MorphMLP: An Efficient MLP-Like Backbone for Spatial-Temporal Representation Learning

David Junhao Zhang[1⋆†], Kunchang Li[3,4⋆], Yali Wang[3⋆], Yunpeng Chen[2], Shashwat Chandra[1], Yu Qiao[3,5], Luoqi Liu[2], Mike Zheng Shou[1✉]

[1]National University of Singapore [2]Meitu, Inc
[3]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences [5]Shanghai AI Laboratory

Abstract. Recently, MLP-Like networks have been revived for image recognition. However, whether it is possible to build a generic MLP-Like architecture on video domain has not been explored, due to complex spatial-temporal modeling with large computation burden. To fill this gap, we present an efficient self-attention free backbone, namely MorphMLP, which flexibly leverages the concise Fully-Connected (FC) layer for video representation learning. Specifically, a MorphMLP block consists of two key layers in sequence, i.e., $\texttt{MorphFC}_\texttt{s}$ and $\texttt{MorphFC}_\texttt{t}$, for spatial and temporal modeling respectively. $\texttt{MorphFC}_\texttt{s}$ can effectively capture core semantics in each frame, by progressive token interaction along both height and width dimensions. Alternatively, $\texttt{MorphFC}_\texttt{t}$ can adaptively learn long-term dependency over frames, by temporal token aggregation on each spatial location. With such multi-dimension and multi-scale factorization, our MorphMLP block can achieve a great accuracy-computation balance. Finally, we evaluate our MorphMLP on a number of popular video benchmarks. Compared with the recent state-of-the-art models, MorphMLP significantly reduces computation but with better accuracy, e.g., MorphMLP-S only uses 50% GFLOPs of VideoSwin-T but achieves 0.9% top-1 improvement on Kinetics400, under ImageNet1K pretraining. MorphMLP-B only uses 43% GFLOPs of MViT-B but achieves 2.4% top-1 improvement on SSV2, even though MorphMLP-B is pretrained on ImageNet1K while MViT-B is pretrained on Kinetics400. Moreover, our method adapted to the image domain outperforms previous SOTA MLP-Like architectures. Code is available at https://github.com/MTLab/MorphMLP.

Keywords: MLP, Video Classification, Representation Learning

1 Introduction

Since the seminal work of Vision Transformer (ViT) [14], attention-based architectures have shown the great power in a variety of computer vision tasks, ranging from image domain [40, 13, 65, 73] to video domain [3, 46, 41, 45, 75]. However,

⋆ Contribute equally. †Work is done during internship at Meitu, Inc.

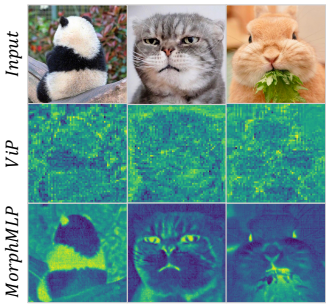

Fig. 1: Visualization of spatial feature in 3rd layer.

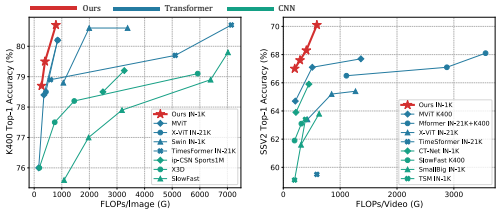

Fig. 2: Our MorphMLP vs. other SOTA Transformers and CNNs for video classification. Left: Kinetics400 [6]; Right: SthV2 [21].

recent studies have demonstrated that, self-attention maybe not critical and it can be replaced by simple Multiple Layer Perceptron (MLP) [55]. Following this line, a number of MLP-Like architectures have been developed on image-domain tasks with promising results [24, 8, 39, 72, 56, 55].

