NVIDIA

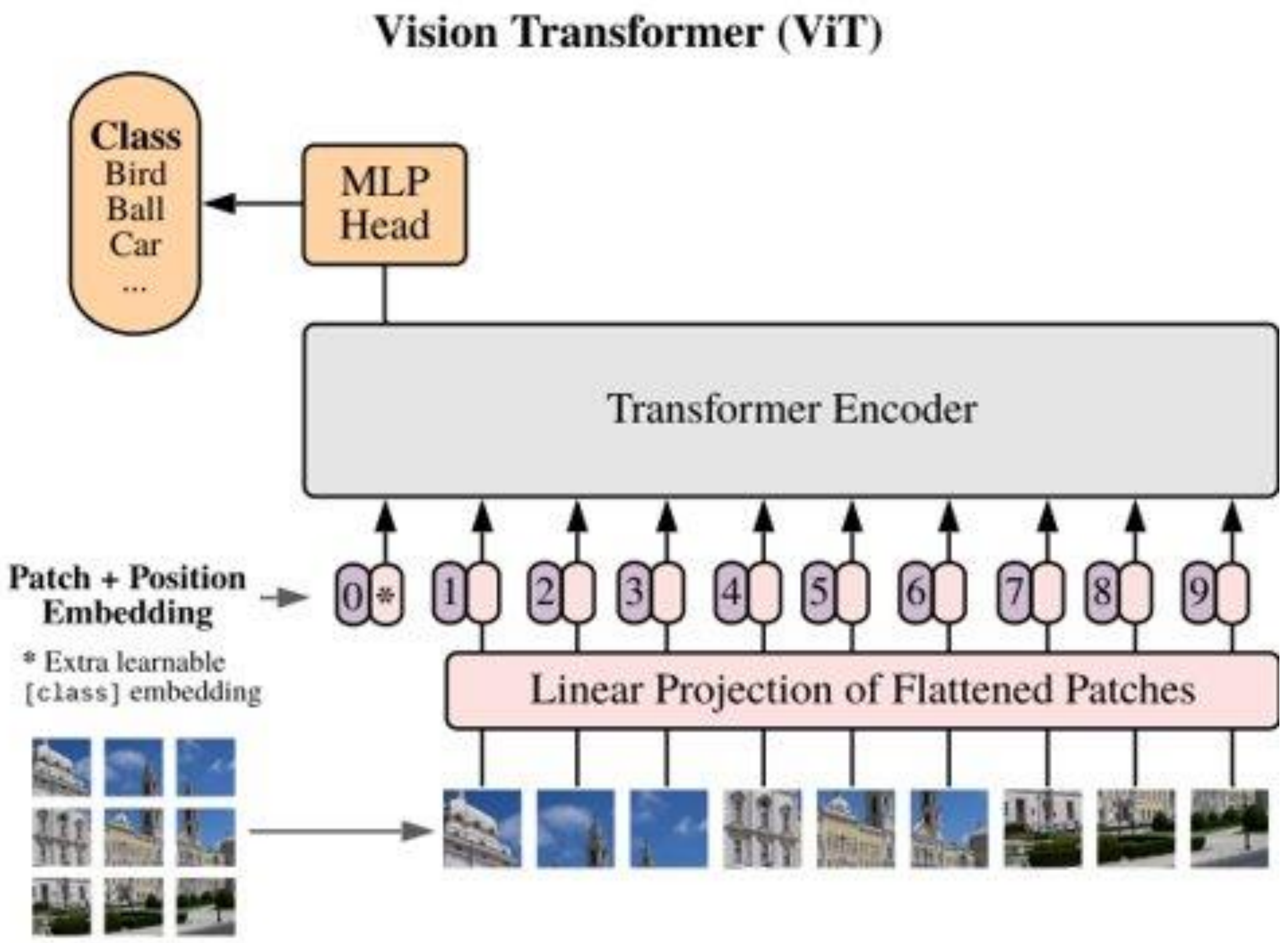
FasterViT: Fast Vision Transformers With Hierarchical Attention

