ICPR2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

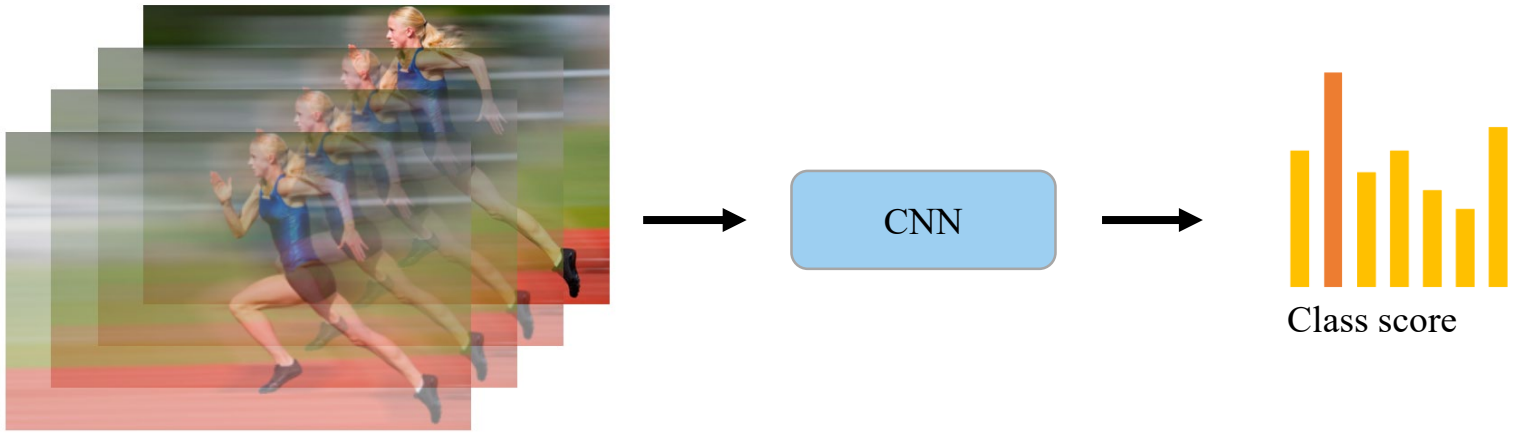
# Motion Complementary Network for Efficient Action Recognition

Ke Cheng (Presenter), Yifan Zhang, Chenghua Li, Jian Cheng, Hanqing Lu



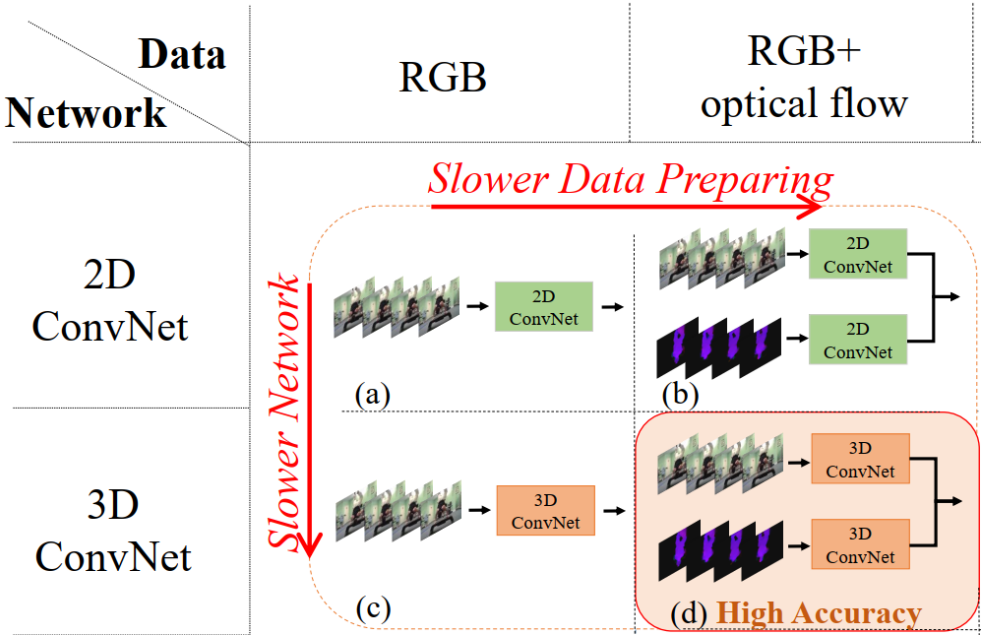
# Preliminaries

- Video-based action recognition is one of the most popular research fields in computer vision.
- The key difference between action recognition and image recognition is *temporal sequence modeling*.



