Video Swin Transformers

Seminar ADQ 413

Presented By

DON SABU SJC21AD024

Guided By

Ms. RASHMI ANNAMMA GEORGE
Asst. Prof. AD Department

September 27, 2024

Outline

- Introduction
- Standard Transformers
- Visual Transformers and its Limitations
- SWIN Transformers
- Architecture of SWIN Transformers
- Advantages Over Standard Transformers
- Limitations in Temporal Modelling
- Video SWIN Transformers
- Architecture Overview
- Performance and Accuracy Advantages
- Conclusion

Introduction

- Evolution from Standard Transformers architecture to Video Swin Transformers.
- Transformers are originally designed for natural language processing (NLP) tasks.
- To extend the capability of Transformers to adapt to vision, Vision Transformers (ViT) was introduced
- To overcome the limitations of ViT such as scalability and efficiency, shifted window based Swin Transformers was introduced.
- Video Swin Transformers were developed to address the need to handle videos, extending the Swin architecture to the spatiotemporal domain.

Standard Transformers [2]

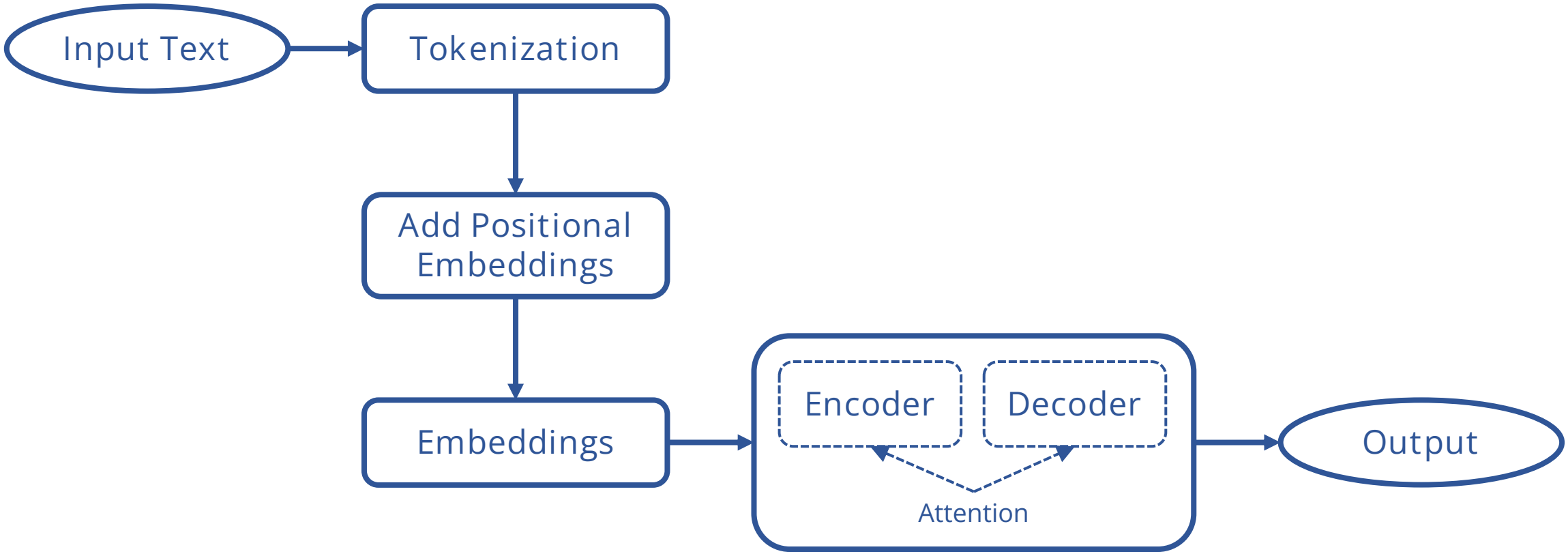

Fig 1. Simple Transformer Architecture

Vision Transformers [3]

