الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

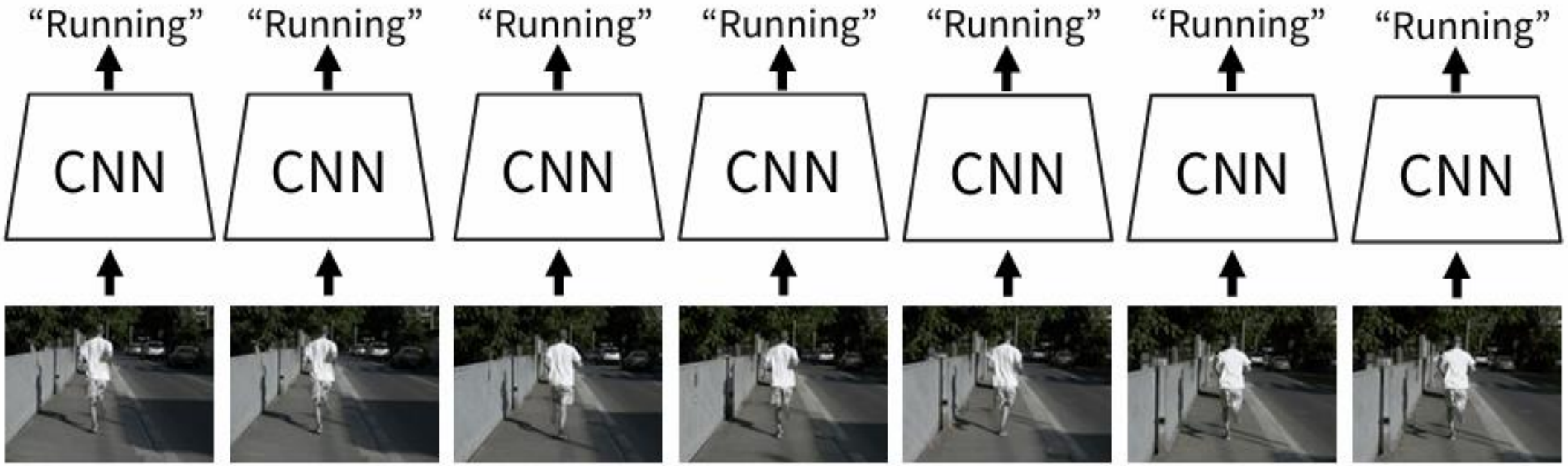
Week 11	كلية الهندسة	الذكاء الصنعي العملي

