Gate-Shift Networks for Video Action Recognition

Swathikiran Sudhakaran[1] Sergio Escalera[2,3] Oswald Lanz[1]

[1]Fondazione Bruno Kessler, Italy [2]Computer Vision Center, Spain

[3]Universitat de Barcelona, Spain

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Motivation

Video action recognition requires spatio-temporal reasoning

Putting something similar to other things that are already on the table

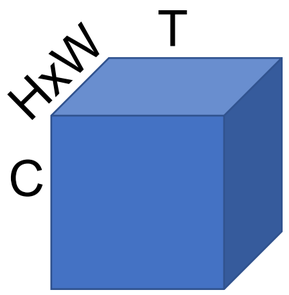
Taking one of many similar things on the table

Contribution

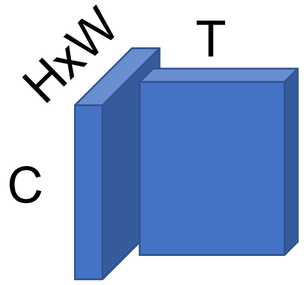
Large number of parameters in 3D CNNs require **large scale annotated data** for training

Existing approaches address this problem by a **hard-wired** decomposition of the 3D kernels which is **suboptimal**

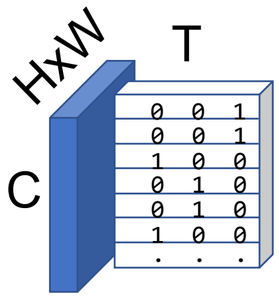
GSM leverages spatial gating for **adaptive feature propagation**



