Part 2
Video Models

CNNs (ResNet50)

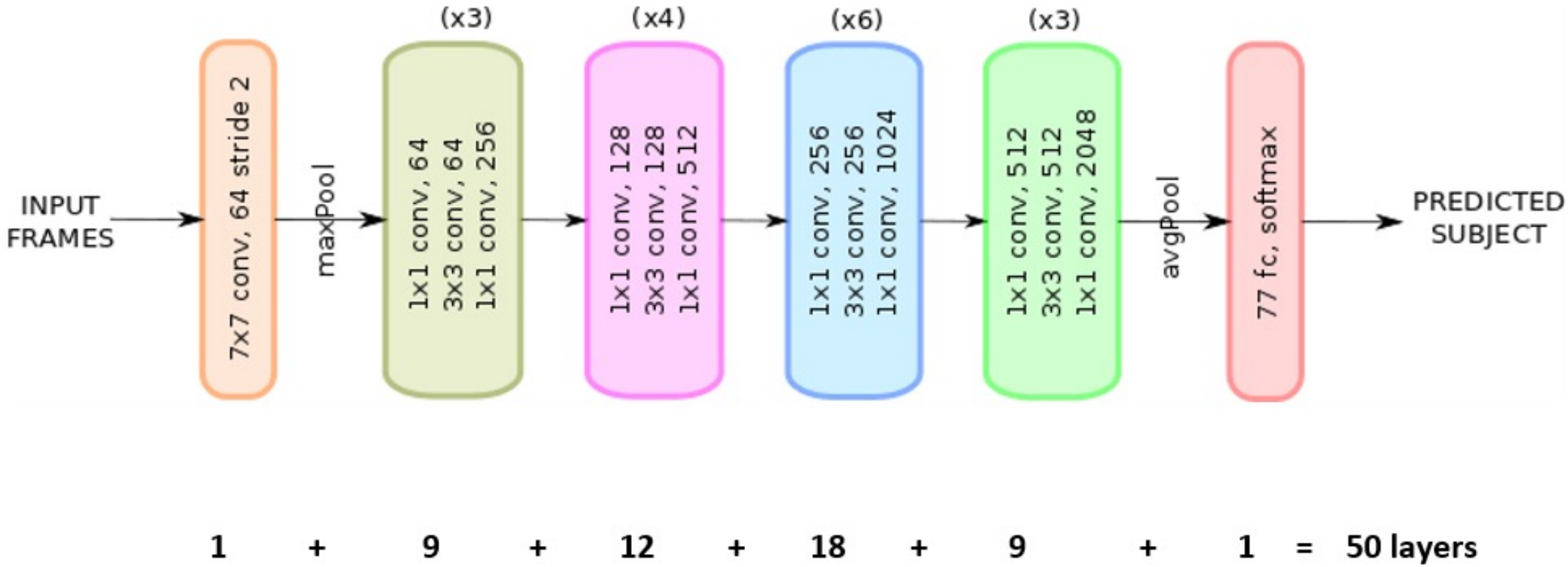
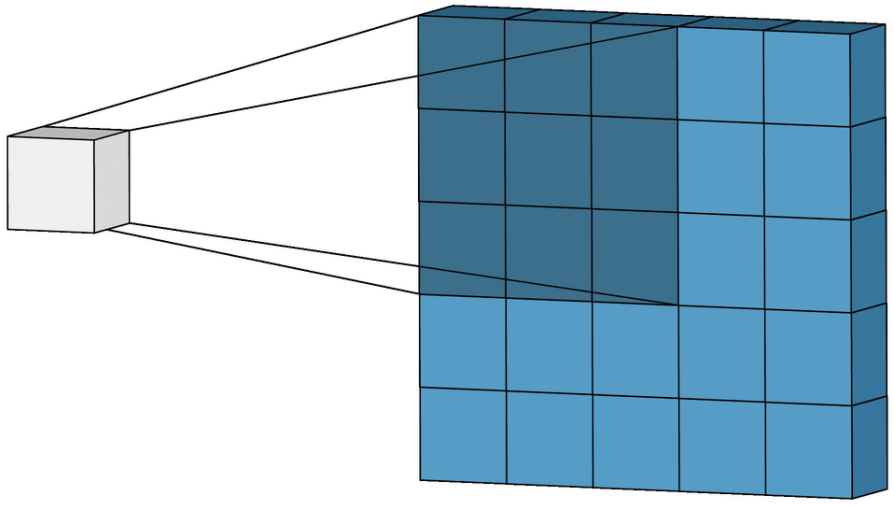
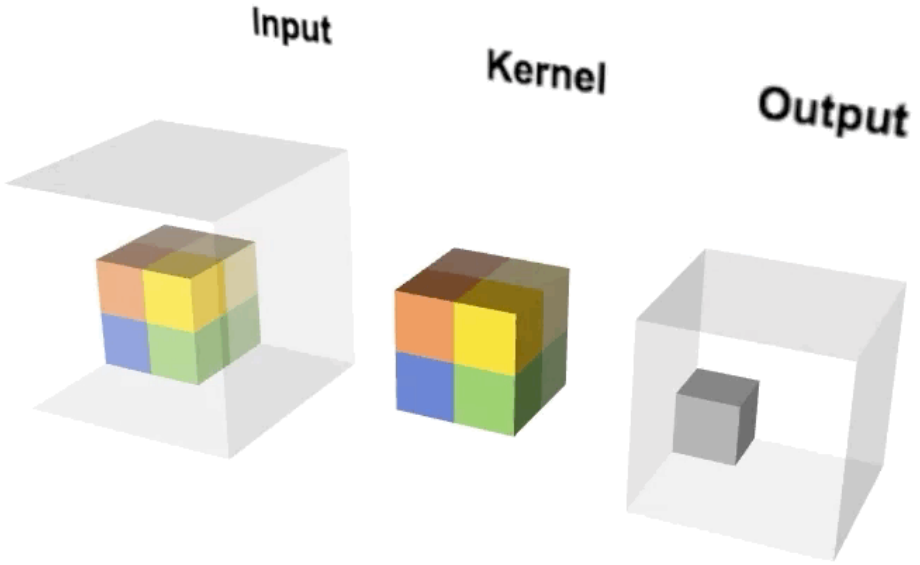


