



# **Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**

Доклад по статье

Курченко Лилия, Безрукова Анастасия, Голобородько Ира  
БПМИ 191

НИС Машинное обучение и приложения, 05.10.2022

---

# Обзор статьи

# Задача

Хотим сделать трансформер универсальной моделью для CV

**Classification**



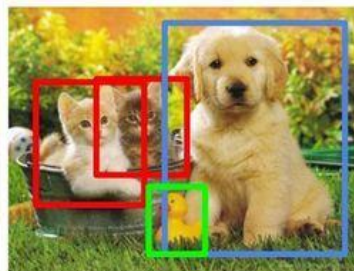
CAT

**Classification  
+ Localization**



CAT

**Object Detection**



CAT, DOG, DUCK

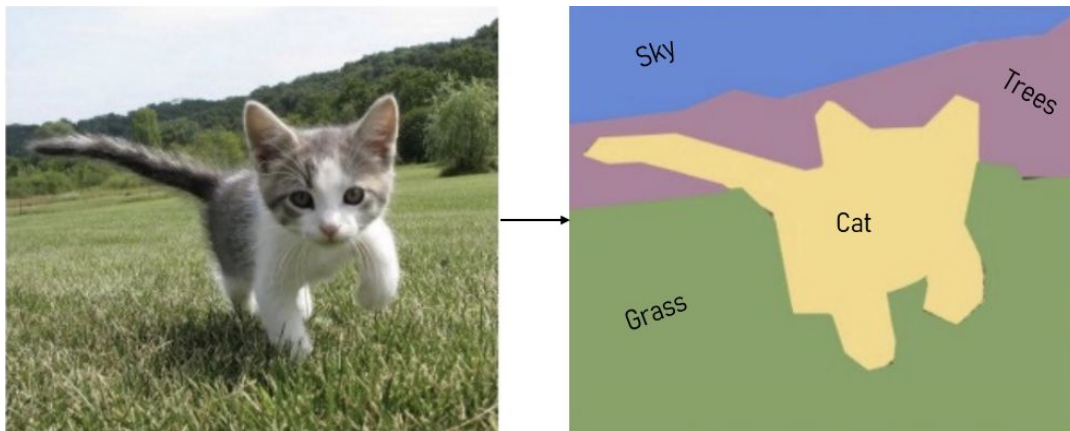
**Instance  
Segmentation**



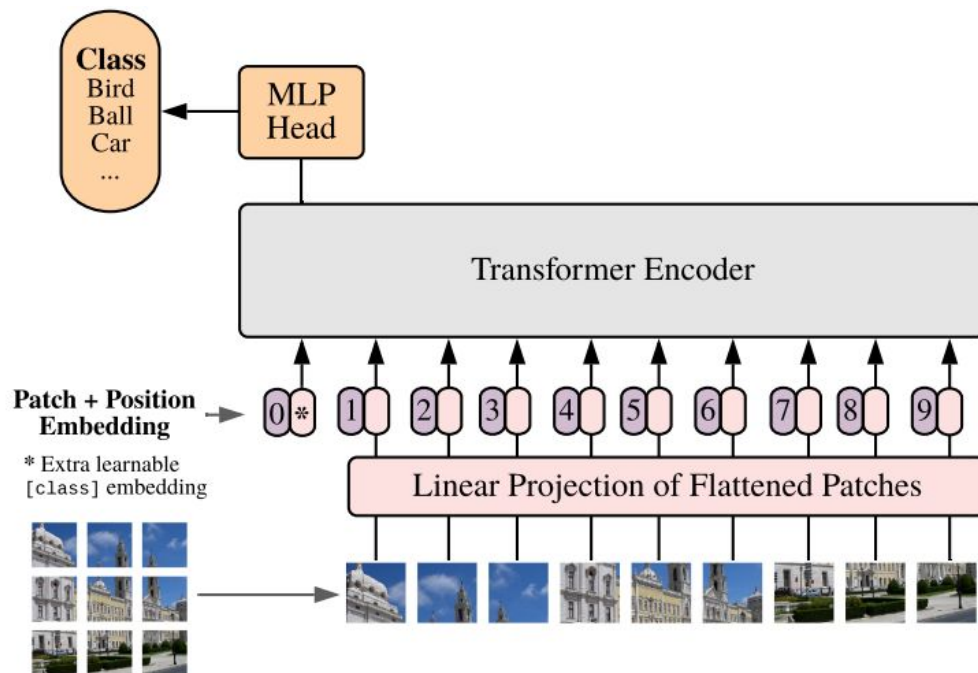
CAT, DOG, DUCK

# Трансформеры в CV

- Различие в относительном масштабе элементов на картинке
- Огромное число токенов (пиксели в задаче сегментации)



