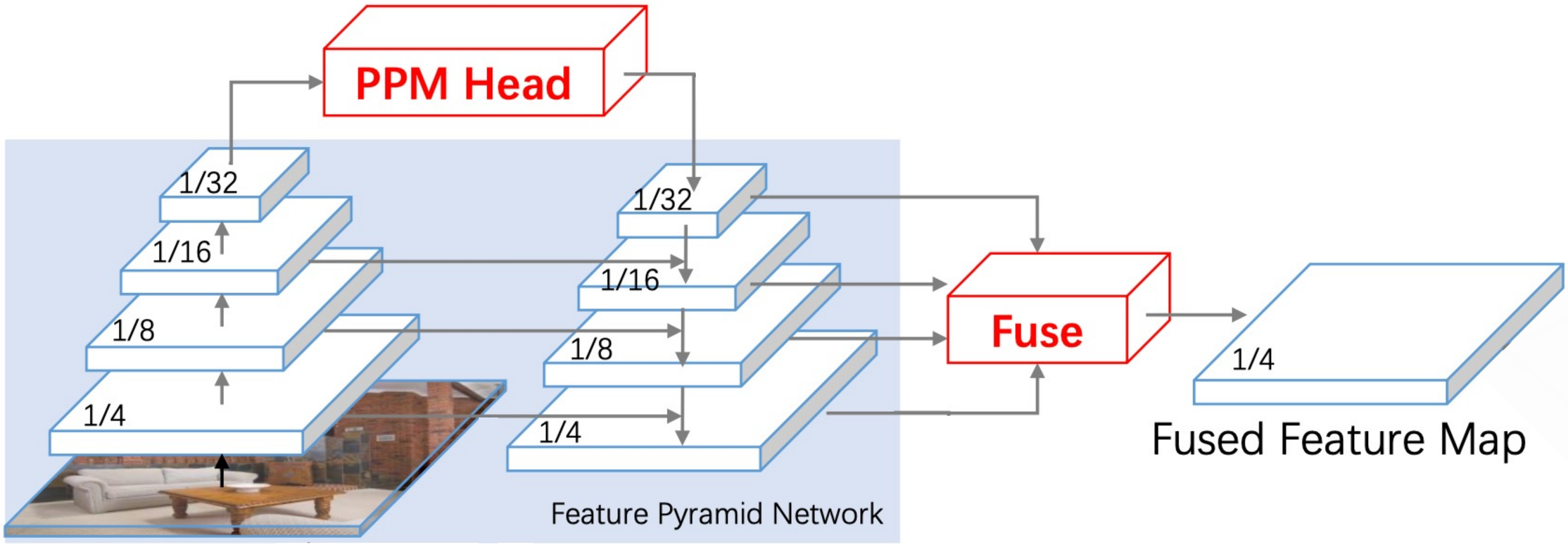
Сегментация



Очень напоминает сверточную сеть

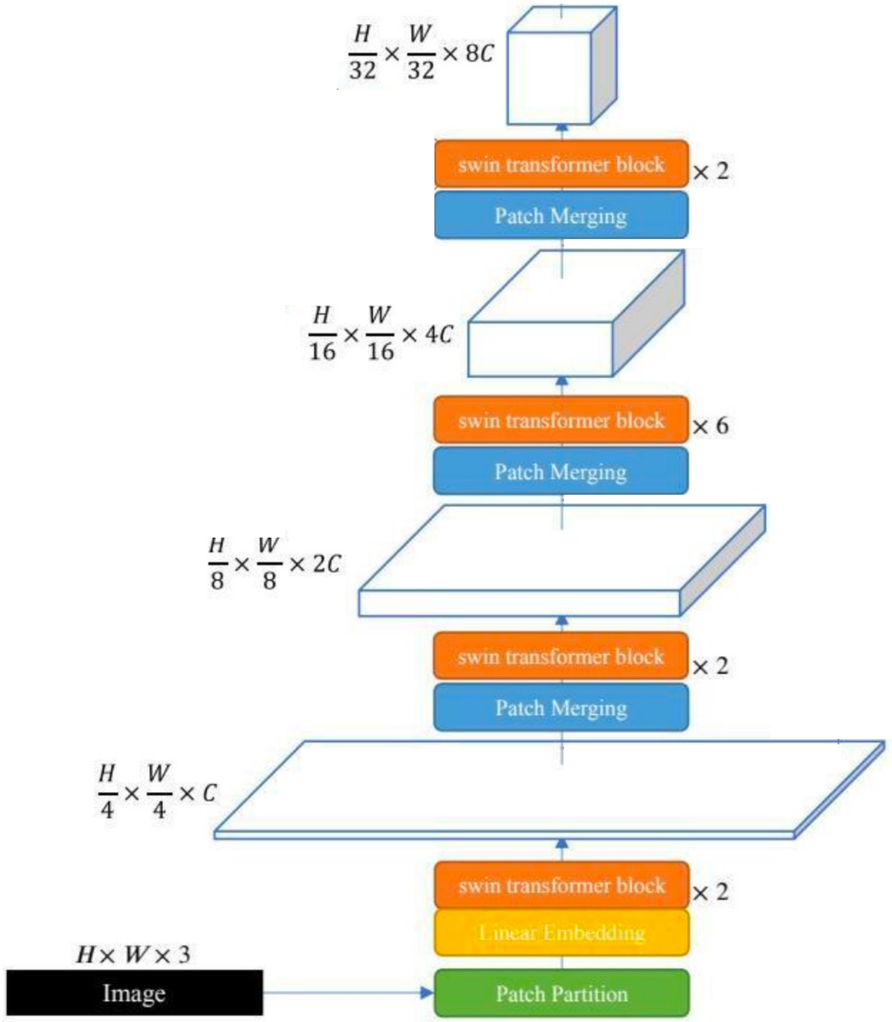
https://arxiv.org/abs/2208.02034

Сегментация – UperNet



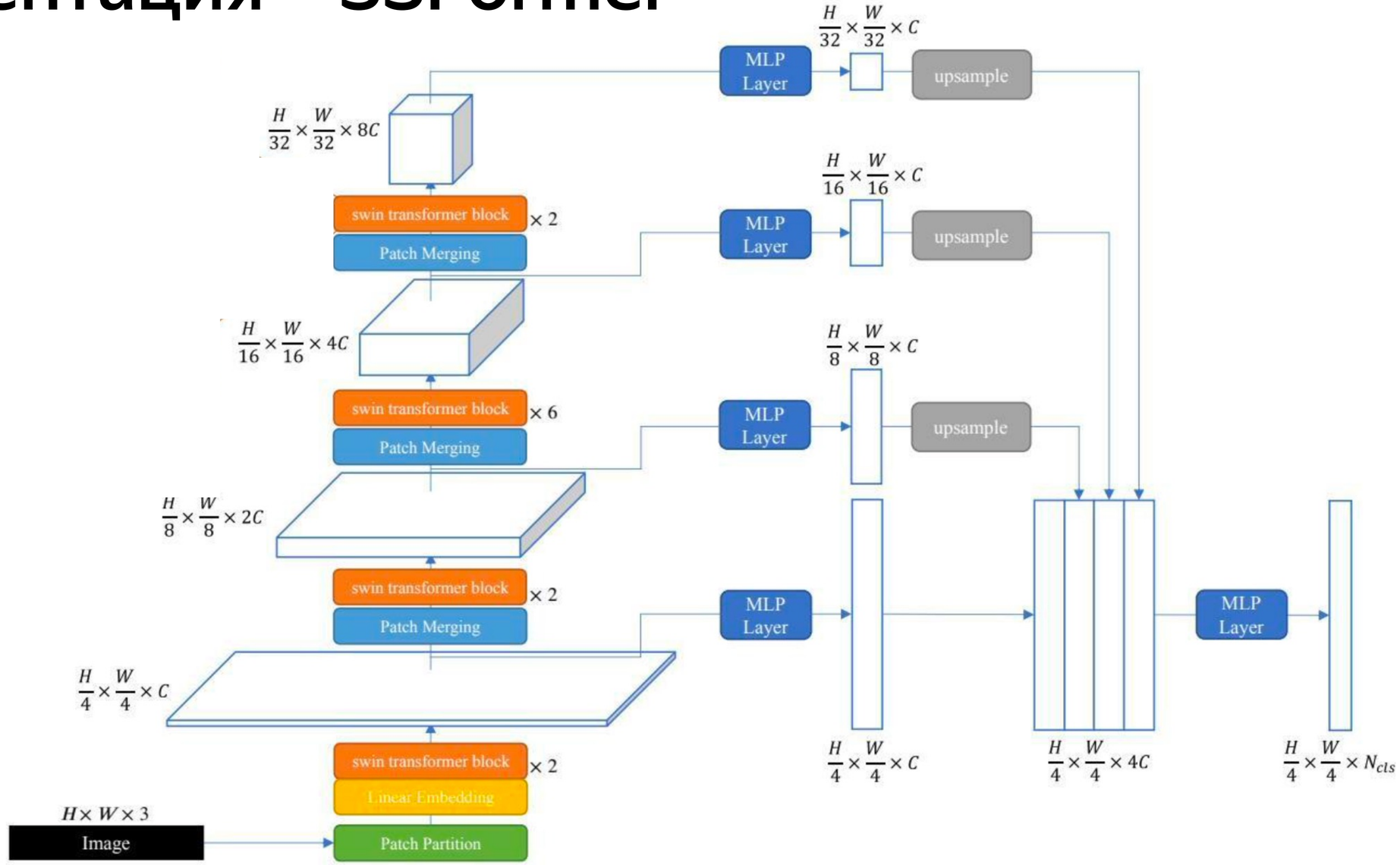
https://arxiv.org/pdf/1807.10221.pdf

Сегментация – SSFormer



https://arxiv.org/abs/2208.02034

Сегментация – SSFormer



https://arxiv.org/abs/2208.02034

Сегментация – ADE 20K

Method	Iteration	Params	Flops*	mIoU
FCN	160000	68.6M	275.7G	39.91
DeepLabV3+	160000	62.7M	255.1G	45.47
DMnet	160000	72.3M	273.6G	45.42
PSPNet	160000	68.1M	256.4G	44.39
PSANet	160000	73.1M	272.5G	43.74
SETR-PUP	160000	317.3M	362.1G	48.24
Swin-T	160000	121.3M	297.2G	50.31
SSformer	160000	87.5M	91.01G	47.71

*The Flops are calculated with the input images of 512×512 resolution.

