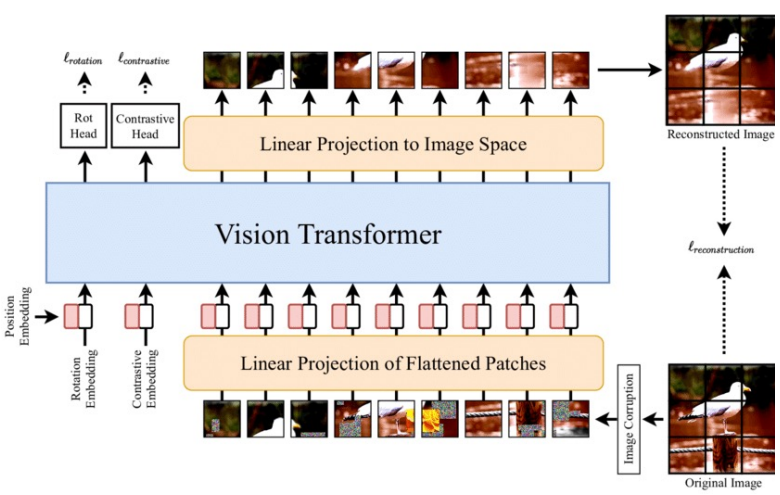
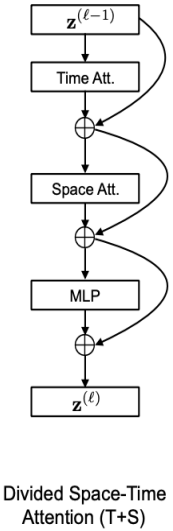
Action Recognition

Minsak (6800151)
Deeptanshu (6783151)
Rutwik (6771154)

Timesformer - Is Space-Time Attention All You Need for Video Understanding?





Divided Space-Time Attention (T+S)

Divided Space-Time Attention (T+S)

Input (Clip of F RGB Frames of Size H × W)

Decomposition into Patches (N Non-overlapping P × P Patches)

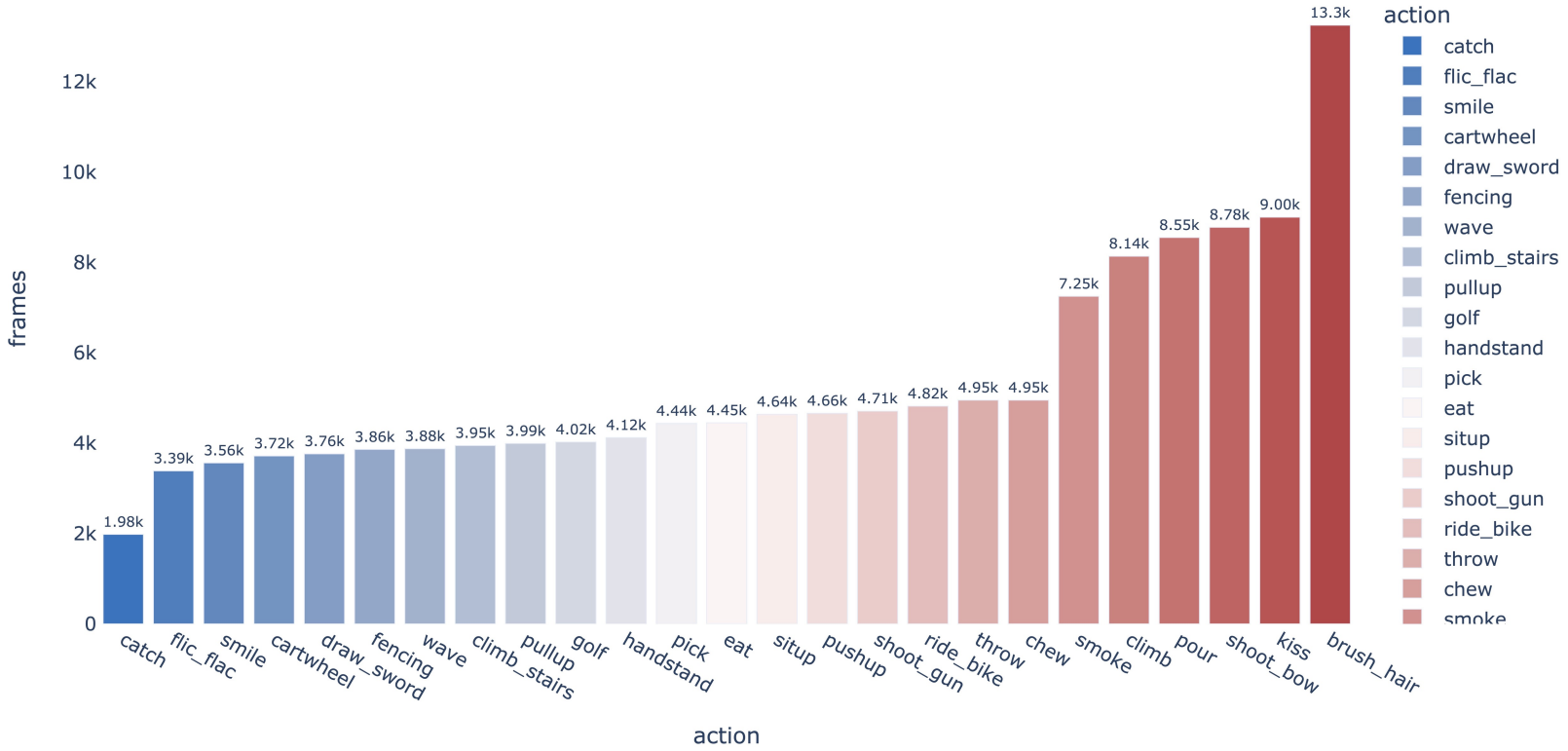
Linear Embedding (Linear Mapping of Patches to Embedding Vectors)