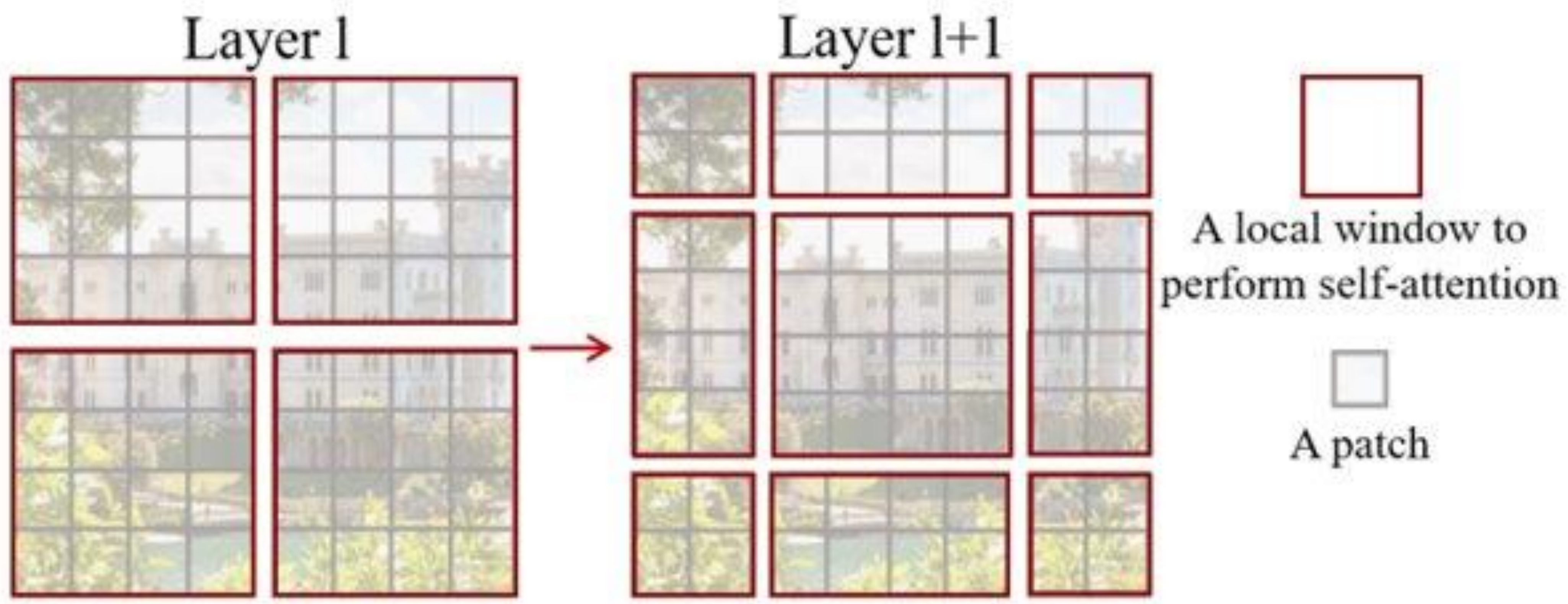
Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov

Vision Transformers

Strengths

- Vision Transformers (ViTs) have gained popularity for various vision tasks
  - Great capability in modeling long-range dependencies.
  - Scalability for large-scale training.
  - SOTA performance on downstream tasks such as classification, detection, etc.