Video Classification

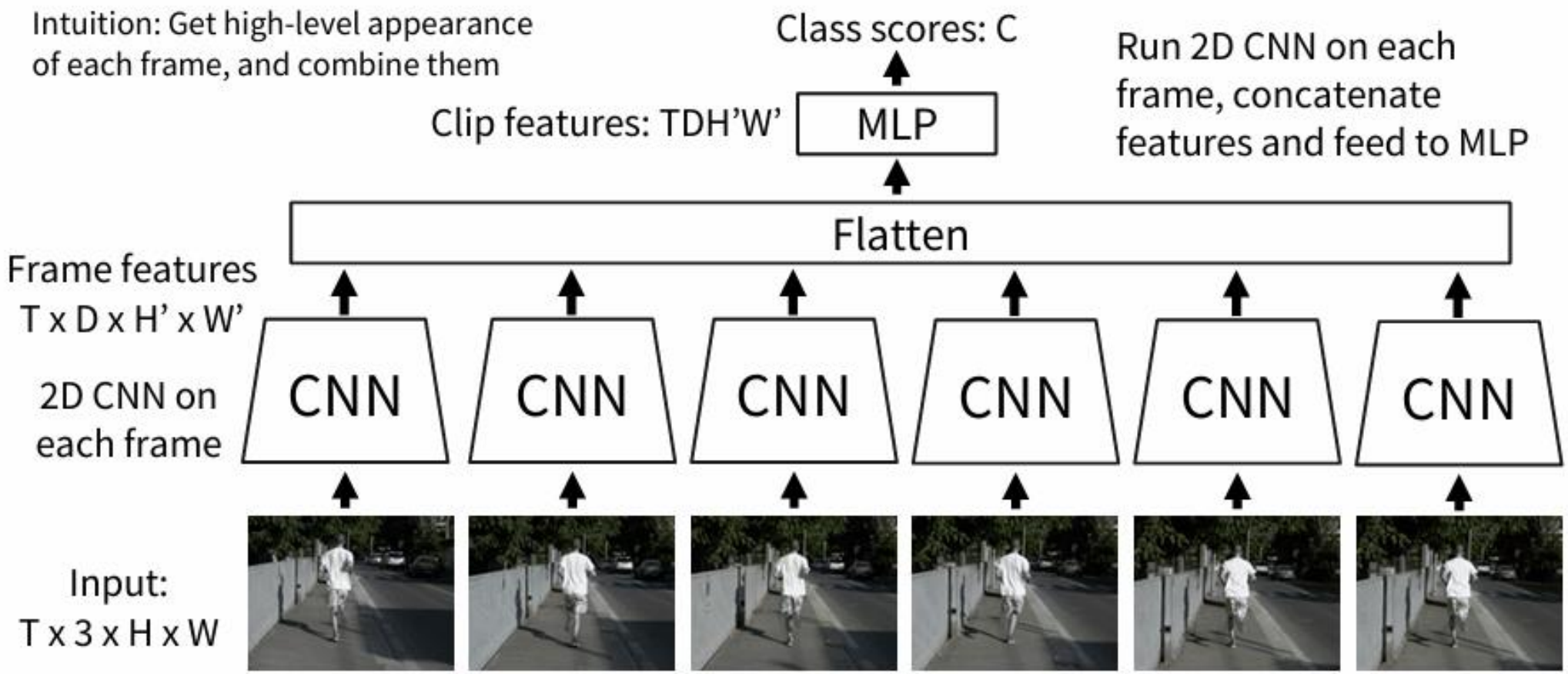
د. رياض سنبل

Single-Frame CNN

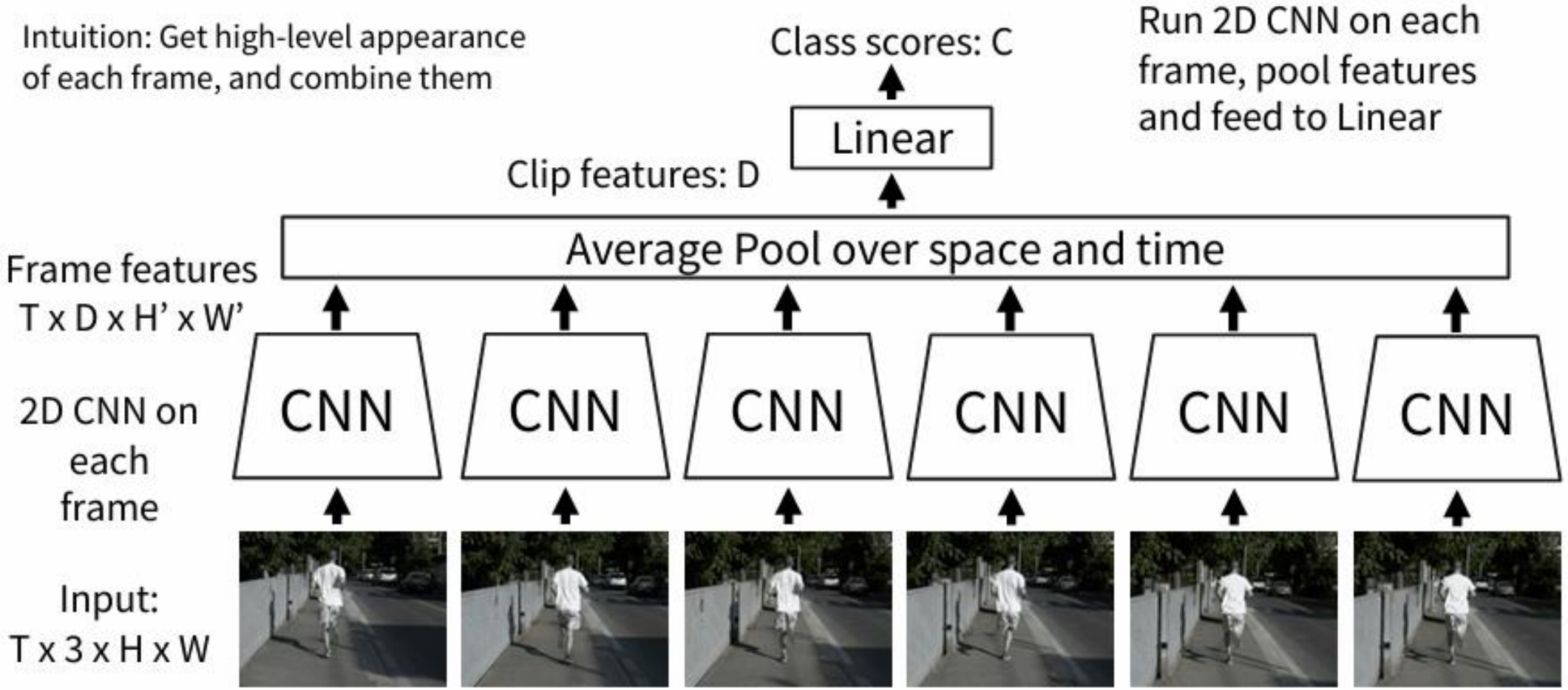
- Simple idea: train normal 2D CNN to classify video frames independently!
- (Average predicted probs at test-time)
- Often a very strong baseline for video classification