Image credit: https://srsapireddy.medium.com/resnet-50-introduction-b5435fdba66f

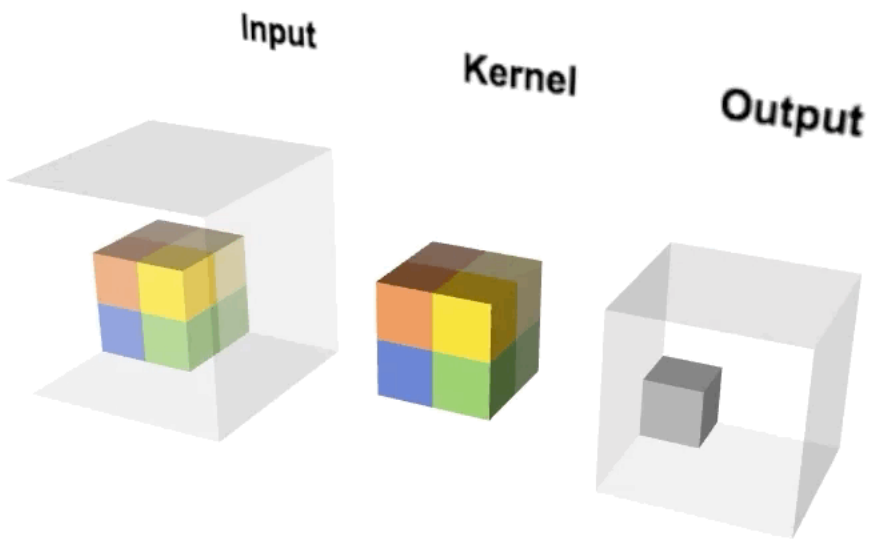
2D to 3D Convolutions

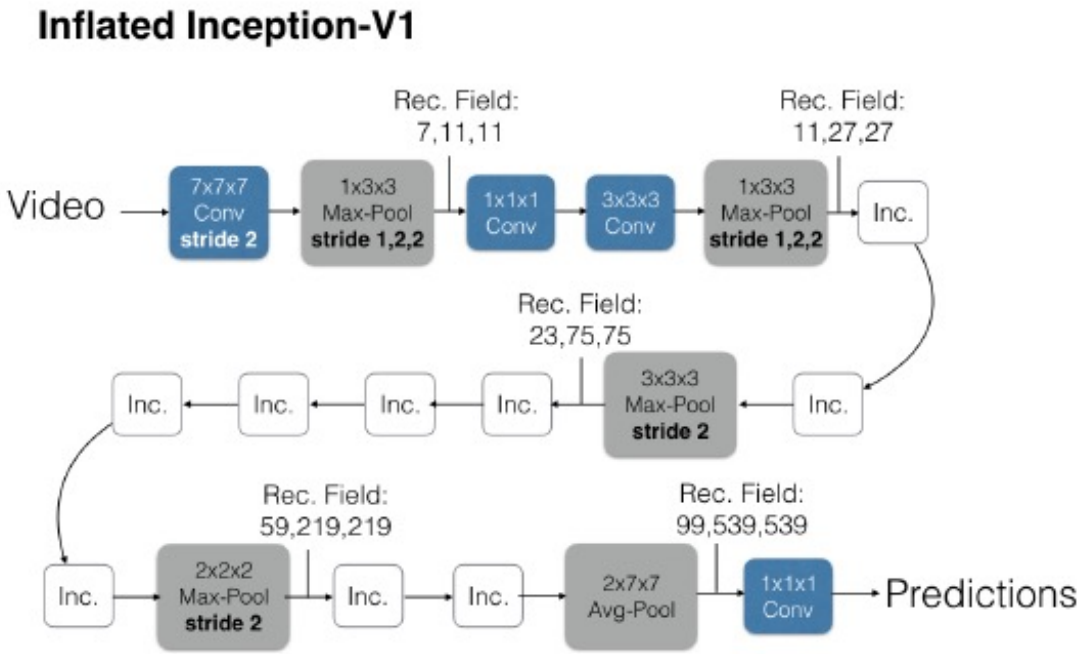




Tran, et al. Learning Spatiotemporal Features with 3D Convolutional Networks. ICCV 2015
Carreira and Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017
Image credits (2D Conv): https://miro.medium.com/v2/resize:fit:2340/1*Fw-ehcNBR9byHtho-Rxbtw.gif
Image credits (3D Conv): https://thomelane.github.io/convolutions/3DConv.html