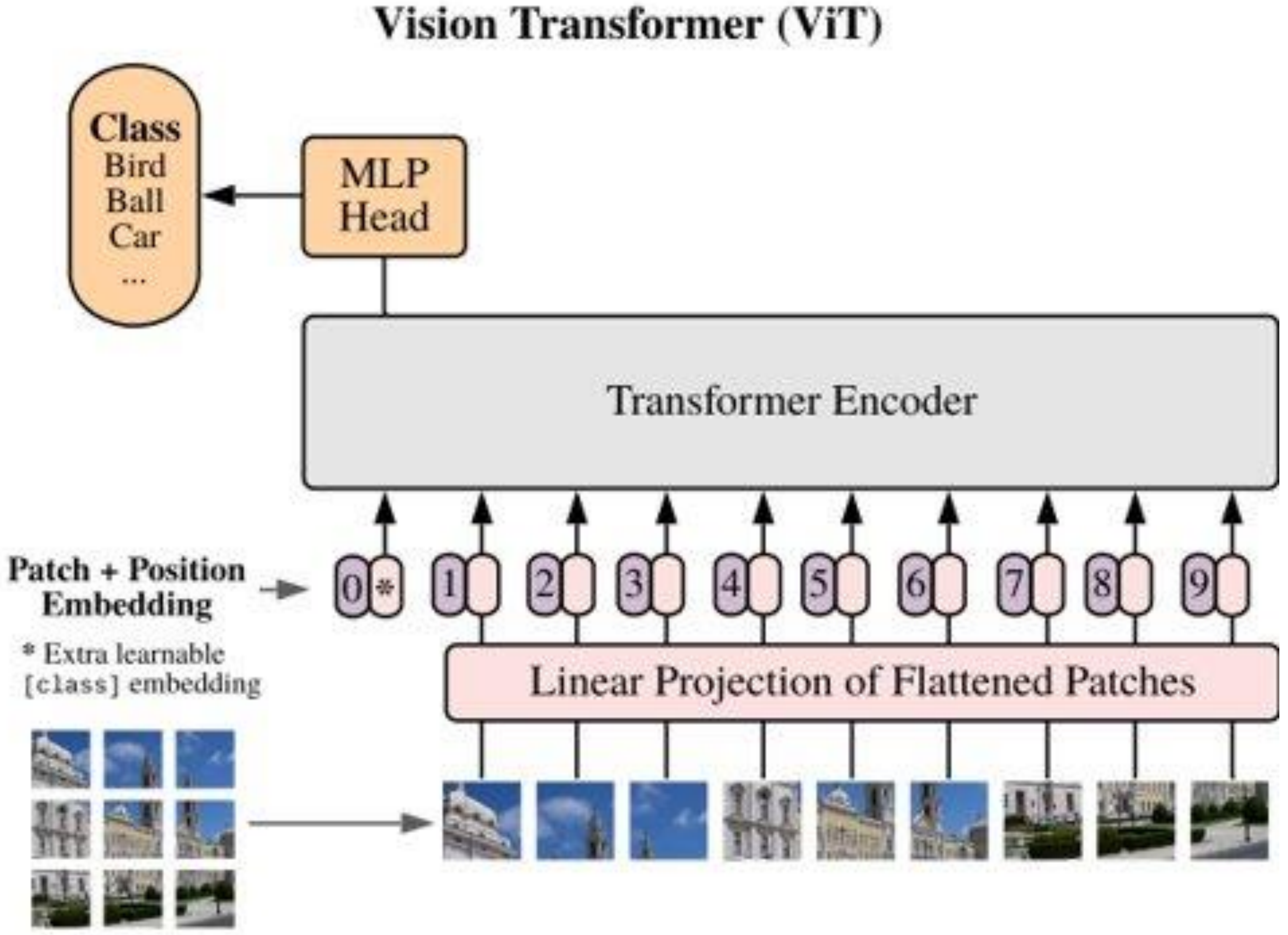


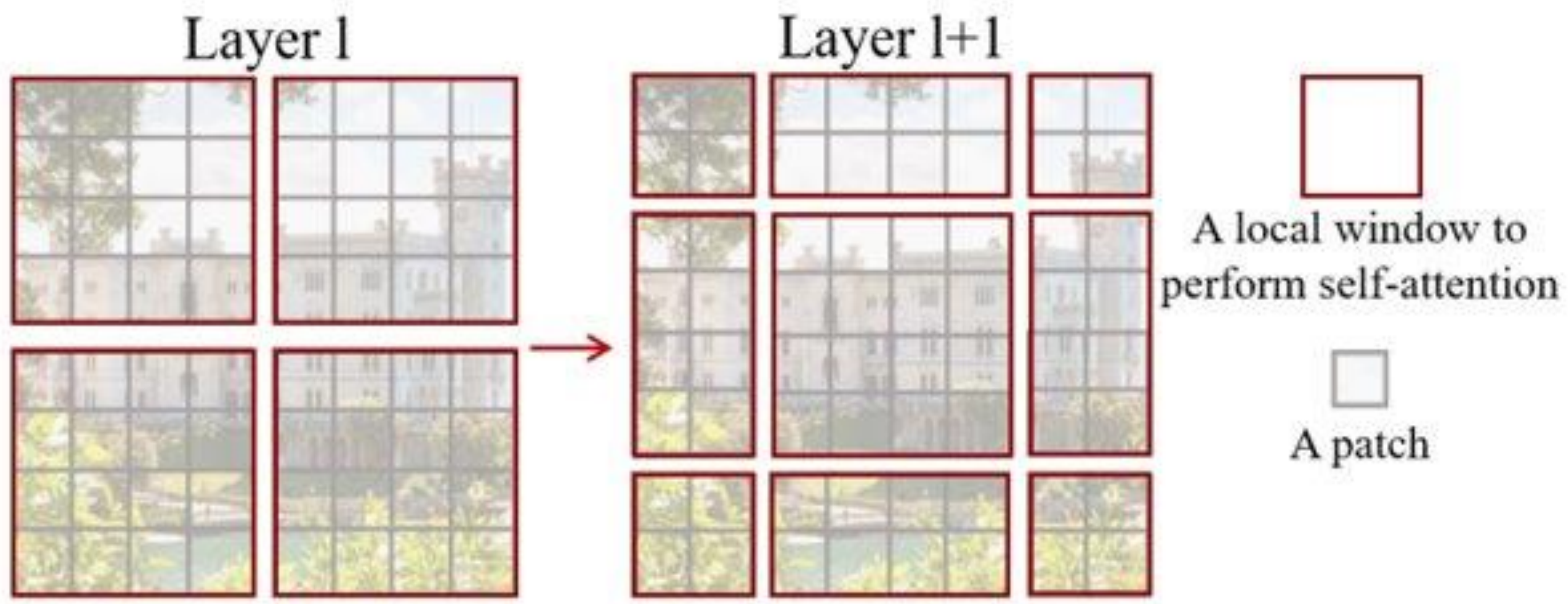


Vision Transformers

Weaknesses

- Despite their popularity, ViTs' are still **NOT efficient** to deploy
  - **Quadradic complexity** of self-attention is **expensive** (both time and memory).
  - Vanilla ViTs need **a lot of data** for training (lack of inductive bias).