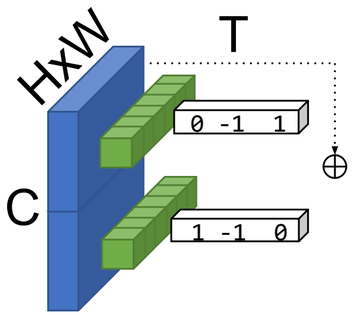
C3D



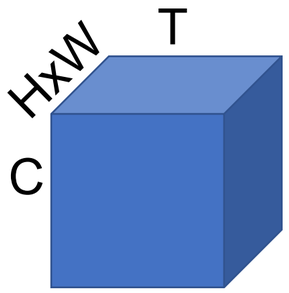
S3D /
R(2+1)D



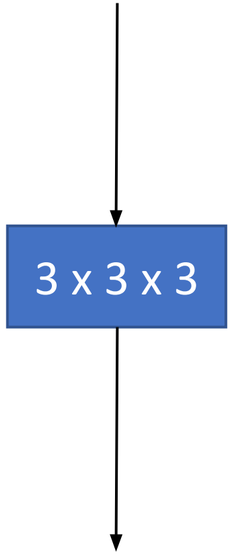
TSM

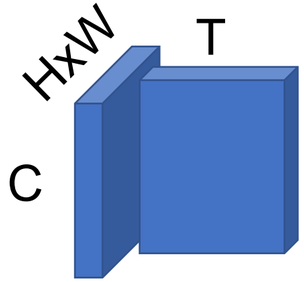


GSM

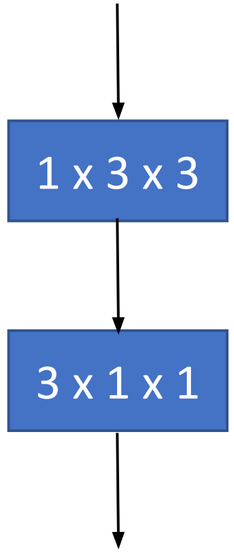


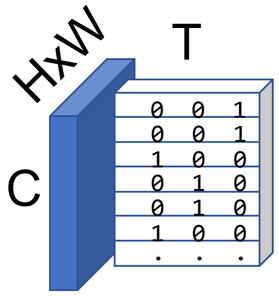
C3D



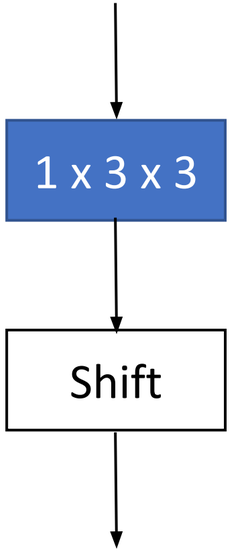


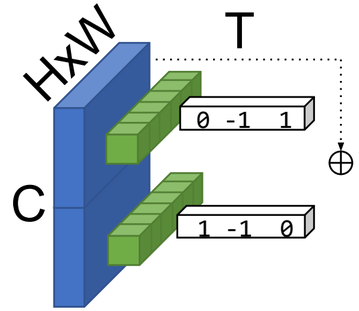