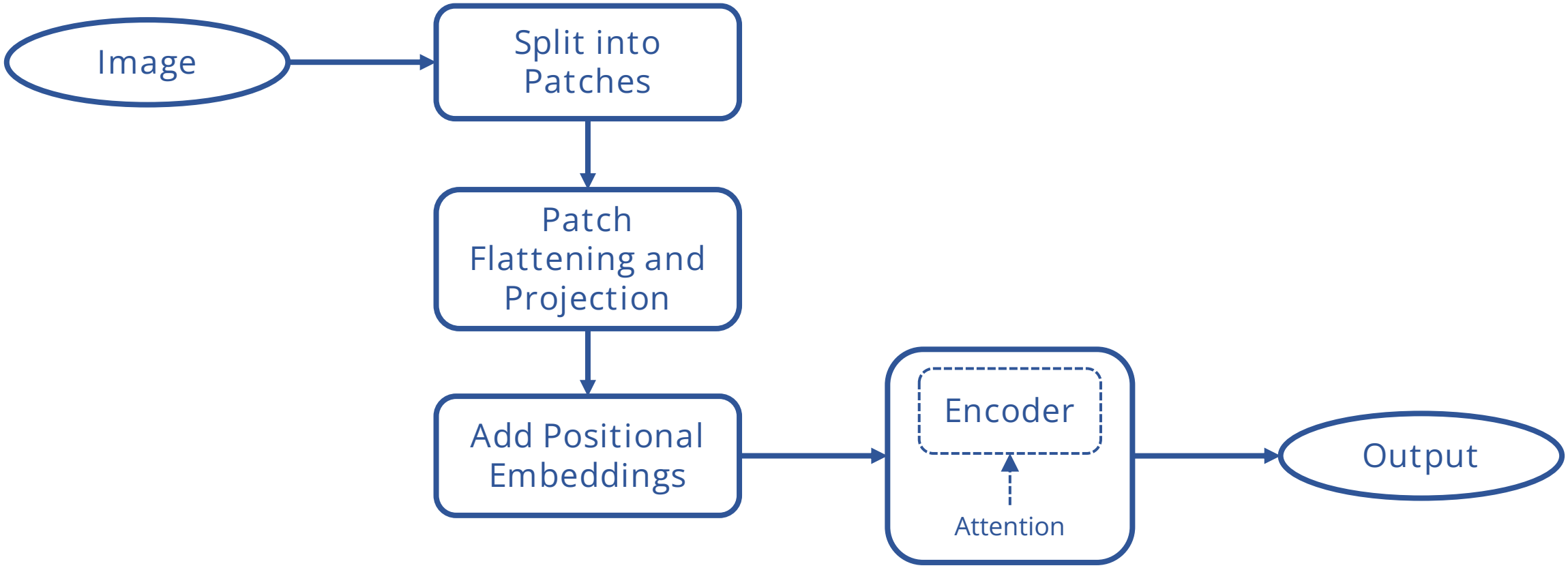

Fig 2. Simple Vision Transformer Architecture

Vision Transformers

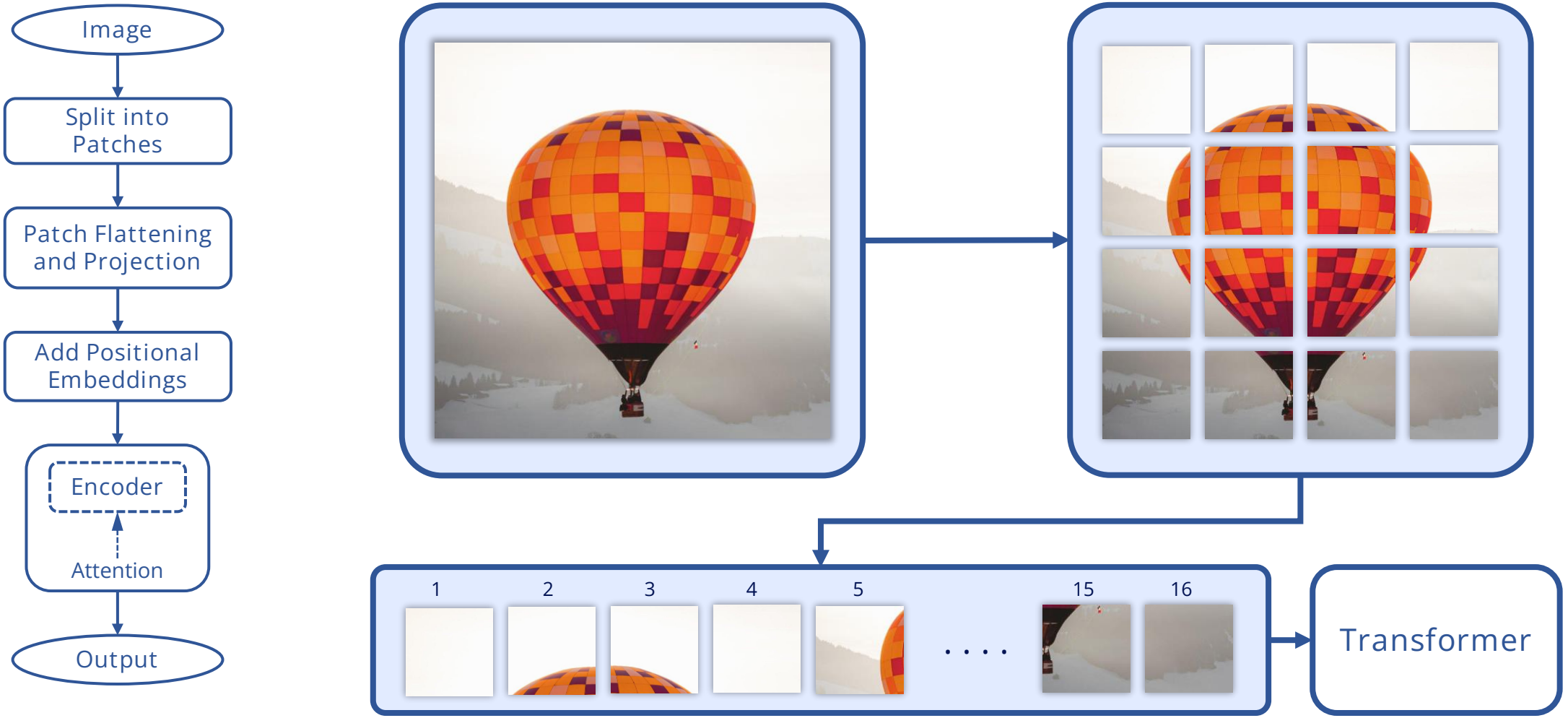

Fig 3. Illustration of patches in Vision Transformers

Limitation of Vision Transformers

- High Computational Cost
- Fixed Patch Size
- Limited Local Context Understanding