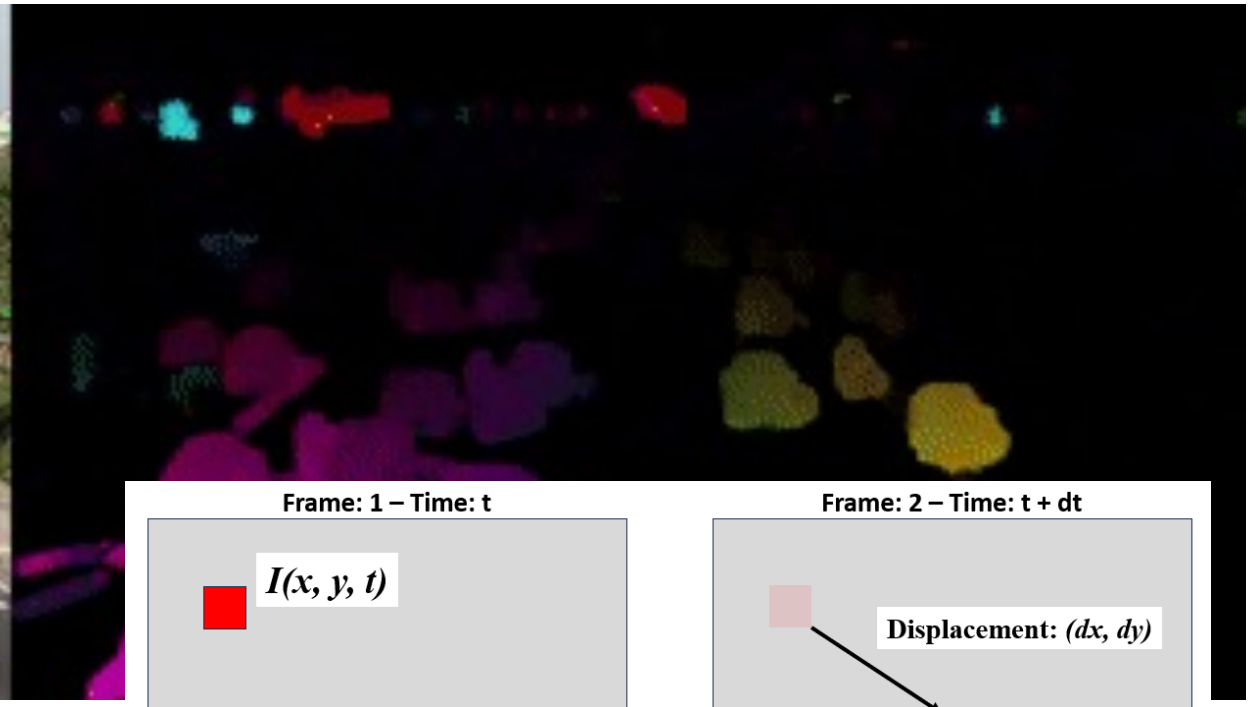
I3D Network





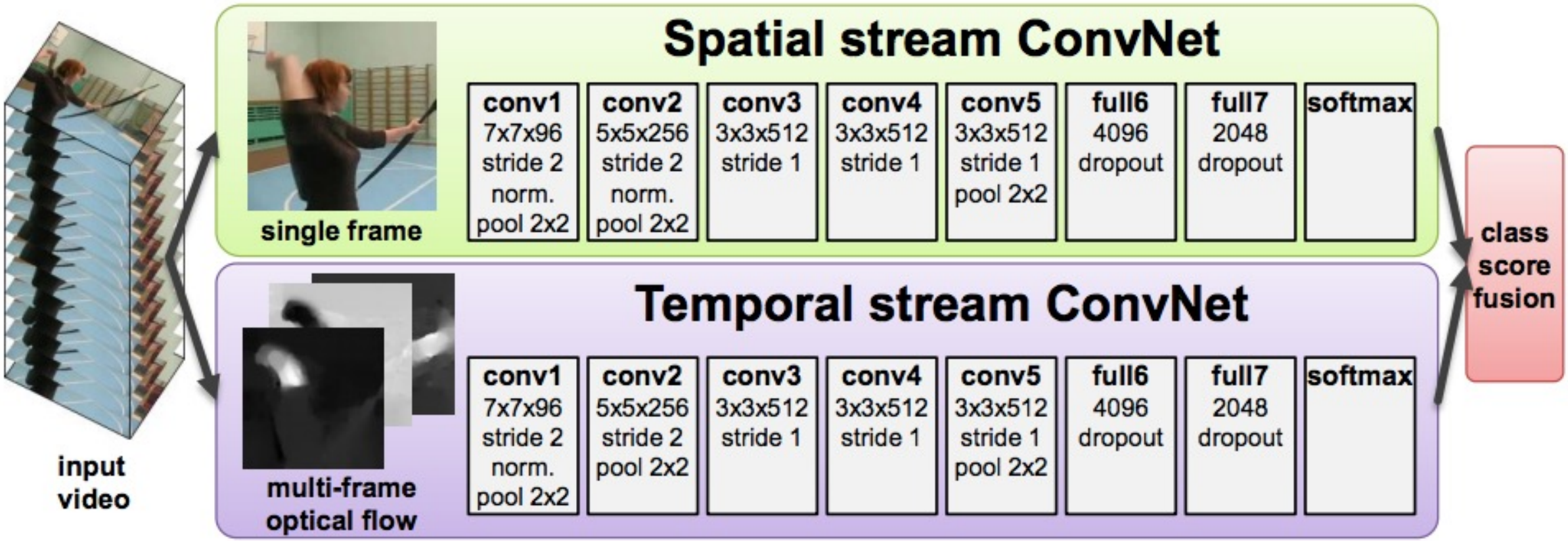
Carreira and Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017

Optical Flow (review)



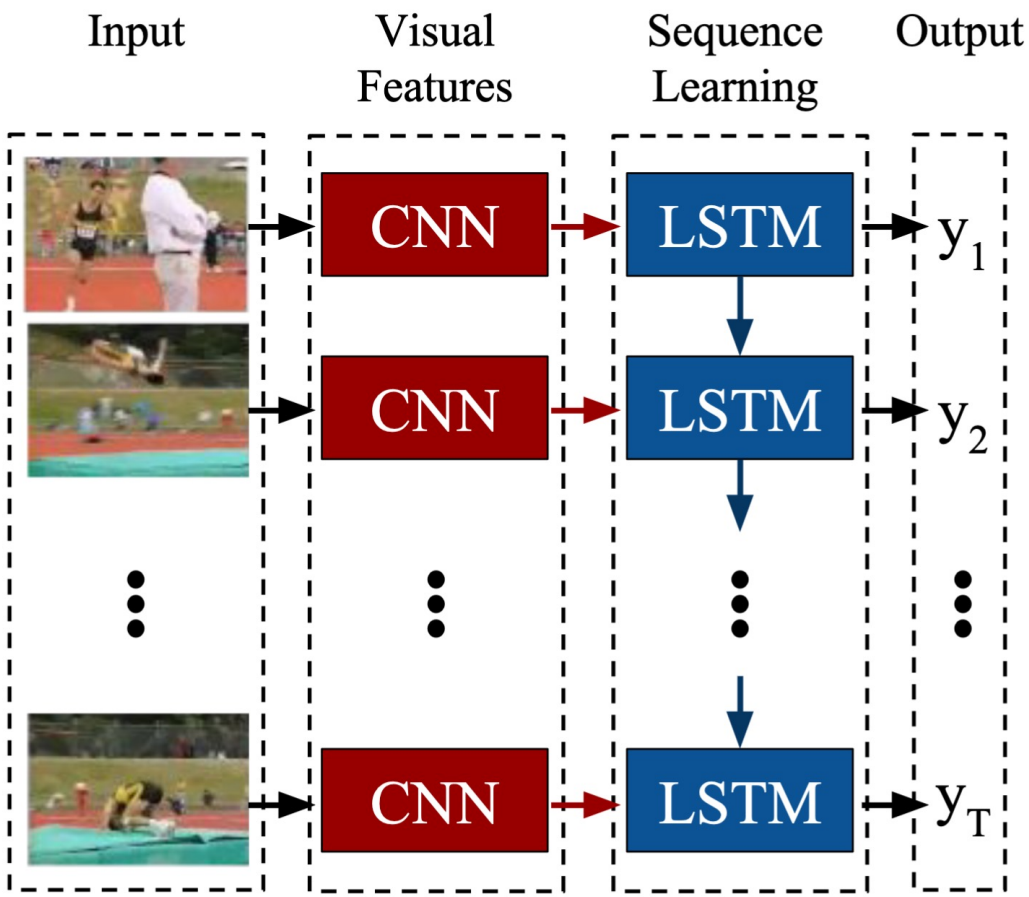
Video credit: https://nanonets.com/blog/optical-flow/
Image credit: https://towardsdatascience.com/optical-flow-with-raft-part-1-f984b4a33993