  - Certain operations are not supported in high performance inference engines like **NVIDIA TensorRT.**



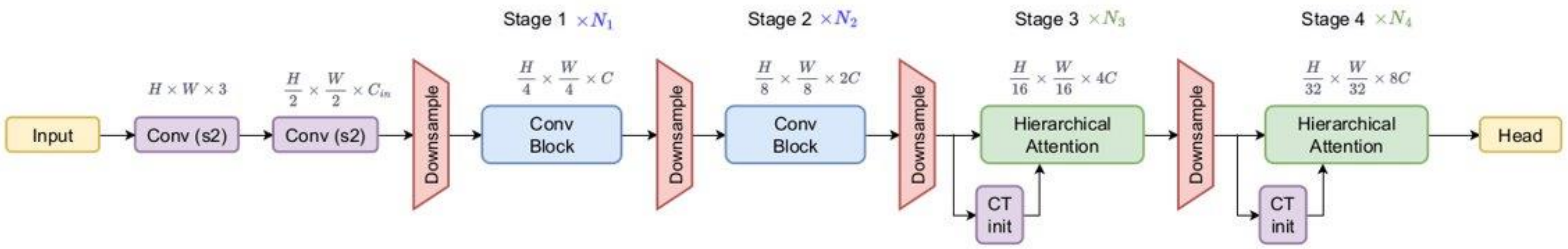


FasterViT

Fast and Scalable ViTs

- Motivated to address these issues, we introduce FasterViT which is a novel hybrid vision transformer architecture designed for an optimal trade-off between performance and image throughput.

