# Vision Transformer & Swin Transformer

## [CSED490X] Recent Trends in ML: A Large-Scale Perspective

Jungtaek Kim
jtkim@postech.ac.kr

POSTECH
Pohang 37673, Republic of Korea
https://jungtaek.github.io

April 27, 2022

# Table of Contents

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

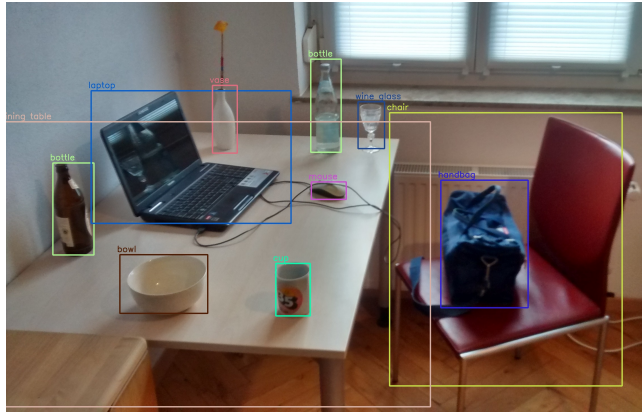
# Today's Lecture

- Variants of the Transformer model [Vaswani et al., 2017] for vision tasks will be covered.
- Vision Transformer and Swin Transformer will be introduced.
- Unlike the language models covered in the previous lectures, it solves a task related to visual information, e.g., image classification.
- Image is generally represented as $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, where $H$ and $W$ are height and width of image, respectively, and $C$ is channel size. If $\mathbf{x}$ is a grayscale image, $C = 1$, and if $\mathbf{x}$ is a colored image, $C = 3$.

[Vaswani et al., 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, pages 5998–6008, Long Beach, California, USA, 2017.

# Why Is Computer Vision Impactful?

- Object detection




Taken from Wikipedia.

# Why Is Computer Vision Impactful?

- Object detection


Taken from Wikipedia.

# Why Is Computer Vision Impactful?

- Video tracking

Taken from Wikipedia.

# An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

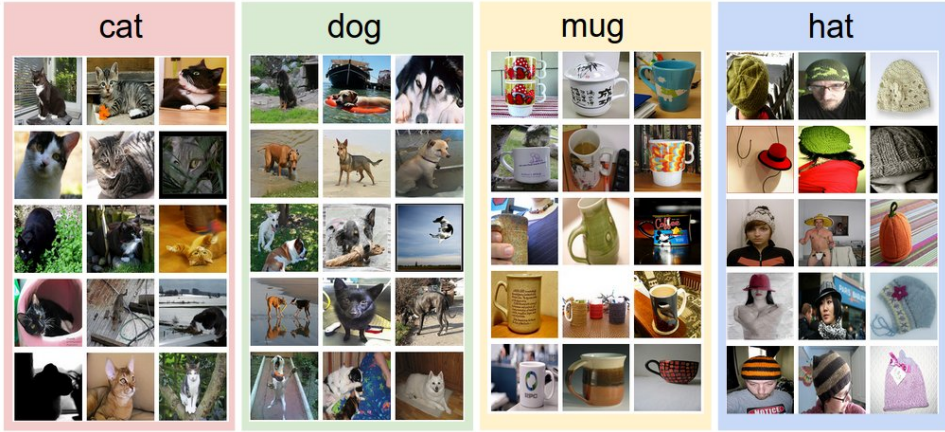
# Vision Transformer (ViT)

- While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited.
- In vision, convolutional neural networks are widely used, but the reliance on convolutional neural networks is not necessary.
- A pure transformer applied directly to sequences of image patches can perform very well on image classification tasks.
- When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks, Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional neural networks.

---

[Dosovitskiy et al., 2021] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 2021.

# Tasks & Datasets




Taken from https://cs231n.github.io.

# ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)

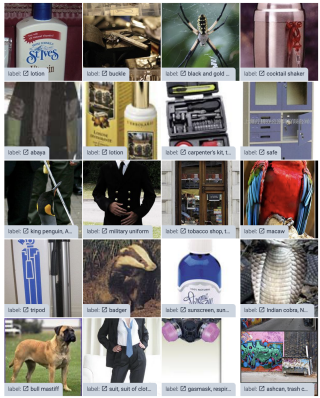



Figure 1: Examples of ILSVRC2012.

