Video Classification with Deep Learning

1. SARTHAK SUNIL DHANKE
2. URSULA GUO
3. IRIS HUANG
4. DAVID WEI

Content

- MOTIVATION
- DATASET – UCF 101
- EDA
- SPATIAL VS SPATIOTEMPORAL CONV
- (2 + 1)D MODEL
- (2 + 1)D + 2D CONV MODEL
- TRADEOFF & RESULTS
- APPLICATION TO STYLE TRANSFER
- CONCLUSION & FUTURE DIRECTION

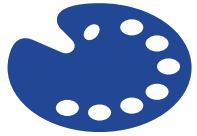
Motivation

🎥 Why Video Matters ?

- **Explosion of video data** in everyday life (social media, surveillance, entertainment).
- **Text/image alone is not enough** — video adds **temporal context** and **richer insights**.

Key Applications:

- 🔒 **Surveillance & Monitoring**
- 📊 **Business Trend Analysis**
- 🎬 **Media Tagging & Highlights**
- 🚗 **Real-Time Systems** (e.g., autonomous driving, AR)



🎨 Why Style Transfer?

A **powerful creative tool** in video generation and editing.

Real-World Use Cases:

- 🎮 **Game/Animation/Film Production**
- 🕰️ **Restoring Classic Films** with modern graphics
- 🧠 **Experimenting with artistic expression** in generative models

UCF – 101

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions

UCF101 includes total number of 101 action classes which we have divided into five types: Human-Object Interaction, Body-Motion Only, Human Interaction, Playing Musical Instruments, Sports.

CATEGORY	DETAILS
Actions	101
Clips	13,320
Groups per Action	25
Clips per Group	4 - 7
Mean Clip Length	7.21 sec
Total Duration	1,600 mins
Min Clip Length	1.06 sec
Max Clip Length	71.04 sec
Frame Rate	25 fps
Resolution	320 × 240
Audio	Yes (51 actions)

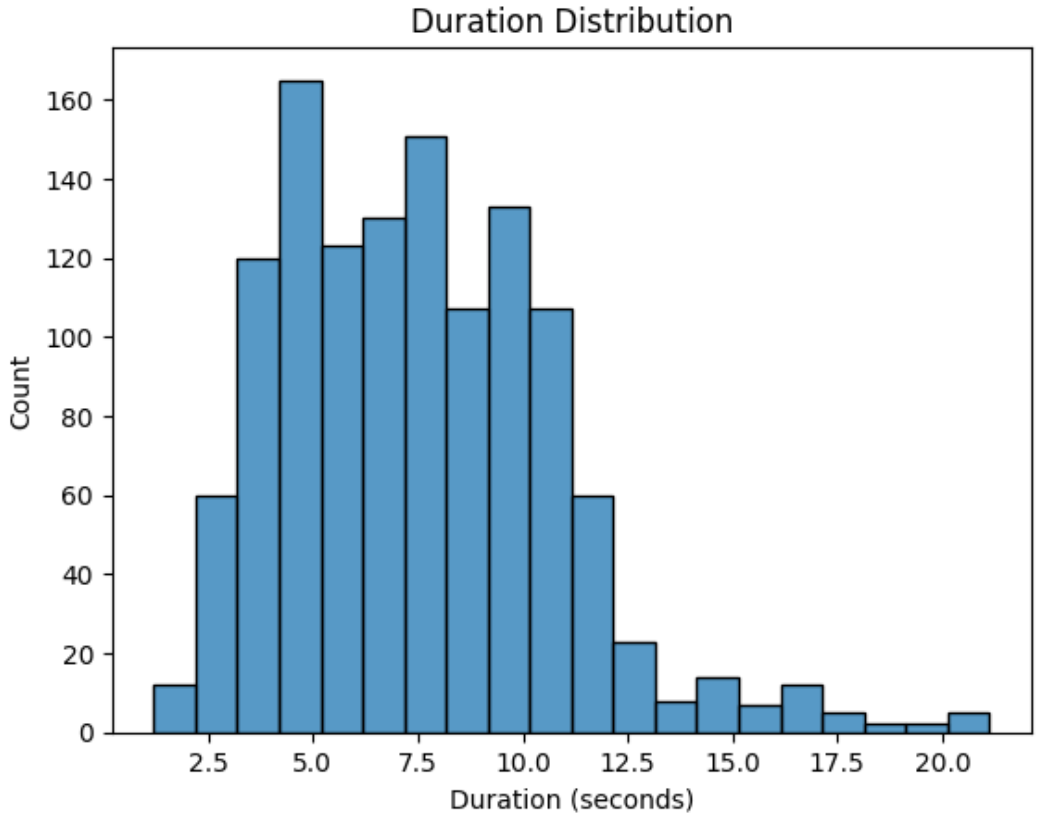
Subset used for this study

- For this project, we selected a 9-class subset representing diverse activity types while remaining **computationally feasible**.
- This subset retains **variability across motion types, lighting conditions,** and environments(indoor/outdoor), allowing meaningful model evaluation despite the reduced scope.

Data Distribution

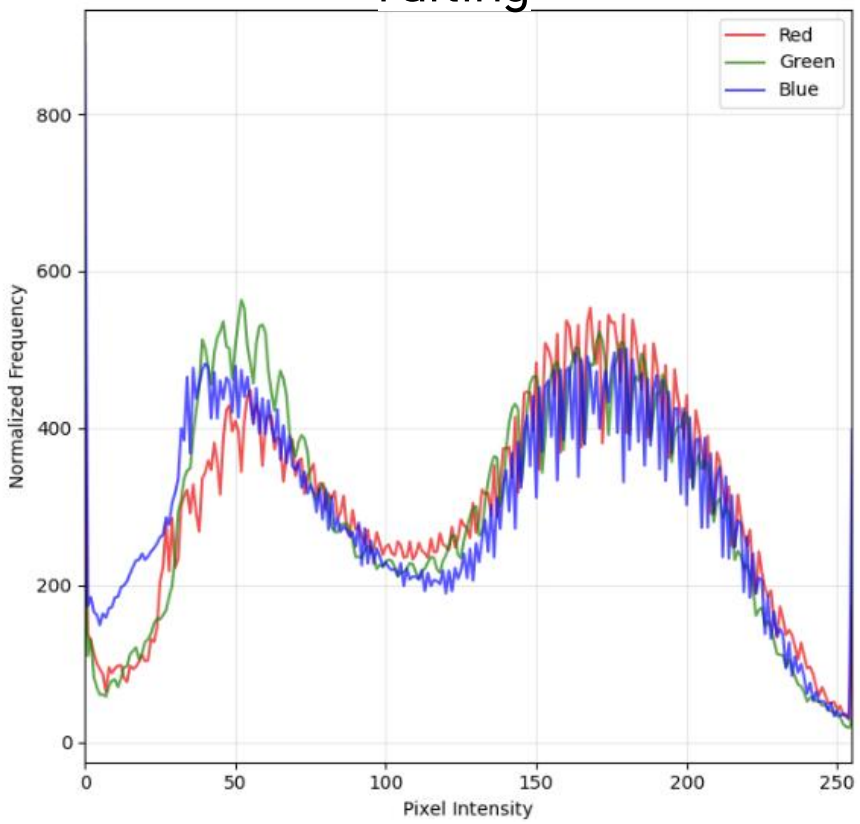
- Overview:
  - Videos are evenly distribution for each class
    - 9 classes in total
  - Original resolution: 320 x 240
  - 25 frames per second

- Duration Distribution:
  - Most videos are short, typically 2–10 seconds.
  - The right-skewed distribution suggests a few longer clips, but overall the dataset supports fixed-length sampling for action recognition.