https://arxiv.org/abs/2208.02034

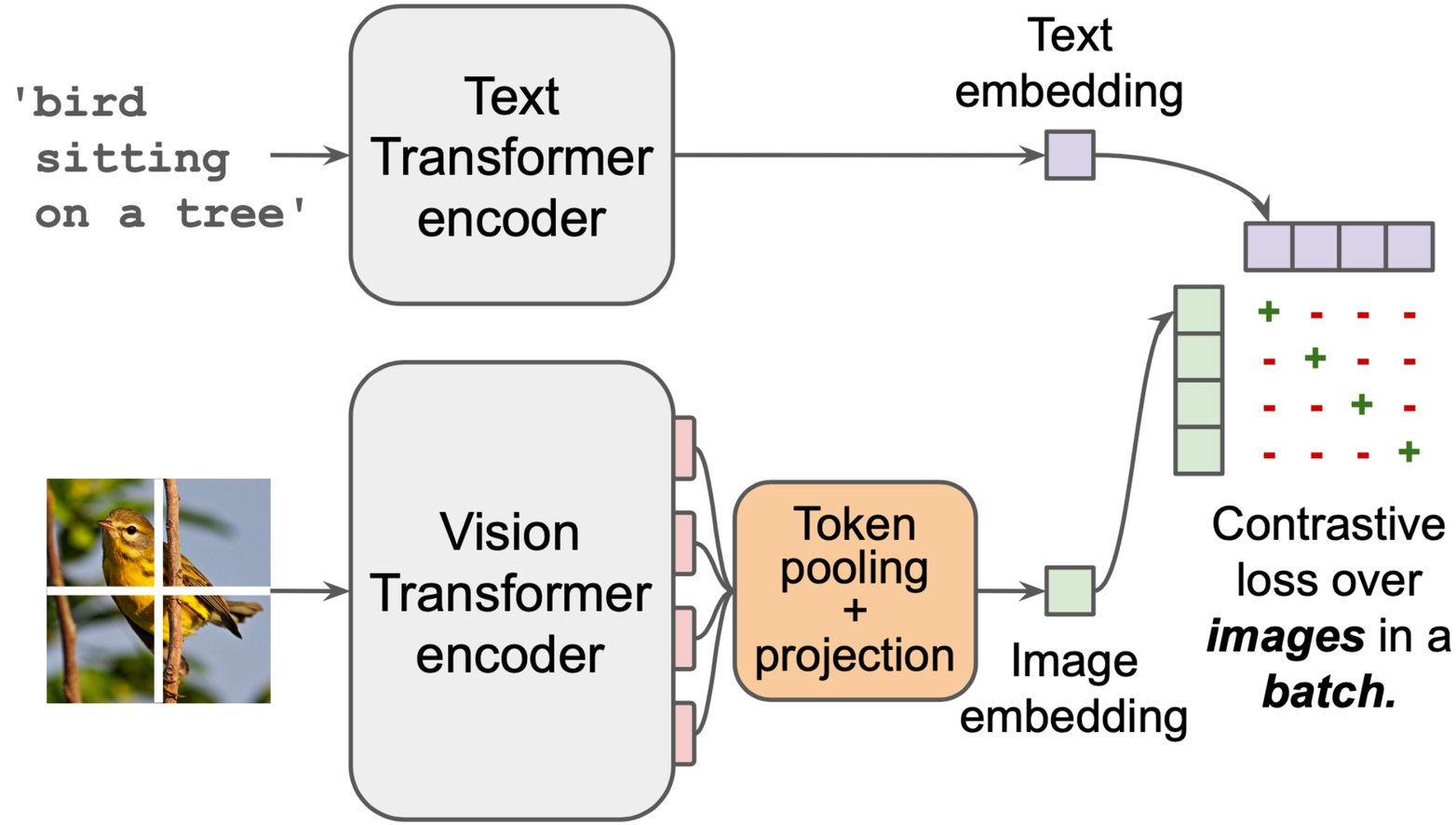
Intro

Swin Transformer

Как применять для разных задач

Zero Shot и One Shot детекция