Late Fusion (with FC layers)

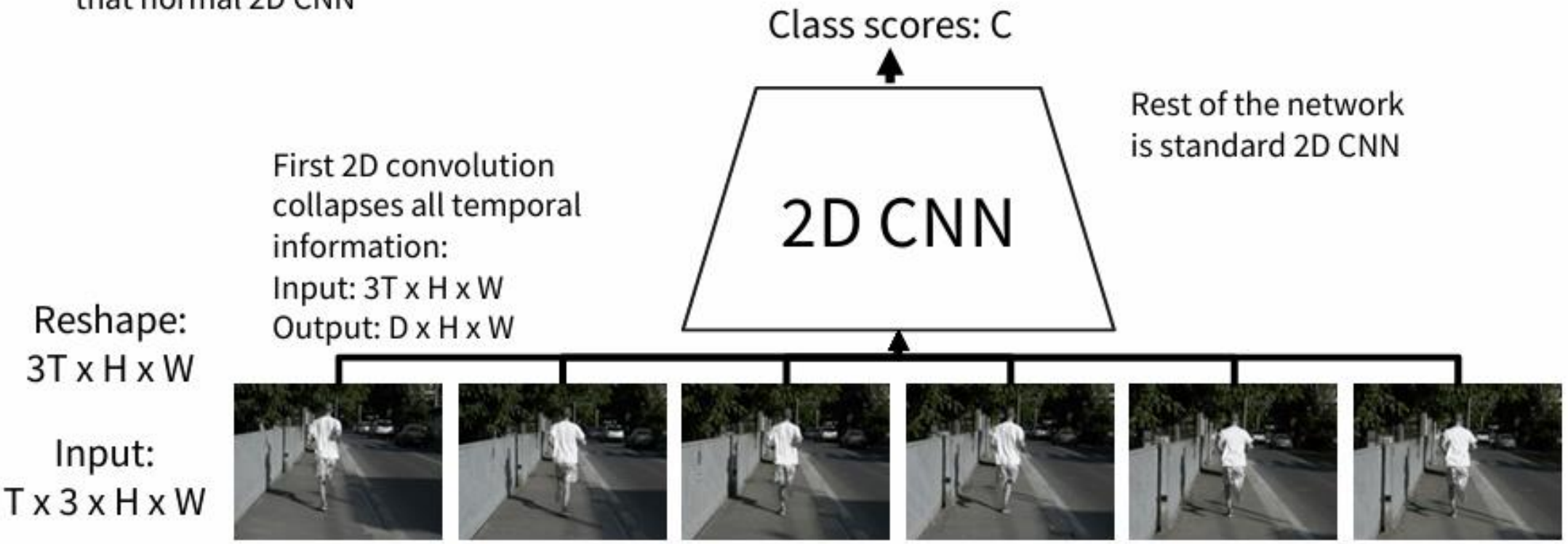


Late Fusion (with pooling)



