SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

ECCV TEL AVIV 2022

# Dynamic Spatio-Temporal Specialization Learning for Fine-Grained Action Recognition

Tianjiao Li[1]*, Lin Geng Foo[1]*, Qiuhong Ke[2], Hossein Rahmani[3], Anran Wang[4], Jinghua Wang[5], and Jun Liu[1]

[1] ISTD Pillar, Singapore University of Technology and Design

[2] Department of Data Science & AI, Monash University

[3] School of Computing and Communications, Lancaster University

[4] ByteDance

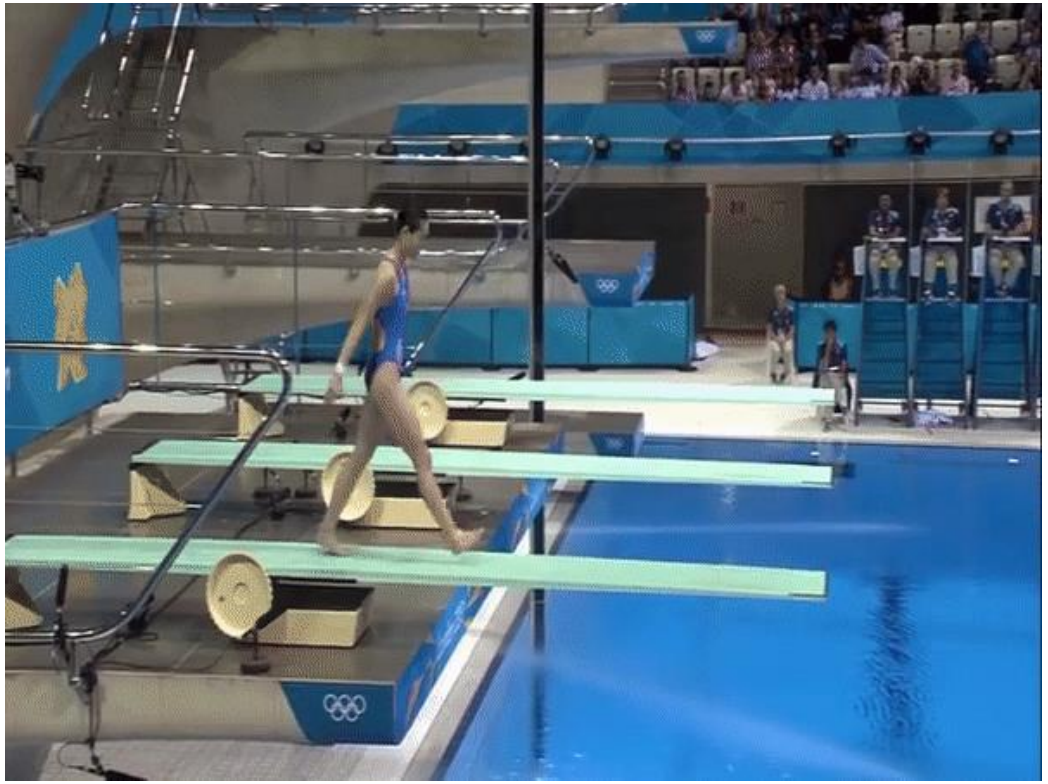
[3] School of Computer Science and Technology, Harbin Institute of Technology

* Equal Contribution

# Fine-Grained Action Recognition

['Forward', '15som', 'NoTwis', 'PIKE']

['Reverse', '15som', '25Twis', 'FREE']