S3D / R(2+1)D



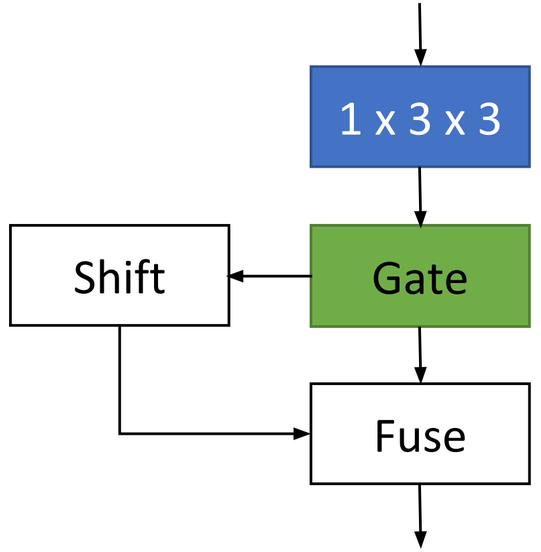


TSM

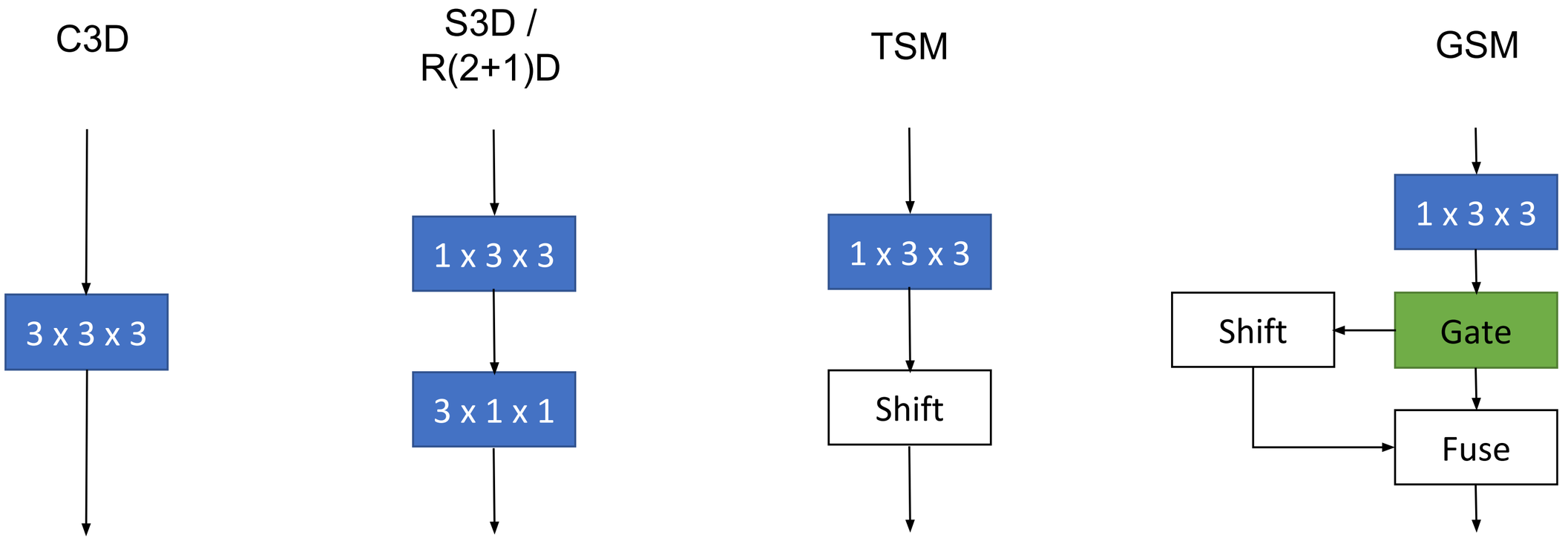




GSM

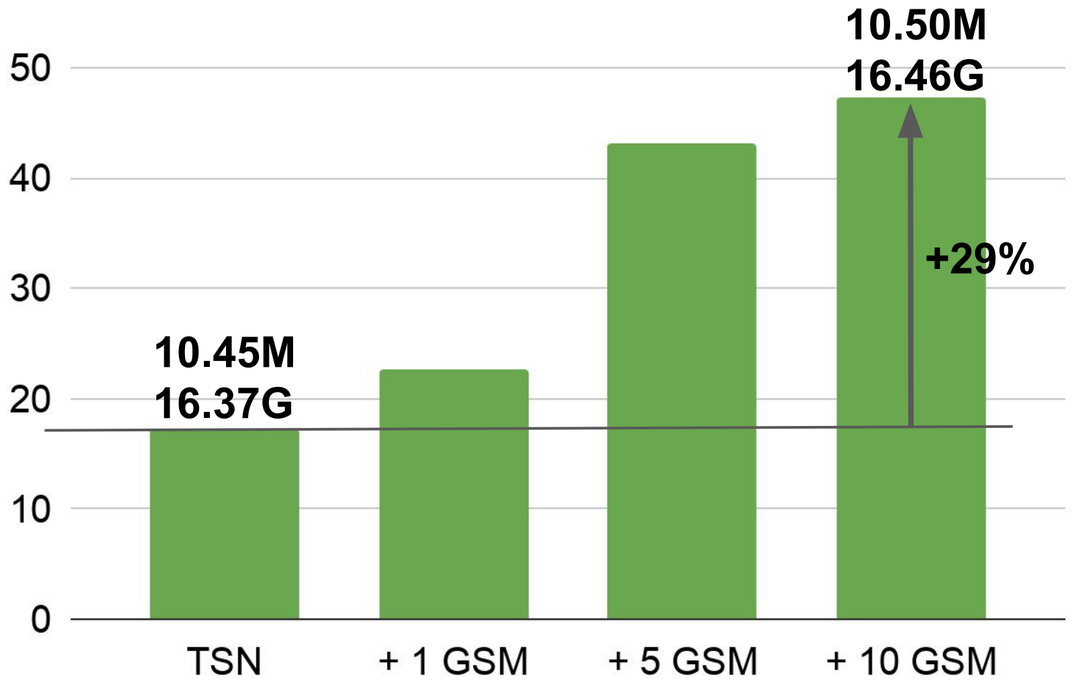


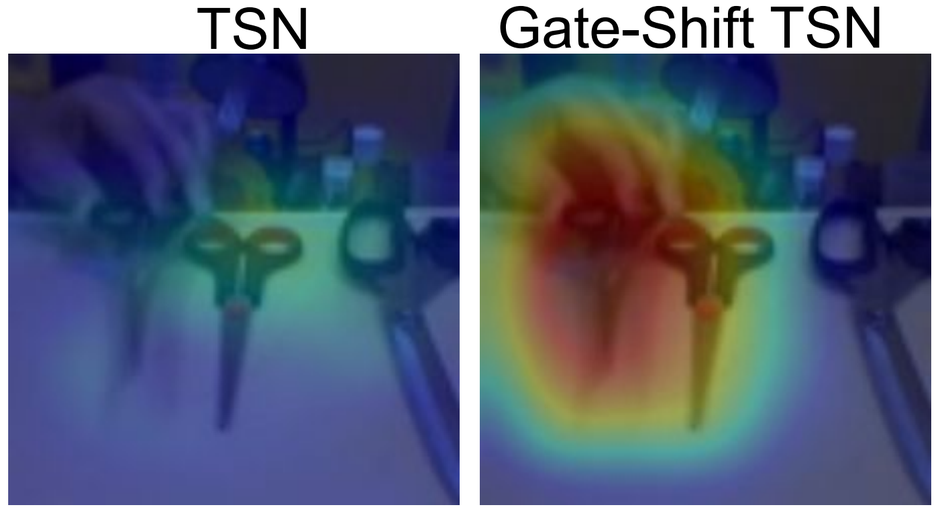
GSM develops a **flexible** and **data dependent decomposition** of 3D kernels with **reduced parameters** and **computational overhead**



