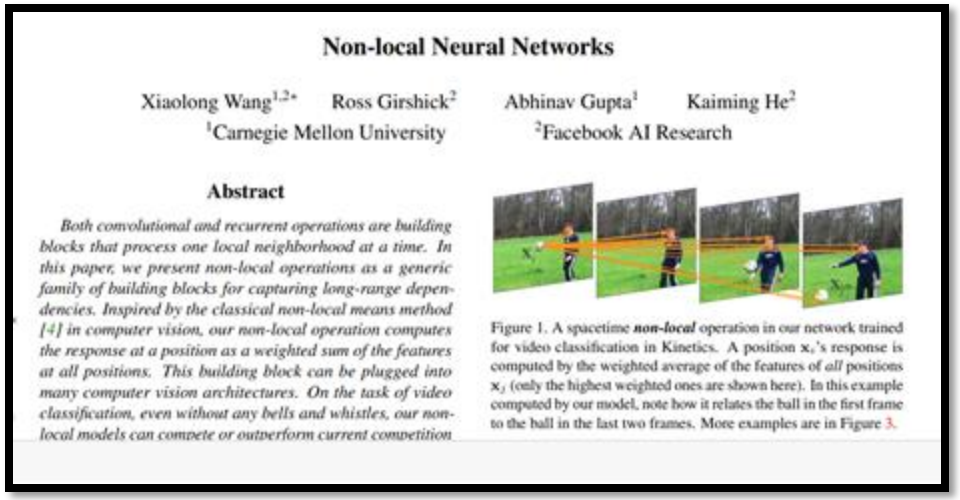
Advancements in Video Understanding

TSM Temporal Shift Module



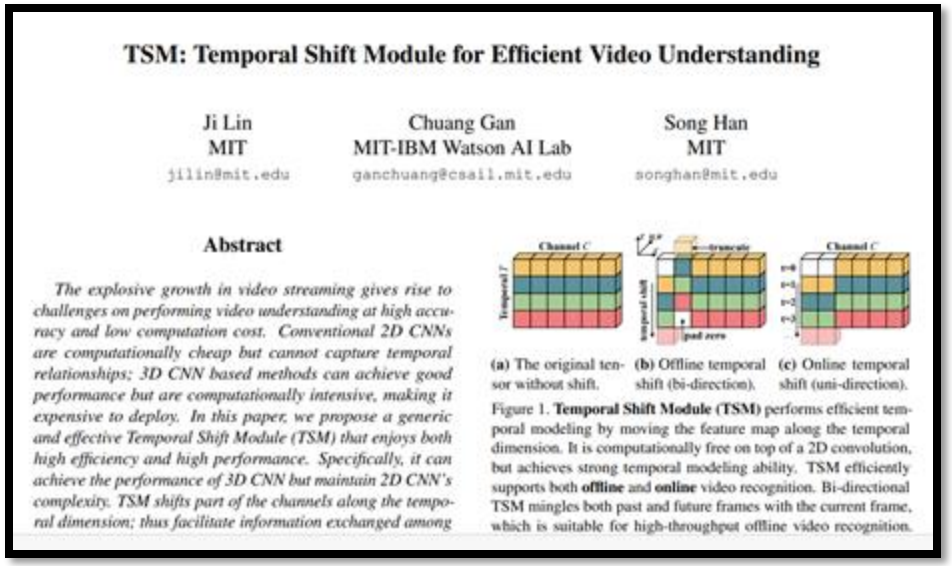
Non-local Neural Networks

Xiaolong Wang[1,2*] Ross Girshick[2] Abhinav Gupta[1] Kaiming He[2]
[1]Carnegie Mellon University [2]Facebook AI Research

Abstract

Both convolutional and recurrent operations are building blocks that process one local neighborhood at a time. In this paper, we present non-local operations as a generic family of building blocks for capturing long-range dependencies. Inspired by the classical non-local means method [4] in computer vision, our non-local operation computes the response at a position as a weighted sum of the features at all positions. This building block can be plugged into many computer vision architectures. On the task of video classification, even without any bells and whistles, our non-local models can compete or outperform current competition

Figure 1. A spacetime ***non-local*** operation in our network trained for video classification in Kinetics. A position $\mathbf{x}_i$'s response is computed by the weighted average of the features of *all* positions $\mathbf{x}_j$ (only the highest weighted ones are shown here). In this example computed by our model, note how it relates the ball in the first frame to the ball in the last two frames. More examples are in Figure 3.