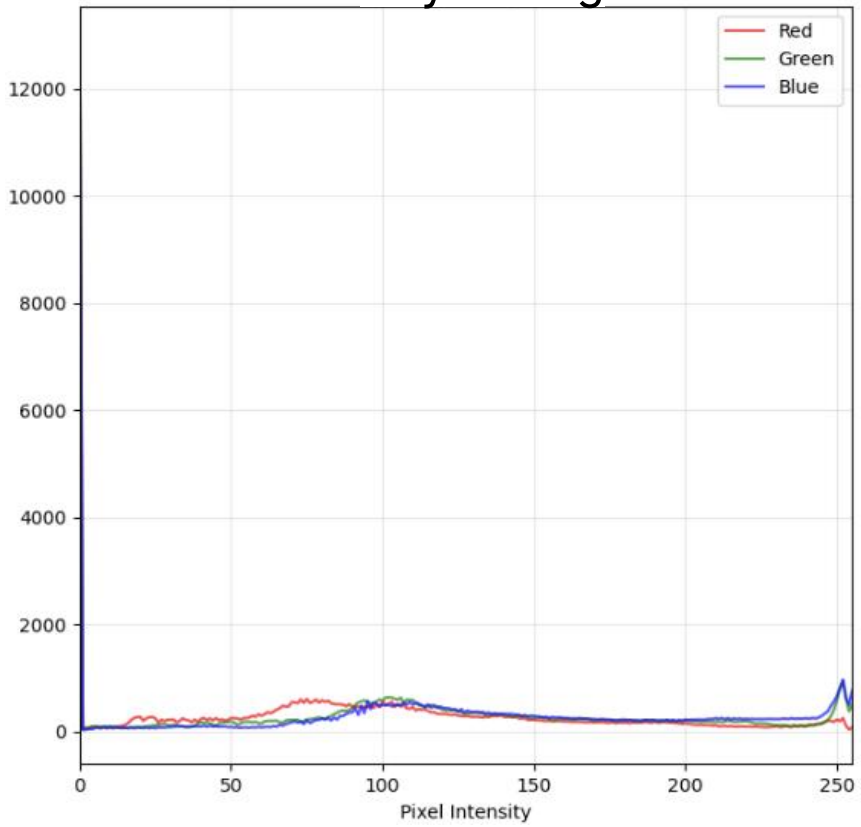


RGB Pixel Intensity Diagram

rafting



Sky Diving



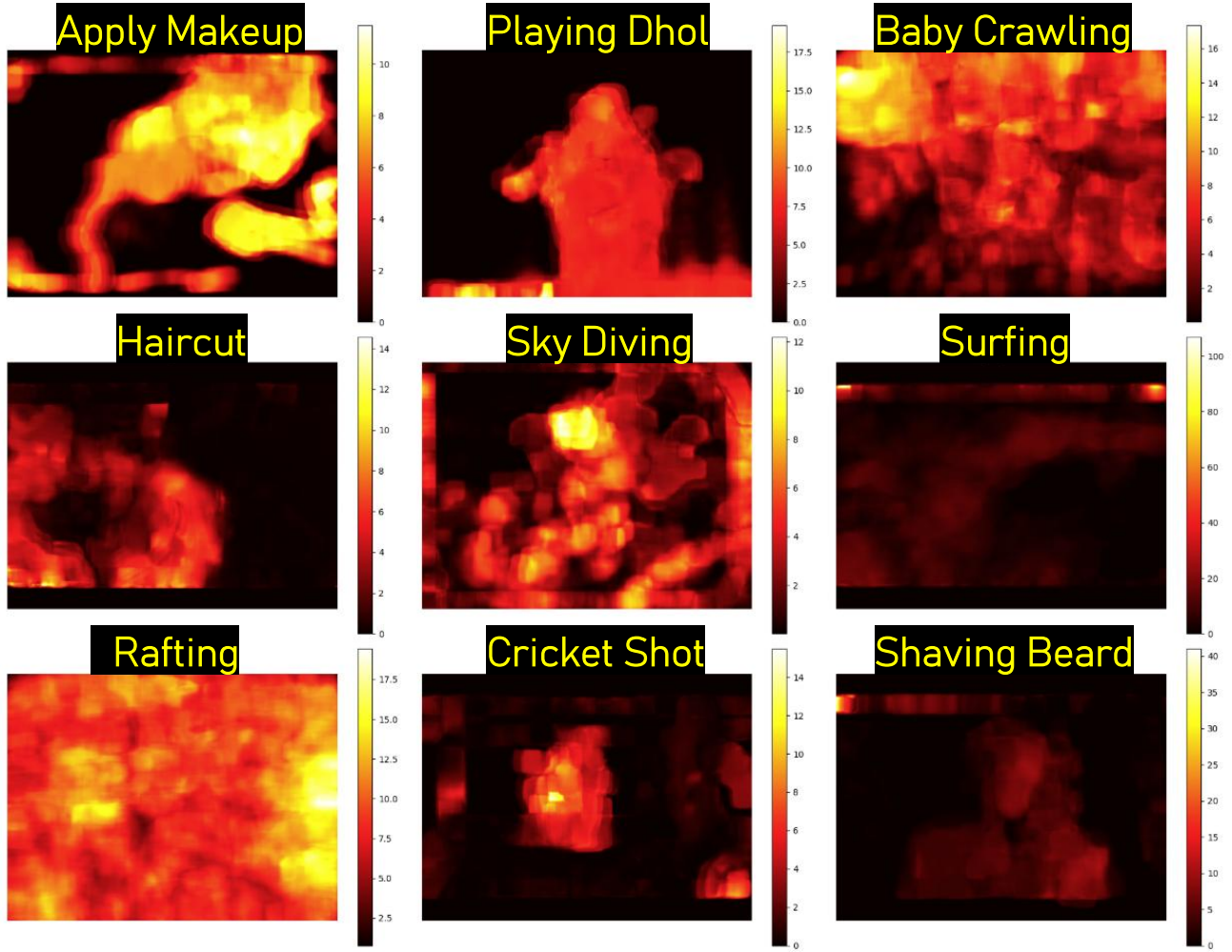
RGB Histogram comparison between rafting and sky diving.

- Showing significant difference in color distribution based on the activity
- Rafting more color, sky diving blue/green peak

Optical Flow

Optical Flow magnitude diagram showing the typical amount of motion for each of the classes/activities.

We see there are some classes that have much lower amount of motion, while others are changing intensely (especially rafting, which also shows most saturation in color across RGB too)

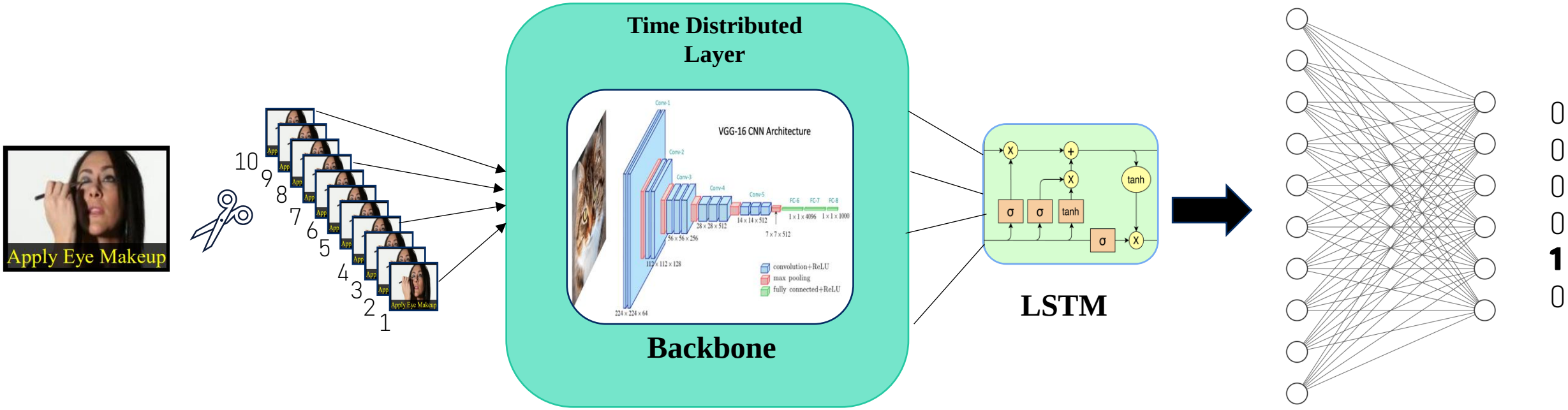


Sequential vs. Joint Spatiotemporal Models

*We chose to **experiment with both** because they emphasize different aspects of temporal modelling and help us understand how design decisions impact learning.*

Aspect	2D CNN + LSTM	(2+1)D Conv	*2D* Depthwise Conv + (2+1)D Conv
Separation of Tasks	Spatial and temporal learning are split (CNN + LSTM)	Spatial and temporal features are learned jointly	Spatial and channel-wise operations are explicitly separated
Temporal Modeling	**Sequence-aware via LSTM memory**	**Temporal filters capture short-term patterns only**	**Captures short temporal patterns depending on the frames.**
Compute Efficiency	**Slower (sequential LSTM updates)**	**Faster (fully convolutional and parallelizable)**	**Faster and more efficient than 3d convolution**
Model Flexibility	Works with variable-length inputs	Fixed-length input window (clip size matters)	Fixed-length input window
Interpretability	Easier to debug (modular structure)	Harder to disentangle spatial vs temporal learning	More interpretable than 3D CNN but harder to debug.
Best Used When	**Temporal dependencies are long or subtle**	**Motion is fast and localized (e.g., actions in short clips)**	**When channel matters, this approach generalize better**