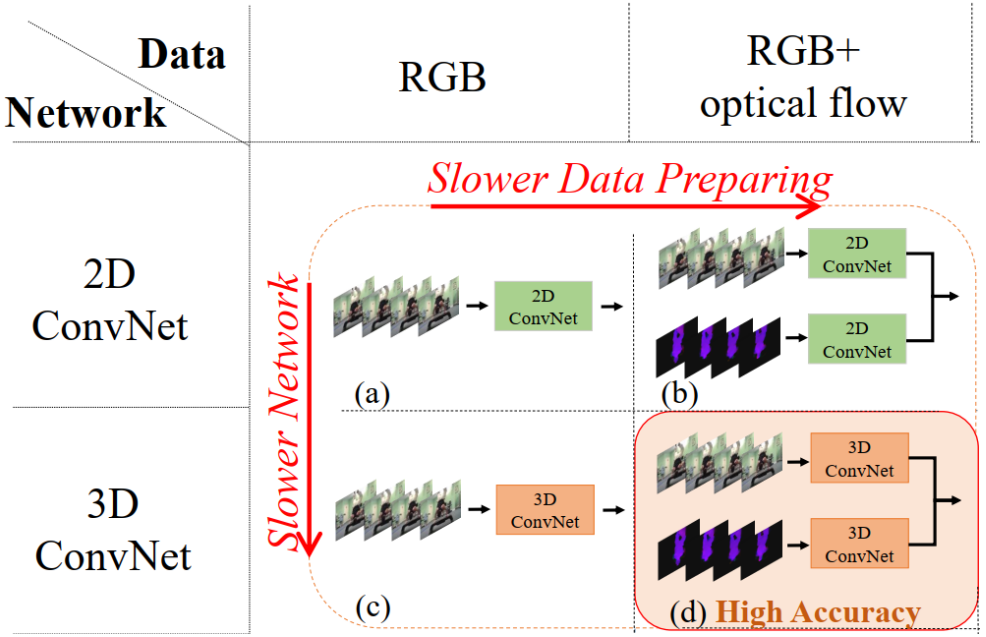
# Motivation

- There are two main methods for *temporal sequence modeling*: two-stream ConvNet and 3D ConvNet.
- Notably, combining two-stream strategy and 3D convolution strategy achieves obviously better performance than any single strategy.
- This phenomenon shows that these two methods are complementary to each other.