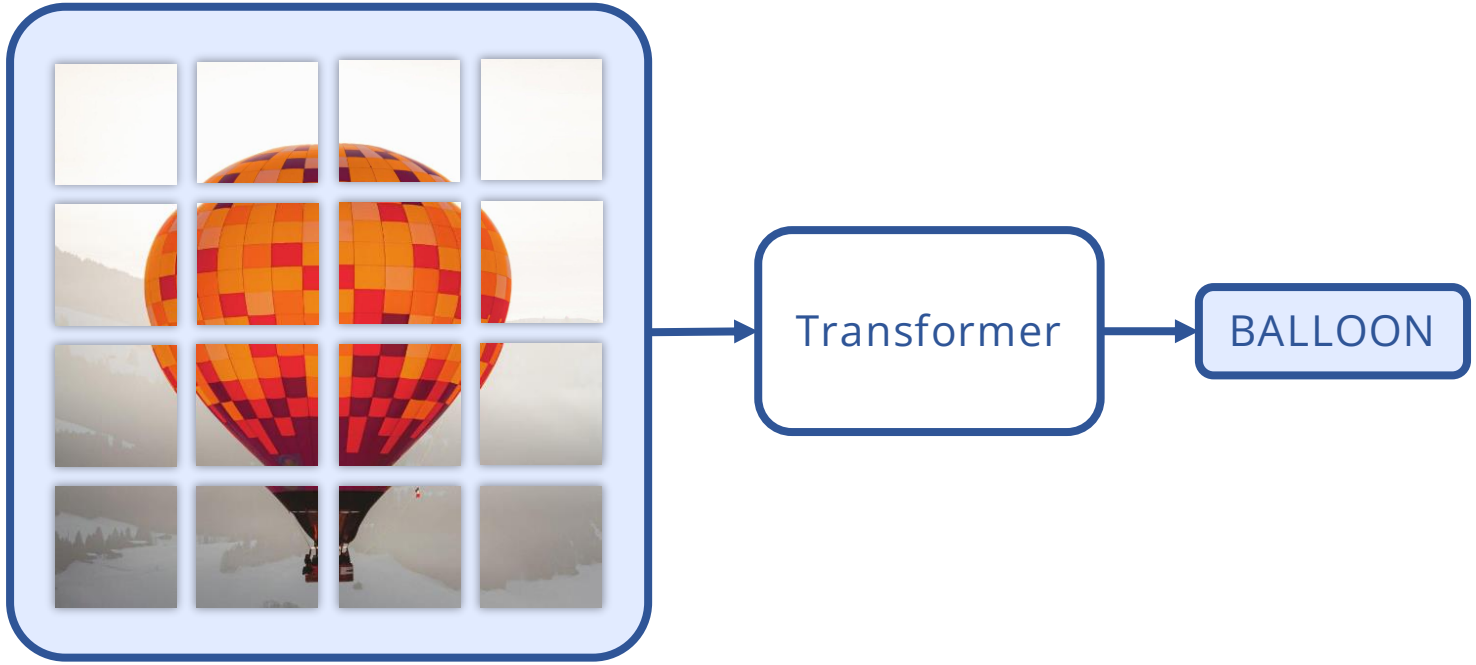

Fig 4. Limitation of Vison Transformers

Limitation of Vision Transformers

- High Computational Cost
- Fixed Patch Size
- Limited Local Context Understanding

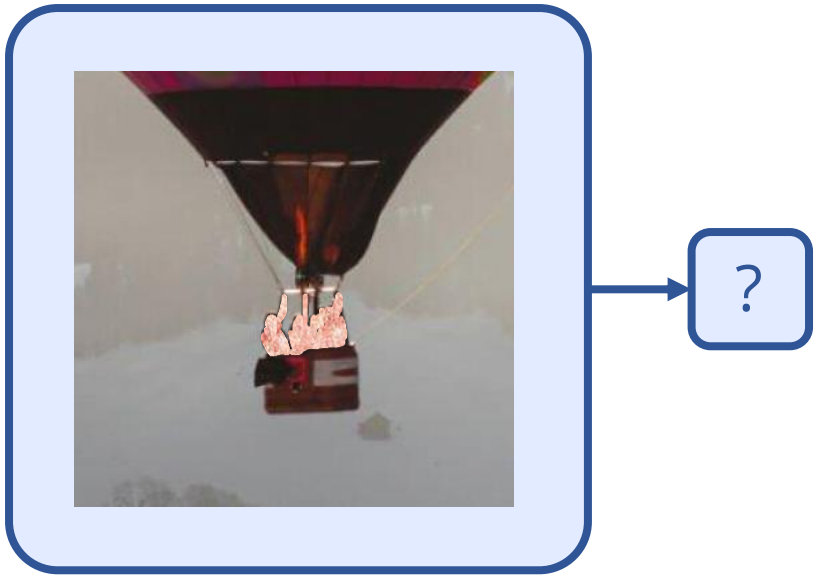

Fig 5. Limitation of Vison Transformers for pixel level identification

HD image (1080p) / 16x16 patches = 4500+ Tokens

Every Pixel as a Token 256x256 image => 65000+ Tokens

Every Pixel as a Token HD image => 2M+ Tokens

Every Pixel as a Token 4K image => 8M+ Tokens

Token Limit of GPT-4 32,768

Swin Transformers [4]

- "Swin" stands for Shifted WINdows.
- Shifted Window-based Self-Attention
- Hierarchical Feature Representation
- Efficient Handling of High-Resolution Images
- Uses patches in the same way as ViT.
- Starts with smaller patches and merges them in the hidden transformer layers