Two-stream networks



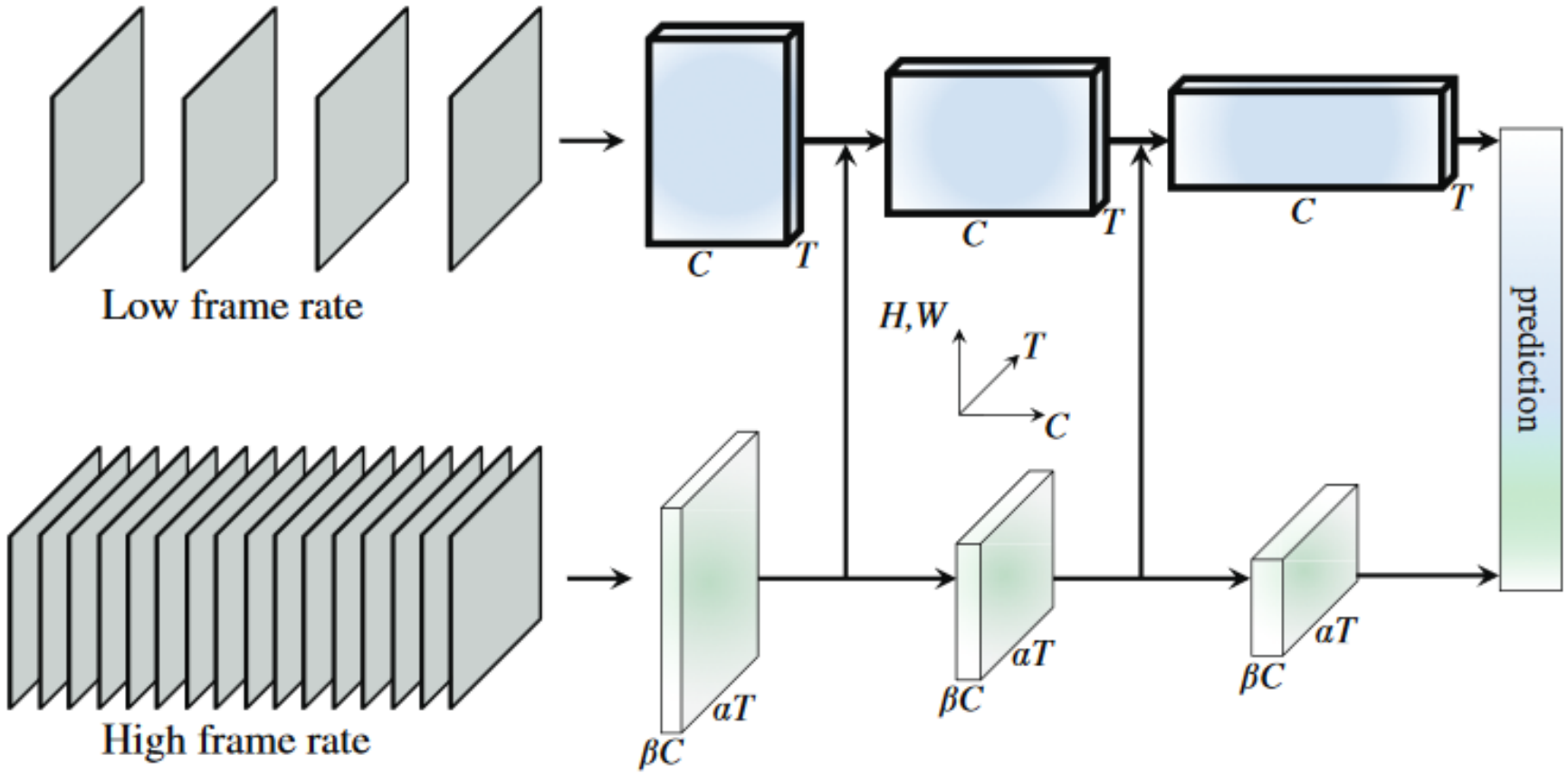
Simonyan and Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. NeurIPS 2014

Convolutional + Recurrent Neural Networks



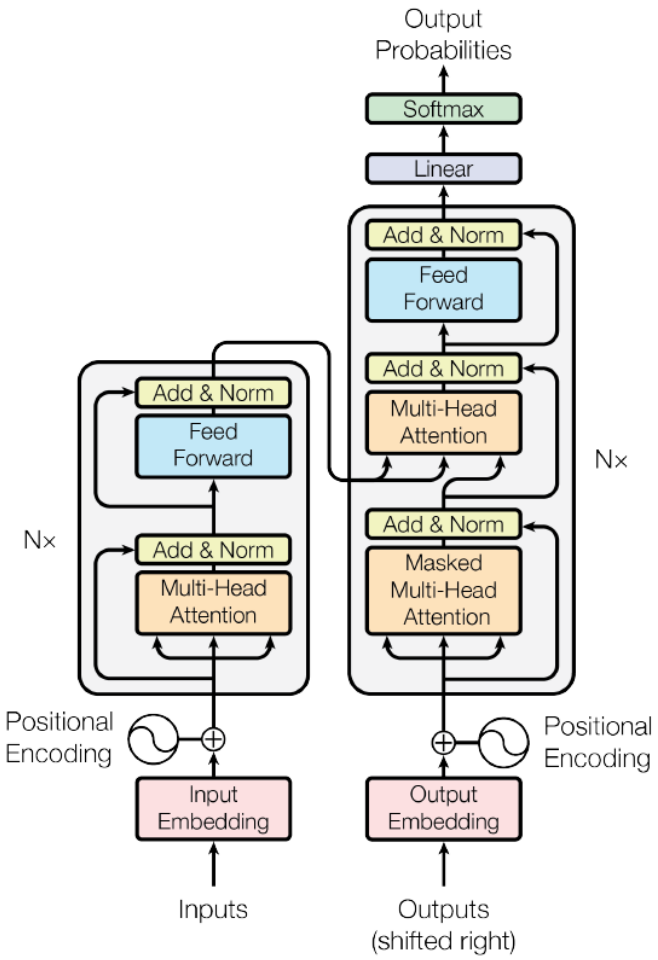
Donahue, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015, PAMI 2016