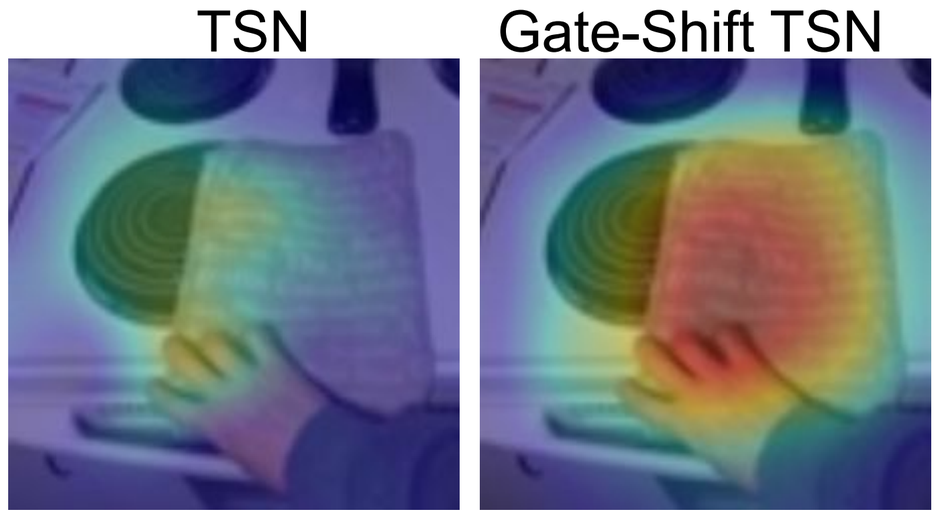
Effectiveness of GSM

Ablation study on Sth-V1

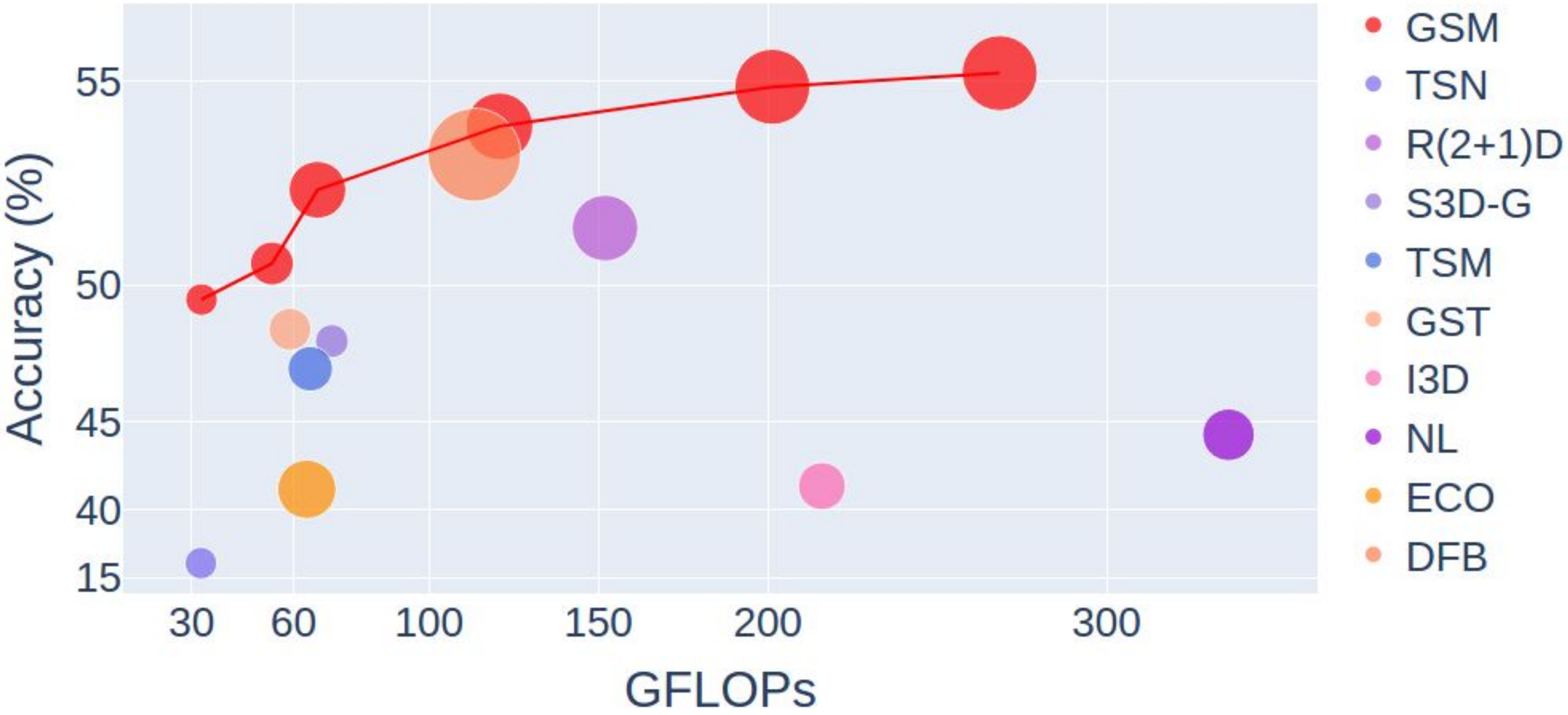




Putting sth similar to other things that are already on the table



Unfolding sth



State-of-the-art recognition accuracy of **55%** on Something Something-V1

Gate-Shift Networks for Video Action Recognition

Swathikiran Sudhakaran[1] Sergio Escalera[2,3] Oswald Lanz[1]

[1]Fondazione Bruno Kessler, Italy [2]Computer Vision Center, Spain

[3]Universitat de Barcelona, Spain

CVPR SEATTLE WASHINGTON JUNE 16-18 2020