2D CNN + LSTM Model Pipeline

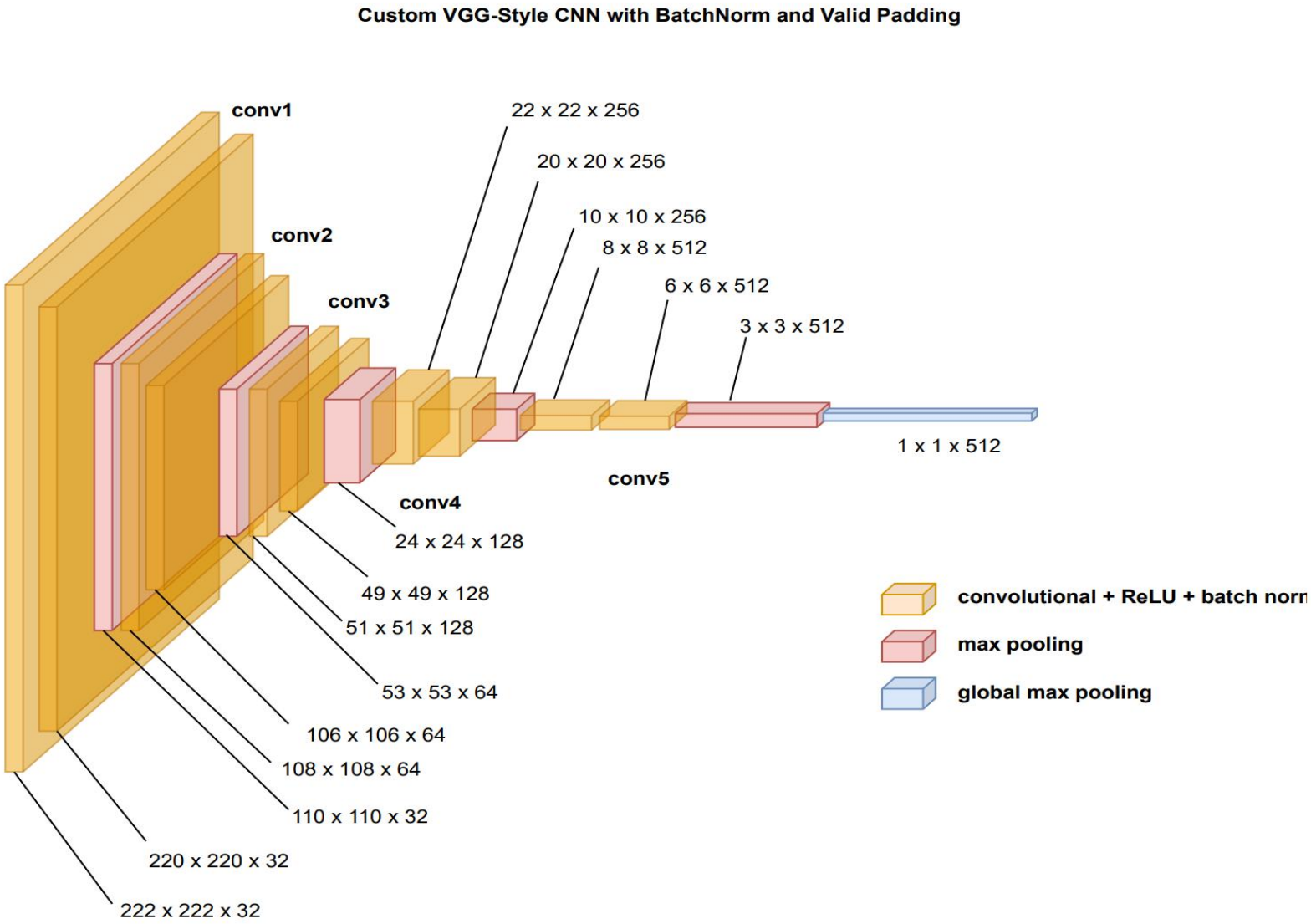


Video Processing Layers (Sampling & Batching Frames)

Frame Sequence processing layers (2D Conv & LSTM)

Fully Connected Layers (Classification Head)

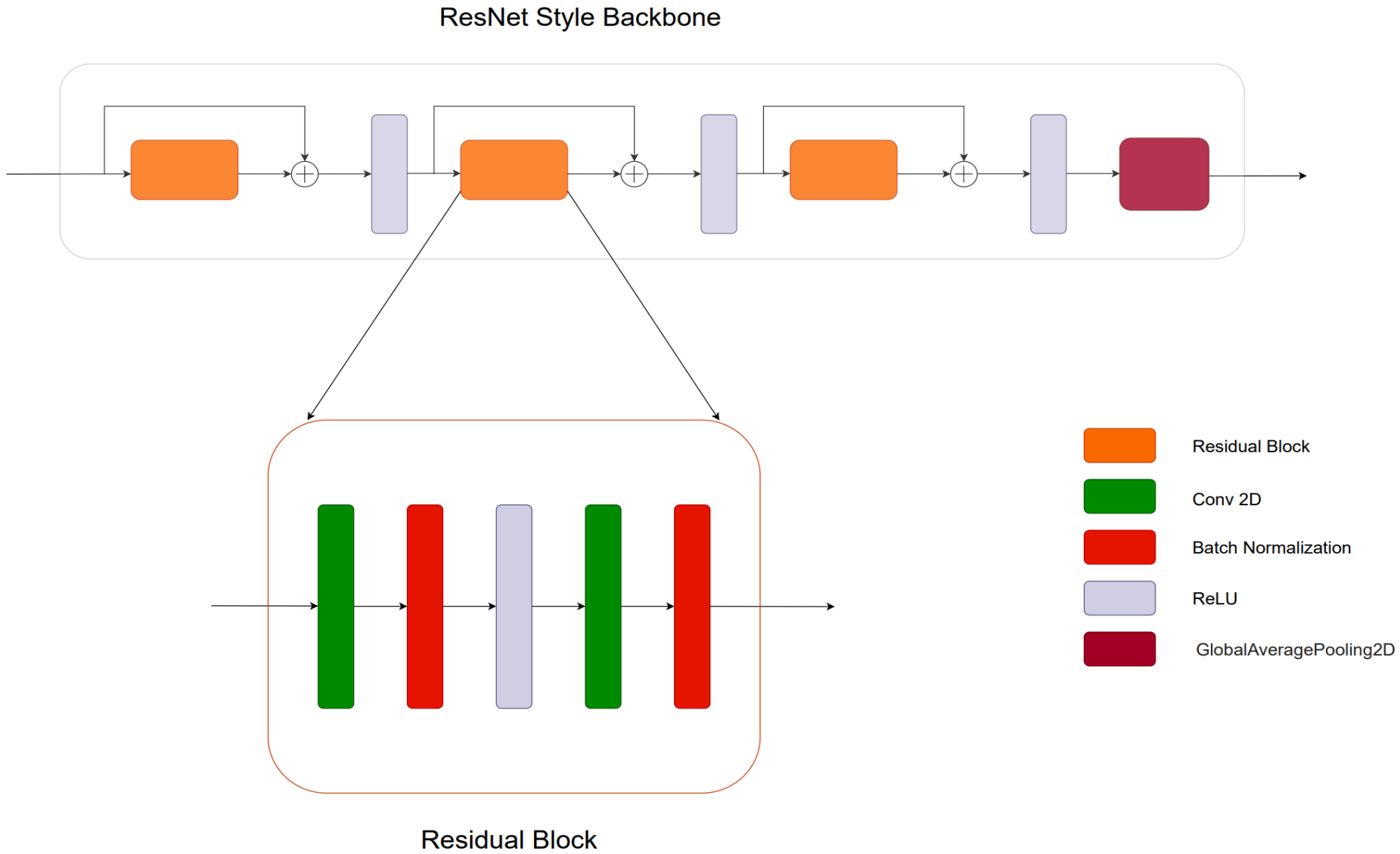
Backbone Models – (Custom VGG-Style)



Continued...