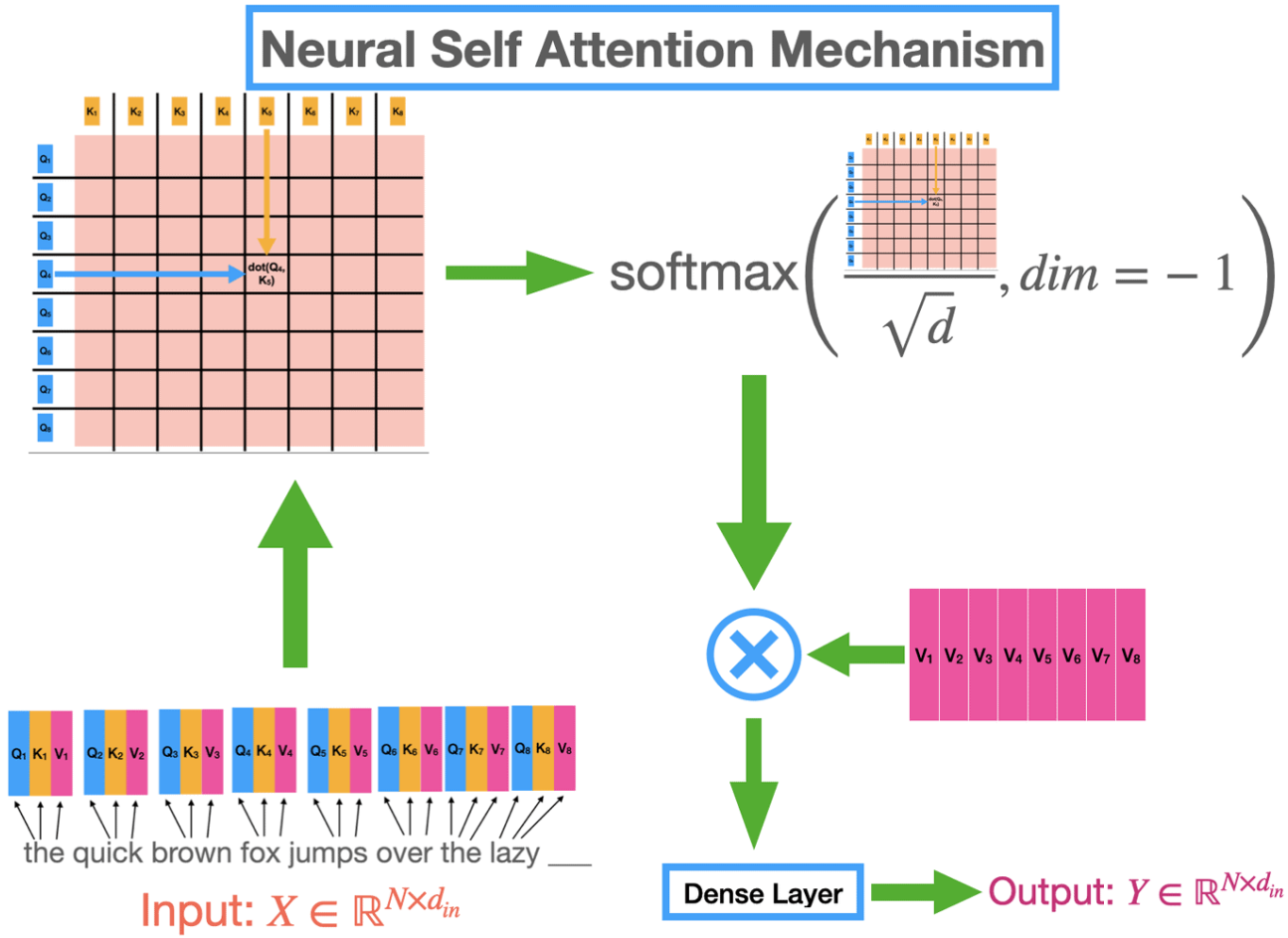
SlowFast Networks



Feichtenhofer, et al. SlowFast Networks for Video Recognition. CVPR 2019

Transformer (review)

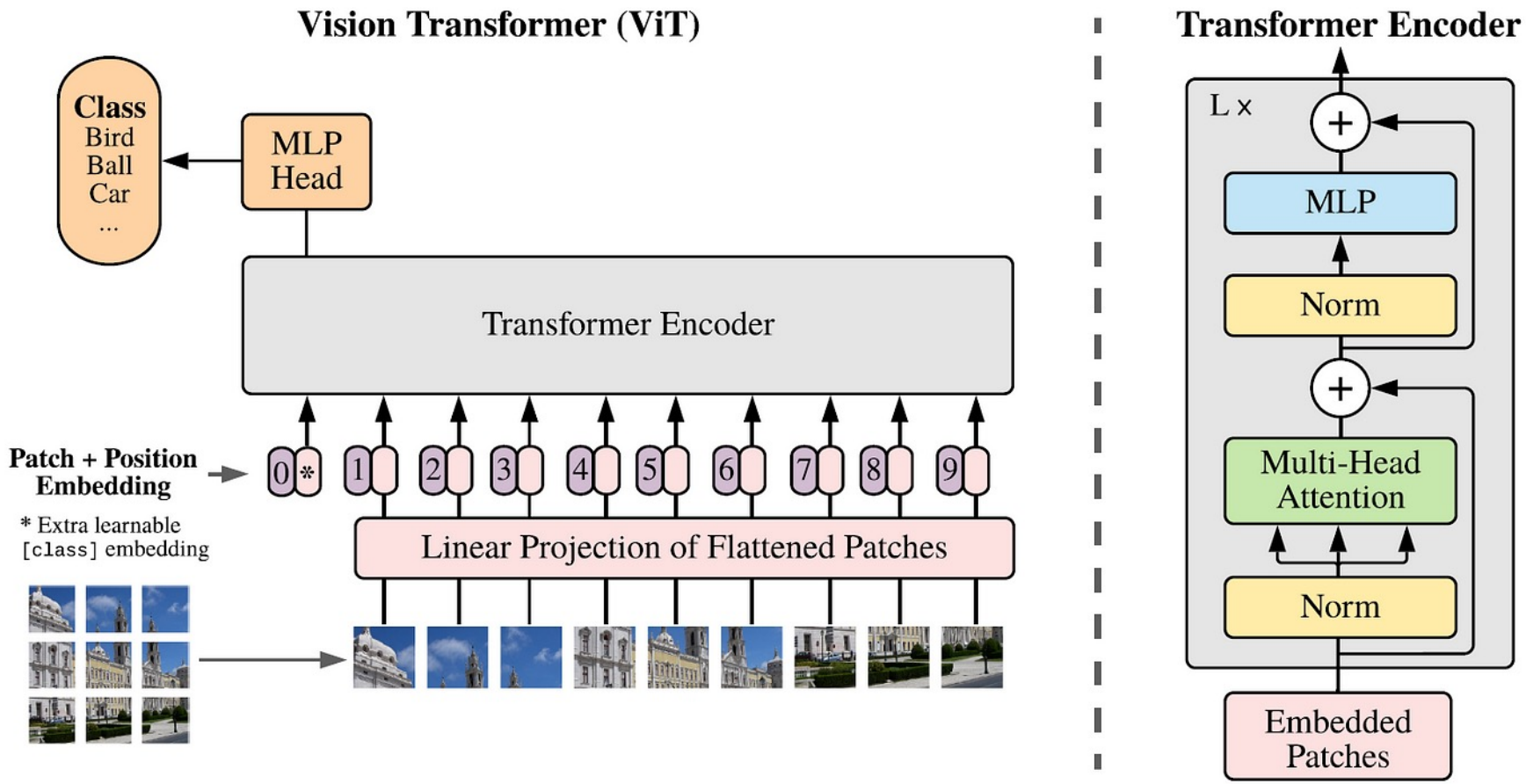




Vaswani, et al. Attention Is All You Need. NeurIPS 2017

Image credit: https://learnopencv.com/attention-mechanism-in-transformer-neural-networks/