Transformer Encoder Blocks:
- Query-Key-Value Computation
- Self-Attention Computation
- Encoding
- Residual Connections

Classification Embedding (Final Clip Embedding)
- MLP Projection
- Classification Prediction

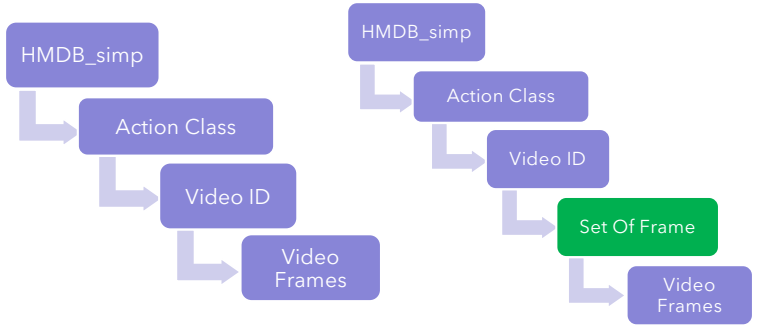
Space-Time Self-Attention Models:
- Joint Space-Time Attention
- **Divided Space-Time Attention**
- Sparse Local Global Attention
- Axial Attention

Data Analysis – Frames per category



Data Processing

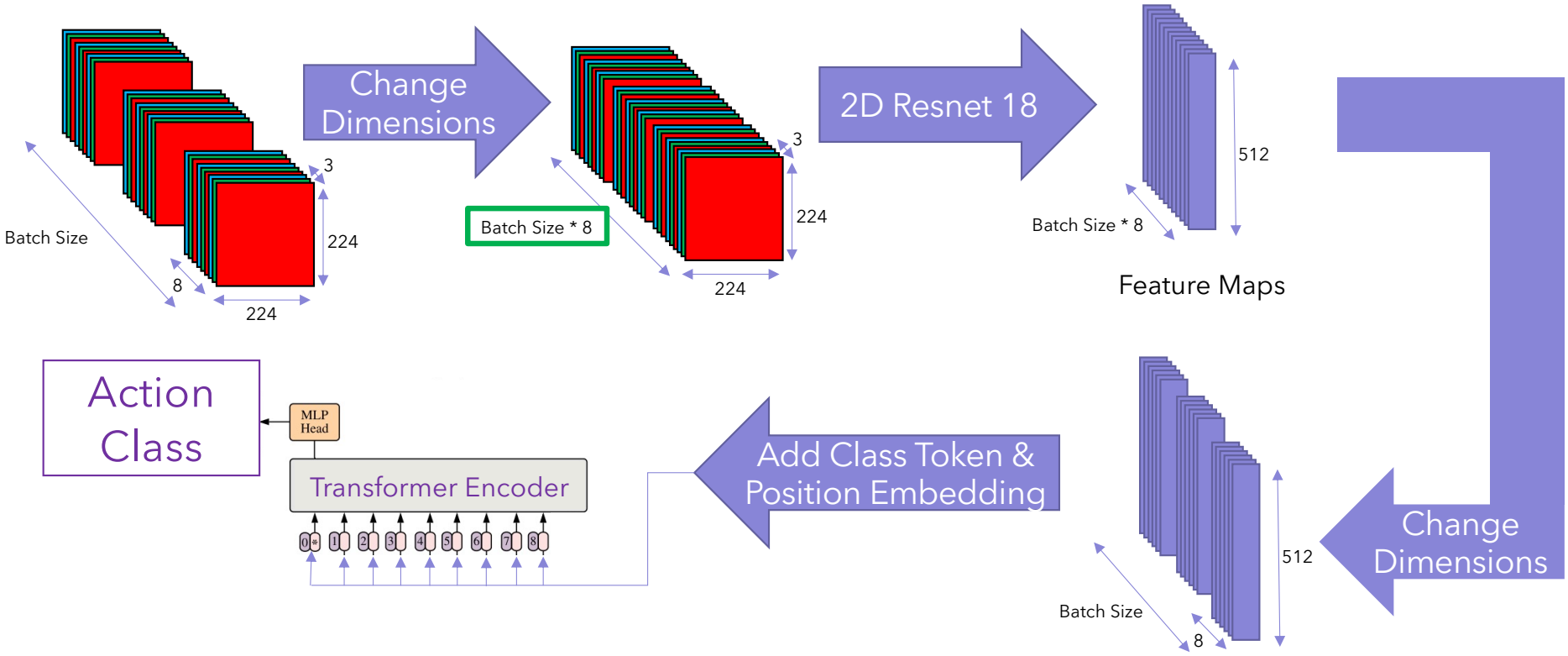
- Issues Faced
  - Out Of Memory Error
  - Random Split Issue



Old Structure New Structure

- Solutions
  - Changed The Data Storage Structure
  - Created Random Split on first instance the saved in CSV

2D Resnet_18 With Attention

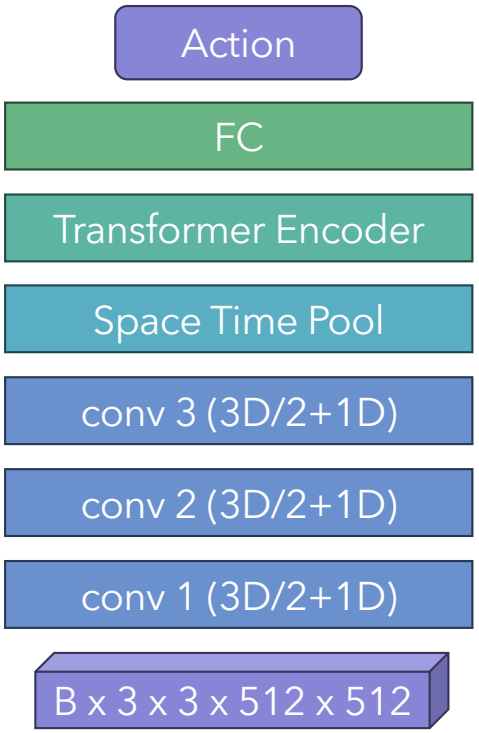


Spatio-temporal Convolutions with Attention
Convolution + contextual information(from transformers)