A natural question is that, is it possible to design a generic MLP-Like architecture for video domain? Unfortunately, it has not been explored in the literature, to our best knowledge. Motivated by this fact, we analyze the main challenges of using MLP on spatial-temporal representation learning. **First**, from the *spatial* perspective, we find that the current MLP-Like models lack progressive understanding of semantic details. This is mainly because that, they often operate MLP globally on all the tokens in the space, while ignoring hierarchical learning of visual representation. For illustration, we visualize the feature map of the well-known MLP-like model (i.e., ViP [24]) in Fig.1. Clearly, it suffers from difficulty in capturing key details, even in the shallow layer. Hence, how to discover semantics in each frame is important for designing spatial operation of MLP-like video backbone. **Second**, from the *temporal* perspective, the critical challenge is to learn long-range dependencies over frames. As shown in Fig. 2, the current video-based transformers can leverage self-attention to achieve this goal, but with huge computation cost. Hence, how to efficiently replace self-attention for long-range aggregation is important for designing temporal operation of MLP-like video backbone.

To tackles these challenges, we propose an effective and efficient MLP-like architecture, namely MorphMLP, for video representation learning. Specifically, it consists of two key layers, i.e., $\texttt{MorphFC}_\texttt{s}$ and $\texttt{MorphFC}_\texttt{t}$, which leverage the concise FC operations on spatial and temporal modeling respectively. Our $\texttt{MorphFC}_\texttt{s}$ can effectively capture core semantics in the space, as shown in Fig. 1. The main reason is that, we gradually expand the receptive field of visual tokens along both height and width dimensions as shown in Fig. 3. Such progressive token design brings two advantages in spatial modeling, compared with the existing MLP-like models, e.g., ViP [24]. First, it can learn hierarchical token interactions

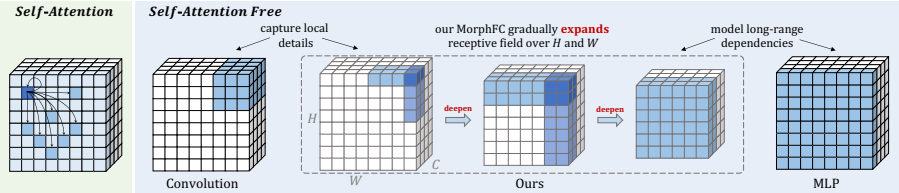

Fig. 3: Overview of progressive token construction in MorphMLP.

to discover the discriminative details, by operating FC from small to big spatial regions. Second, such small-to-big token construction can effectively reduce computation of FC operation for spatial modeling.

Moreover, our $\texttt{MorphFC}_\texttt{t}$ can adaptively capture long-range dependencies over frames. Instead of exhausting token comparison in self-attention, we concatenate the features of each spatial location across all frames into a temporal chunk. In this way, each temporal chunk can be processed efficiently by FC, which adaptively aggregates token relations in the chunk to model temporal dependencies. Finally, we build up a MorphMLP block by arranging $\texttt{MorphFC}_\texttt{s}$ and $\texttt{MorphFC}_\texttt{t}$ in sequence, and stack these blocks into our generic MorphMLP backbone for video modeling. On one hand, such hierarchical manner can enlarge the cooperative power of $\texttt{MorphFC}_\texttt{s}$ and $\texttt{MorphFC}_\texttt{t}$ to learn complex spatial-temporal interactions in videos. On the other hand, such multi-scale and multi-dimension factorization allows our MorphMLP to achieve a preferable balance between accuracy and efficiency.

To our best knowledge, we are the first to build efficient MLP-Like architecture for video domain. Compared with the recent state-of-the-art video models, MorphMLP significantly reduces computation but with better accuracy.

We further apply our architecture to an image classification task on ImageNet-1K[12] and a semantic segmentation task on ADE20K[76], by simply removing the temporal dimension of the video. Our method adapted to the image domain achieves competitive results compared to previous SOTA MLP-Like architectures.

