International Journal of Engineering in Computer Science

E-ISSN: 2663-3590
P-ISSN: 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2025; 7(1): 119-123
Received: 10-02-2025
Accepted: 12-03-2025

Onkar Tiwari
Department of Computer Science and Engineering, Shri Krishna University, Chhatarpur, Madhya Pradesh, India

Krishan Kumar
Department of Computer Science and Engineering, Shri Krishna University, Chhatarpur, Madhya Pradesh, India

Anurag Tiwari
Department of Computer Application and Information Technology, AKS University, Satna, Madhya Pradesh, India

Pinki Sharma
Department of Computer Application and Information Technology, AKS University, Satna, Madhya Pradesh, India

Corresponding Author:
Pinki Sharma
Department of Computer Application and Information Technology, AKS University, Satna, Madhya Pradesh, India

Comparative analysis of spatial filtering and temporal filtering in convolutional neural networks

Onkar Tiwari, Krishan Kumar, Anurag Tiwari and Pinki Sharma

DOI: https://www.doi.org/10.33545/26633582.2025.v7.i1b.167

Abstract
Convolutional neural networks, or CNNs, have transformed machine learning, especially in the interpretation of images and videos. CNNs use spatial filtering to extract static features from images, while temporal filtering allows them to also extract dynamic data, like video sequences. Spatial and temporal filtering are compared in this research, which also examines their theoretical foundations, applications, advantages, disadvantages, and use cases. By contrasting different filtering mechanisms, we hope to shed light on their uses and help researchers choose the best filtering methods for a range of tasks.

Keywords: Convolutional neural networks (CNNS), recurrent neural networks (RNNS), deep learning, action recognition, spatial filtering, temporal filtering

1. Introduction

The foundation of CNNs is spatial filtering, which lets one find spatial characteristics such edges, textures, and objects in pictures. Contrarily, temporal filtering deals with the temporal dimension by identifying variations between frames in sequential data, such as time-series or films. Understanding the interactions and distinctions between spatial and temporal filtering is essential given the increasing complexity of jobs utilizing spatiotemporal data. With an emphasis on how they enhance one another in contemporary deep learning, this paper explores their functions, mathematical formulations, and implementations [2].

2. Theoretical Background

2.1 Spatial Filtering: The spatial dimensions (height as well as width) of an input are used in spatial filtering. Convolutional kernels use translation invariance to detect features independent of their position as they move over these dimensions to extract local patterns. In CNNs, the technique of employing convolutional kernels to analyze and extract characteristics from the spatial dimensions of input data-usually images-is known as "spatial filtering." These kernels are tiny, learnable filters that recognize local patterns like corners, edges, and textures by sliding (or convolving) over the input image. Important characteristics:

2.2 Spatial Dimensions: The input data's width and height are taken into account. These two dimensions are the focus of the convolutional operation.

A. **Translation Invariance:** This characteristic allows the kernel to identify a particular feature (such an edge) in the input image, independent of where it is located.

B. **Mathematical Representation:** The equation:

$$Y(i,j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X(i+m, j+n) \cdot W(m,n)$$

- $Y(i,j)$: Output at position (i,j).
- $X(i+m, j+n)$: Input values around position (i,j), covered by the kernel.
- $W(m,n)$: Kernel (or filter) values.
- k: Size of the kernel $(e.g., 3 \times 3, 5 \times 5)$.

C. Goal: To record localized patterns and characteristics that are essential for image analysis jobs like object detection, segmentation, and classification.

CNNs use this filtering technique as their fundamental mechanism, which allows them to efficiently process visual input and create hierarchical feature representations [1].

2.3 Temporal Filtering: By examining successive inputs, temporal filtering takes the time dimension into account. Temporal filters, which are frequently used to video data, record motion patterns, temporal dependencies, and changes across time. Analyzing sequential data over the time dimension is known as temporal filtering. Spatial filtering, on the other hand, works with static images or spatial information. For jobs involving sequential data, such video analysis or time-series data, temporal filtering is particularly pertinent. Examples include identifying trends in temporal datasets or motion patterns in films.

A. Mathematical Representation: In order to capture patterns that change over time, temporal filtering employs kernels that span time steps. For example, the following formula:

$$Y(t) = \sum_{\tau=-T}^{T} X(t+\tau) \cdot W(\tau)$$

Captures

- $Y(t)$: The output at a specific time t.
- *$X(t + \tau)$:* The input data at a neighboring time step *$(t + \tau)$.*
- *$W(\tau)$:* The filter weights applied over the temporal window.
- *T:* The range of time steps considered.