Vision Transformer, ViT (review)

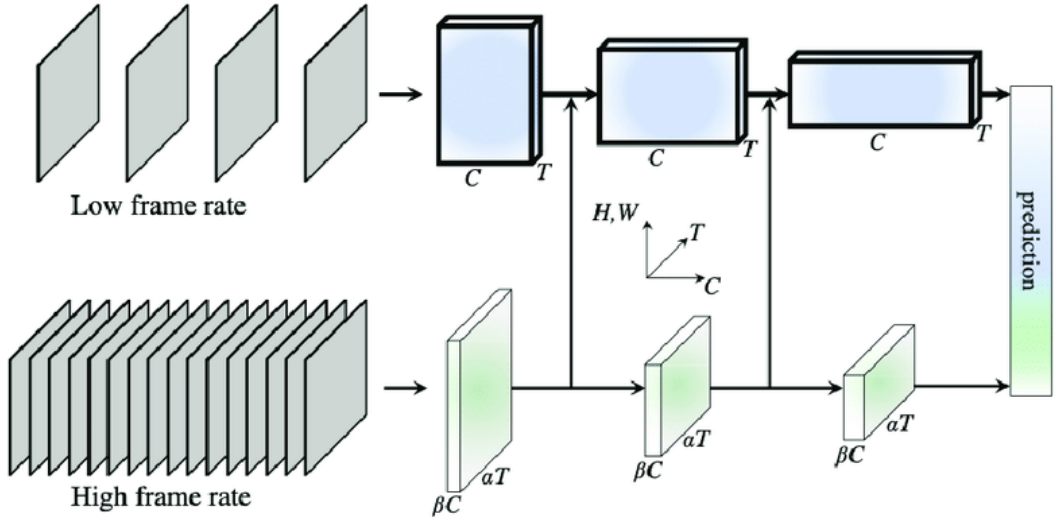


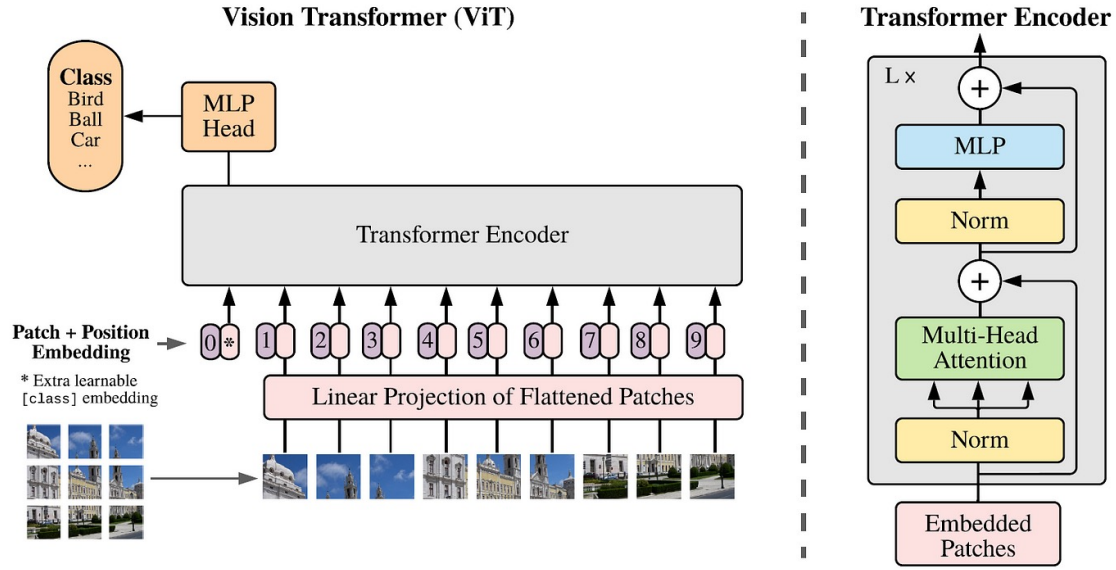
Dosovitskiy, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