# Motivation

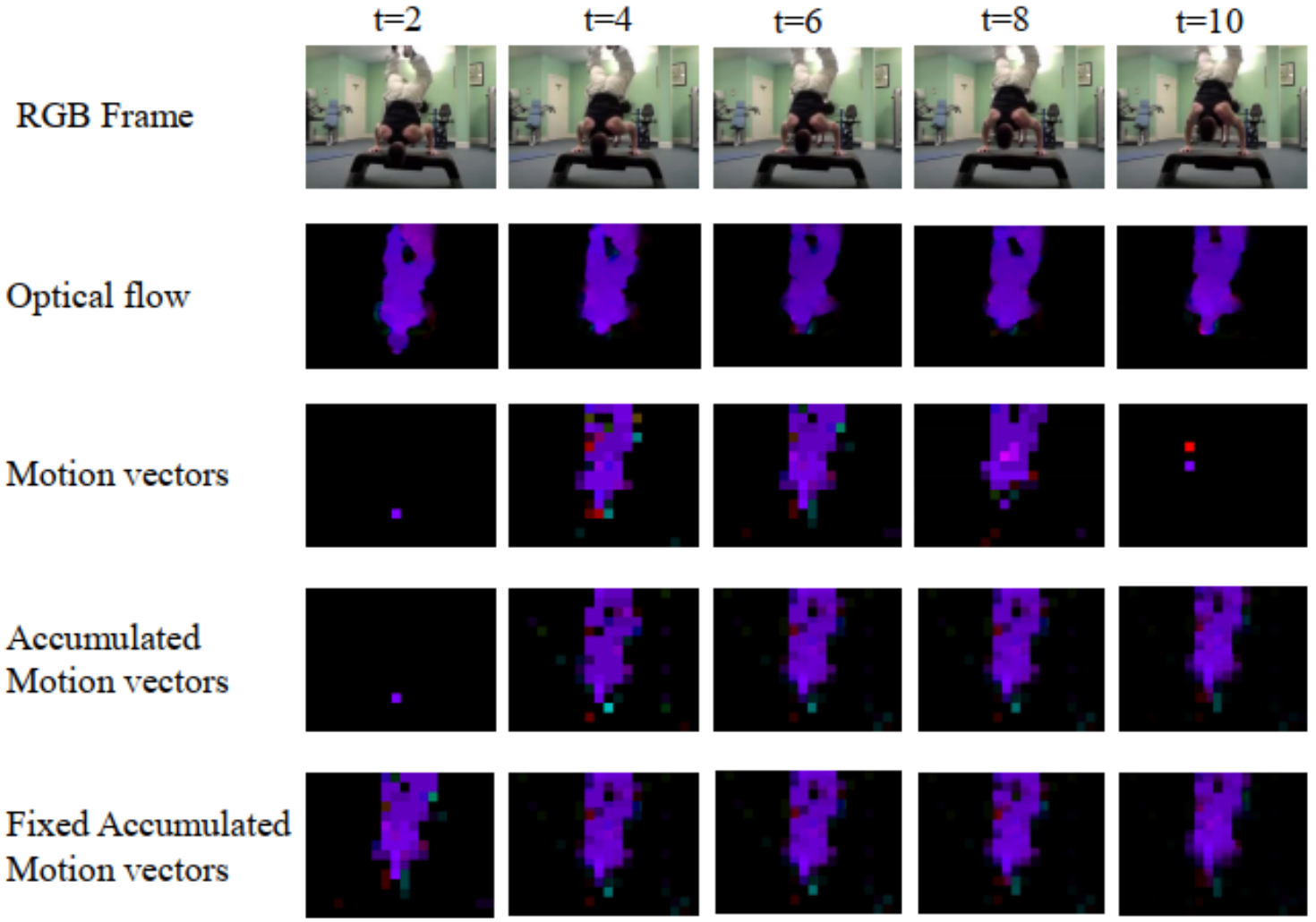
- However, both of these two methods are computationally expensive:
  - Computing optical flow is slow, which is not suitable for real-time application;
  - 3D convolution with a kernel size of $k \times k \times k$ need $k^3$ multiply-adds, which is much larger than 2D convolution with $k \times k$ kernel.
- Therefore, although 3D two-stream ConvNet achieves high accuracy, it is not efficient for deployment.