B. Purpose: It detects motion patterns, such as the movement of objects between video frames, or records temporal dependencies, such as forecasting future trends from historical data.

This idea is essential for problems like video action identification, where precise predictions require an awareness of both spatial and temporal dynamics [1].

3. Implementation Approaches in CNN and Comparison

Convolutional layers with backpropagation-optimized kernels are used to create spatial filters. While dilated convolutions increase the receptive field, variations such as depth-wise separable convolutions increase efficiency [2]. On the other hand, temporal filters frequently make use of 3D convolutions, whose kernels cover both temporal and spatial dimensions. CNNs are also used with Transformer topologies and Recurrent Neural Networks (RNNs) to manage temporal relationships [1].

Both spatial and temporal filtering play different roles in data analysis and are essential methods in deep learning and signal processing. Because it concentrates on obtaining spatial patterns and information from still pictures, spatial filtering is very useful for object detection and image classification. The study of sequential data, on the other hand, requires temporal filtering in order to capture motion patterns and temporal dependencies, which are critical for time-series analysis and video categorization.

With effective shared weights, spatial filtering is excellent at identifying both low- and high-level spatial characteristics, but it has trouble with temporal dynamics. On the other hand, temporal filtering is essential for tasks like action identification and comprehending sequential patterns, despite being computationally demanding and prone to overfitting in settings with little data. When combined, these methods strengthen the models' capacity to efficiently process temporal and geographical input.

Table 1: Comparison between Spatial Filtering and Temporal Filtering

Factors	Spatial Filtering	Temporal Filtering
Strengths	A. Highly effective for static image analysis.	A. Essential for tasks involving sequential data, e.g., action recognition.
	B. Efficient with shared weights and sparse connectivity.	B. Captures temporal relationships and motion patterns.
	C. Can detect low- and high-level spatial features.	
Limitations	A. Ineffective for capturing temporal dynamics in videos or sequential data.	A. Computationally intensive due to additional temporal dimensions.
	B. Limited context when features span multiple frames.	B. Prone to overfitting in limited data scenarios.
Use Cases	A. In image classification, detecting spatial patterns in images.	A. Image classification is not applicable.
	B. In video classification, extracting spatial features per frame.	B. In video classification, capturing motion across frames.
	C. Initial feature extraction in action recognition.	C. Understanding motion and sequence both in action recognition.
	D. Limited application in Time-Series Analysis	D. Essential for temporal trends Time-Series Analysis

4. Experimental Comparison

When experimented on benchmark datasets, such as UCF101 for spatiotemporal tasks and CIFAR-10 for spatial tasks. The findings show that while temporal filtering is essential for video-based applications, spatial filtering performs best in static picture categorization.

4.1 The CIFAR-10 dataset [12]: There are 60000 32x32 color images in 10 classes, with 6000 images in each class, in the CIFAR-10 dataset [7]. Ten thousand test photos and fifty thousand training images are included. Five training batches and one test batch, each containing 10,000 photos, make up the dataset. There are precisely 1000 randomly chosen photos from each class in the test batch. The remaining photographs are arranged randomly in the training batches, albeit certain training batches could include more images from one class than another. The training batches are made up of precisely 5000 photos from each class. The dataset's classes and ten randomly selected photos from each are displayed here in Fig 1.

4.2 The UCF101 Dataset [7]: This action recognition data set consists of 101 action categories from realistic action videos that were gathered from YouTube. The UCF50 data set, which has 50 action categories, is expanded upon by this data set.

With 13320 videos across 101 action categories, UCF101 offers the greatest diversity in terms of actions. It is also the most difficult data set to date because of the wide range of camera motion, object appearance and pose, object scale, viewpoint, cluttered background, lighting conditions, and other factors. By learning and investigating new realistic action categories, UCF101 seeks to promote more action recognition research, since the majority of the available action recognition data sets are staged by actors and are not realistic.

The videos in the 101 activity categories are divided into 25 groups, with four to seven videos of one action per group. Videos from the same group could have some things in common, like a similar background or point of view.

There are five categories into which the action categories can be separated: 1) Human-Item Communication 2) Just Body Motion 3) Interaction between Humans 4) Performing on an Instrument 5) Athletics. The action categories for UCF101 data set are: shown in Fig 2.