CNNs vs Transformer models

- Convolutions
- Pooling layers
- Local receptive fields

- Patches as tokens
- Self-attention across tokens
- Modeling across the image

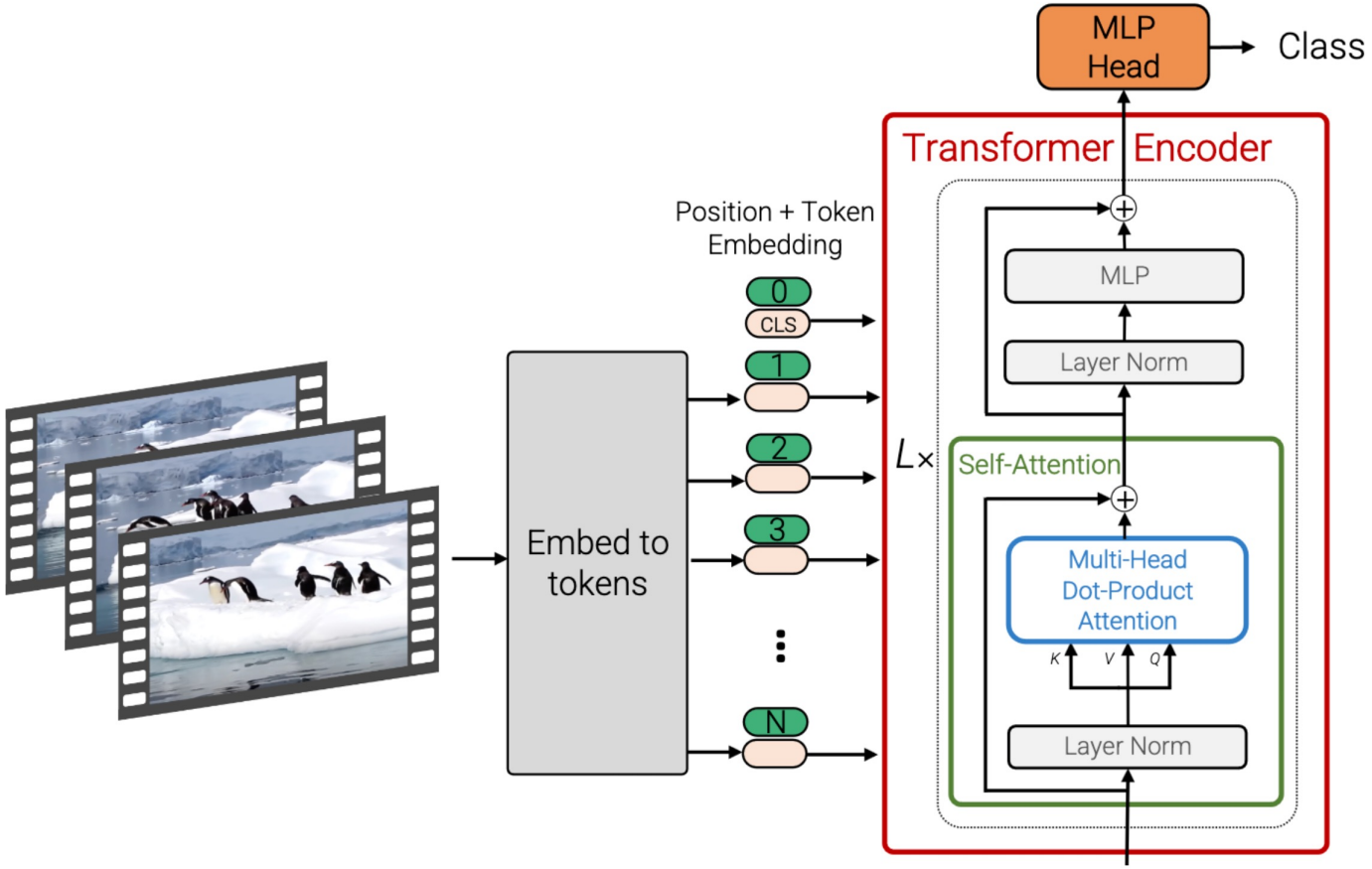




Feichtenhofer, et al. SlowFast Networks for Video Recognition. ICCV 2019

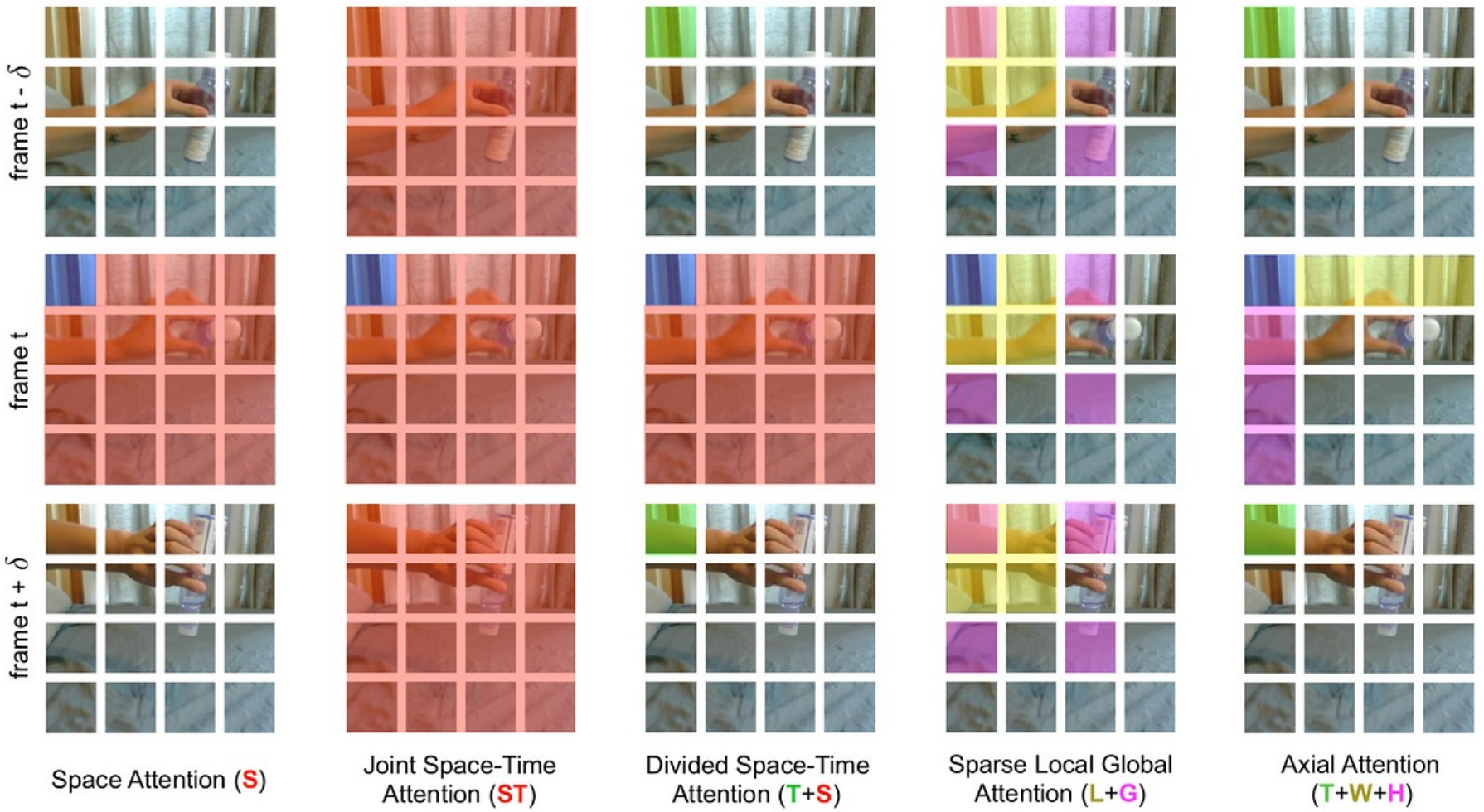
Dosovitskiy, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Video Vision Transformer