TSM: Temporal Shift Module for Efficient Video Understanding

Ji Lin
MIT
jilin@mit.edu

Chuang Gan
MIT-IBM Watson AI Lab
ganchuang@csail.mit.edu

Song Han
MIT
songhan@mit.edu

Abstract

The explosive growth in video streaming gives rise to challenges on performing video understanding at high accuracy and low computation cost. Conventional 2D CNNs are computationally cheap but cannot capture temporal relationships; 3D CNN based methods can achieve good performance but are computationally intensive, making it expensive to deploy. In this paper, we propose a generic and effective Temporal Shift Module (TSM) that enjoys both high efficiency and high performance. Specifically, it can achieve the performance of 3D CNN but maintain 2D CNN's complexity. TSM shifts part of the channels along the temporal dimension; thus facilitate information exchanged among

(a) The original tensor without shift. (b) Offline temporal shift (bi-direction). (c) Online temporal shift (uni-direction).

Figure 1. **Temporal Shift Module (TSM)** performs efficient temporal modeling by moving the feature map along the temporal dimension. It is computationally free on top of a 2D convolution, but achieves strong temporal modeling ability. TSM efficiently supports both **offline** and **online** video recognition. Bi-directional TSM mingles both past and future frames with the current frame, which is suitable for high-throughput offline video recognition.

Presenter: Giorgio Roffo

Temporal Shift Module

Proceedings of the IEEE/CVF International Conference on Computer Vision.

(ICCV) **2019**.

TSM: Temporal Shift Module for Efficient Video Understanding

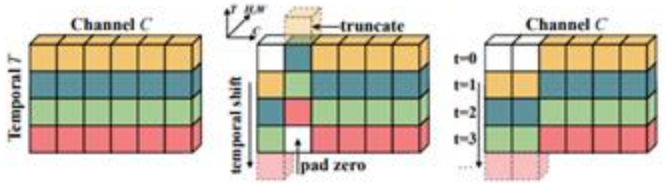
Ji Lin
MIT
jilin@mit.edu

Chuang Gan
MIT-IBM Watson AI Lab
ganchuang@csail.mit.edu

Song Han
MIT
songhan@mit.edu

Abstract

The explosive growth in video streaming gives rise to challenges on performing video understanding at high accuracy and low computation cost. Conventional 2D CNNs are computationally cheap but cannot capture temporal relationships; 3D CNN based methods can achieve good performance but are computationally intensive, making it expensive to deploy. In this paper, we propose a generic and effective Temporal Shift Module (TSM) that enjoys both high efficiency and high performance. Specifically, it can achieve the performance of 3D CNN but maintain 2D CNN's complexity. TSM shifts part of the channels along the temporal dimension; thus facilitate information exchanged among



(a) The original tensor without shift. **(b)** Offline temporal shift (bi-direction). **(c)** Online temporal shift (uni-direction).

Figure 1. **Temporal Shift Module (TSM)** performs efficient temporal modeling by moving the feature map along the temporal dimension. It is computationally free on top of a 2D convolution, but achieves strong temporal modeling ability. TSM efficiently supports both **offline** and **online** video recognition. Bi-directional TSM mingles both past and future frames with the current frame, which is suitable for high-throughput offline video recognition.