- It is to solve tasks for classification, classification with localization, and fine-grained classification.
- #Classes: 1,000
- #Training: 1,281,167
- #Validation: 50,000
- #Test: 100,000
- Details can be found in this link.

Figure 1 is taken from `https://knowyourdata-tfds.withgoogle.com/#tab=STATS&dataset=imagenet2012`.

# Tasks & Datasets

- ImageNet-21k and JFT-300M datasets are used to pre-train the ViT model.
- Diverse datasets for image classification are tested.
- In particular, Visual Task Adaptation Benchmark (VTAB) is used to evaluate models.

# Visual Task Adaptation Benchmark (VTAB)

- VTAB contains the following 19 tasks:

  Caltech101, CIFAR-100, CLEVR distance prediction, CLEVR counting, Diabetic Rethinopathy, Dmlab Frames, dSprites orientation prediction, dSprites location prediction, Describable Textures Dataset (DTD), EuroSAT, KITTI distance prediction, 102 Category Flower Dataset, Oxford IIIT Pet dataset, PatchCamelyon, Resisc45, Small NORB azimuth prediction, Small NORB elevation prediction, SUN397, SVHN.

- This benchmark expects a pre-trained model as an input.

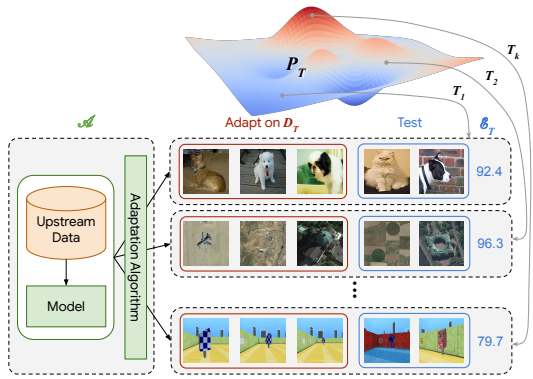
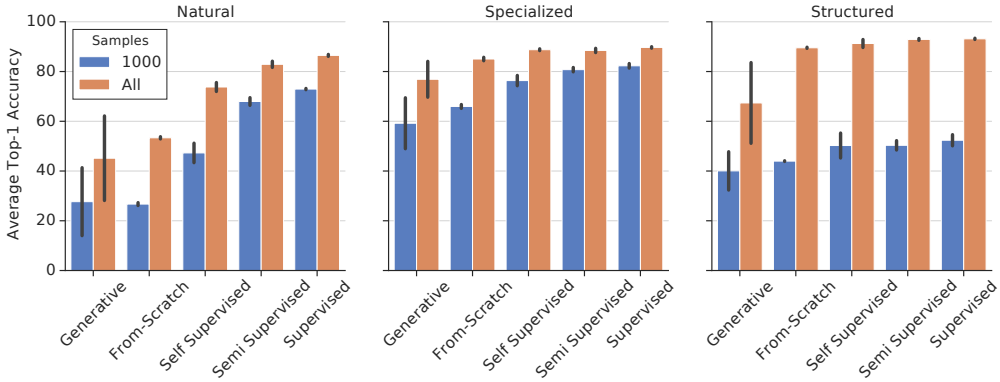


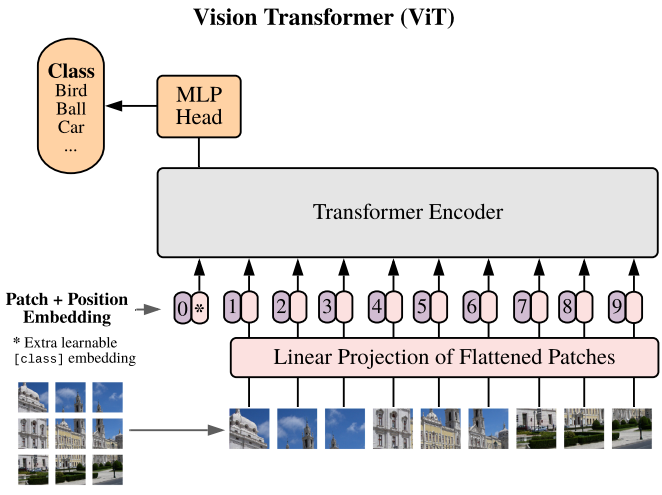

Figure 2: VTAB protocol.