# Method

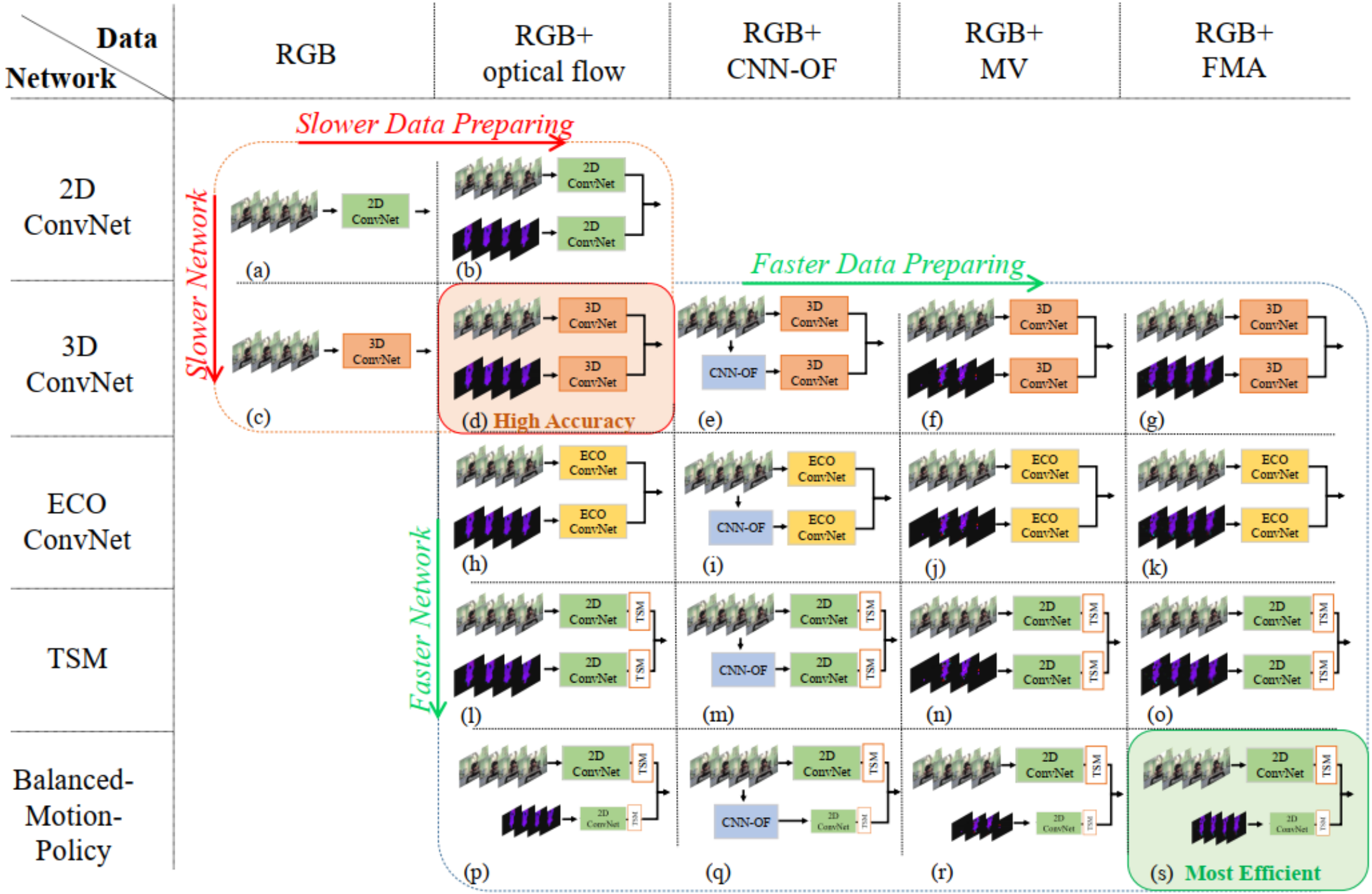
- To construct an efficient action recognition model, we analyze the time cost in action recognition and divide it into the data preparing time (Data Time) and the network computation time (Network Time).
- To accelerate the data preparing process, we need to:
  - sample fewer frames;
  - speed up the optical flow calculation.
- Thus, in our paper:
  - only sample 5-8 frames from an action;
  - use motion vector as an alternative optical flow.

# Method


t=2
t=4
t=6
t=8
t=10
RGB Frame
Optical flow
Motion vectors
Accumulated Motion vectors
Fixed Accumulated Motion vectors