TSM: Temporal Shift Module for Efficient Video Understanding

Ji Lin
MIT
jilin@mit.edu

Chuang Gan
MIT-IBM Watson AI Lab
ganchuang@csail.mit.edu

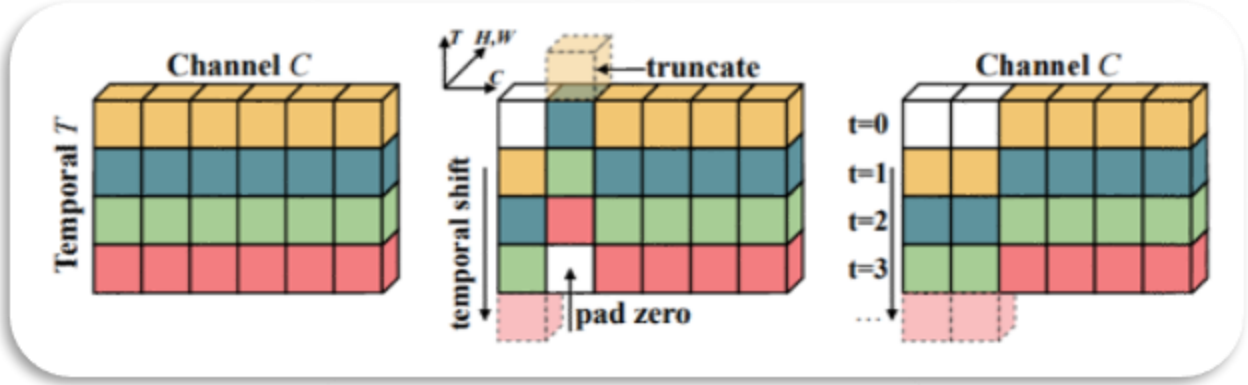
Song Han
MIT
songhan@mit.edu

Motivation

- Conventional 2D CNNs are computationally cheap but cannot capture temporal relationships.
- 3D CNN based methods can achieve good performance but are computationally intensive, making it expensive to deploy.
- **Convert any off-the-shelf 2D CNN model into a pseudo-3D model.**

Temporal Shift Module (TSM)

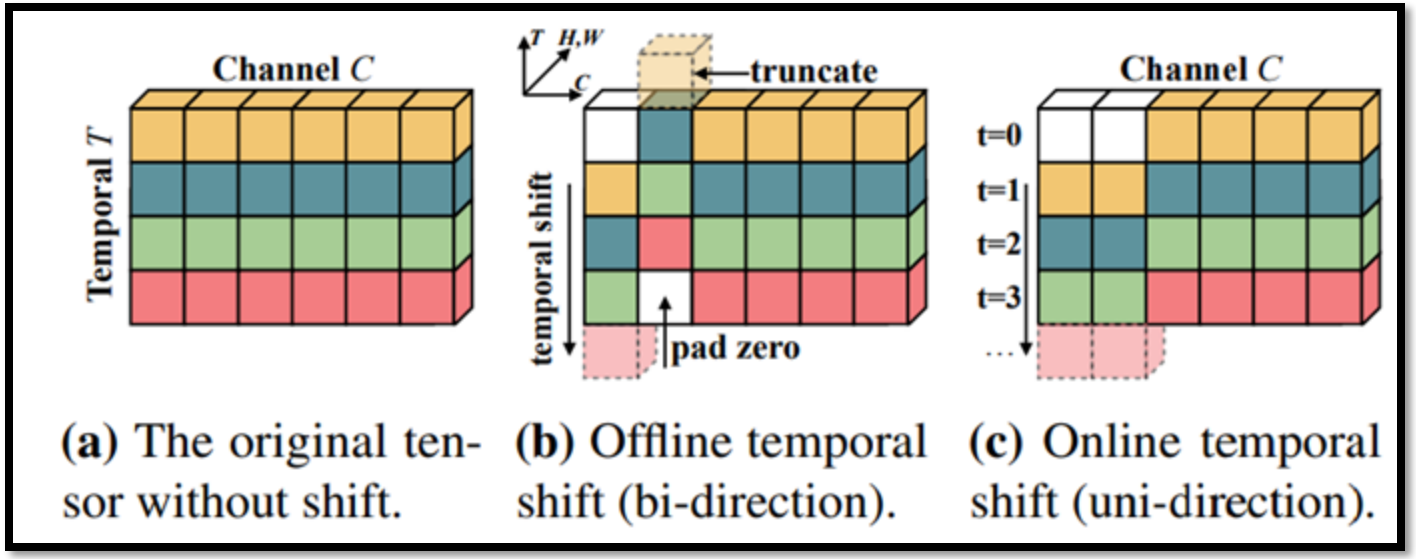
- C2Ds operate independently over the dimension T, thus no temporal modeling takes effects.
- TSM can achieve the performance of 3D CNN but maintain C2D's complexity.
- TSM can be inserted into C2Ds to achieve temporal modeling ***at zero computation*** **and zero parameters**.
- TSM **shifts part of the channels along the temporal dimension**, thus facilitate information exchanged among neighboring frames.