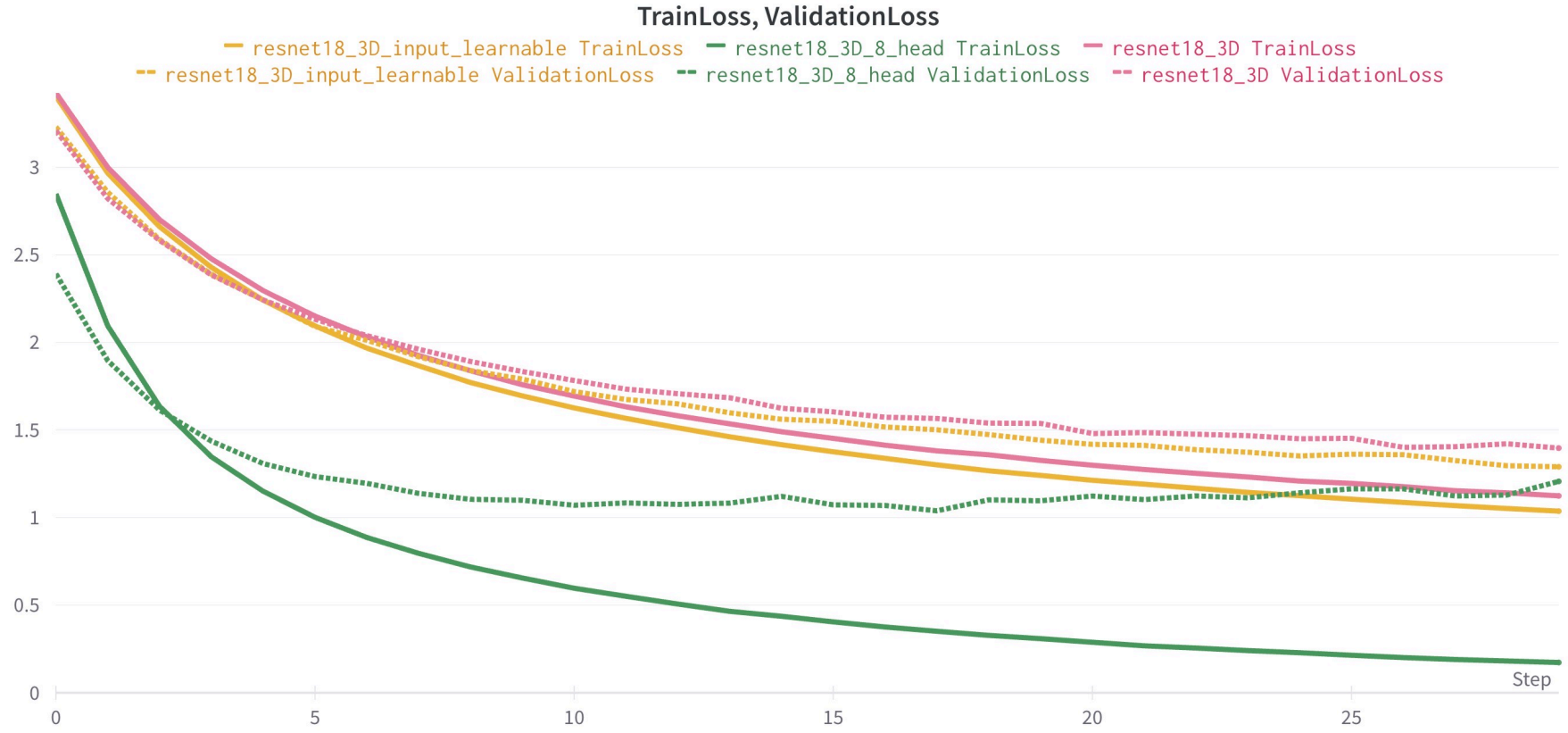
- **Backbone Networks**
- **3D** convolution
  - Extracts temporal and spatial information together

  2+1D convolution
  - Extracts temporal and spatial information separately | different kernels