Early Fusion

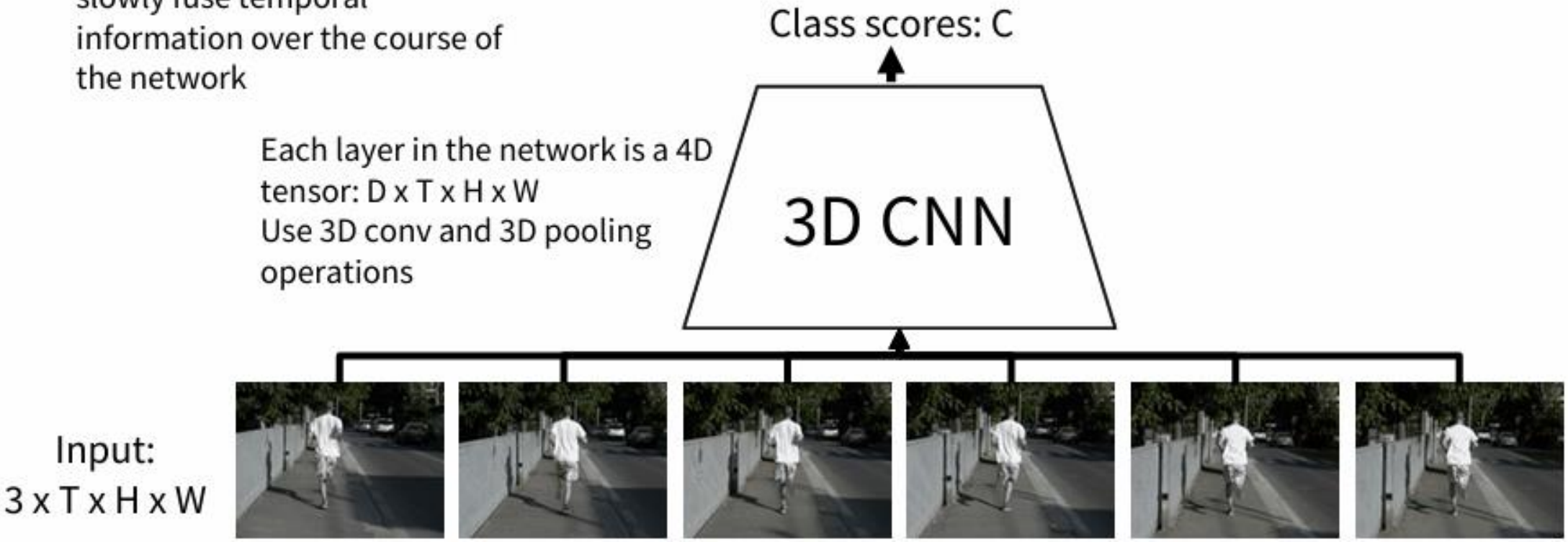
Intuition: Compare frames with very first conv layer, after that normal 2D CNN