- FasterViT (SOTA for Top-1 vs image throughput)
  - Tailored to optimize throughput and GPU utilization.
  - Hierarchical Attention for efficient and scalable modeling of high-resolution images.
  - Outperforms FastViT and EfficientNetV2 by a large margin.



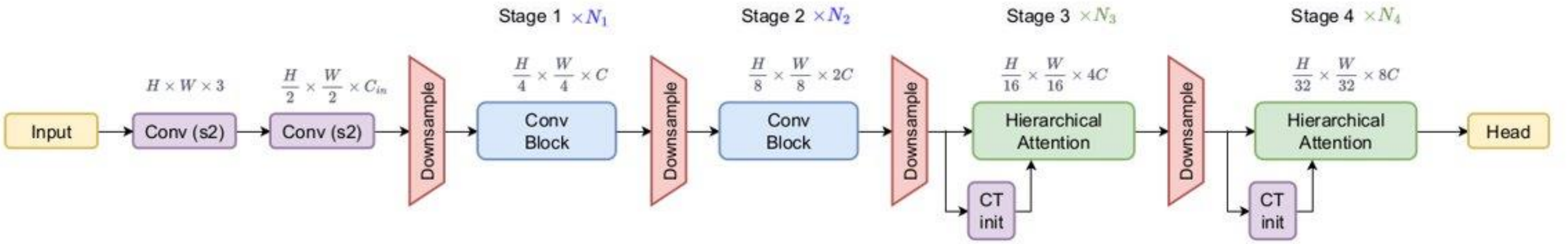
Fast Vision Transformers with Hierarchical Attention

FasterViT

- FasterViT comprises of hybrid architecture (CNN + ViT) with 4 different stages.

- CNN-based stages are used to extract features in an efficient way.
  - In these stage, low-level features are mainly captured.

- ViT-based stages learn high-level feature via our proposed **hierarchical self-attention.**