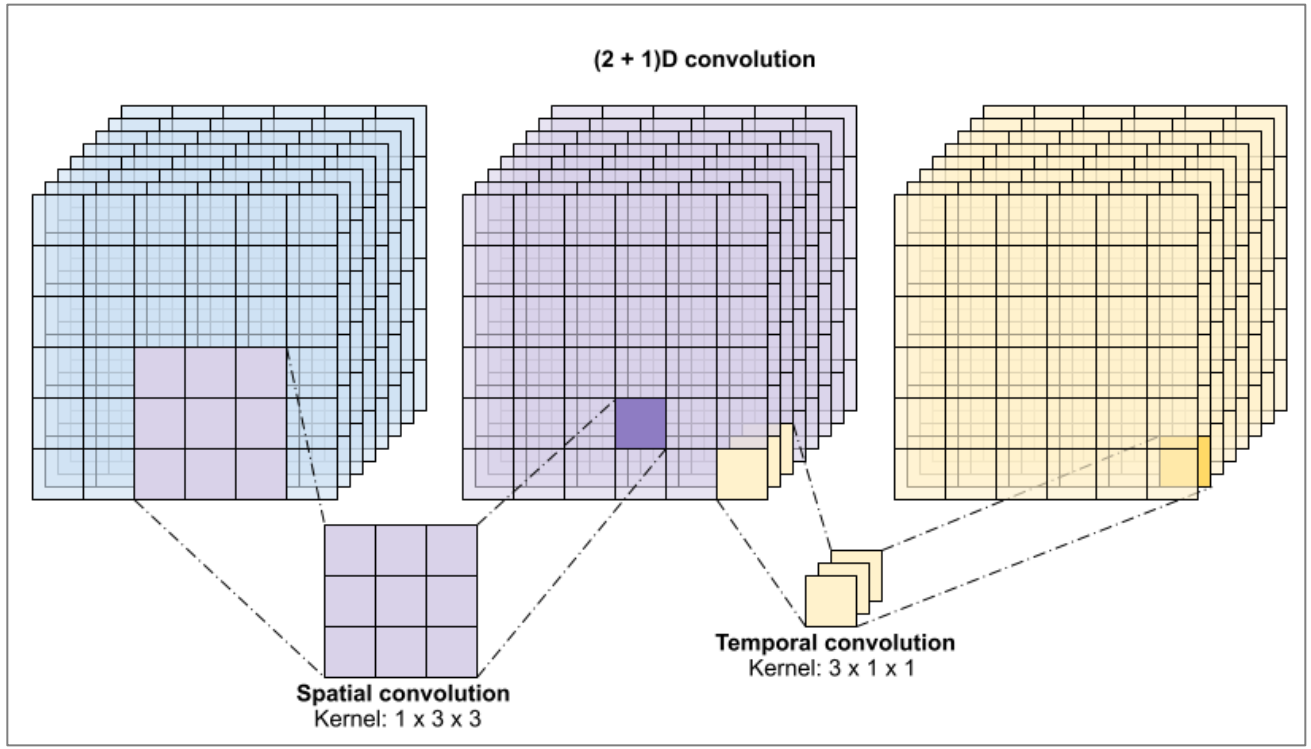
BLOCK	CONV LAYERS	OUTPUT SHAPE AFTER MAXPOOLING
Conv-1	Conv2D(32) → Conv2D(32)	110 × 110 × 32
Conv-2	Conv2D(64) → Conv2D(64)	53 × 53 × 64
Conv-3	Conv2D(128) → Conv2D(128)	24 × 24 × 128
Conv-4	Conv2D(256) → Conv2D(256)	10 × 10 × 256
Conv-5	Conv2D(512) → Conv2D(512)	3 × 3 × 512
GAP	GlobalAveragePooling2D	512

ResNet Style Backbone



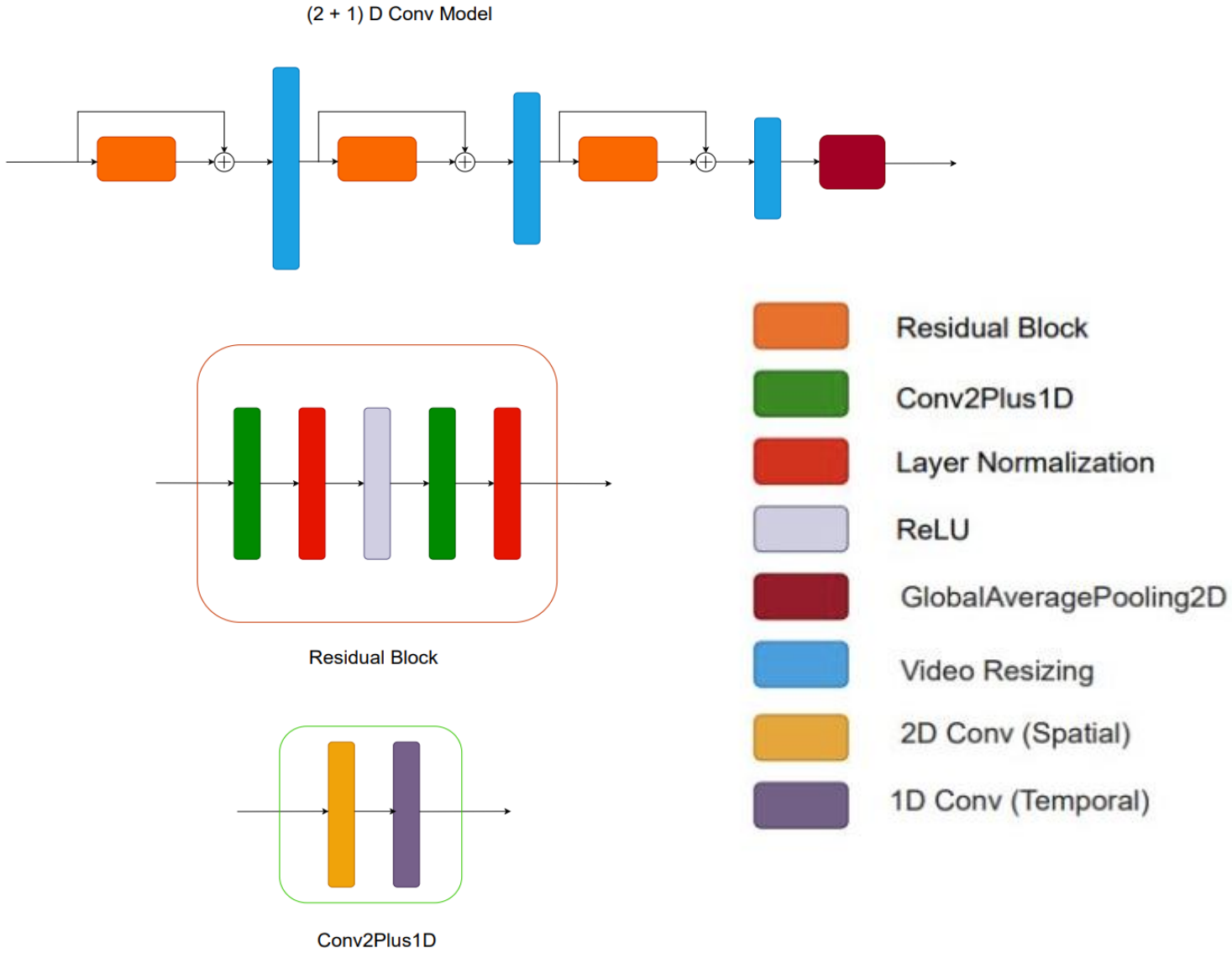
(2 + 1) D Convolution *

- Factorises 3-D conv → 2-D (HxW) + 1-D (T)
- Fewer parameters & FLOPs than full 3-D conv
- Keeps temporal modelling, improves training speed

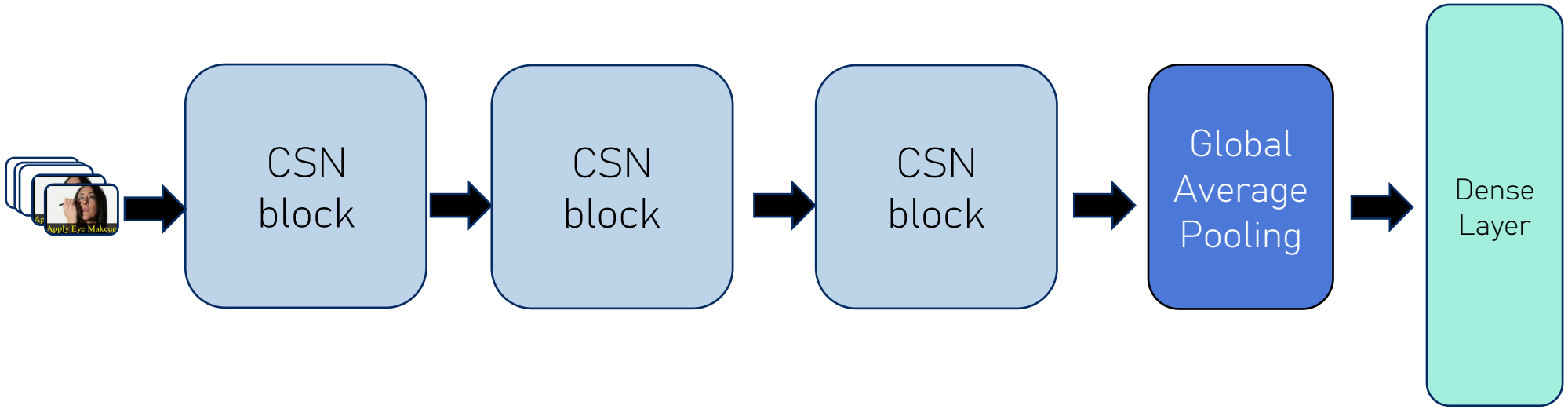


* Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition, 2018

(2 + 1) D Conv + ResNet Model



CSN Model Pipeline*



5d tensor

(batch size,
frames,
height,
width,
channels)

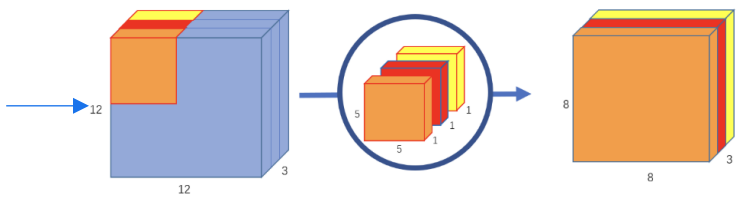
*: Tran, D., Wang, H., Torresani, L., & Feiszli, M. (2019). *Video classification with channel-separated convolutional networks* (arXiv:1904.02811). arXiv. https://doi.org/10.48550/arXiv.1904.02811

CSN block consists of ...