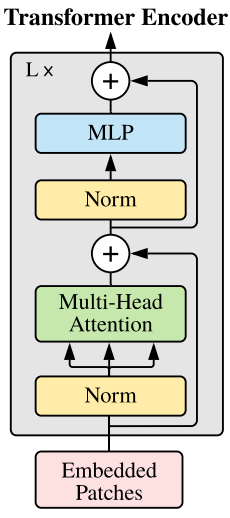
Figure 2 is taken from this link.

# Visual Task Adaptation Benchmark (VTAB)



# Vision Transformer (ViT)



# Vision Transformer (ViT)

- The standard Transformer receives as input a 1D sequence of token embeddings.
- To handle 2D images, the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is reshaped into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 C)}$, where $(H, W)$ is the resolution of image, $C$ is the number of channels, $(P, P)$ is the resolution of each image patch.
- Note that $N = HW/P^2$.
- The Transformer uses constant latent vector size $D$ through all of its layers, so the flattened patches are mapped to $D$ dimensions with a trainable linear projection.
- Similar to BERT's `[class]` token, it prepends a learnable embedding to the sequence of embedded patches ($\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$), whose state at the output of the Transformer encoder ($\mathbf{z}_L^0$) serves as the image representation $\mathbf{y}$.
- Both during pre-training and fine-tuning, a classification head is attached to $\mathbf{z}_L^0$.

# Inductive Bias of ViT

- ViT has much less image-specific inductive bias than convolutional neural networks.
- In convolutional neural networks, locality, two-dimensional neighborhood structure, and translation equivariance are baked into each layer throughout the whole model.
- On the contrary, in ViT, only MLP layers are local and translationally equivariant, while the self-attention layers are global.
- The two-dimensional neighborhood structure is used very sparingly; in the beginning of the model by cutting the image into patches and at fine-tuning time for adjusting the position embeddings for images of different resolution.
- The position embeddings at initialization time carry no information about the 2D positions of the patches and all spatial relations between the patches have to be learned from scratch.

# Fine-Tuning and Higher Resolution

- ViT is pre-trained on large datasets and fine-tuned to (smaller) downstream tasks.
- The pre-trained prediction head is removed and attached a zero-initialized $D \times K$ feedforward layer, where $K$ is the number of downstream classes.
- When feeding images of higher resolution, it keeps the patch size the same, which results in a larger effective sequence length.
- Since the pre-trained position embeddings may no longer be meaningful, 2D interpolation of the pre-trained position embeddings is performed.

# Details of ViT Model Variants

Table 1: Details of Vision Transformer model variants.

| Model | Layers | Hidden size | MLP size | Heads | Parameters |
|---|---|---|---|---|---|
| ViT-Base | 12 | 768 | 3072 | 12 | 86M |
| ViT-Large | 24 | 1024 | 4096 | 16 | 307M |
| ViT-Huge | 32 | 1280 | 5120 | 16 | 632M |

# A Learning Algorithm

- An Adam optimizer [Kingma and Ba, 2015] is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.
- A batch size is 4096.
- A linear learning rate warmup and decay are used.

# Experimental Results

| | Ours-JFT (ViT-H/14) | Ours-JFT (ViT-L/16) | Ours-I21k (ViT-L/16) | BiT-L (ResNet152x4) | Noisy Student (EfficientNet-L2) |
|---|---|---|---|---|---|
| ImageNet | **88.55** ± 0.04 | 87.76 ± 0.03 | 85.30 ± 0.02 | 87.54 ± 0.02 | 88.4/88.5* |
| ImageNet ReaL | **90.72** ± 0.05 | 90.54 ± 0.03 | 88.62 ± 0.05 | 90.54 | 90.55 |
| CIFAR-10 | **99.50** ± 0.06 | 99.42 ± 0.03 | 99.15 ± 0.03 | 99.37 ± 0.06 | – |
| CIFAR-100 | **94.55** ± 0.04 | 93.90 ± 0.05 | 93.25 ± 0.05 | 93.51 ± 0.08 | – |
| Oxford-IIIT Pets | **97.56** ± 0.03 | 97.32 ± 0.11 | 94.67 ± 0.15 | 96.62 ± 0.23 | – |
| Oxford Flowers-102 | 99.68 ± 0.02 | **99.74** ± 0.00 | 99.61 ± 0.02 | 99.63 ± 0.03 | – |
| VTAB (19 tasks) | **77.63** ± 0.23 | 76.28 ± 0.46 | 72.72 ± 0.21 | 76.29 ± 1.70 | – |
| TPUv3-core-days | 2.5k | 0.68k | 0.23k | 9.9k | 12.3k |