# ViT

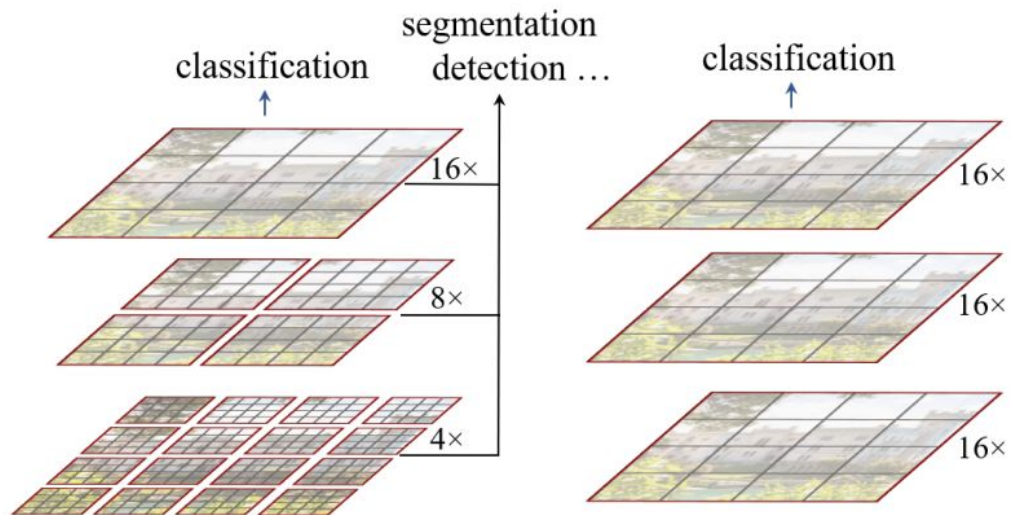


# SWIN-transformer

- Всё изображение бьётся на локальные окошки
- Иерархия feature maps

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

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



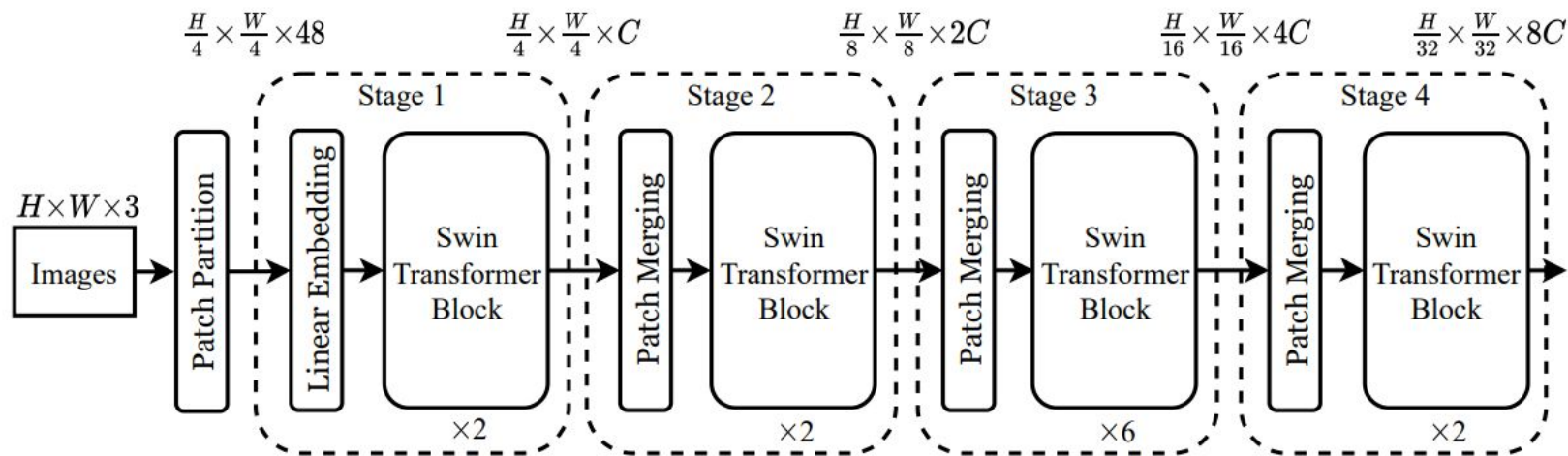
SWIN vs ViT

## Shifted window

- Окна в последующем разбиении пересекают границу предыдущего  $\Rightarrow$  связь между окошками