- Conventional convolution operation consists of shift and multiply-accumulate.

- **Offline**: TSM shifts in the time dimension by ±1 and folds the multiply accumulate from time dimension to channel dimension.

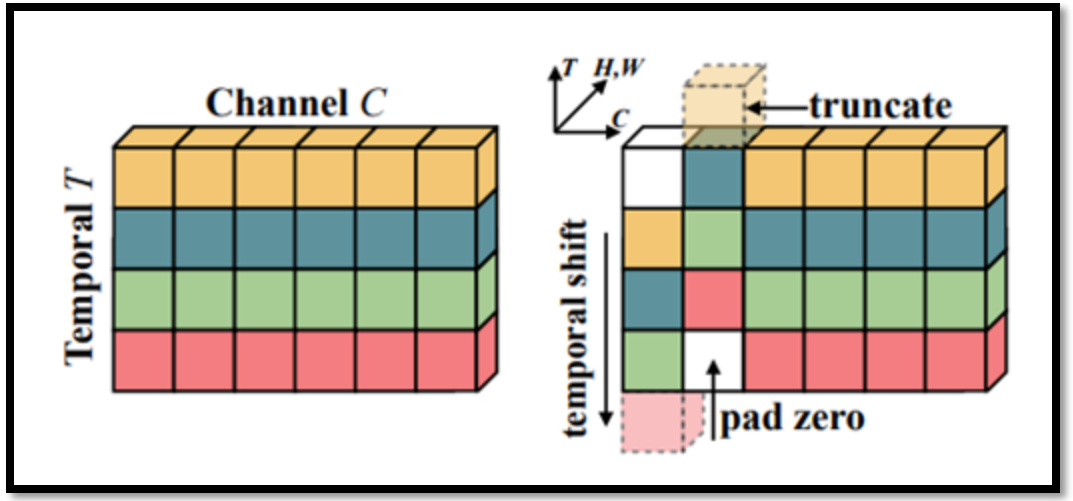
- **Online**: real-time online video understanding, future frames cannot get shifted to the present, so TSM is applied in a uni-directional way.



(a) The original tensor without shift. **(b)** Offline temporal shift (bi-direction). **(c)** Online temporal shift (uni-direction).