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

# Experimental Results

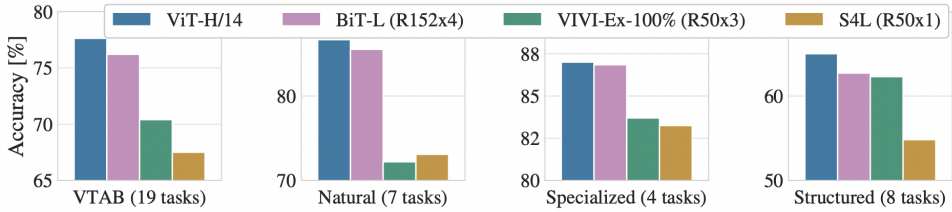



Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

# Experimental Results

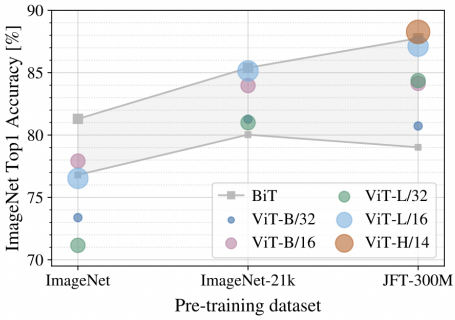



Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

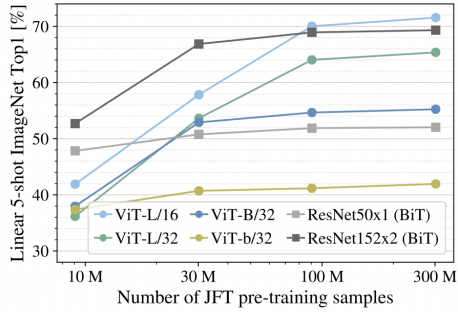



Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

# Experimental Results

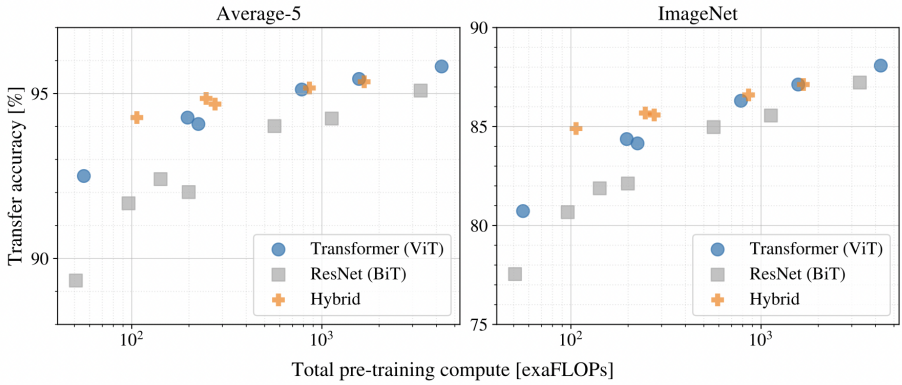



Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

# Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

# Swin Transformer

- This paper presents a new Vision Transformer, called Swin Transformer, that capably serves as a general-purpose backbone for computer vision.
- The authors propose a hierarchical Transformer whose representation is computed with **S**hifted **win**dows.
- The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection.
- This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to image size.

[Liu et al., 2021] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the International Conference on Computer Vision (ICCV), pages 10012–10022, Virtual, 2021.