Два последовательных слоя SWIN

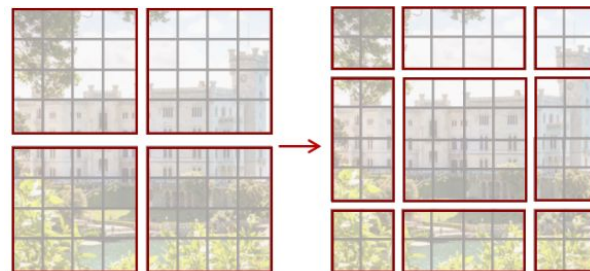
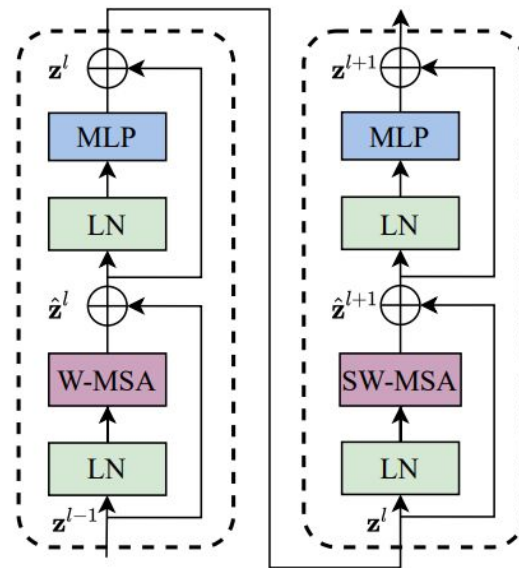


Архитектура SWIN-T

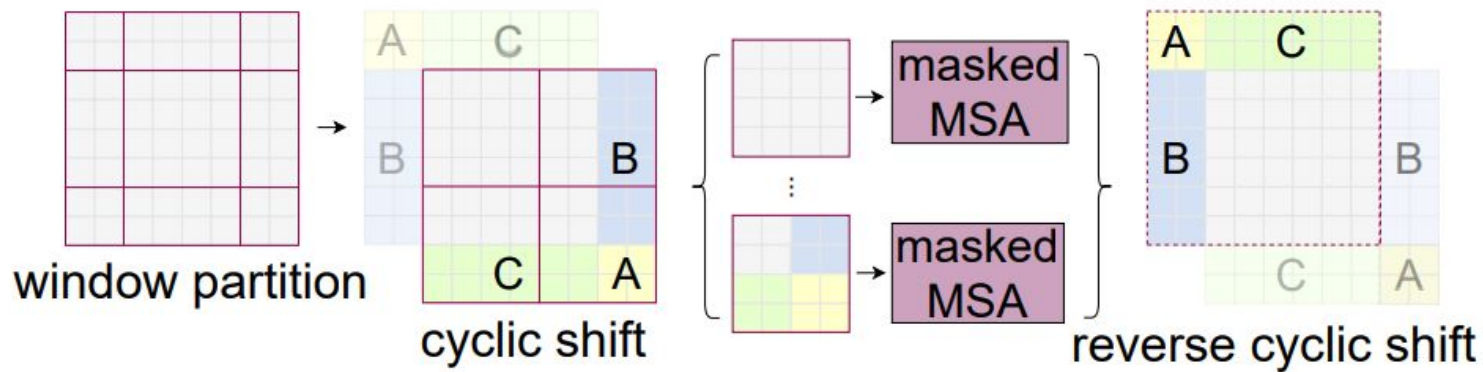


# SWIN-block

- LN == LayerNorm
- W-MSA = Multi-head self-attention, regular window
- SW-MSA = Multi-head self-attention, shifted window