5d tensor

(batchsize, time, height, width, channel)

2D depthwise spatial convolution (3*3 per frame)



Batch Norm

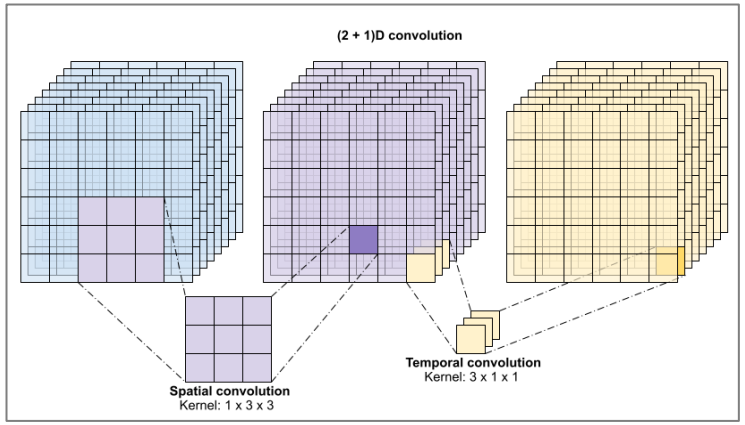
ReLU

5d tensor

(batchsize, time, height, width, channel)

ReLU

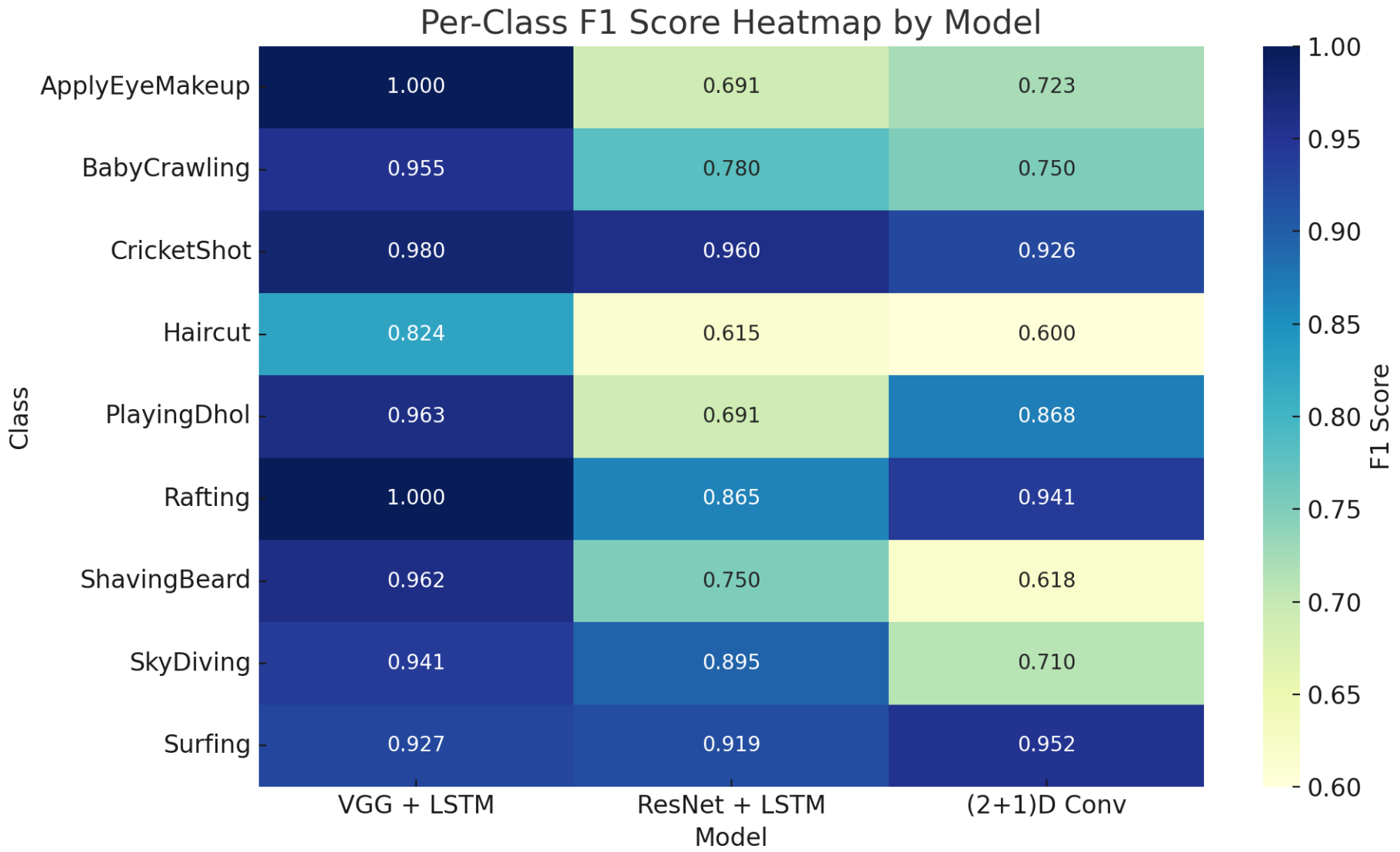
Batch Norm



Results & Trade - Offs

Aspect	2D CNN + LSTM (VGG)	2D CNN + LSTM (ResNet)	(2+1)D Conv	*2D* Depthwise Conv + (2+1)D Conv
Test Accuracy	**90 %**	77%	79%	18%
Parameters	5.5 M	0.77 M	**0.44 M**	0.5M
Size	21 M	2.95 MB	**1.69 MB**	**0.15 MB**
Training Time	0.5 hrs	2 hrs	**0.5 hrs**	3 hrs