# Tasks & Datasets

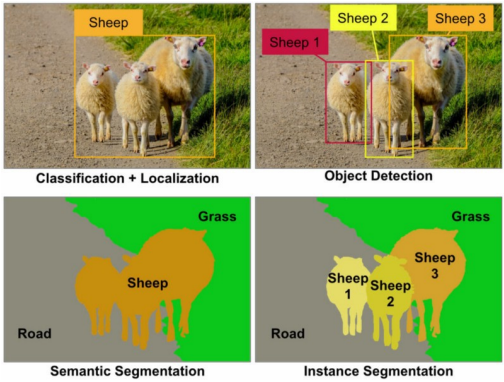



Figure 3: Comparisons of classification + localization, object detection, semantic segmentation, and instance segmentation.

- Three tasks are solved in this paper:
  - image classification on ImageNet-1K;
  - object detection on COCO 2017;
  - semantic segmentation on ADE20K.

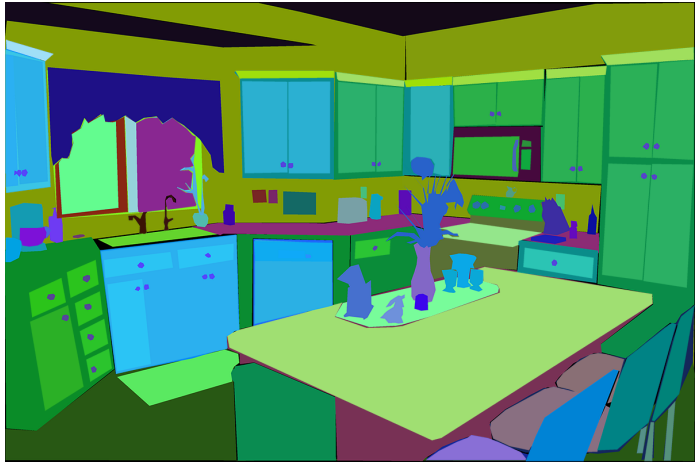
Figure 3 is taken from this link.

# Object Detection on COCO 2017



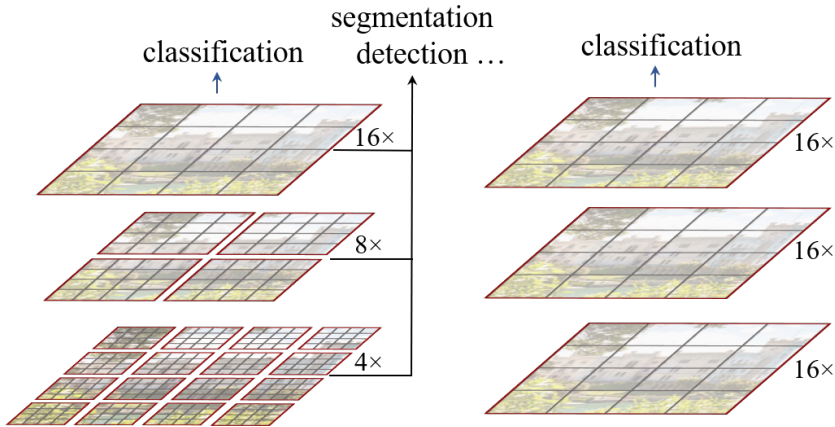
Taken from this link.

# Semantic Segmentation on ADE20K

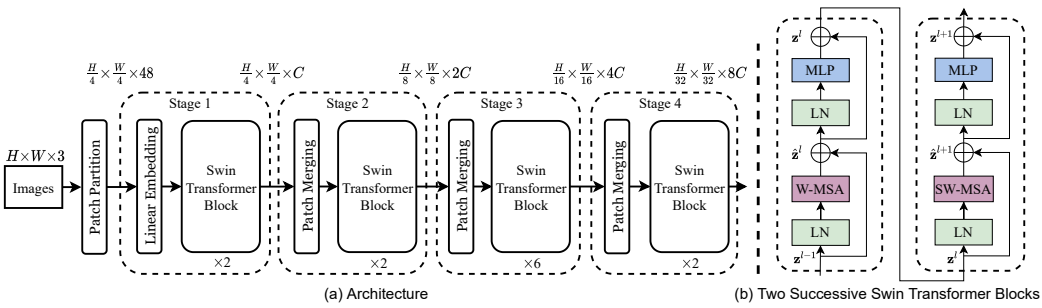


Taken from this link.