Swin Transformers Architecture

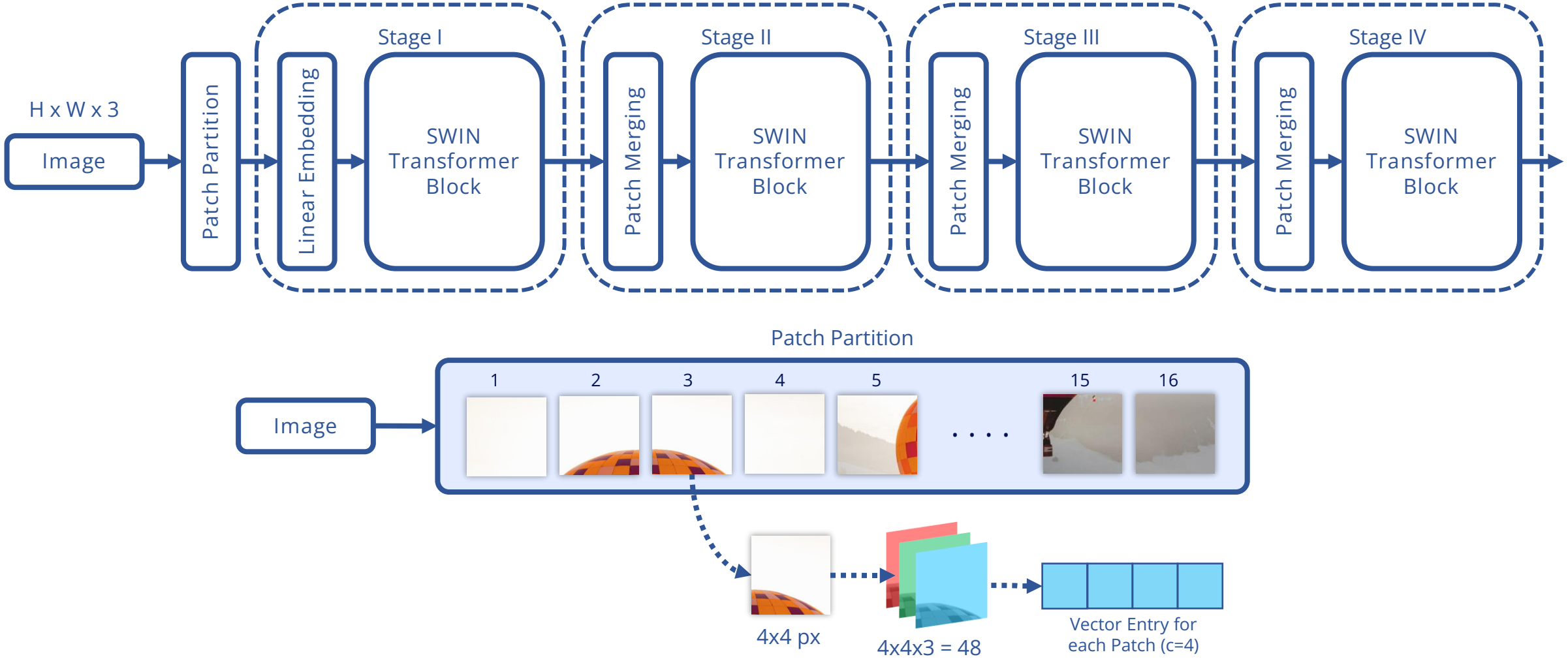

Fig 6. SWIN Transformer Architecture

Swin Transformers Block

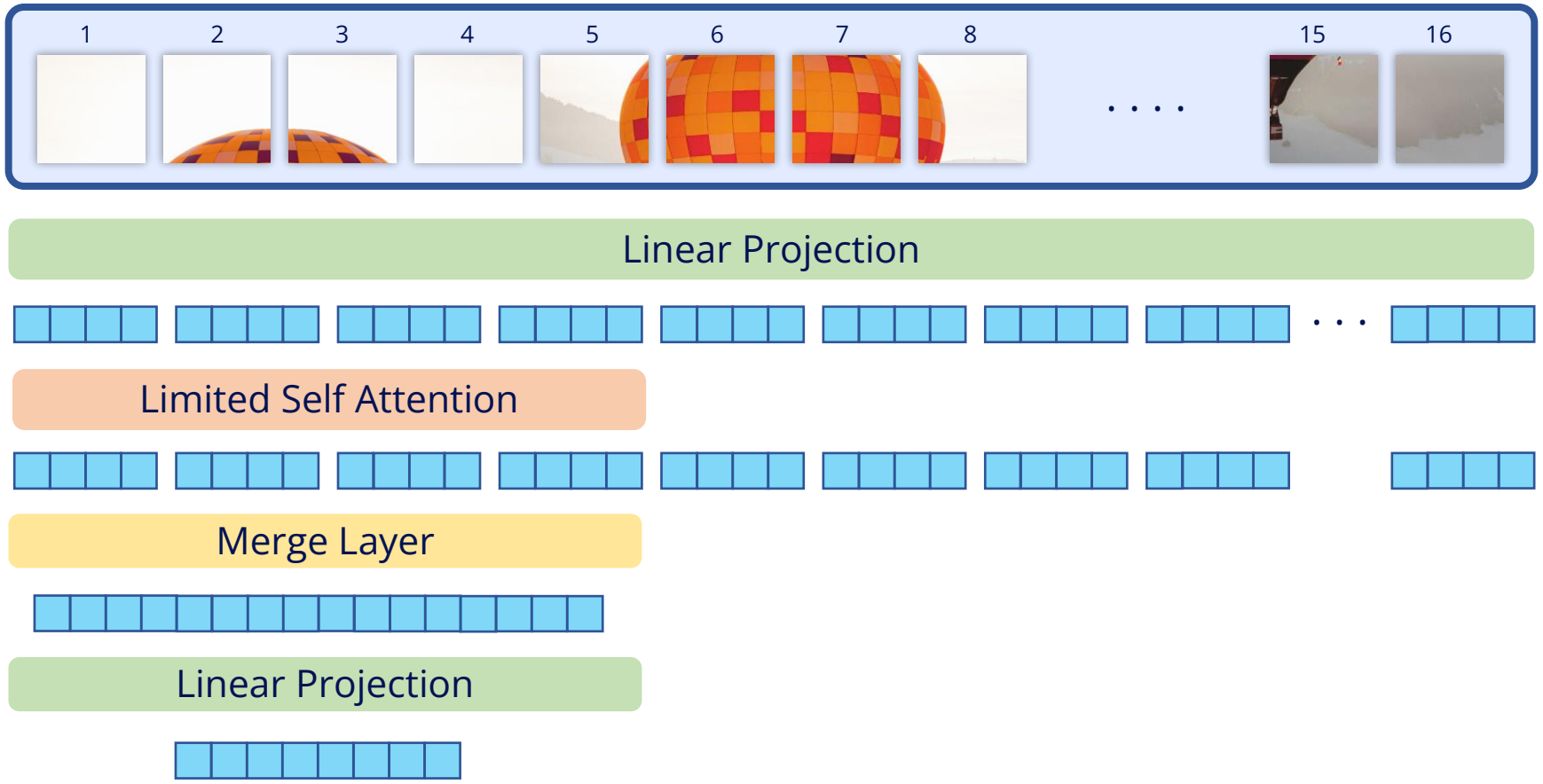

Fig 7. Working of Swin Transformer Block and shited windows

Swin Transformers Block

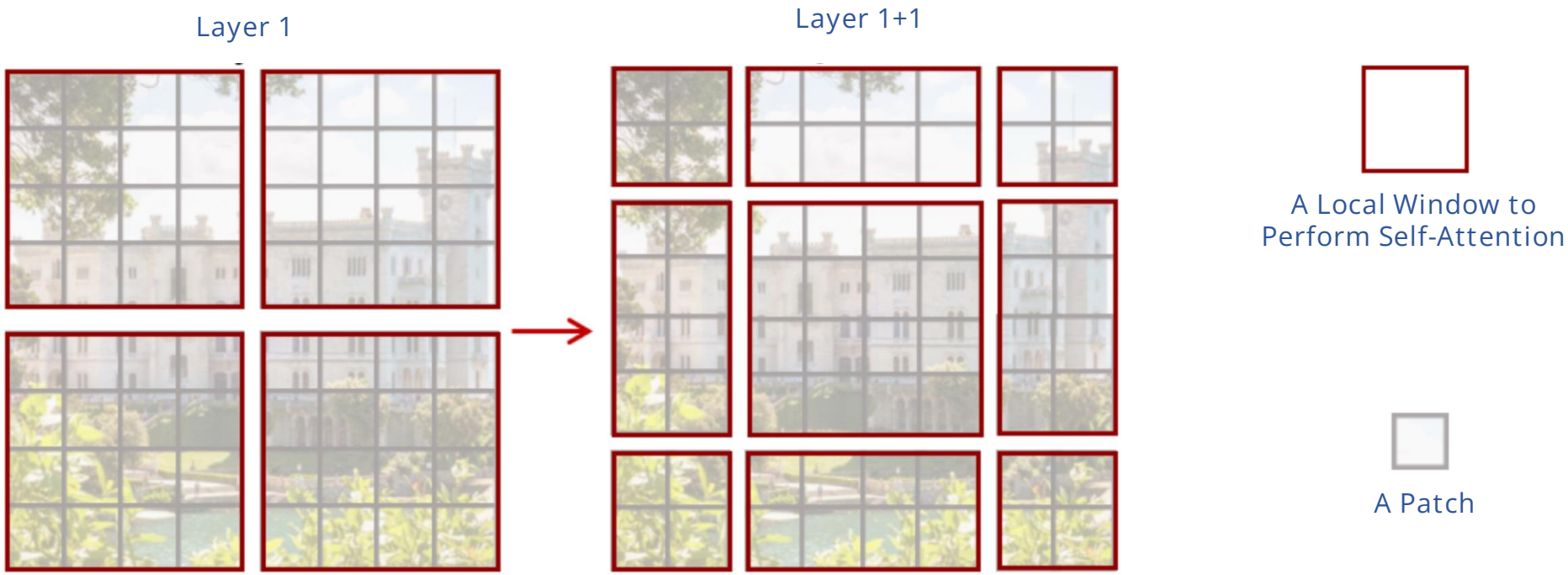

Fig 8. An illustration of the shifted window approach

Advantages of Swin Transformers

- Hierarchical Feature Representation
- Reduce complexity from n^2 to m*n
- Improved Performance
- Better Local and Global Context Modelling

Limitations of Swin Transformers

- Lack of Temporal Awareness
- No Built-in Mechanism for Sequence Learning

Video Swin Transformers [1]

- Video Swin Transformer is an extension of the Swin Transformer model, designed specifically for handling video data.
- It captures spatial and temporal features in videos, making it well-suited for tasks like video classification and action recognition.
- Instead of just 2D image patches, Video Swin Transformers operate on 3D patches that include both spatial dimensions (height and width) and the temporal dimension (time).
- The shifted window-based self-attention mechanism used in Swin Transformers is extended to 3D, meaning that attention is applied not just within spatial windows but also across adjacent time frames.