GIFs from Diving48 Dataset

# Spatial vs Temporal Fine-grained Differences

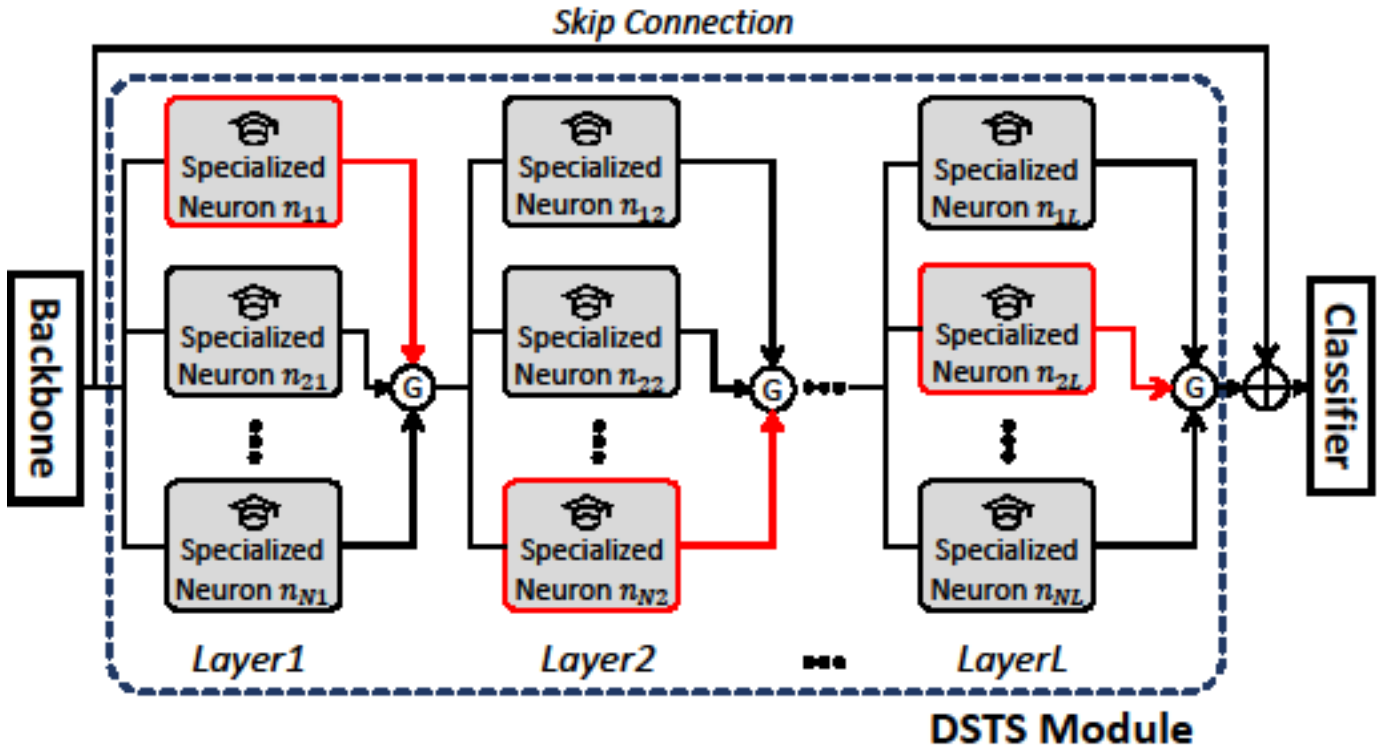

Fine-grained differences can exist in spatial or temporal aspects. A greater emphasis on the important aspect w.r.t the input video can improve performance

# Dynamic Spatio-Temporal Specialization

**Overview:** We design a DSTS module to handle fine-grained differences.

There are **L layers** within the DSTS module, each comprising **N specialized neurons**.

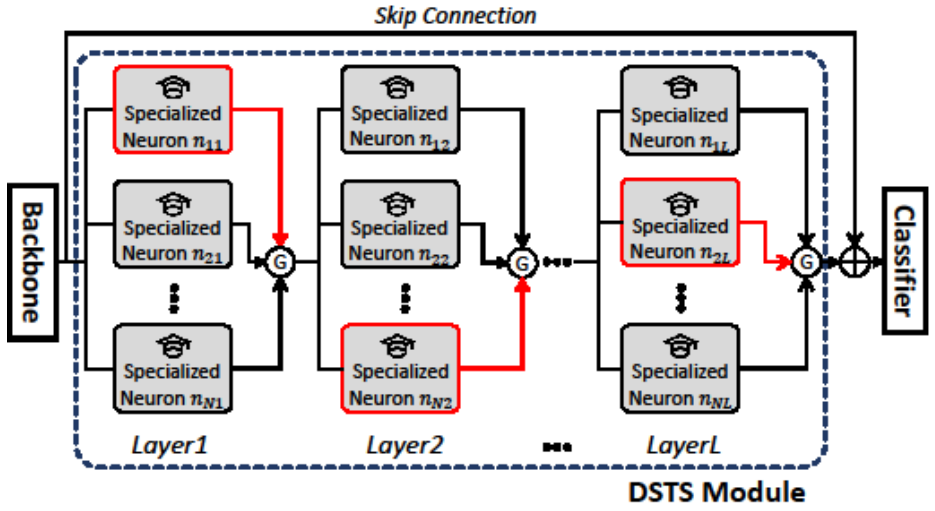
In the forward pass, specialized neurons are dynamically activated **based on the input**.