# Swin Transformer



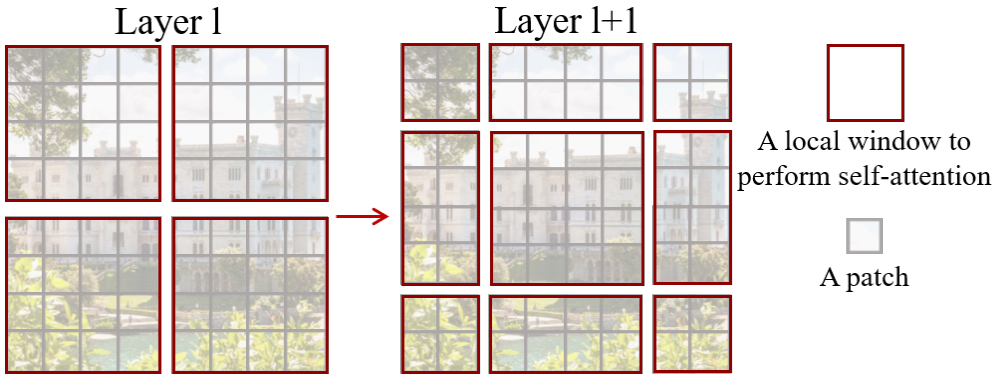
# Swin Transformer



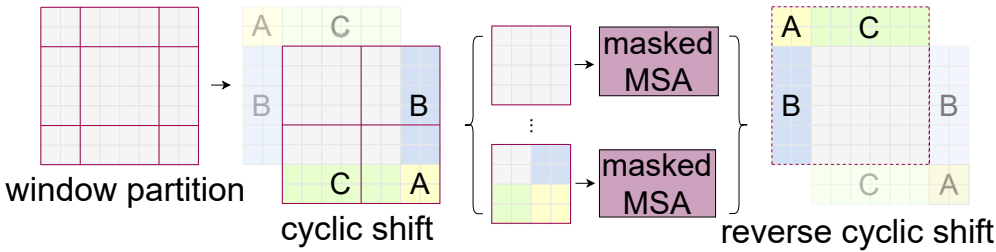

(a) Architecture

(b) Two Successive Swin Transformer Blocks

# Swin Transformer



# Swin Transformer



# Experimental Results

**(a) Regular ImageNet-1K trained models**

| method | image size | #param. | FLOPs | throughput (image / s) | ImageNet top-1 acc. |
|---|---|---|---|---|---|
| RegNetY-4G [48] | $224^2$ | 21M | 4.0G | 1156.7 | 80.0 |
| RegNetY-8G [48] | $224^2$ | 39M | 8.0G | 591.6 | 81.7 |
| RegNetY-16G [48] | $224^2$ | 84M | 16.0G | 334.7 | 82.9 |
| EffNet-B3 [58] | $300^2$ | 12M | 1.8G | 732.1 | 81.6 |
| EffNet-B4 [58] | $380^2$ | 19M | 4.2G | 349.4 | 82.9 |
| EffNet-B5 [58] | $456^2$ | 30M | 9.9G | 169.1 | 83.6 |
| EffNet-B6 [58] | $528^2$ | 43M | 19.0G | 96.9 | 84.0 |
| EffNet-B7 [58] | $600^2$ | 66M | 37.0G | 55.1 | 84.3 |
| ViT-B/16 [20] | $384^2$ | 86M | 55.4G | 85.9 | 77.9 |
| ViT-L/16 [20] | $384^2$ | 307M | 190.7G | 27.3 | 76.5 |
| DeiT-S [63] | $224^2$ | 22M | 4.6G | 940.4 | 79.8 |
| DeiT-B [63] | $224^2$ | 86M | 17.5G | 292.3 | 81.8 |
| DeiT-B [63] | $384^2$ | 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 |