Video Swin Transformers Architecture

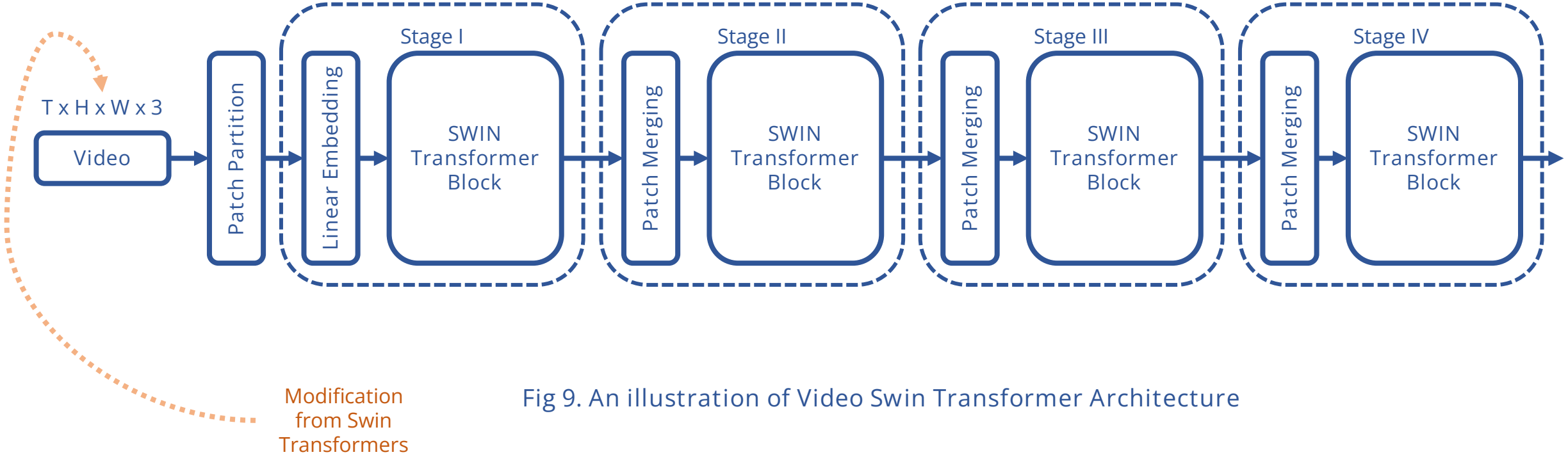

Fig 9. An illustration of Video Swin Transformer Architecture

Video Swin Transformers Architecture

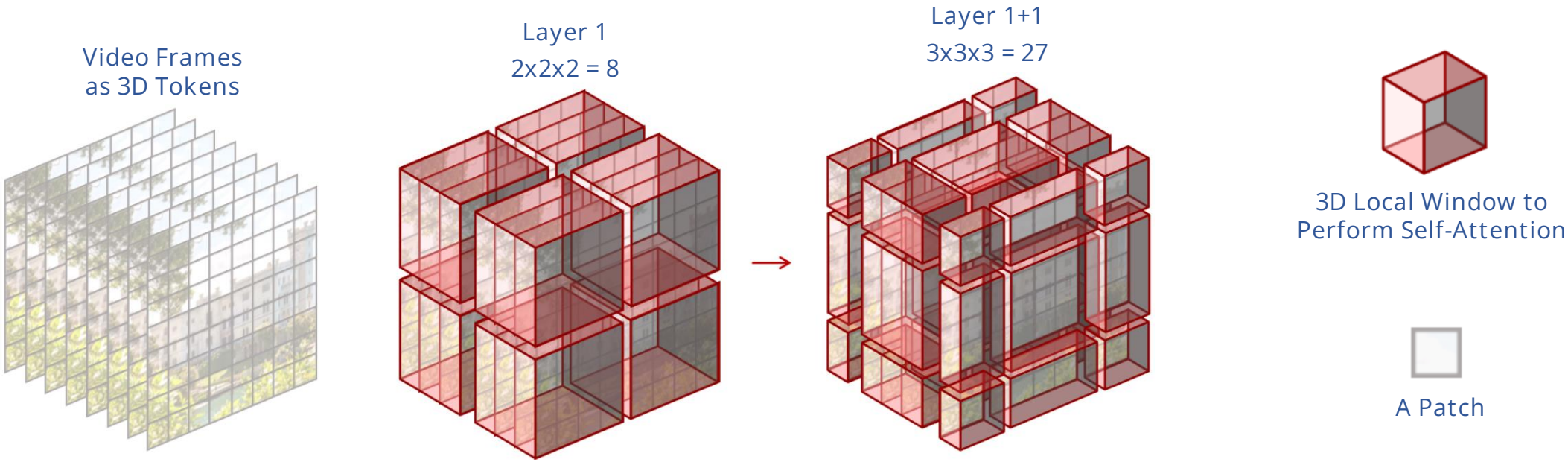

Fig 10. An illustration of 3D shifted window approach

Advantages of Video Swin Transformers

- State-of-the-Art Accuracy on Video Recognition Benchmarks.
- Efficient Spatiotemporal Modelling.
- Strong Performance in Temporal Modeling
- Smaller Model Size and Lower Computational Cost
- Improved Local and Global Feature Learning

Conclusion

- Explored the evolution of Transformers from NLP to computer vision.
- Introduced Visual Transformers and the need for new approaches.
- Swin Transformers are efficient in handling high-resolution images and dense prediction tasks.
- Limitations in temporal modelling, leading to the development of Video Swin Transformers.
- Video Swin Transformers utilizes 3D shifted window-based attention for spatiotemporal modelling.
- It provides a better speed-accuracy trade-off for video tasks.

References

[1] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video Swin Transformer. Retrieved from https://arxiv.org/abs/2106.13230

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Retrieved from https://arxiv.org/abs/1706.03762

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Retrieved from https://arxiv.org/abs/2010.11929

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Retrieved from https://arxiv.org/abs/2103.14030

Questions

Thank You