## Циклический сдвиг батчей





## Bias

- Информация о позиции – обучаемая
- Параметризуем  $B$  через меньшую матрицу


$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d} + B)V$$

## Эксперименты

(a) Regular ImageNet-1K trained models


method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
RegNetY-4G [48]	224 <sup>2</sup>	21M	4.0G	1156.7	80.0
RegNetY-8G [48]	224 <sup>2</sup>	39M	8.0G	591.6	81.7
RegNetY-16G [48]	224 <sup>2</sup>	84M	16.0G	334.7	82.9
EffNet-B3 [58]	300 <sup>2</sup>	12M	1.8G	732.1	81.6
EffNet-B4 [58]	380 <sup>2</sup>	19M	4.2G	349.4	82.9
EffNet-B5 [58]	456 <sup>2</sup>	30M	9.9G	169.1	83.6
EffNet-B6 [58]	528 <sup>2</sup>	43M	19.0G	96.9	84.0
EffNet-B7 [58]	600 <sup>2</sup>	66M	37.0G	55.1	84.3
ViT-B/16 [20]	384 <sup>2</sup>	86M	55.4G	85.9	77.9
ViT-L/16 [20]	384 <sup>2</sup>	307M	190.7G	27.3	76.5
DeiT-S [63]	224 <sup>2</sup>	22M	4.6G	940.4	79.8
DeiT-B [63]	224 <sup>2</sup>	86M	17.5G	292.3	81.8
DeiT-B [63]	384 <sup>2</sup>	86M	55.4G	85.9	83.1
Swin-T	224 <sup>2</sup>	29M	4.5G	755.2	81.3
Swin-S	224 <sup>2</sup>	50M	8.7G	436.9	83.0
Swin-B	224 <sup>2</sup>	88M	15.4G	278.1	83.5
Swin-B	384 <sup>2</sup>	88M	47.0G	84.7	84.5

Классификация, датасет ImageNet-1K



Method	mini-val		test-dev		#param. FLOPs	
	AP <sup>box</sup>	AP <sup>mask</sup>	AP <sup>box</sup>	AP <sup>mask</sup>		
RepPointsV2* [12]	-	-	52.1	-	-	-
GCNet* [7]	51.8	44.7	52.3	45.4	-	1041G
RelationNet++* [13]	-	-	52.7	-	-	-
SpineNet-190 [21]	52.6	-	52.8	-	164M	1885G
ResNeSt-200* [78]	52.5	-	53.3	47.1	-	-
EfficientDet-D7 [59]	54.4	-	55.1	-	77M	410G
DetectoRS* [46]	-	-	55.7	48.5	-	-
YOLOv4 P7* [4]	-	-	55.8	-	-	-
Copy-paste [26]	55.9	47.2	56.0	47.4	185M	1440G
X101-64 (HTC++)	52.3	46.0	-	-	155M	1033G
Swin-B (HTC++)	56.4	49.1	-	-	160M	1043G
Swin-L (HTC++)	57.1	49.5	57.7	50.2	284M	1470G
Swin-L (HTC++)*	<b>58.0</b>	<b>50.4</b>	<b>58.7</b>	<b>51.1</b>	284M	-

COCO object detection



ADE20K		val	test	#param.	FLOPs	FPS
Method	Backbone	mIoU	score			
DANet [23]	ResNet-101	45.2	-	69M	1119G	15.2
DLab.v3+ [11]	ResNet-101	44.1	-	63M	1021G	16.0
ACNet [24]	ResNet-101	45.9	38.5	-		
DNL [71]	ResNet-101	46.0	56.2	69M	1249G	14.8
OCRNet [73]	ResNet-101	45.3	56.0	56M	923G	19.3
UperNet [69]	ResNet-101	44.9	-	86M	1029G	20.1
OCRNet [73]	HRNet-w48	45.7	-	71M	664G	12.5
DLab.v3+ [11]	ResNeSt-101	46.9	55.1	66M	1051G	11.9
DLab.v3+ [11]	ResNeSt-200	48.4	-	88M	1381G	8.1
SETR [81]	T-Large <sup>‡</sup>	50.3	61.7	308M	-	-
UperNet	DeiT-S <sup>†</sup>	44.0	-	52M	1099G	16.2
UperNet	Swin-T	46.1	-	60M	945G	18.5
UperNet	Swin-S	49.3	-	81M	1038G	15.2
UperNet	Swin-B <sup>‡</sup>	51.6	-	121M	1841G	8.7
UperNet	Swin-L <sup>‡</sup>	<b>53.5</b>	<b>62.8</b>	234M	3230G	6.2