Fast Vision Transformers with Hierarchical Attention

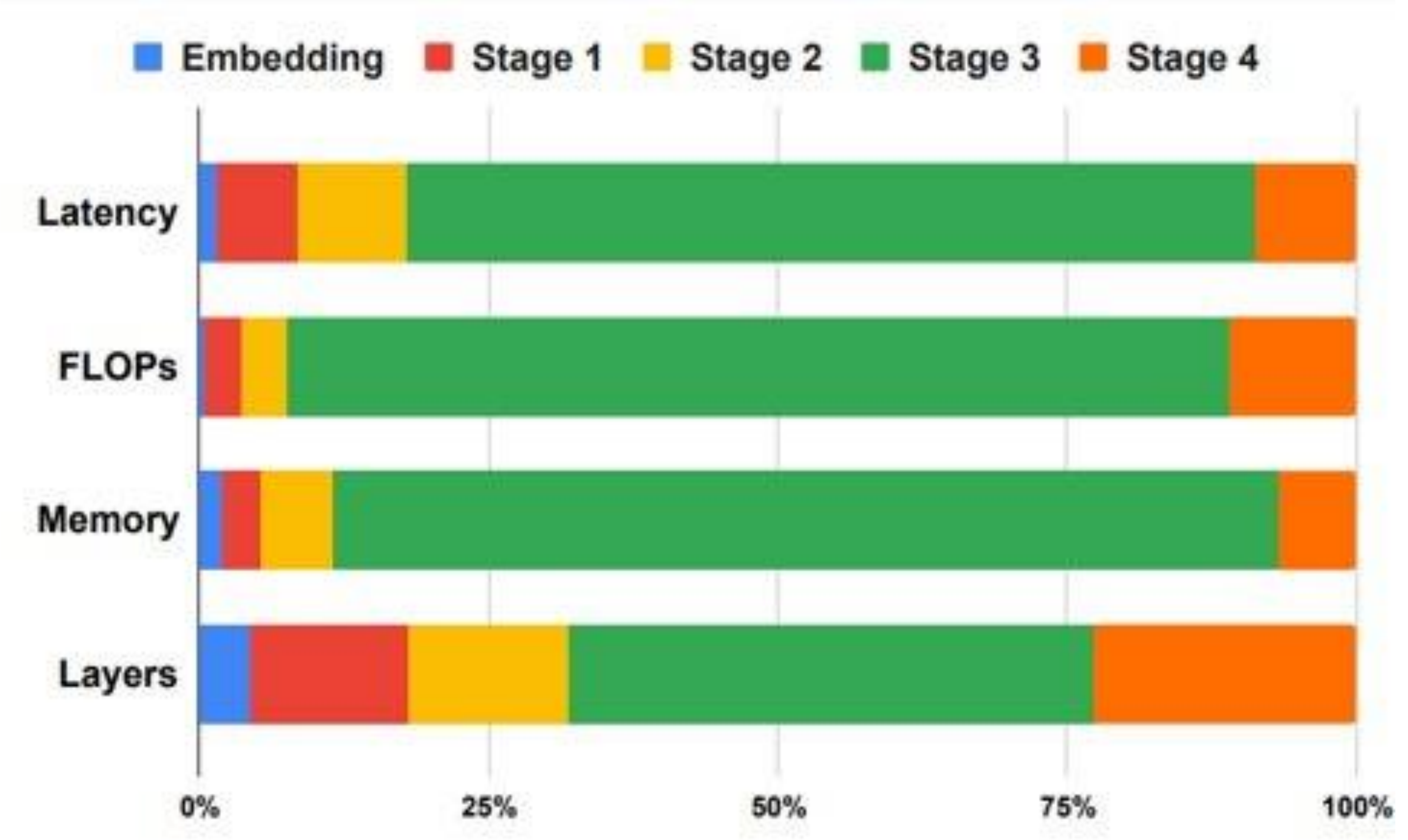
Hierarchical Self-Attention

- Hierarchical attention is a scalable self-attention block.
  - Recursively learns a summary of each window region via carrier tokens.
  - Performs cross-window interaction to capture long-range dependencies.