# Dynamic Spatio-Temporal Specialization

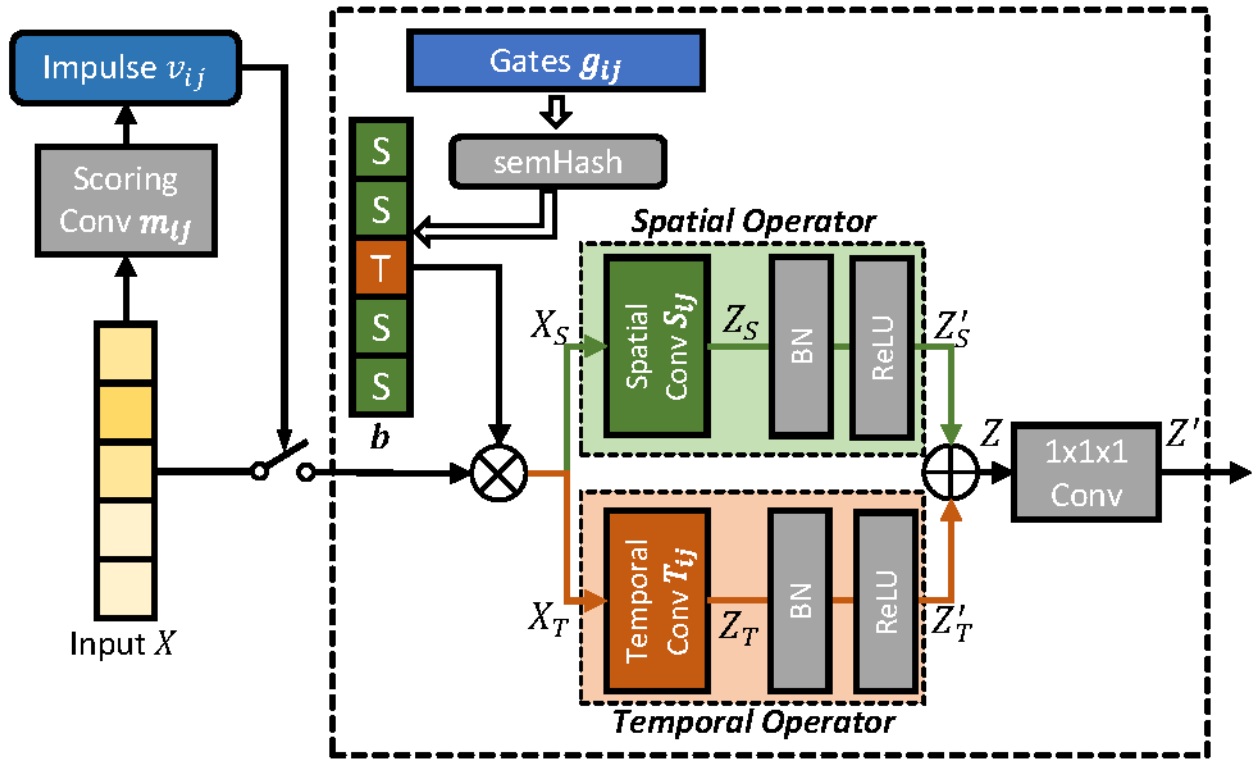
A **Synapse Mechanism** dynamically activates each specialized neuron only on a subset of samples that are highly similar, such that only fine-grained differences exist between them.

During training, in order to distinguish among that subset of similar samples, the loss will push the specialized neurons to **focus on exploiting the fine-grained differences** between them.



# Dynamic Spatio-Temporal Specialization

We also design a **Spatio-Temporal Specialization** method that additionally provides *specialized neurons* with *spatial* or *temporal* specializations, allowing them to have to higher sensitivity towards the fine-grained differences in those aspects.