Задача сегментации, датасет ADE20K



## Итог

- Разбиение картинки на окна с фиксированным числом патчей  $\Rightarrow$  линейная сложность
- Shifted window  $\Rightarrow$  связь между соседними окошками
- Patch merge  $\Rightarrow$  иерархическое признаковое представление

Profit: эффективный трансформер, универсальный для задач CV

---

# Рецензия





## Данные о публикации

Дата публикации:

- 25/03/2021 (v1)
- 17/08/2021 (v2)

Авторы: Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo (Microsoft Research Asia)

Представление: ICCV 2021



## Предыстория: развитие идеи

12 Jun 2017 - выходит статья "Attention is All You Need"

2017 - 2020 гг - эксперименты с применением трансформеров к изображениям.

22 Oct 2020 - выходит статья "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"

25 Mar 2021 - выходит статья "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows"



## На чем основана работа

Список наиболее цитируемых статей в работе:

- 1) "Training data-efficient image transformers & distillation through attention" (Touvron et al.) [15]
- 2) "An image is worth 16x16 words: Transformers for image recognition at scale" (Dosovitskiy et al.) [12]
- 3) "Rethinking model scaling for convolutional neural networks" (M. Tan, Q. Le) [9]

Работы, которые авторы называют базовыми:

- 1) "An image is worth 16x16 words: Transformers for image recognition at scale" (Dosovitskiy et al.)
- 2) "Training data-efficient image transformers & distillation through attention" (Touvron et al.)
- 3) "Tokens to-token vit: Training vision transformers from scratch on imagenet" (Yuan et al.)
- 4) "Do we really need explicit position encodings for vision transformers?" (Chu et al.)
- 5) "Transformer in transformer" (Han et al.)



## Работы, вышедшие в то же время

Критерий: соответствующая тематика, публикация в период с 25 марта 2021 по 17 августа 2021

- 1) "Focal Self-attention for Local-Global Interactions in Vision Transformers" (Yang et al., июль 2021)
- 2) "XCiT: Cross-Covariance Image Transformers" (El-Nouby et al., июнь 2021)



## В каких работах цитируется

В сумме "Swin Transformer" была процитирована в 404 (scite\_) работах за 1,5 года

Избранные работы:

- В которых развивается идея применения Swin Transformer:
  - "Video Swin Transformer" (Liu et al., июнь 2021)
  - "Swin Transformer V2: Scaling Up Capacity and Resolution" (Liu et al., апрель 2022 года)
  - "Swin Transformer for Fast MRI" (Huang et al., апрель 2022)
  - "Swin-Pose: Swin Transformer Based Human Pose Estimation" (Xiong et al., июнь 2022)
- В которых описываемая архитектура сравнивается с Swin Transformer:
  - "Vision Transformers with Hierarchical Attention" (Sun et al., июнь 2022)
  - "MixFormer: Mixing Features across Windows and Dimensions" (Chen et al., апрель 2022),
  - "K-Net: Towards Unified Image Segmentation" (Zhang et al., ноябрь 2021)
  - "nnFormer: Interleaved Transformer for Volumetric Segmentation" (Zhou et al. февраль 2022)



## Сильные/слабые стороны работы

- + Качество написания работы
- + Теоретические выкладки, используемые в работе
- + Наличие подтверждающих и подкрепляющих выводы результатов экспериментов
- + Новизна работы
- + Актуальность работы
  
- Неконкурентность (устаревание)

---

# Ссылки

- <https://arxiv.org/pdf/2103.14030.pdf>
- <https://towardsdatascience.com/swin-vision-transformers-hacking-the-human-eye-4223ba9764c3>
- <https://youtu.be/vrSCD88X8BU?t=18692>