Model Performance by Class

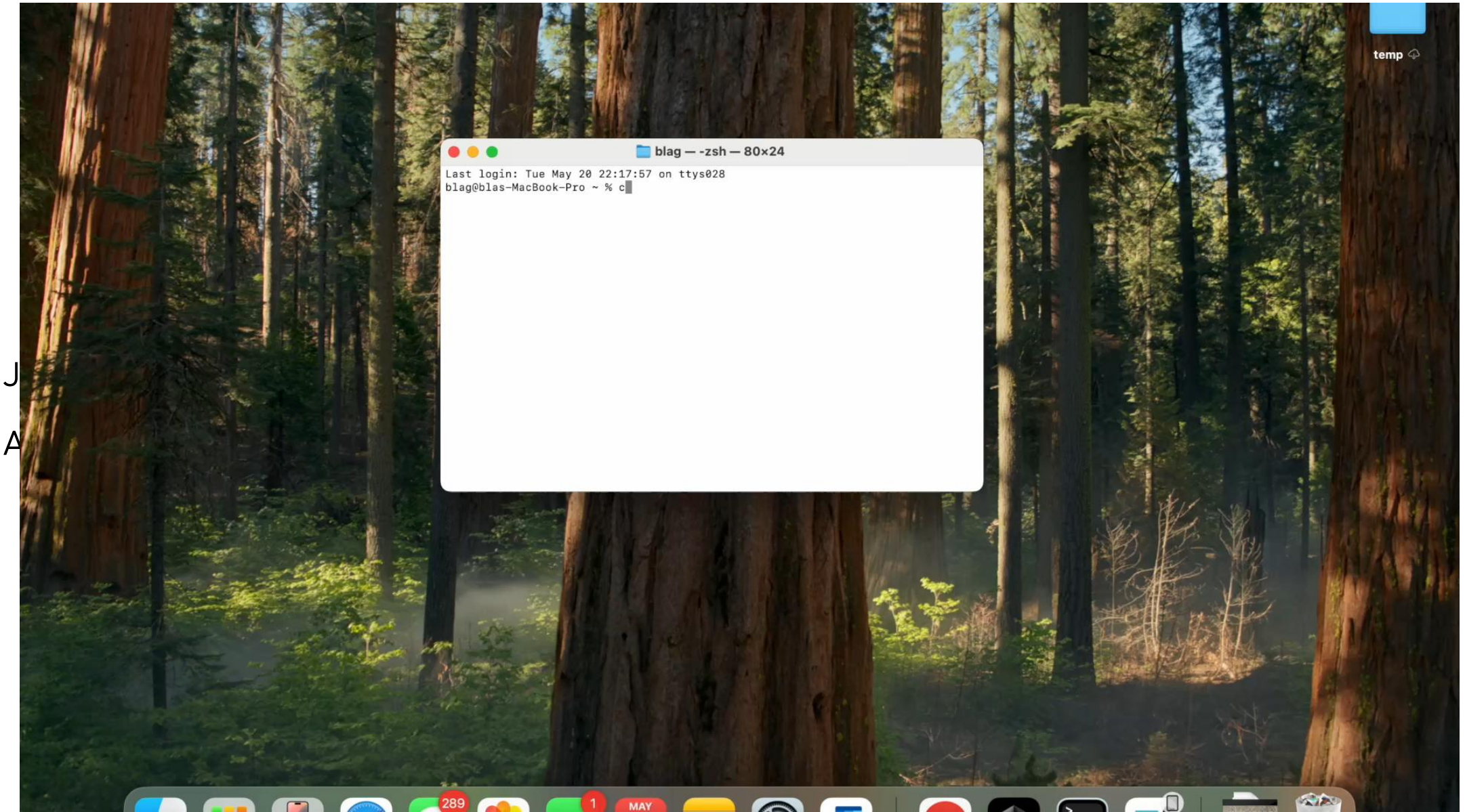


STYLE TRASFER

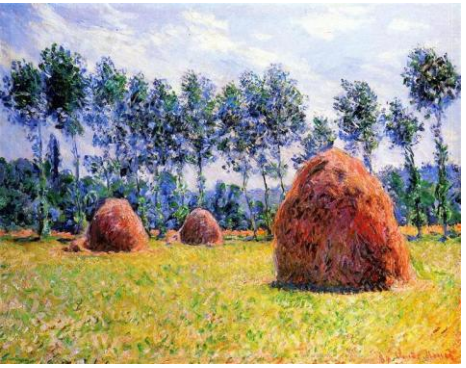
Application to Style Transfer

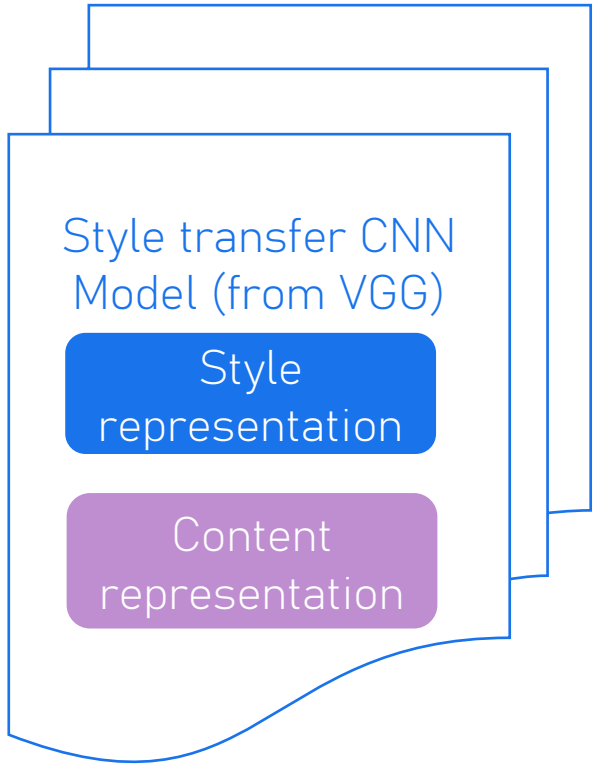
J

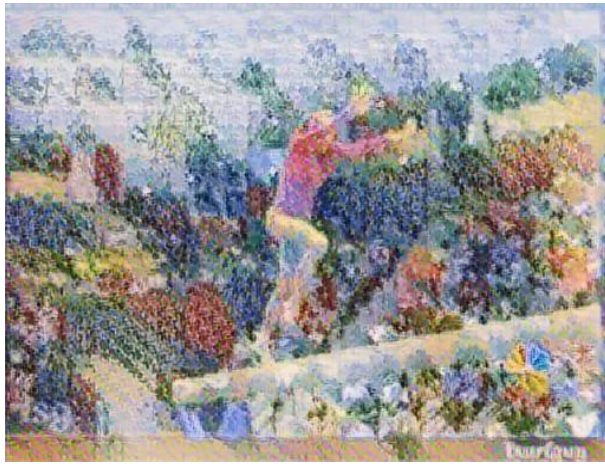
A



Steps*







*: *Neural style transfer | TensorFlow Core*. (n.d.). TensorFlow. Retrieved May 22, 2025, from https://www.tensorflow.org/tutorials/generative/style_transfer

CycleGAN Steps

CycleGAN pre-trained ukiyo-e Generator

CycleGAN pre-trained model and architecture from https://junyanz.github.io/CycleGAN/

Comparison

Feature	Classic Neural Style Transfer (VGG)	CycleGAN
Goal	Stylize one image with the style of another image	Translate images between two visual domains (e.g., photo ↔ painting)
Input Requirement	One content image + one style image	A dataset of **unpaired** images from two domains (X and Y)
Training	**Not much training (optimization-based, per image)**	**Requires full training of two generators + two discriminators**
Architecture	**Uses pretrained VGG (e.g., VGG-19) for feature extraction**	**Custom CNNs for generators and discriminators (GAN-based)**
Loss Functions	- Content loss (VGG) - Style loss (Gram matrix) - Total variation loss	- Adversarial loss - Cycle-consistency loss - Identity loss (optional)
Output	Stylized image combining input content and style	Domain-translated image (e.g., photo styled like a Monet)
Speed	Fast because doesn't require too much training	Slow in training, but fast after training, inference is real-time
Generalization	**Can apply arbitrary styles, but might need to fine tune**	**Learns to map between two domains; cannot do arbitrary styles without retraining**
Style Control	Direct control via style image	Indirect control via domain training

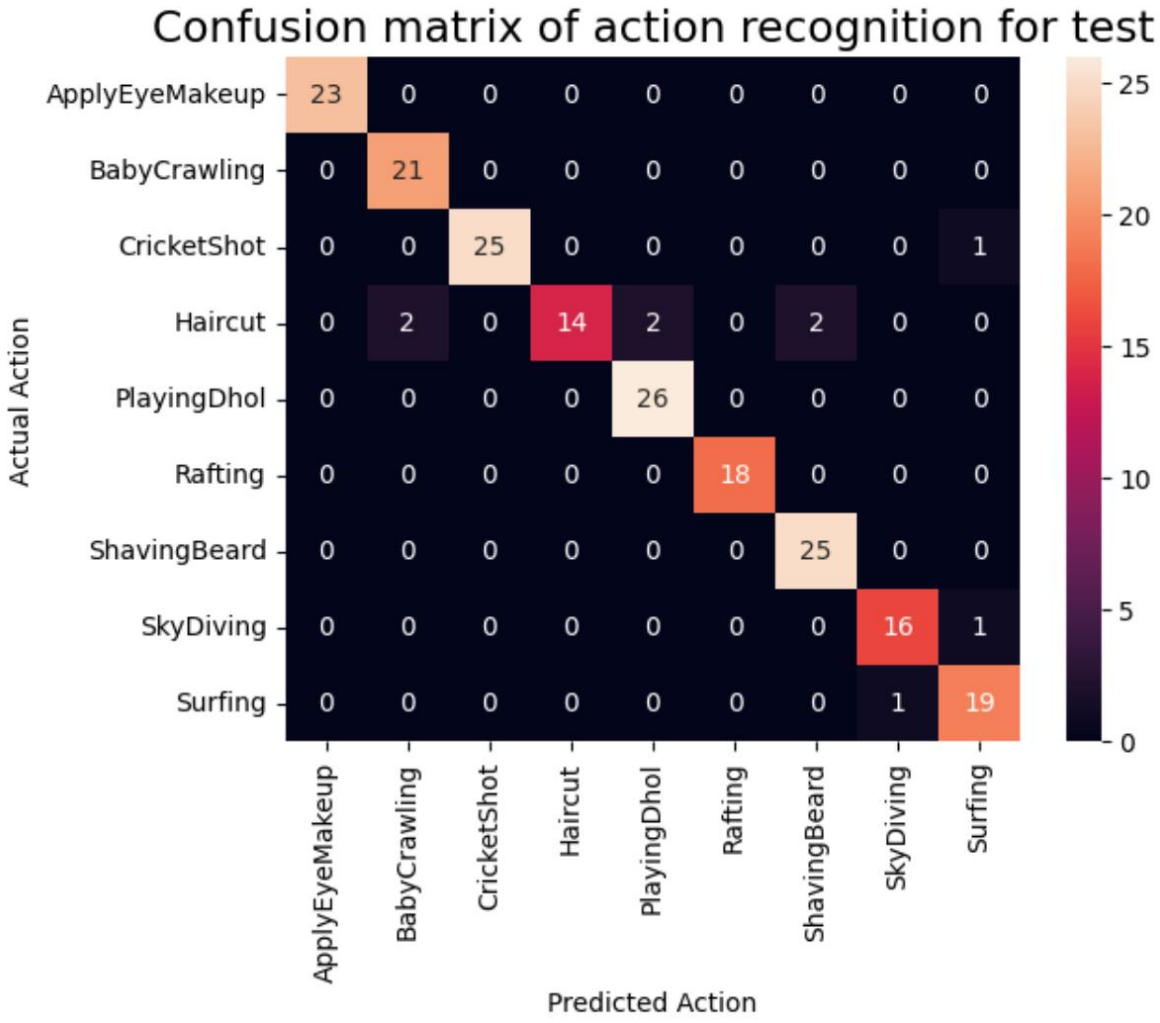
Conclusion & Future Scope

Potential directions include:

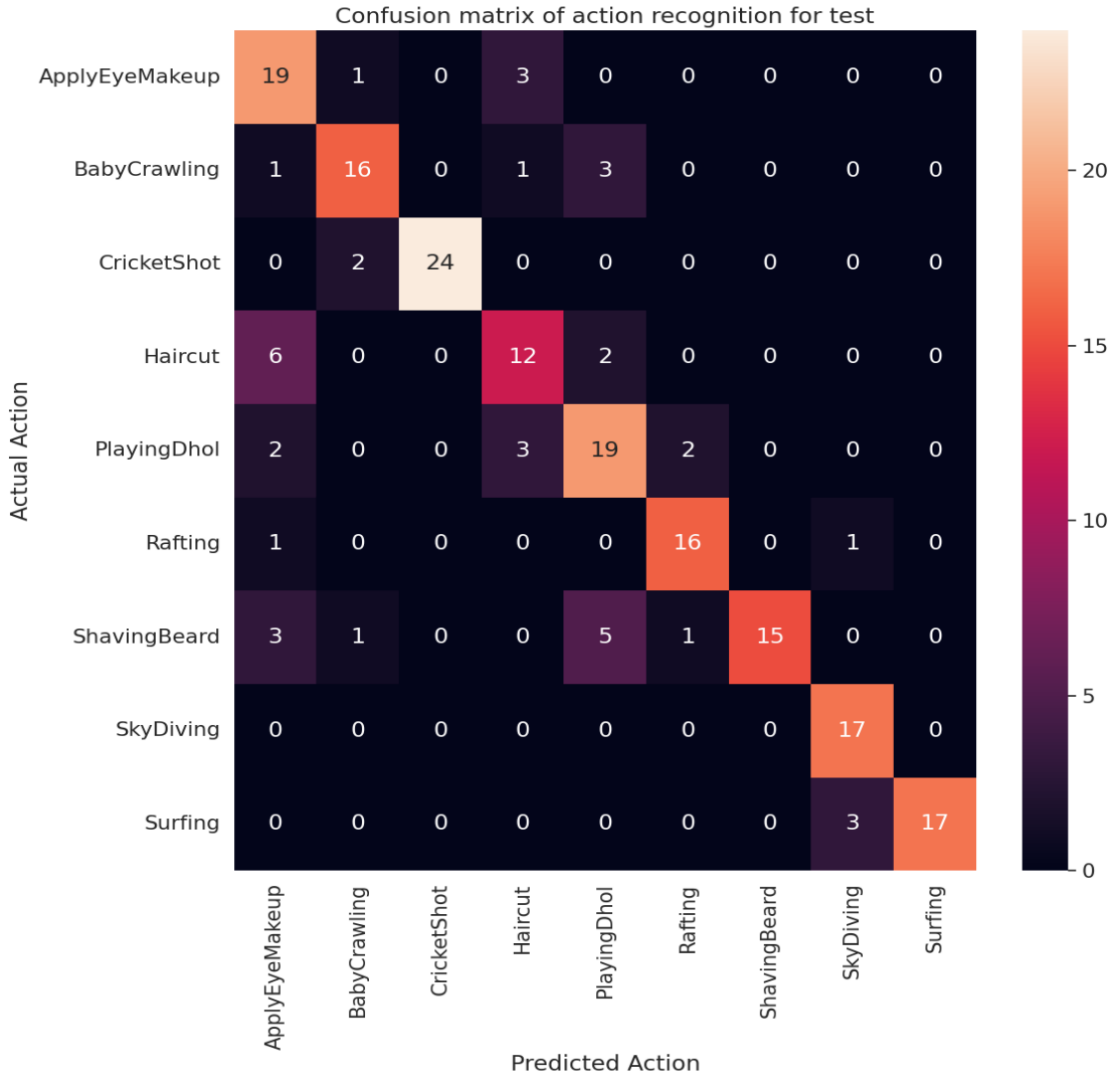
- Training with the full UCF101 dataset to improve generalization,
- Testing on fine-grained or ambiguous action categories,
- Real-time inference on mobile devices,
- And extending the classification framework to more creative applications such as video segmentation or lip reading.

Appendix

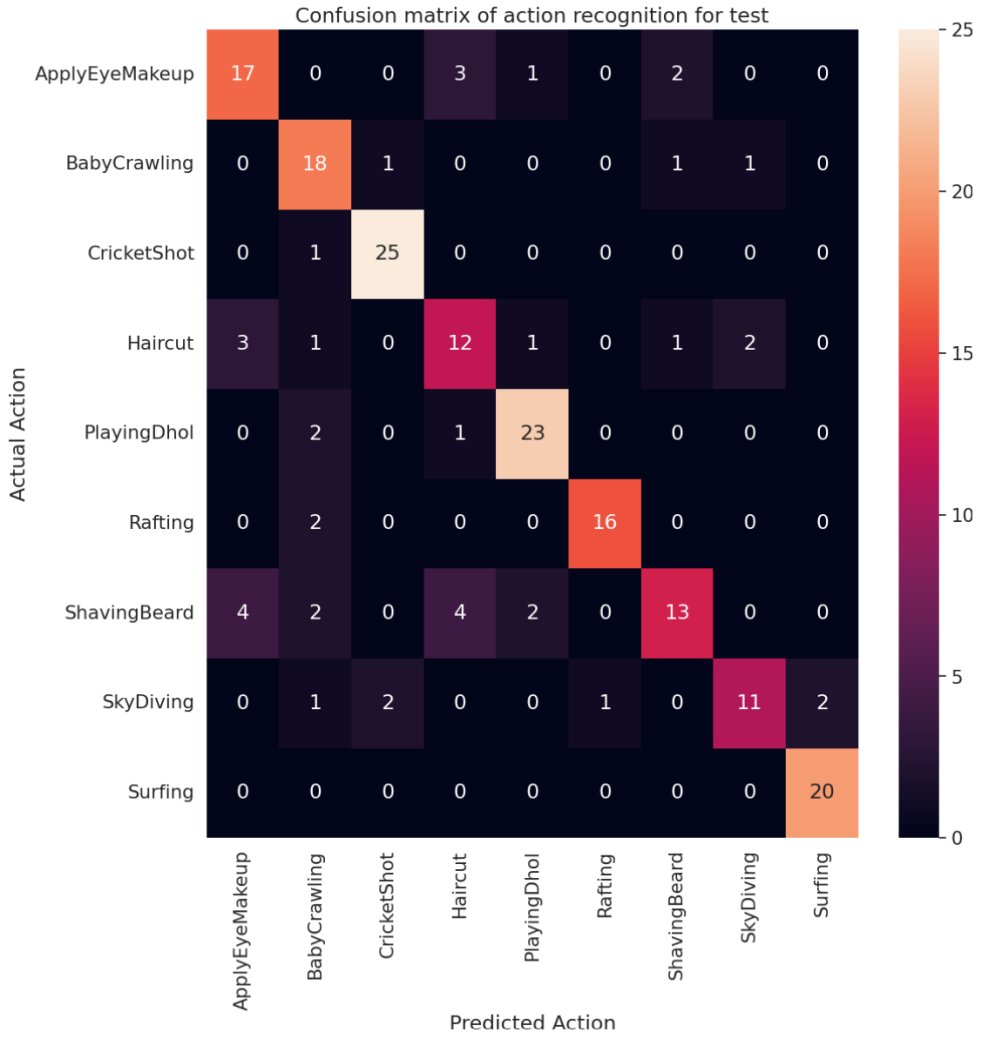
Confusion Matrix 2D Conv + LSTM (VGG Style)