*Specialized neuron* with Spatio-Temporal Specialization

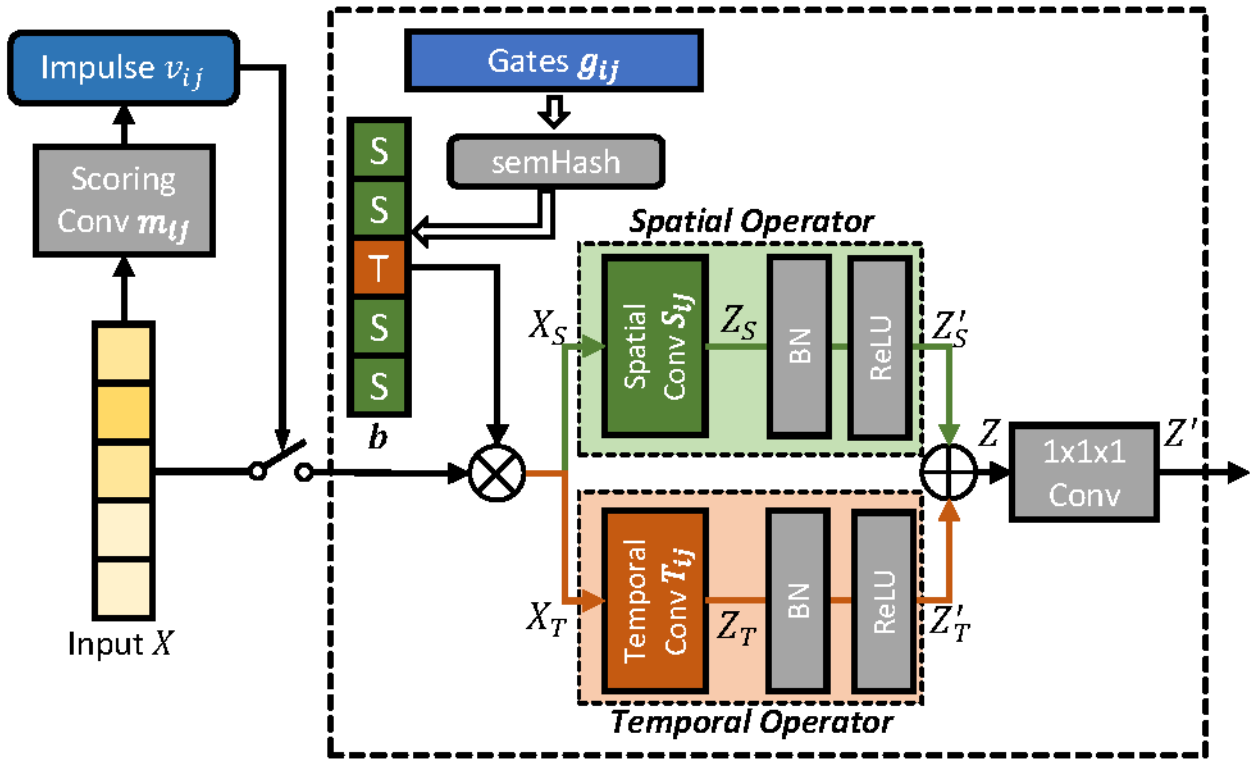
# Dynamic Spatio-Temporal Specialization

**Highlights of the Design**

- Each *specialized neuron* is designed with *a Spatial Operator* and a *Temporal Operator* in each channel

- *Gate* parameters in each channel control the choice of operator, and are optimized during training, adapting the architecture of the *specialized neuron*

- The set of *neurons* will have diversified architectures, which collectively are capable of handling a large variety of *spatial* and *temporal* fine-grained differences