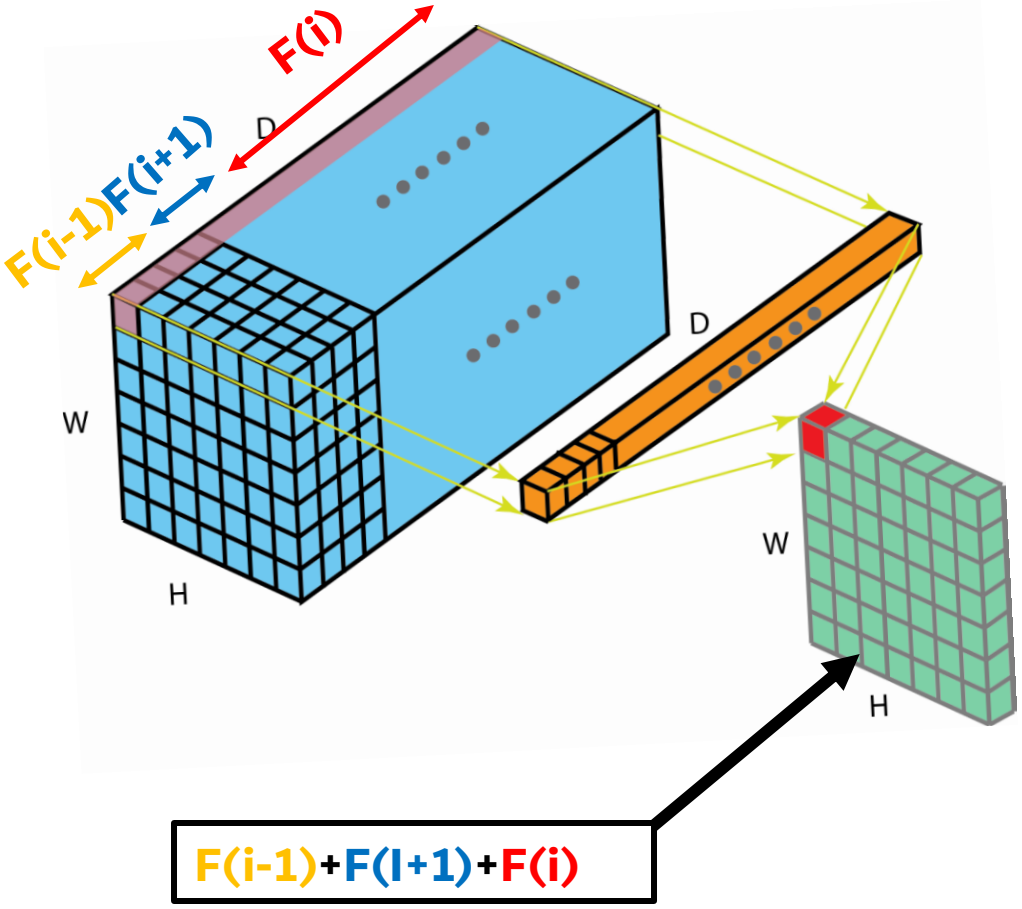
Offline Models with Bi-directional TSM

- Given a video V , we first sample T frames Fi , (F1, ..., FT) from the video.
- After frame sampling, TSM processes each of the frames individually (just like a 2D CNN).
- The difference is that TSM is inserted for each residual block, which **enables temporal information fusion at no computation.**



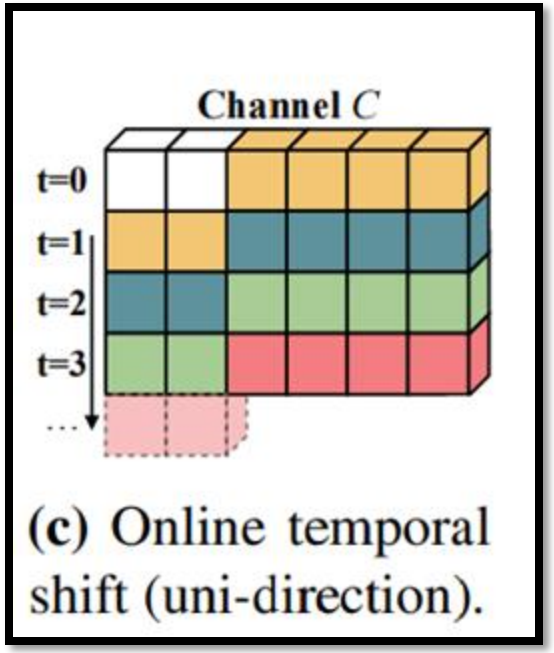
Offline Models with Bi-directional TSM

- Given a video V , we first sample T frames Fi , (F1, ..., FT) from the video.
- After frame sampling, TSM processes each of the frames individually (just like a 2D CNN).
- The difference is that TSM is inserted for each residual block, which enables temporal information fusion at no computation.
- For each inserted temporal shift module, the **temporal receptive field** will be **enlarged by 2**, as if running a convolution with the **kernel size of 3 along the temporal dimension**.
- TSM model has a very large temporal receptive field.
- Only 1/4 (1/8 for each direction) features are shifted.



Online Models with Uni-directional TSM

- Given a video V , we first sample T frames Fi , (F1, ..., FT) from the video.
- Offline TSM requires features from future frames to replace the features in the current frame.
- Online recognition with uni-directional TSM only **shifts the feature from previous frames to current frames**.



(c) Online temporal shift (uni-direction).