Fast Vision Transformers with Hierarchical Attention

Design Insights

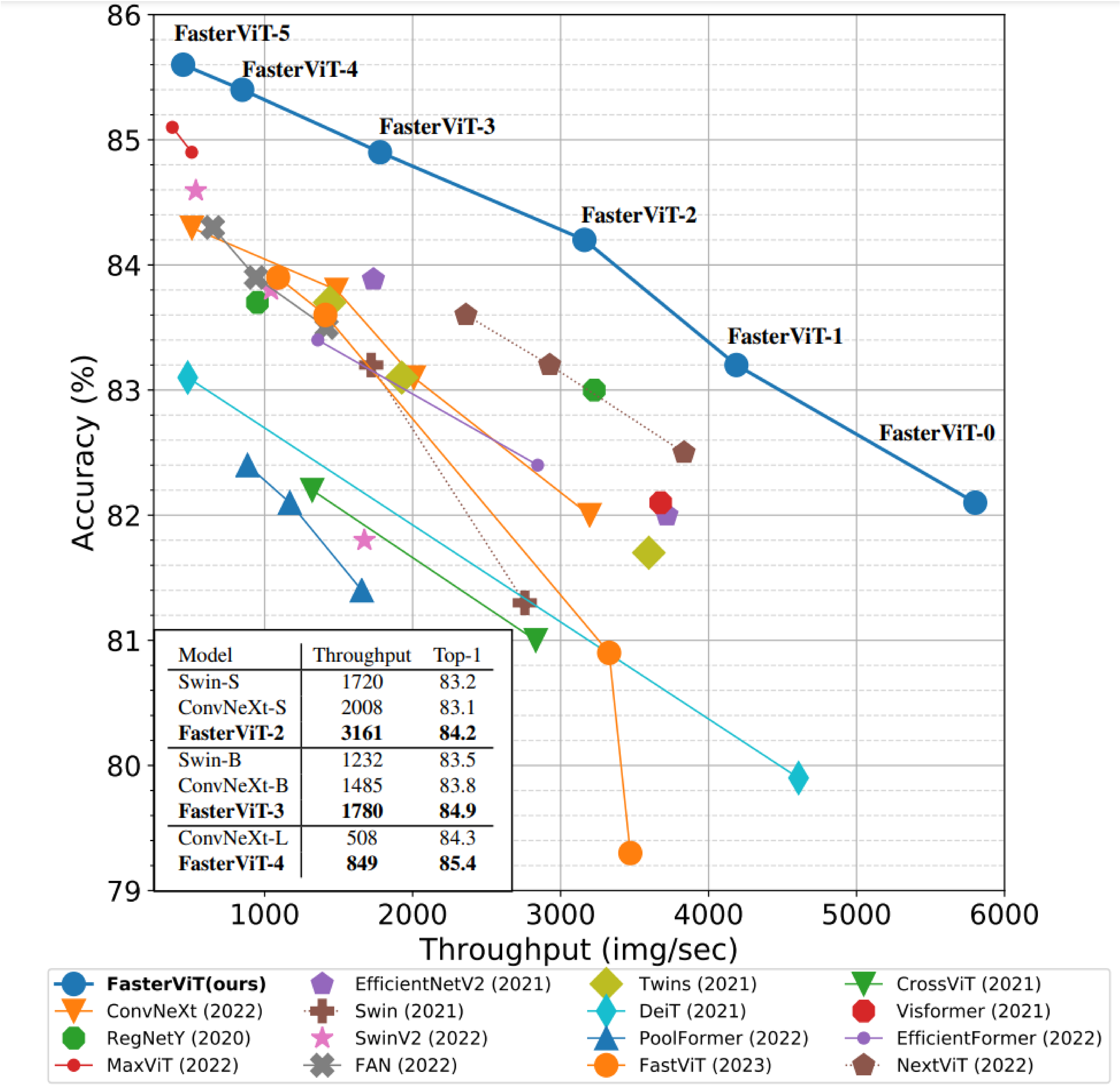
- Stages 1, 2 (CNN-based) are memory bound.
  - We employ dense conv-based layers.
    - Although more parameter-heavy than depth-wise but better GPU utilization and throughput (e.g. FP16, INT8) .
  - We utilize BN layers (foldable in TensorRT) which are faster than LN.
    - Conv-BN-ReLU are not used due training instabilities.
- Stages 3, 4 (ViT-based) are math-bound.
  - We use LN for training stability and GELU for better performance.



Fast Vision Transformers with Hierarchical Attention

Results

FasterViT achieves new Pareto Fronts (Top1 vs. throughput) on ImageNet-1K dataset



4x Faster for Classification

Model	Top1	Throughput (Image/Sec)
Swin Transformer (Microsoft)	83.8	168
FasterViT	**84.0**	**605**

High-resolution (512 × 512) ImageNet Benchmarks.