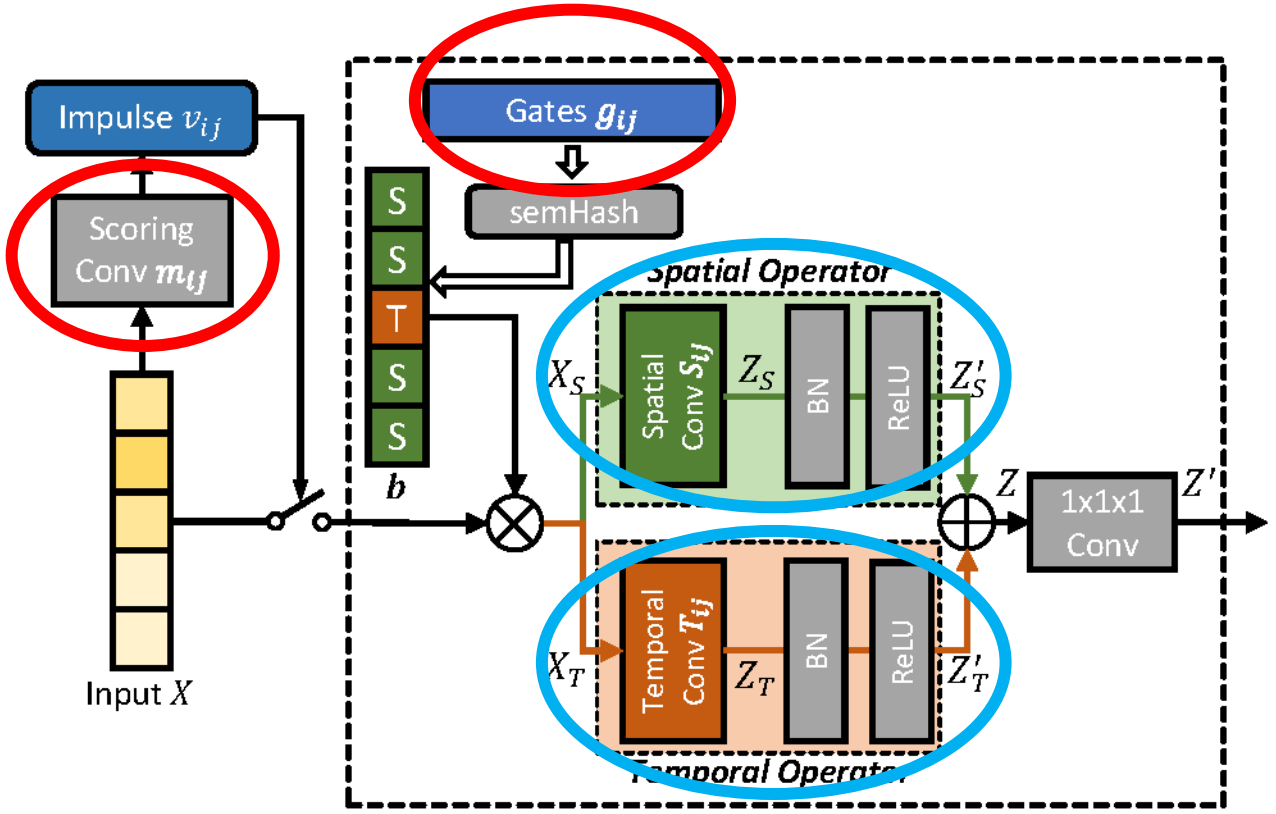

*Specialized neuron* with Spatio-Temporal Specialization

# Upstream-Downstream Learning

**Motivations of UDL**

- Our Upstream-Downstream Learning (UDL) algorithm better optimizes the model parameters involved in making dynamic decisions.

- *This is because upstream parameters* that make dynamic decisions and *downstream parameters* that process input, are jointly trained during our end-to-end training, which can be challenging as upstream parameters themselves also affect the training of downstream parameters.



# Experiments

**Results on SSv2**

| Method | Type | Top-1 | Top-5 |
|---|---|---|---|
| SlowFast [7] | C | 63.1 | 87.6 |
| TPN [41] | C | 64.7 | 88.1 |
| ViViT-L [1] | T | 65.4 | 89.8 |
| TSM (Two-stream) [19] | C | 66.6 | 91.3 |
| MViT-B [5] | T | 67.7 | 90.9 |
| Swin-B [20] | T | 69.6 | 92.7 |
| TPN w/ DSTS | C | 67.2 | 89.2 |
| Swin-B w/ DSTS | T | **71.8** | **93.7** |

**Results on Diving48**

| Method | Type | Top-1 | Class-wise Acc |
|---|---|---|---|
| I3D [3] | C | 48.3 | 33.2 |
| TSM (Two-stream) [19] | C | 52.5 | 32.7 |
| GST [22] | C | 78.9 | 69.5 |
| TQN [45] | T | 81.8 | 74.5 |
| Swin-B [20] | T | 80.5 | 69.7 |
| TPN [41] | C | 86.2 | 76.0 |
| Swin-B w/ DSTS | T | 83.0 | 71.5 |
| TPN w/ DSTS | C | **88.4** | **78.2** |

# Ablations on Diving48

## Spatio-Temporal Specialization

| Method | Top-1 | Class-wise Acc |
|---|---|---|
| DSTS w/o STS | 87.2 | 76.5 |
| DSTS w/o Gates | 87.3 | 76.7 |
| DSTS w/ STS | 88.4 | 78.2 |

## Synapse Mechanism

| Method | Top-1 | Class-wise Acc | Model Size |
|---|---|---|---|
| Baseline TPN | 86.2 | 76.0 | 63M |
| w/o Synapse Mechanism | 86.5 | 76.4 | 75M |
| w/ Synapse Mechanism | 88.4 | 78.2 | 75M |