Online Models with Uni-directional TSM

- Given a video V , we first sample T frames Fi , (F1, ..., FT) from the video.
- Offline TSM requires features from future frames to replace the features in the current frame.
- Online recognition with uni-directional TSM only **shifts the feature from previous frames to current frames**.
- For each frame, the first 1/8 feature maps of each residual block are cached in the memory.
- For the next frame, the first 1/8 of the current feature maps are replaced with the cached feature maps.
    - For ResNet-50, only needed 0.9MB memory cache

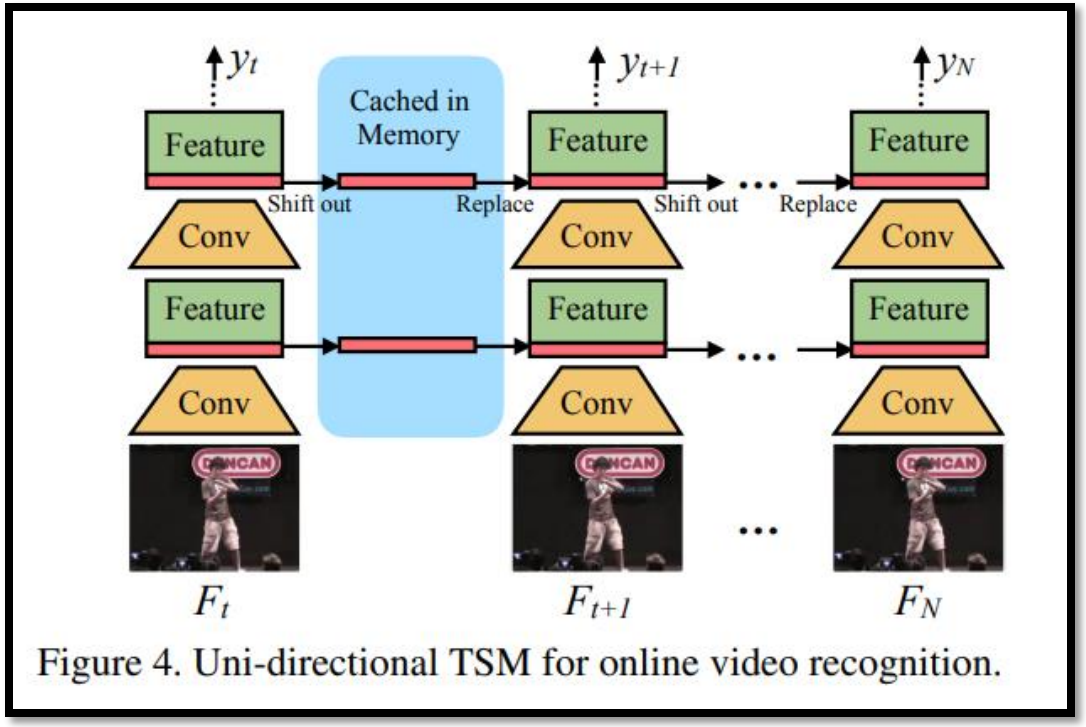
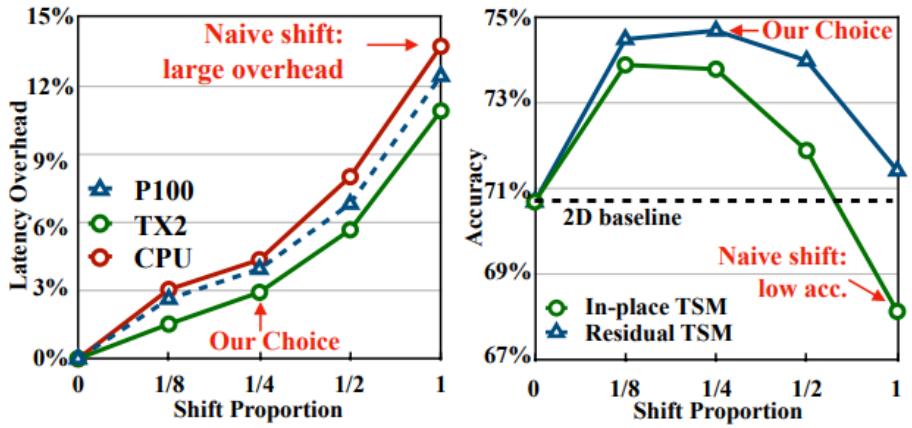


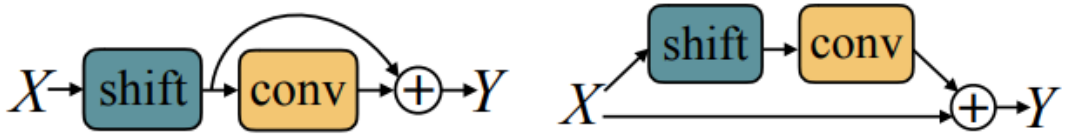
Figure 4. Uni-directional TSM for online video recognition.