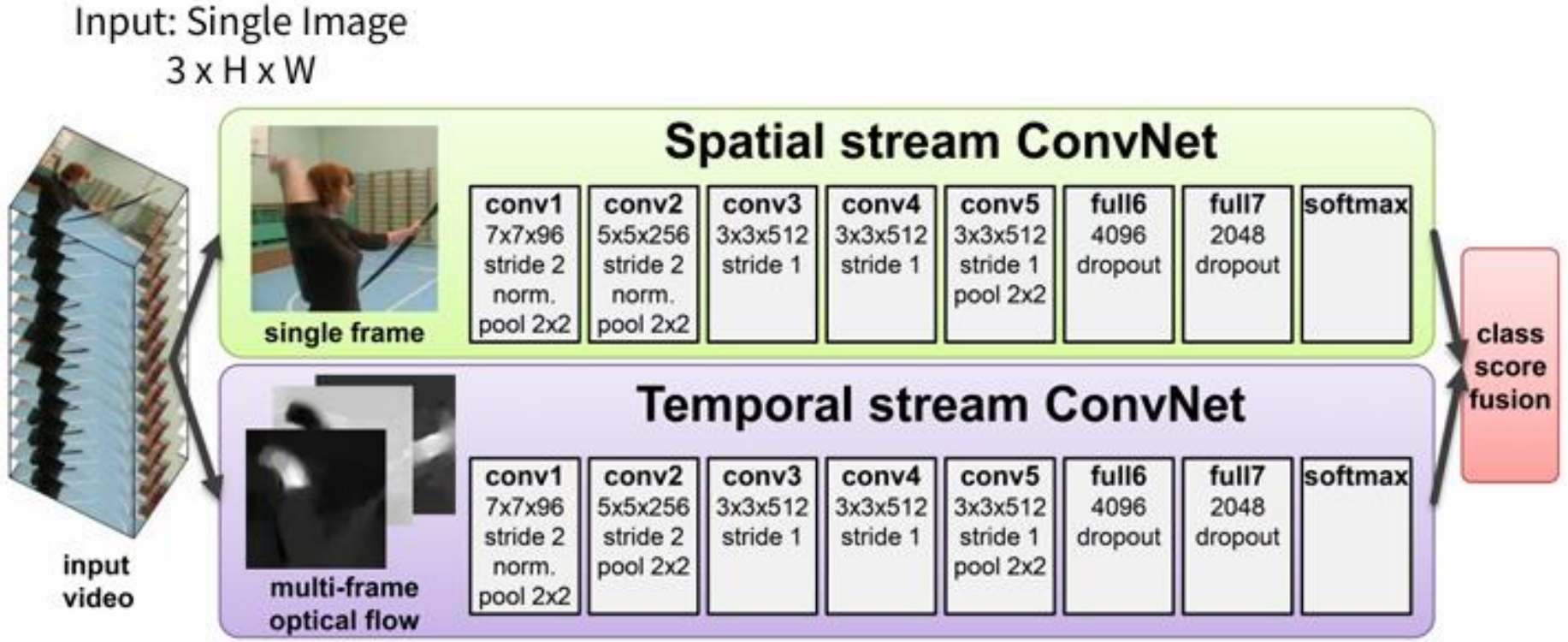
3D CNN

Intuition: Use 3D versions of convolution and pooling to slowly fuse temporal information over the course of the network

Each layer in the network is a 4D tensor: D x T x H x W
Use 3D conv and 3D pooling operations