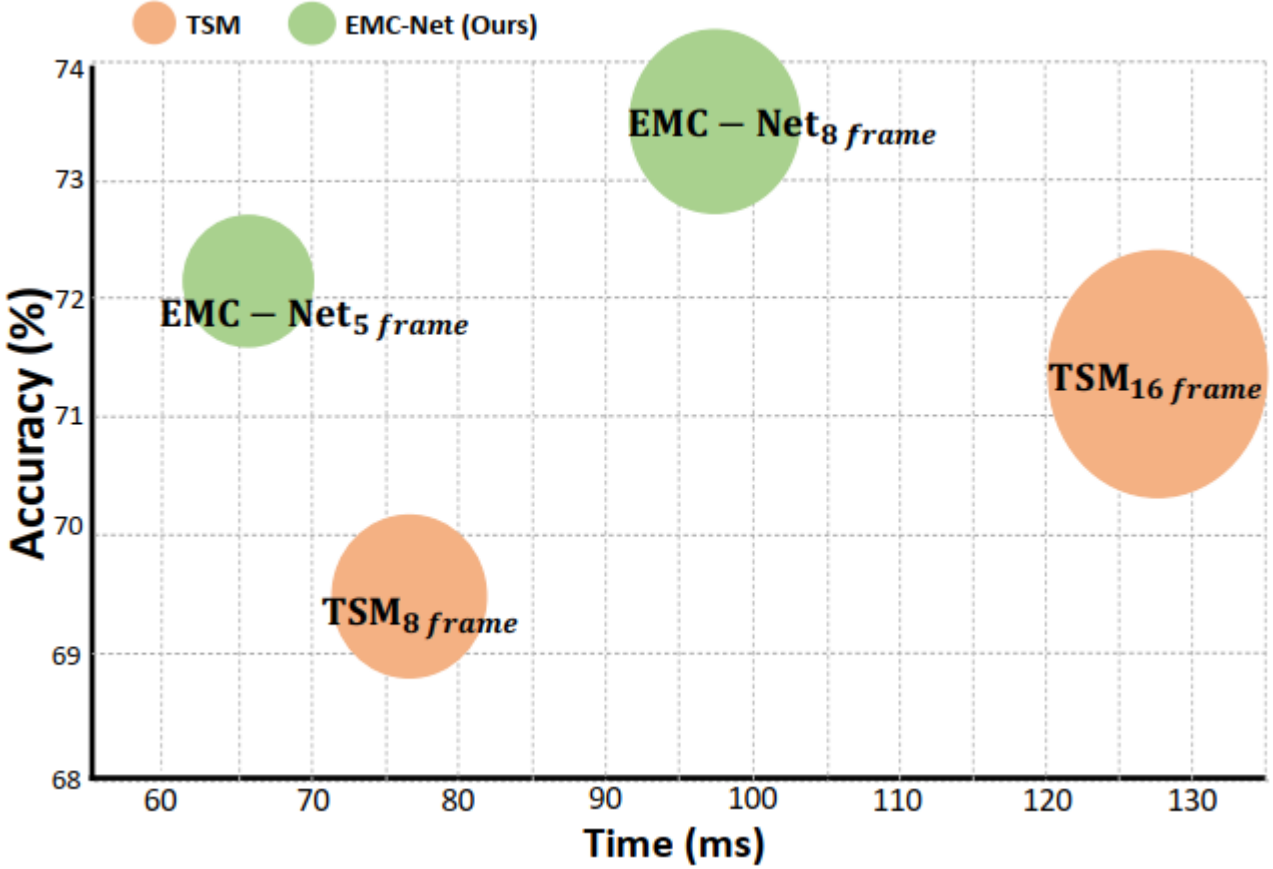
# Method

- The network time (Network Time) is determined by network structure and input spatial resolution.
- In our paper, both RGB pathway and MV pathway adopt TSM (Lin et al., 2019) network as backbone.
- If we simply use the same network on both pathways, the total computational cost will double.
- Our key insight is that motion vectors contain macro block ranging from $8 \times 8$ pixels to $16 \times 16$ pixels and only describe block-level motion information.
- Therefore, we use a larger resolution on RGB pathway and a smaller resolution on MV pathway to achieve a trade-off between accuracy and efficiency. We call it balanced-motion-policy (BMP).


Data
Network
RGB
RGB+ optical flow
RGB+ CNN-OF
RGB+ MV
RGB+ FMA
2D ConvNet
3D ConvNet
ECO ConvNet
TSM
Balanced-Motion-Policy
Slower Data Preparing
Faster Data Preparing
Slower Network
Faster Network
2D ConvNet
3D ConvNet
ECO ConvNet
CNN-OF
TSM
(a)
(b)
(c)
(d) High Accuracy
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)
(r)
(s) Most Efficient

# Result

- We conduct extensive experiments on Kinetics, UCF101, and Jester datasets.
- On Kinetics dataset, we achieve 2.6% better performance than TSM (Lin et al., 2019) with 1.4× fewer FLOPs and 10ms faster on K80 GPU.



# Thanks for your attention!

**Motion Complementary Network for Efficient Action Recognition**