# Experimental Results

| (b) ImageNet-22K pre-trained models | | | | | |
|---|---|---|---|---|---|
| method | image size | #param. | FLOPs | throughput (image / s) | ImageNet top-1 acc. |
| R-101x3 [38] | $384^2$ | 388M | 204.6G | - | 84.4 |
| R-152x4 [38] | $480^2$ | 937M | 840.5G | - | 85.4 |
| ViT-B/16 [20] | $384^2$ | 86M | 55.4G | 85.9 | 84.0 |
| ViT-L/16 [20] | $384^2$ | 307M | 190.7G | 27.3 | 85.2 |
| Swin-B | $224^2$ | 88M | 15.4G | 278.1 | 85.2 |
| Swin-B | $384^2$ | 88M | 47.0G | 84.7 | 86.4 |
| Swin-L | $384^2$ | 197M | 103.9G | 42.1 | 87.3 |

# Experimental Results

**(a) Various frameworks**

| Method | Backbone | $AP^{box}$ | $AP^{box}_{50}$ | $AP^{box}_{75}$ | #param. | FLOPs | FPS |
|---|---|---|---|---|---|---|---|
| Cascade Mask R-CNN | R-50 | 46.3 | 64.3 | 50.5 | 82M | 739G | 18.0 |
| | Swin-T | **50.5** | **69.3** | **54.9** | 86M | 745G | 15.3 |
| ATSS | R-50 | 43.5 | 61.9 | 47.0 | 32M | 205G | 28.3 |
| | Swin-T | **47.2** | **66.5** | **51.3** | 36M | 215G | 22.3 |
| RepPointsV2 | R-50 | 46.5 | 64.6 | 50.3 | 42M | 274G | 13.6 |
| | Swin-T | **50.0** | **68.5** | **54.2** | 45M | 283G | 12.0 |
| Sparse R-CNN | R-50 | 44.5 | 63.4 | 48.2 | 106M | 166G | 21.0 |
| | Swin-T | **47.9** | **67.3** | **52.3** | 110M | 172G | 18.4 |

**(b) Various backbones w. Cascade Mask R-CNN**

| | $AP^{box}$ | $AP^{box}_{50}$ | $AP^{box}_{75}$ | $AP^{mask}$ | $AP^{mask}_{50}$ | $AP^{mask}_{75}$ | param | FLOPs | FPS |
|---|---|---|---|---|---|---|---|---|---|
| DeiT-S† | 48.0 | 67.2 | 51.7 | 41.4 | 64.2 | 44.3 | 80M | 889G | 10.4 |
| R50 | 46.3 | 64.3 | 50.5 | 40.1 | 61.7 | 43.4 | 82M | 739G | 18.0 |
| Swin-T | **50.5** | **69.3** | **54.9** | **43.7** | **66.6** | **47.1** | 86M | 745G | 15.3 |
| X101-32 | 48.1 | 66.5 | 52.4 | 41.6 | 63.9 | 45.2 | 101M | 819G | 12.8 |
| Swin-S | **51.8** | **70.4** | **56.3** | **44.7** | **67.9** | **48.5** | 107M | 838G | 12.0 |
| X101-64 | 48.3 | 66.4 | 52.3 | 41.7 | 64.0 | 45.1 | 140M | 972G | 10.4 |
| Swin-B | **51.9** | **70.9** | **56.5** | **45.0** | **68.4** | **48.7** | 145M | 982G | 11.6 |

# Experimental Results

| (c) System-level Comparison | | | | | | |
|---|---|---|---|---|---|---|
| Method | mini-val | | test-dev | | #param. | FLOPs |
| | $AP^{box}$ | $AP^{mask}$ | $AP^{box}$ | $AP^{mask}$ | | |
| 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++)* | **58.0** | **50.4** | **58.7** | **51.1** | 284M | - |

# Experimental Results

| ADE20K Method | Backbone | val mIoU | test score | #param. | FLOPs | FPS |
|---|---|---|---|---|---|---|
| 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‡ | 50.3 | 61.7 | 308M | - | - |
| UperNet | DeiT-S† | 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‡ | 51.6 | - | 121M | 1841G | 8.7 |
| UperNet | Swin-L‡ | **53.5** | **62.8** | 234M | 3230G | 6.2 |

# Any Questions?

# References I


A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 2021.

D. P. Kingma and J. L. Ba. ADAM: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California, USA, 2015.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 10012–10022, Virtual, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, Long Beach, California, USA, 2017.