2x Faster for Detection

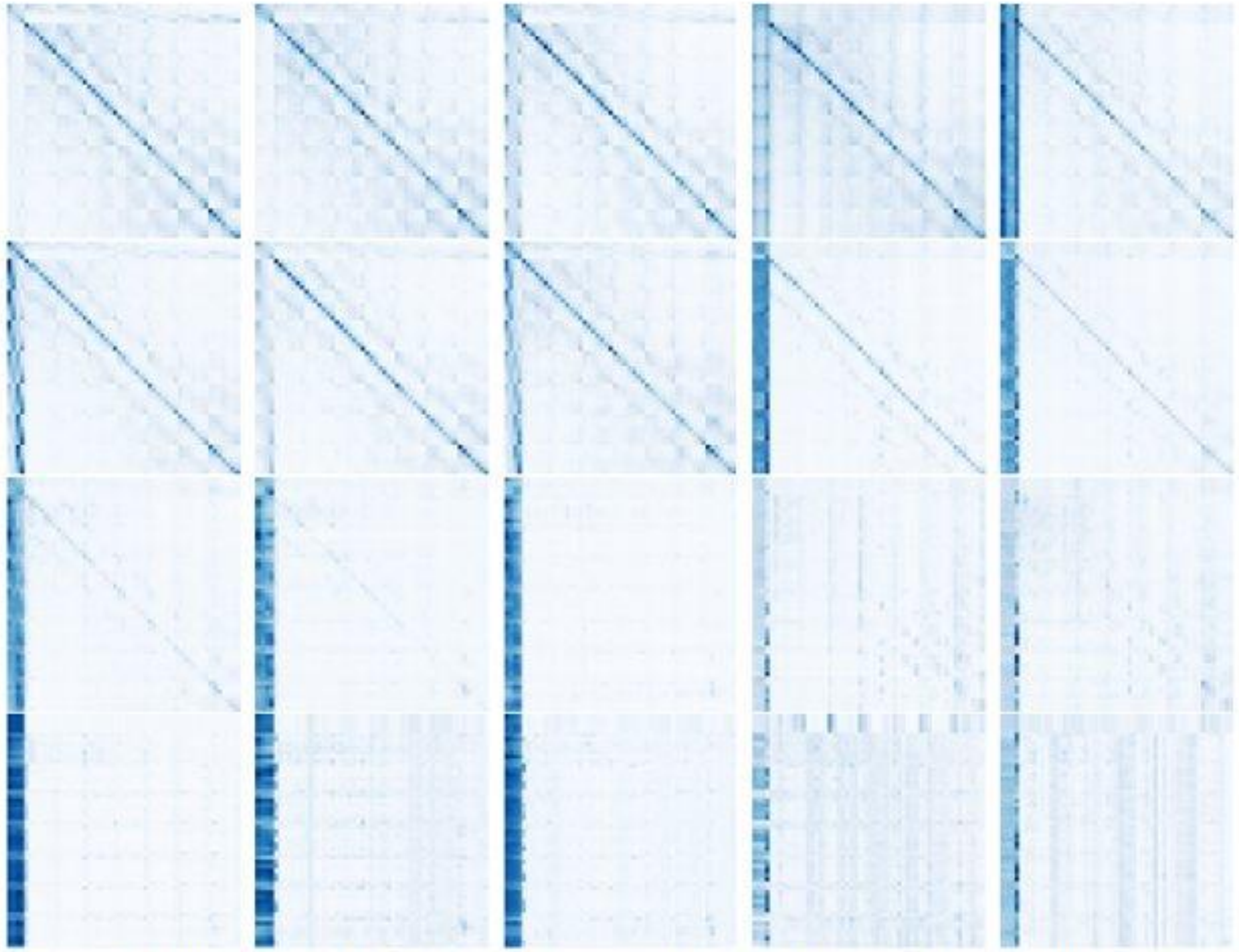
Backbone	Head	AP^{box}	Throughput (Image/Sec)
ConvNeXt (Meta)	MaskRCNN	51.9	127.8
FasterViT	MaskRCNN	**52.1**	**287.3**

High-resolution (800 × 1216) Detection Benchmarks.

Fast Vision Transformers with Hierarchical Attention

Hierarchical Self-Attention

- Dense attention maps demonstrate patterns of learning both local and global interactions with carrier tokens.



(a) FasterViT-0 (b) FasterViT-1 (c) FasterViT-2 (d) FasterViT-3 (d) FasterViT-4

Conclusion

- FasterViT is the current SOTA for Top-1 accuracy vs image throughput.
- Hybrid FasterViT architecture is tailored to maximize GPU utilization and throughput.
- Hierarchical attention is an efficient and scalable mechanism to capture long-range spatial dependencies, especially for high-resolution images.

https://github.com/NVlabs/FasterViT

NVIDIA®