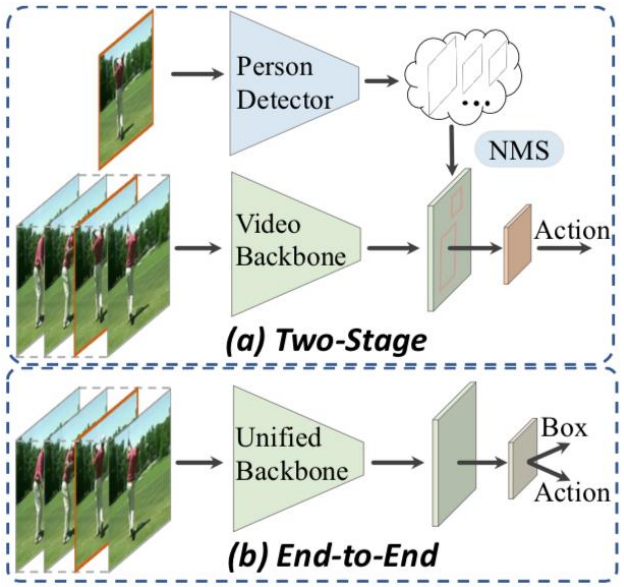
Learning Visual Representation via Neural Architecture Search

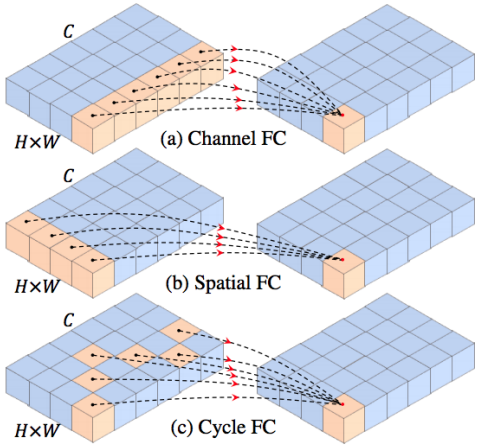
➢ Watch Only Once (WOO) for video action detection.

- We propose an end-to-end framework for video action detection, which directly produces the bounding boxes and action classes simultaneously, given a video clip as input. Our framework does not need an independent person detector.
- We propose a spatial-and-temporal embedding, and an embedding interaction mechanism, which improve discriminative ness of the features for action classification. A spatial-temporal fusion module is further proposed to aggregate features from spatial and temporal dimensions.
- Extensive experiments on AVA and JHMDB demonstrate that the performance of WOO could outperform or on par with previous well-established and complicated two-stage action detectors, while still reducing up to 16.7% FLOPs.



➢ New Neural Network Design.

- We propose a new MLP-like operator, Cycle FullyConnected Layer (Cycle FC), which is a generic, plug-and-play replacement of Spatial FC, enabling MLP like models to work in a scenario where input scales are flexible. Moreover, Cycle FC has a linear computational complexity to input resolution, while the complexity of Spatial FC is quadratic.
- With Cycle FC, we build a family of MLP-like architectures, learning pyramid feature representation for dense prediction tasks. To our knowledge, CycleMLP provides the first comprehensive baselines for both detection and segmentation tasks.
- We conduct extensive experiments on ImageNet classification, COCO object instance detection, and segmentation, and ADE20K semantic segmentation. The experimental results demonstrate that CycleMLP outperforms existing MLP-like models. Furthermore, CycleMLP is comparable to and sometimes better than CNNs and Transformers on dense predictions.



FC	$\mathcal{O}(HW)$	Scale Variable	ImgNet Top-1	COCO AP	ADE20K mIoU
Channel	HW	✓	79.4	35.0	36.3
Spatial	H^2W^2	✗	80.9	✗	✗
Cycle	HW	✓	81.6	41.7	42.4

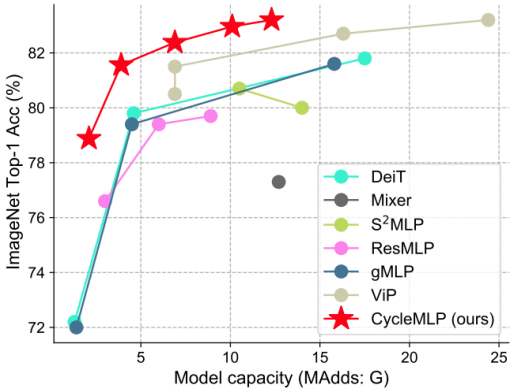


Table 3 The accuracy-FLOPs tradeoff of CycleMLP consistently outperforms existing MLP-like models under a wide range of FLOPs, which we attribute to the effectiveness of our Cycle FC.

➢ Progress.

2021.08.01 – 2021.10.30	Propose a self-supervised representation learning algorithm in videos
2021.11.01 – 2022.03.31	Propose a NAS algorithm to recognize actions in videos

➢ Research outputs.

1. Patent IP01101 一种基于端到端训练测试的动作检测算法，Ping Luo, Shoufa Chen, Lan Ma
2. Patent IP01102 一种端到端的视频时序动作提名生成算法，Ping Luo, Jiannan Wu, Jiajun Shen
3. Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo, CycleMLP: A MLP-like architecture for dense prediction, arXiv:2107.10224, 2021a
4. Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan, Wu, Lan Ma, Jiajun Shen, and Ping Luo, Watch Only Once: An End-to-End Video Action Detection Framework, ICCV 2021