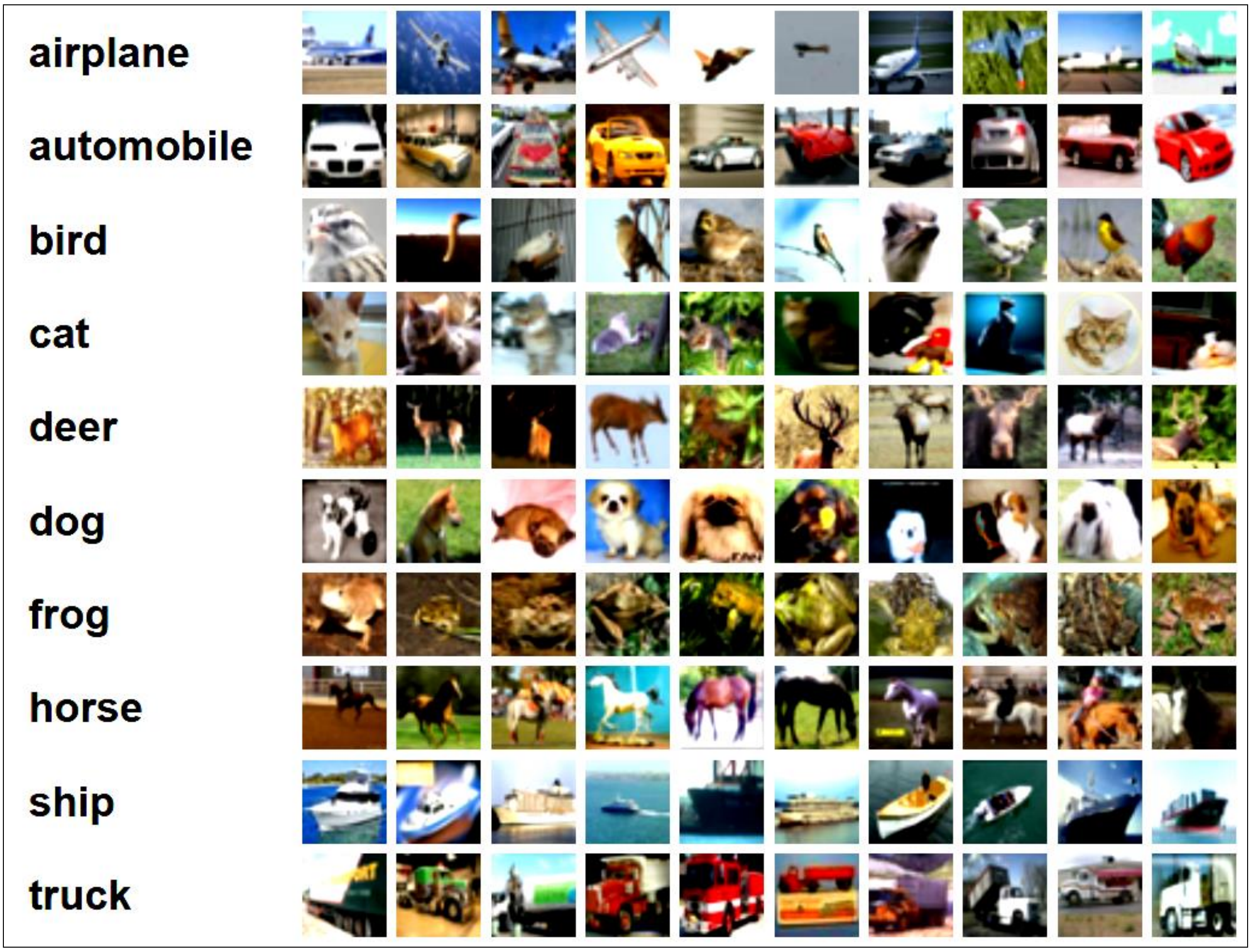

Fig 1: Classes in the dataset and 10 random images from each.