## Upstream-Downstream Learning

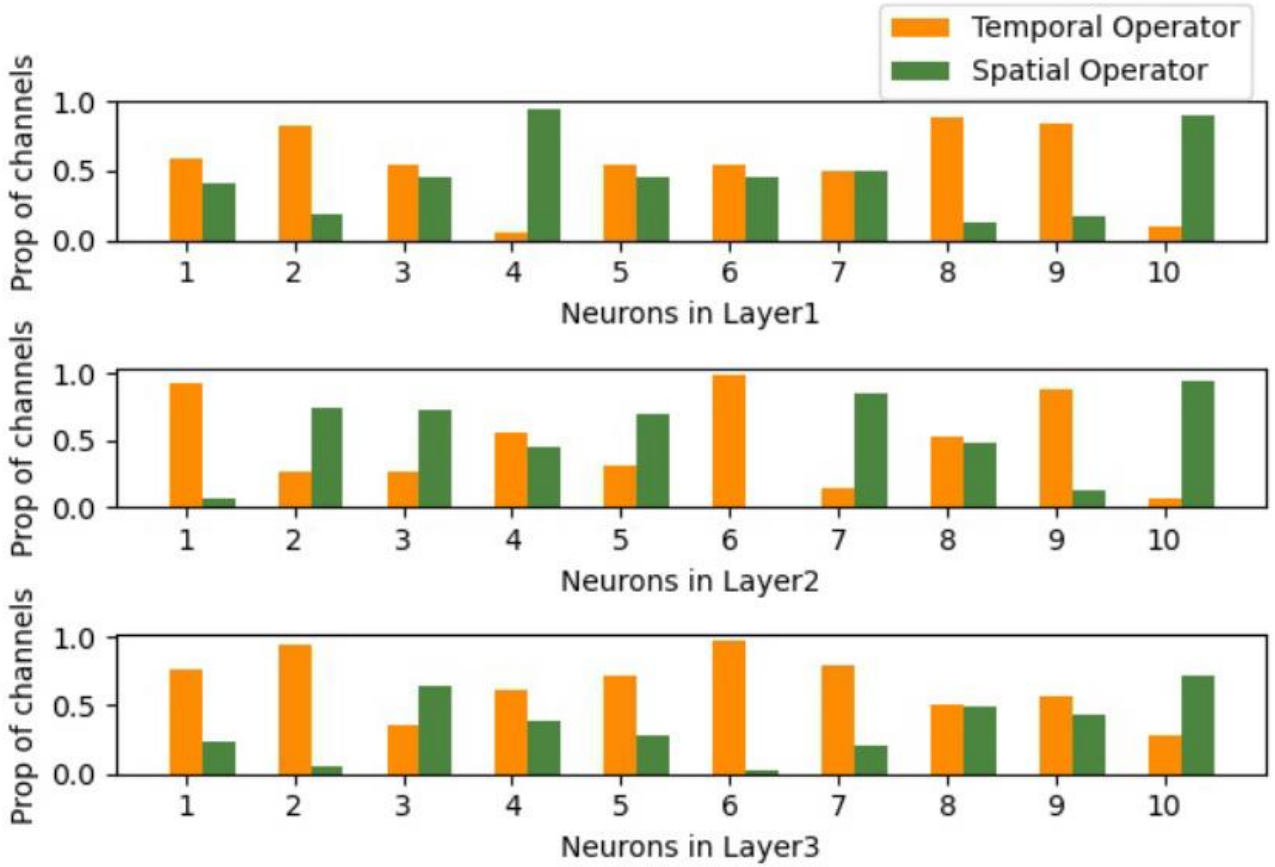
| Method | Top-1 | Class-wise Acc |
|---|---|---|
| DSTS w/o UDL | 87.4 | 76.7 |
| DSTS w/ UDL | 88.4 | 78.2 |

## Impact of N and L

| $N$ | Top-1 | Class-wise Acc |
|---|---|---|
| 5 | 87.3 | 76.2 |
| 10 | 88.4 | 78.2 |
| 15 | 88.3 | 78.2 |

| $L$ | Top-1 | Class-wise Acc |
|---|---|---|
| 1 | 87.5 | 76.8 |
| 3 | 88.4 | 78.2 |
| 5 | 88.2 | 78.2 |

# Visualization of Spatio-Temporal Specialization



# Thank You!