- **Temporal partial shift strategy**
  Fig.2(a). Shifts only a small portion of the channels (1/4) for efficient temporal fusion (*shifting too many channels in a network will significantly hurt the spatial modeling ability*).
- **TSM inside residual branch**
  Fig.2(b). All the information in the original activation is still accessible after temporal shift through identity mapping.



(a) Overhead *vs.* proportion. **(b)** Residual *vs.* in-place.

Figure 2. **(a)** Latency overhead of TSM due to data movement. **(b)** Residual TSM achieve better performance than in-place shift. We choose 1/4 proportion residual shift as our default setting. It achieves higher accuracy with a negligible overhead.



(a) In-place TSM. **(b)** Residual TSM.

Figure 3. Residual shift is better than in-place shift. In-place shift happens before a convolution layer (or a residual block). Residual shift fuses temporal information inside a residual branch.

Experiments and results

- **Datasets**. Something-Something (V1&V2), Charades, and Jester are more focused on modeling the temporal relationships.
- **Something-Something-Vx** is a challenging dataset, as activity cannot be inferred merely from individual frames (e.g., pushing something from right to left rather than left to right).
- TSM achieves the first place on the leaderboard upon publication.

Table 1. Our method consistently outperforms 2D counterparts on multiple datasets at zero extra computation (protocol: ResNet-50 8f input, 10 clips for Kinetics, 2 for others, full-resolution).

	Dataset	Model	Acc1	Acc5	Δ Acc1
Less Temporal	Kinetics	TSN	70.6	89.2	+3.5
		Ours	**74.1**	**91.2**	
	UCF101	TSN	91.7	99.2	+4.2
		Ours	**95.9**	**99.7**	
	HMDB51	TSN	64.7	89.9	+8.8
		Ours	**73.5**	**94.3**	
More Temporal	Something V1	TSN	20.5	47.5	+28.0
		Ours	**47.3**	**76.2**	
	Something V2	TSN	30.4	61.0	+31.3
		Ours	**61.7**	**87.4**	
	Jester	TSN	83.9	99.6	+11.7
		Ours	**97.0**	**99.9**	

Experiments and results (TSM Online)

Table 6. Comparing the accuracy of offline TSM and online TSM on different datasets. Online TSM brings negligible latency overhead.

Model	Latency	Kinetics	UCF101	HMDB51	Something
TSN	4.7ms	70.6%	91.7%	64.7%	20.5%
+Offline	-	74.1%	95.9%	73.5%	47.3%
+Online	4.8ms	74.3%	95.5%	73.6%	46.3%

Table 7. Video detection results on ImageNet-VID.

Model	Online	Need Flow	Latency	mAP			
				Overall	Slow	Medium	Fast
R-FCN [23]	✓		1×	74.7	83.6	72.5	51.4
FGFA [60]		✓	2.5×	75.9	**84.0**	74.4	55.6
Online TSM	✓		1×	**76.3**	83.4	**74.8**	**56.0**

Online Object Detection

Conclusions

- TSM **shifts part of the channels along the temporal dimension (1/4)**
- It can be inserted into 2D CNN backbone to enable joint spatial-temporal modeling at no additional cost.
- TSM can achieve the performance of 3D CNN but maintain C2D's complexity.
- TSM supports both offline and online video recognition.

Summary

- **NL Nets**
  - Non-local operations allows the model to capture long-range dependencies.
  - Adding NL blocks improves C2d and i3D architectures.
  - The best performances are obtained with the NL i3D model with 5 blocks.

  Pros: NL i3D SOTA on activity recognition, object detection, instance segmentation and KeyPoint detection tasks.

  Cons: i3D models are computationally intensive (*).
- **TSM**
  - **TSM converts any 2D CNN model into a pseudo-3D model (*).**
  - TSM achieves the performance of 3D CNN but maintain C2D's complexity.
  - TSM **shifts part of the channels along the temporal dimension (1/4)**
  - TSM supports both **offline** and **online** video recognition.

Thank you