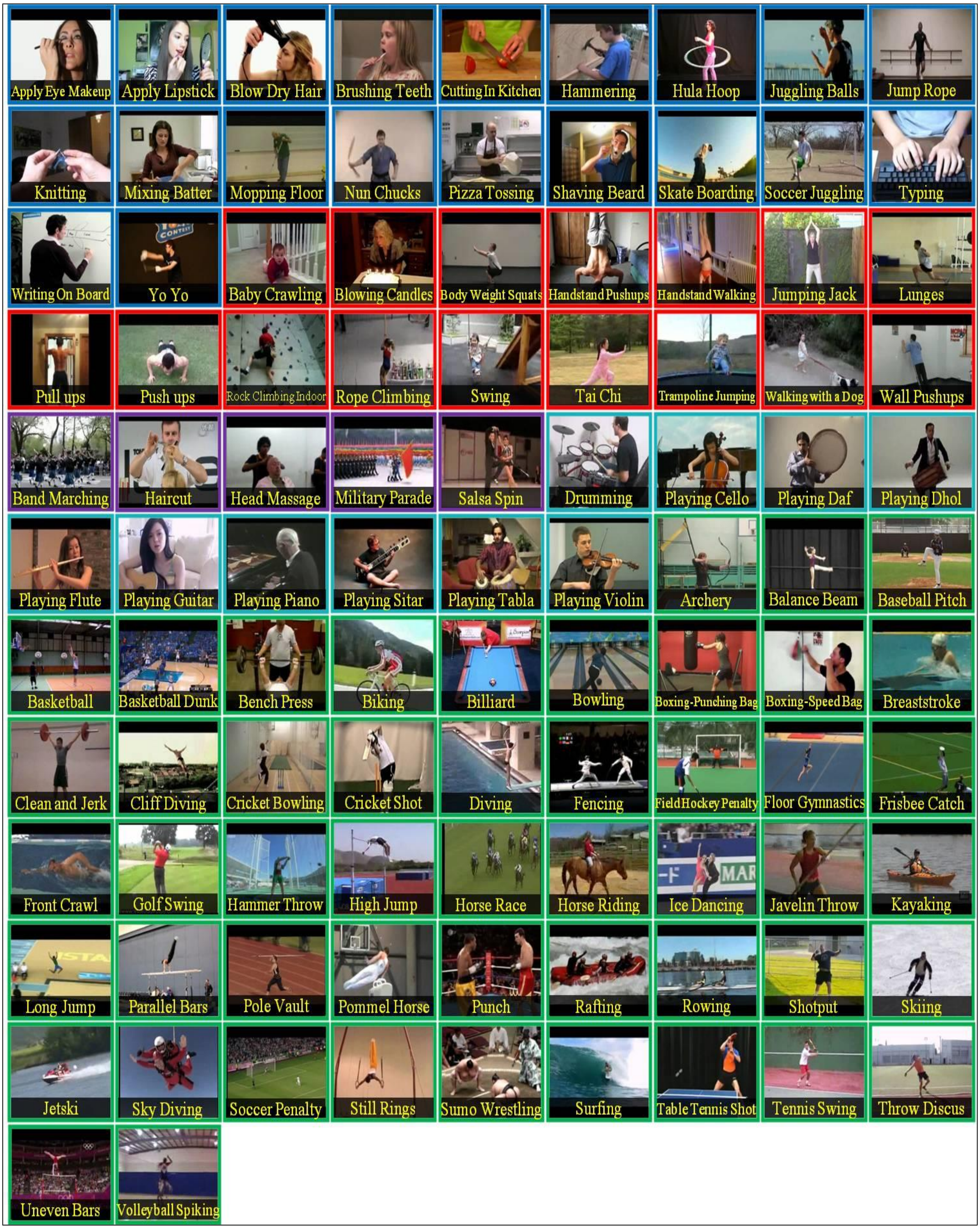

Fig 2: The action categories for UCF101 data set.

5. Result

The study evaluates the effectiveness of Spatial Filtering and Temporal Filtering in different computational contexts, highlighting their strengths, limitations, and use cases.

Factor	Spatial Filtering	Temporal Filtering
Effectiveness	Highly effective for static image analysis.	Crucial for tasks involving sequential data, such as action recognition.
Efficiency	Utilizes shared weights and sparse connectivity efficiently.	Computationally intensive due to additional temporal dimensions.
Feature Detection	Detects both low- and high-level spatial features.	Captures temporal relationships and motion patterns effectively.
Limitations	Ineffective in capturing temporal dynamics, limiting applications	Prone to overfitting when applied to limited datasets.

	in videos or sequential data.	
	Struggles with features spanning multiple frames, reducing contextual understanding.	
Applications	Image classification.	Video classification by capturing motion across frames.
	Extracting spatial features from video frames.	Understanding sequential data in action recognition.
	Initial feature extraction in action recognition.	Analyzing temporal trends in time-series data.
	Limited utility in time-series analysis.	