OWL Detector

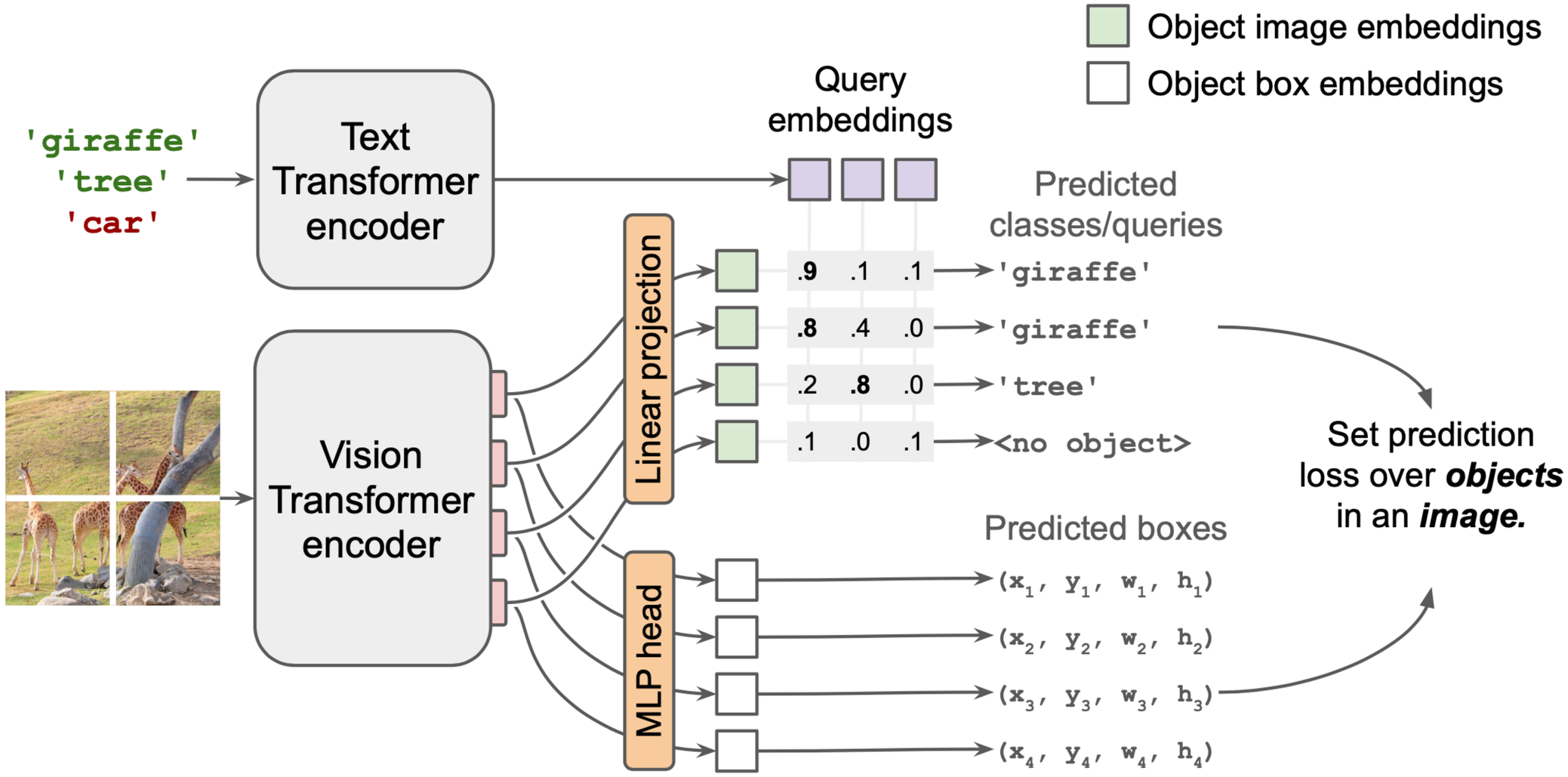


Учим модель сопоставлять текст и изображение

Аналогично CLIP

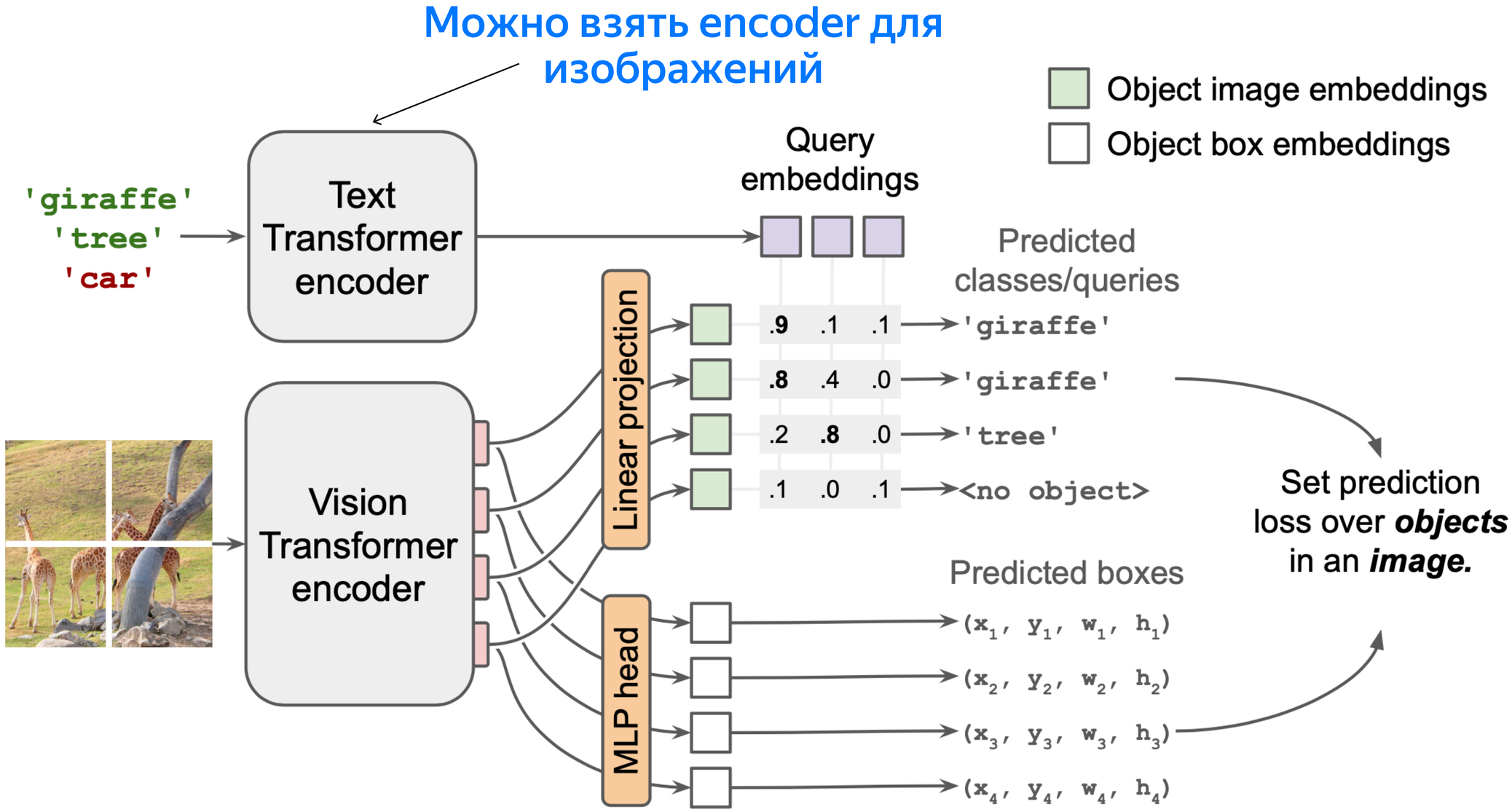
https://arxiv.org/pdf/2205.06230v2.pdf

OWL Detector



https://arxiv.org/pdf/2205.06230v2.pdf

OWL Detector



https://arxiv.org/pdf/2205.06230v2.pdf