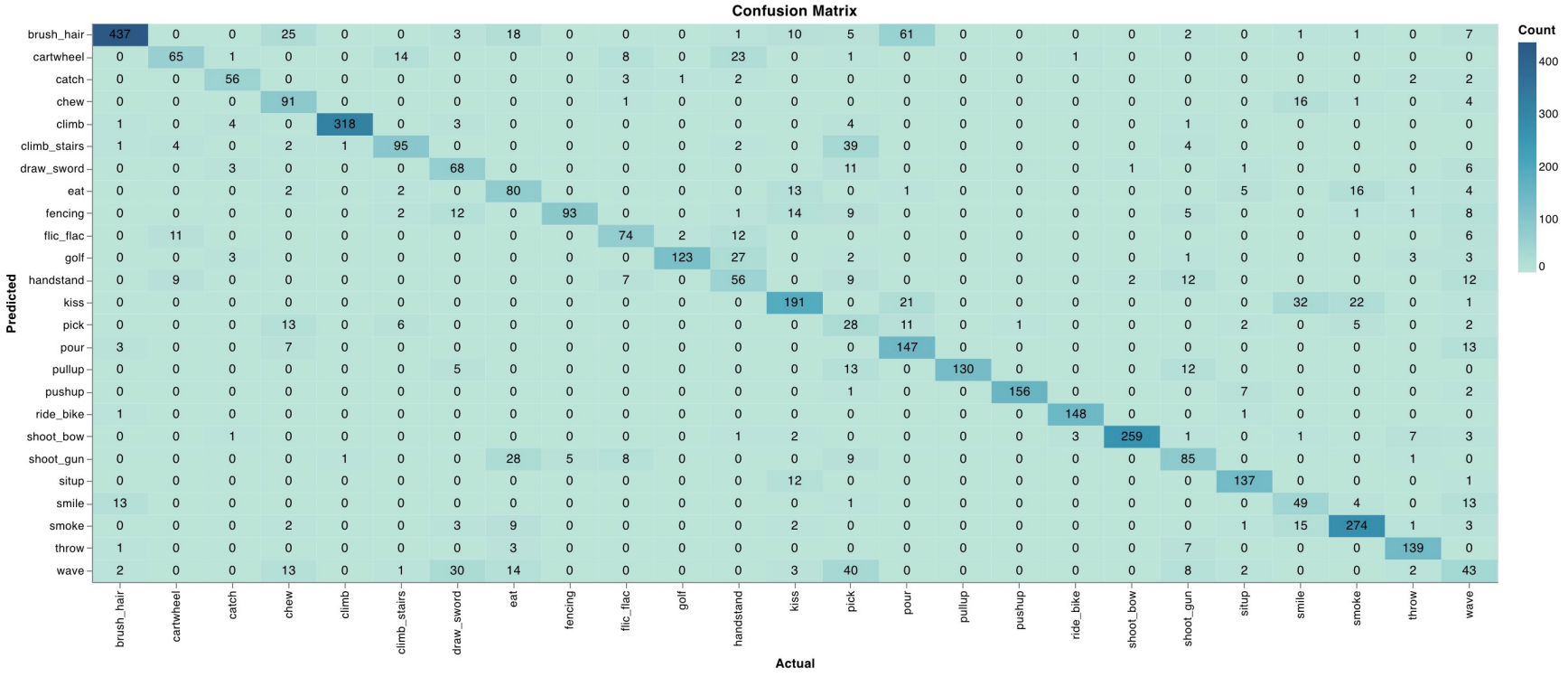
  Experiments with Backbones
  - 5D Input - [B, T, C, H, W] **|** T = 3
  - Backbones pre-trained on Kinetics 400 - **65.32%**
  - **Transformer encoder Layer**
  - Tried different number of attention heads
  - Making convolution Input learnable | Vanishing gradient
  - Modified Adaptive average pooling layer of backbone
    - **More feature output**
    - **Rank 1 – 77.36%**



Plotting the loss with WandB



Confusion Matrix for ResNet3D with Attention



PR Curve for ResNet3D with attention