6. Conclusion

Spatial and temporal filtering are complementary tools in CNNs. Spatial filtering efficiently captures static spatial features, making it ideal for image processing. Temporal filtering extends this capability to sequential data, enabling dynamic feature extraction. By understanding their unique advantages and limitations, researchers can design hybrid architectures to address complex spatiotemporal problems [1].

- Spatial Filtering is optimal for static image processing and initial feature extraction.
- Temporal Filtering is indispensable for dynamic data analysis, particularly in video and time-series applications.
- Future research could explore hybrid approaches integrating both techniques for enhanced performance across various domains.

7. Acknowledgments

We would like to express our sincere gratitude to our institutions for providing the necessary resources and support throughout this research. We extend our appreciation to our mentors, colleagues, and peers for their valuable insights and constructive feedback, which greatly contributed to refining our study. Additionally, we acknowledge the contributions of researchers and scholars whose work laid the foundation for our investigation. Finally, we are thankful to our families and friends for their encouragement and unwavering support during this research journey.

8. References

1. Chen X, Wang Y, Zhang H. Deep analysis of CNN-based spatio-temporal representations for action recognition. In: Proceedings of CVPR 2021 [Internet]. 2021 [cited 2025 Apr 14]. Available from: https://openaccess.thecvf.com
2. Gonzalez R, Woods R. Digital image processing. 4th ed. Pearson; 2018.
3. Donahue J, Hendricks LA, Guadarrama S, *et al.* Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):677-691.
4. Feichtenhofer C, Fan H, Malik J, *et al.* SlowFast networks for video recognition. In: Proceedings of ICCV. 2019. p. 6202-6211.
5. Hara K, Kataoka H, Satoh Y. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of CVPR. 2018. p. 6450-6459.
6. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):221-231.
7. Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01. 2012.
8. Kobayashi Y, Tanaka S, Nakamura T. Spatio-temporal filter analysis improves 3D-CNN for action classification. In: Proceedings of WACV 2024 [Internet]. 2024 [cited 2025 Apr 14]. Available from: https://openaccess.thecvf.com
9. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Adv Neural Inf Process Syst. 2012.
10. Lea C, Flynn MD, Vidal R, *et al.* Temporal convolutional networks: A unified approach to action segmentation. In: Proceedings of CVPR. 2017.
11. Lea C, Vidal R, Reiter A, *et al.* Temporal convolutional networks for action segmentation in videos. In: Proceedings of CVPR. 2017. p. 5695-5704.
12. Krizhevsky A. Learning multiple layers of features from tiny images [Master's thesis]. 2009.
13. Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. In: Proceedings of ICCV. 2019. p. 7083-7093.
14. Qiu Z, Yao T, Mei T. Spatiotemporal joint filter decomposition in 3D convolutional neural networks. In: Adv Neural Inf Process Syst (NeurIPS) [Internet]. 2021 [cited 2025 Apr 14]. Available from: https://proceedings.neurips.cc
15. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Adv Neural Inf Process Syst. 2014.
16. Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE Int Conf on Computer Vision (ICCV). 2015. p. 4489-4497.
17. Tran D, Wang H, Torresani L, *et al.* Video classification with channel-separated convolutional networks. In: Proceedings of ICCV. 2019. p. 5552-61.
18. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.
19. Wang H, Gupta A. Comparative analysis of CNN-based spatiotemporal reasoning in video understanding. Semantics Scholar [Internet]. 2018 [cited 2025 Apr 14]. Available from: https://semanticscholar.org
20. Wang L, Xiong Y, Wang Z, *et al.* Temporal segment networks: Towards good practices for deep action recognition. In: European Conf on Computer Vision (ECCV). 2016.
21. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of CVPR. 2018. p. 7794-803.
22. Xie S, Girshick R, Dollár P, *et al.* Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of ECCV. 2018. p. 305-321.
23. Zhou B, Andonian A, Oliva A, Torralba A. Temporal relational reasoning in videos. In: Proceedings of ECCV. 2018. p. 803-818.
24. Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding. In: Proceedings of ECCV. 2018. p. 695-712.