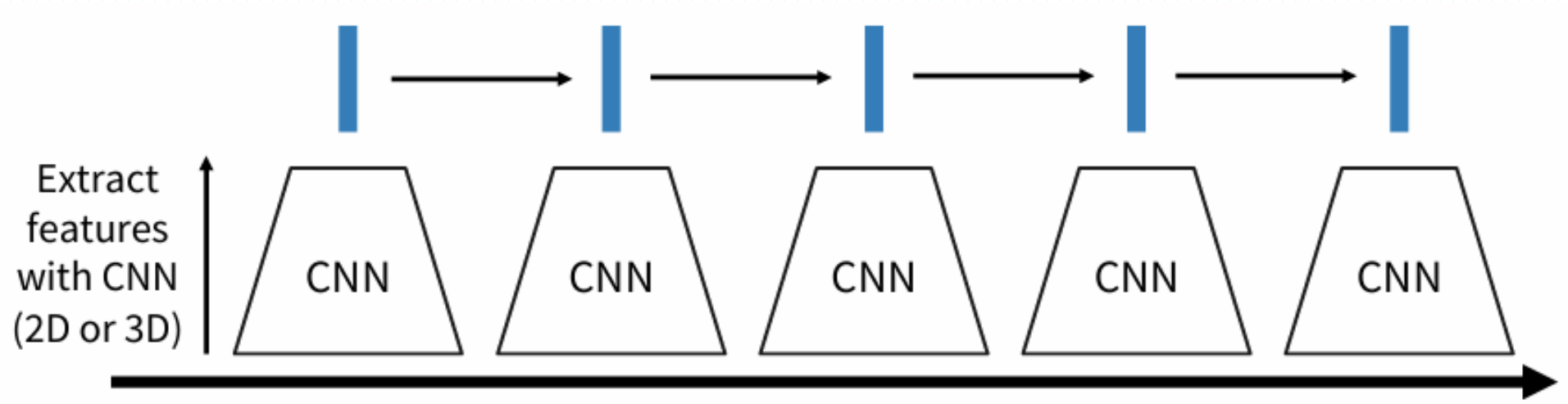


Separating Motion and Appearance: Two-Stream Networks



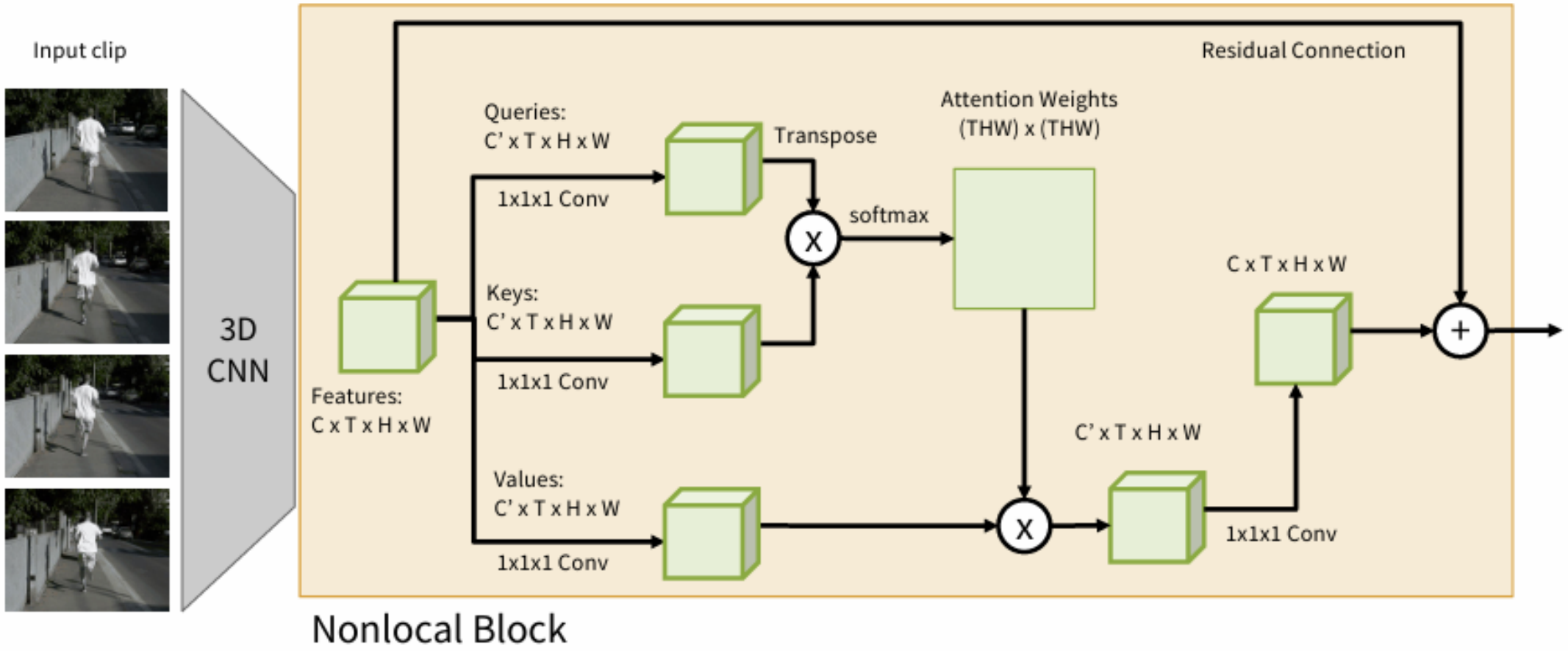
Modeling long-term temporal structure

- Inside CNN: Each value is a function of a ***fixed temporal window*** (local temporal structure) Inside RNN: Each vector is a function of all previous vectors (global temporal structure) Can we merge both approaches?



Baccouche et al, "Sequential Deep Learning for Human Action Recognition", 2011
Donahue et al, "Long-term recurrent convolutional networks for visual recognition and description", CVPR 2015

Spatio-Temporal Self-Attention (Nonlocal Block)



Inflating 2D Networks to 3D (I3D)