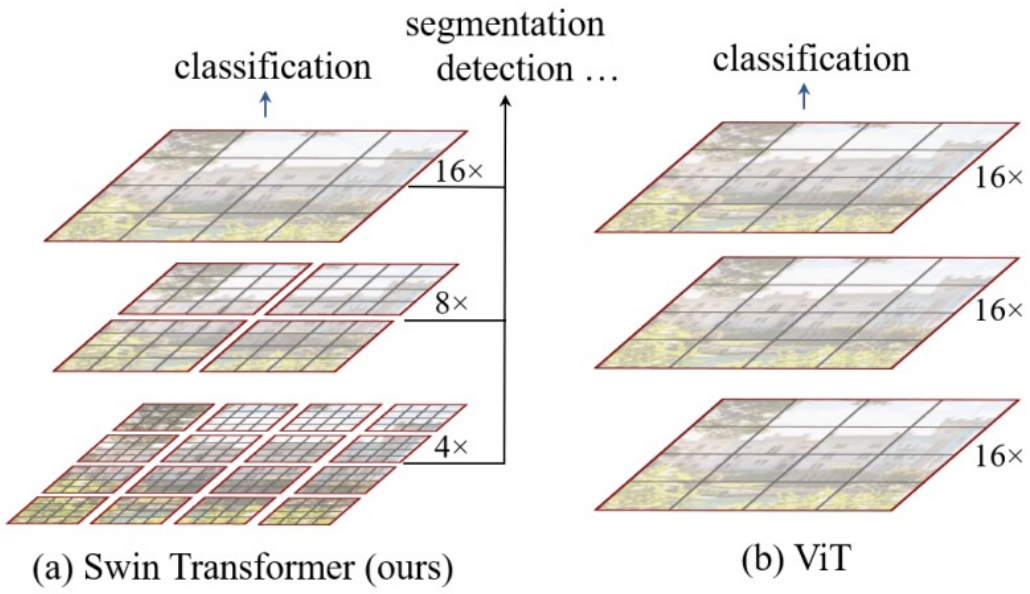
Arnab, et al. ViViT: A Video Vision Transformer. ICCV 2021

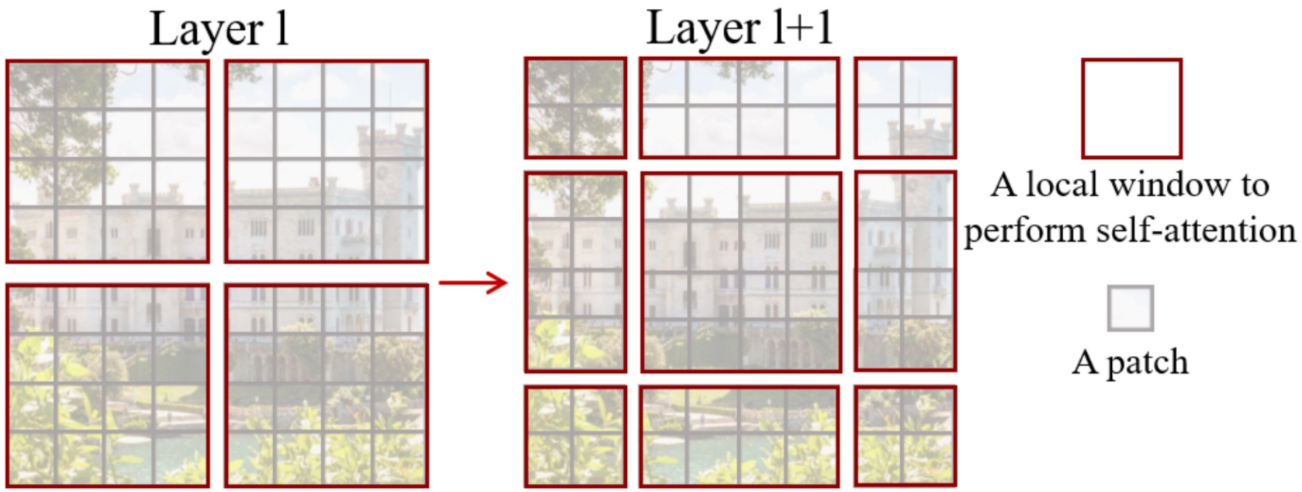
TimeSFormer: Where to attend?



Bertasius, et al. Is Space-Time Attention All You Need for Video Understanding. ICML 2021

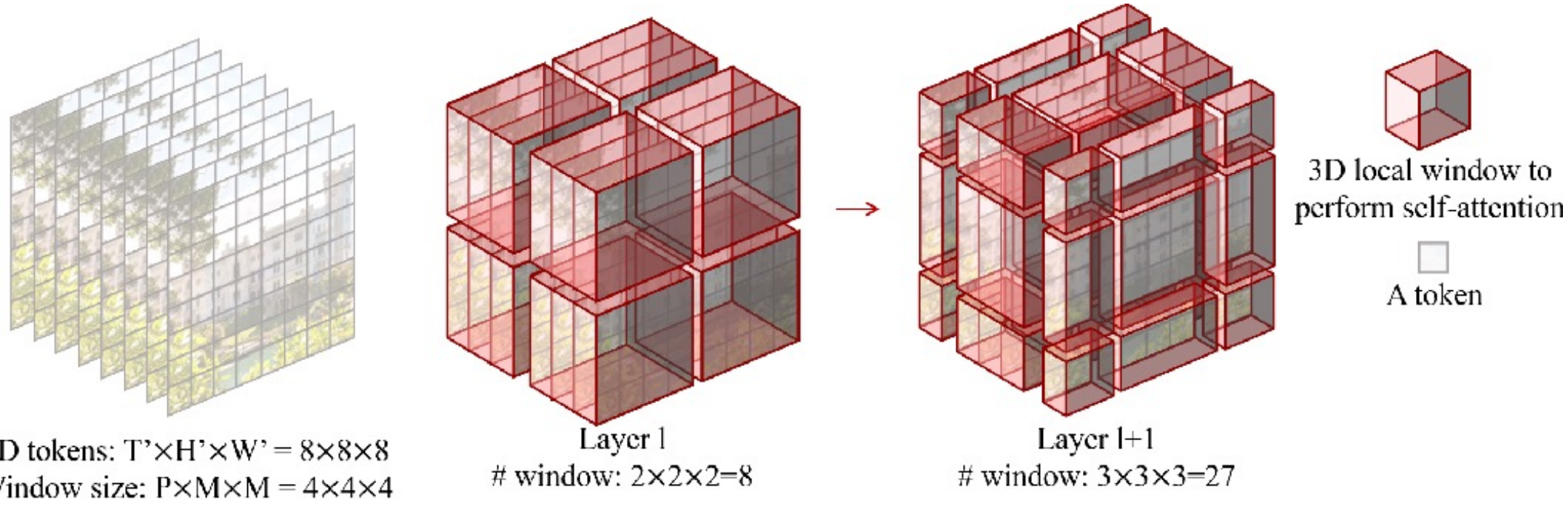
Swin Transformer





Liu, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. CVPR 2021

Video Swin Transformer



Liu, et al. Video Swin Transformer. arXiv 2106.13230

Questions?