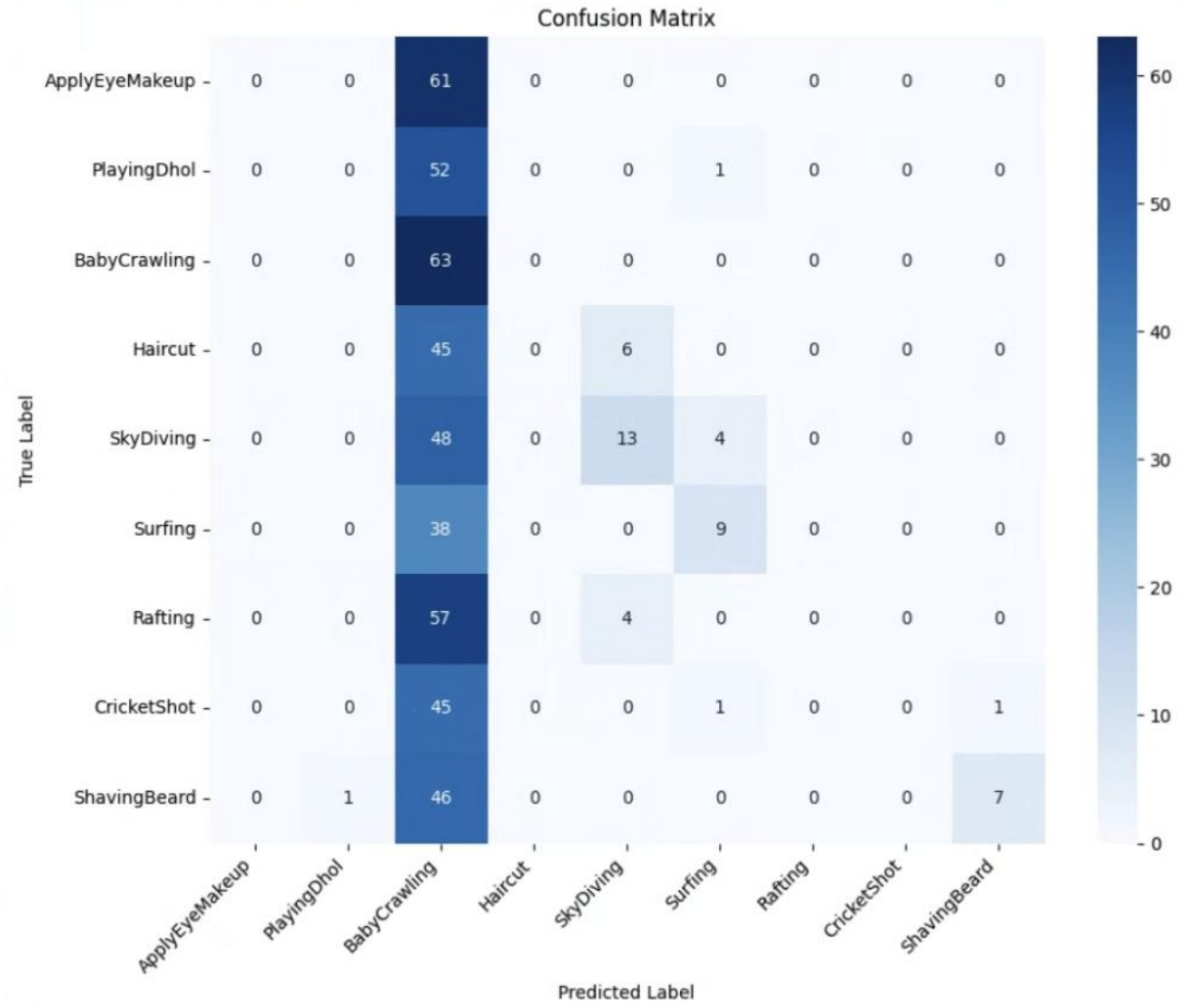
Confusion Matrix 2D Conv + LSTM (ResNet Style)



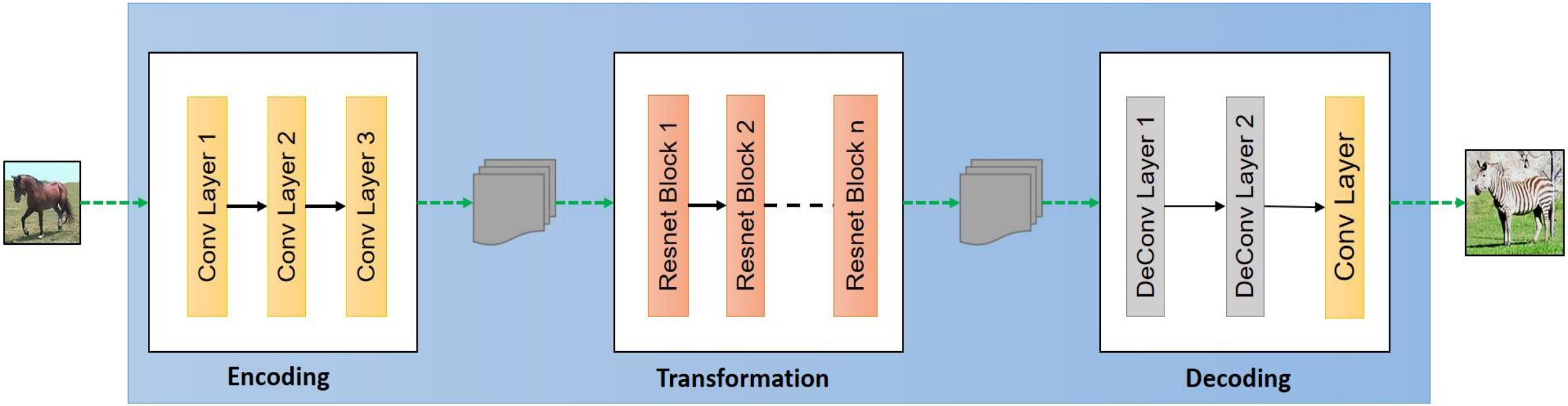
Confusion Matrix (2 + 1)D Conv



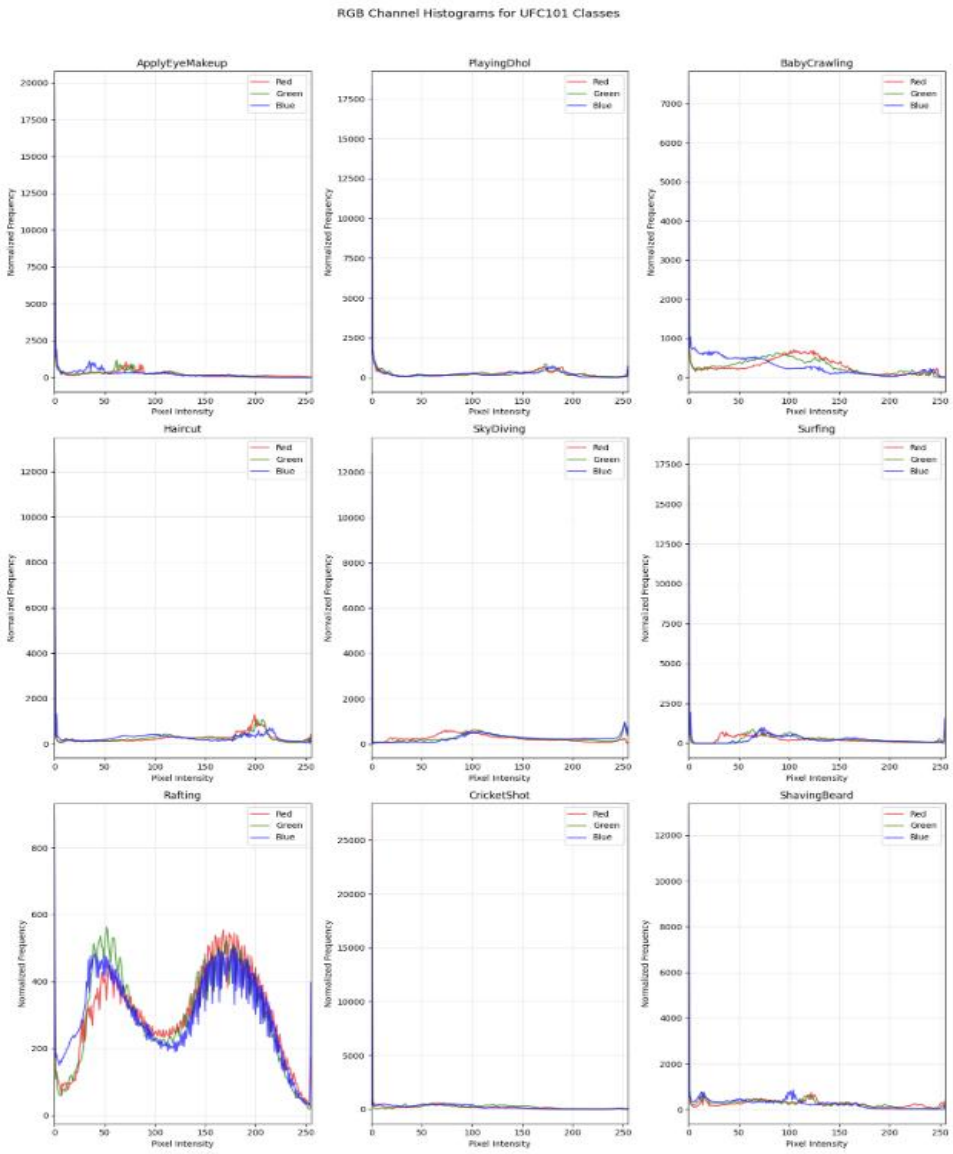
Confusion matrix for CSN



CycleGAN generator in detail



EDA: RGB Histogram for all 9 classes



CycleGAN web application