2 Related Work

Self-Attention based backbones. Vision Transformer (ViT) [14] firstly applies Transformer architecture to a sequence of image tokens. It utilizes multi-head self-attention to capture long-range dependencies, thus achieving surprising results on image classification. Following works[40, 68, 13, 73, 65, 51] make a series of breakthroughs to achieve state-of-art performance on several image tasks, i.e., semantic segmentation [69, 29] and object detection [5, 77]. In video domain, a couple of woks [46, 3, 75, 71, 41, 16] explore space-time self-attention to model spatial-temporal relation and achieve state-of-the-art performance. It seems that

self-attention based architectures have been gradually dominating the computer vision community.

In this paper, we aim to explore a simple yet effective self-attention free architecture, which builds upon the FC layer to extract features. Our comparisons show that MorphMLP can achieve competitive results compared with Transformers not only in images but also in videos without self-attention layers.

CNN based backbones. CNNs [26, 25, 49, 47, 70, 34, 33] have dominated vision tasks in the past few years. In image domain, beginning with AlexNet[31], more effective and deeper networks, VGG[50], GoogleNet[52], ResNet[23], DenseNet[27] and EfficentNet[53] are proposed and achieve great success in computer vision. In the video domain, several works[59, 7, 61, 19] explore how to utilize convolution to learn effective spatial-temporal representation. However, the typical spatial and temporal convolution are so local that they struggle to capture long-range information well even if stacked deeper. A series of works propose efficient modules (e.g., Non-local[66], Double Attention[9]) to enhance local features via integrating long-range relation. The improvement of these methods can not be achieved without the supplement of self-attention layers.

In contrast, we propose the MorphMLP, which is self-attention free but not limited to capture local structure. The FC filter of `MorphFC` operates from small to big spatial regions. Meanwhile, the $\texttt{MorphFC}_\texttt{t}$ can capture long-term temporal information.

MLP-Like based backbones. Recent works[56, 39, 55, 72] try to replace self-attention layer with FC layer to explore the necessity of self-attention in Transformer architecture. But they suffer from dense parameters and computation. [24, 22, 54] apply FC layer along horizontal, vertical, and channel directions, respectively, in order to reduce the number of parameters and computation cost. However, the parameters of FC layer are still determined by the input resolution, so it is hard to handle different image scales. CycleMLP [8] addresses such problem with padding, but it only focuses on global information, ignoring local inductive bias. Meanwhile, the ability of MLP-Like architecture for video modeling has not been explored.

On the contrary, our MorphMLP can cope with diverse scales via splitting the sequence of tokens into chunks. Furthermore, it is able to effectively capture local to global information by gradually expanding chunk length. More importantly, we are the first to build MLP-Like architecture on videos to explore its generalization ability as a new paradigm of versatile backbone.

3 Method

In this section, we present our MorphMLP. We first introduce the two critical components of MorphMLP, $\texttt{MorphFC}_\texttt{s}$ and $\texttt{MorphFC}_\texttt{t}$. Then, we illustrate how to build efficient spatial-temporal MorphMLP block. Finally, the overall spatial-temporal network architecture and its adaption to image domain are provided.

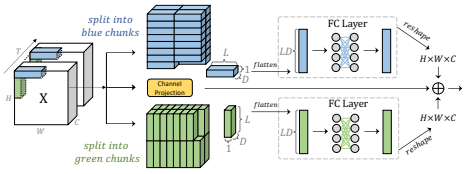

Fig. 4: `MorphFC`$_s$ on the spatial dimension. Note that chunk length L hierarchically expands as network goes deeper.

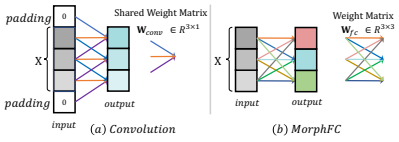

Fig. 5: Comparison with the typical convolution.

3.1 MorphFC for Spatial Modeling

As discussed above, mining core semantics is critical to video recognition. Typical CNN and previous MLP-Like architectures only focus on either local or global information modeling thus they fail to do that. To tackle this challenge, we propose a novel `MorphFC`$_s$ layer that can hierarchical expand the receptive field of FC and make it operate from small to big regions. Our `MorphFC`$_s$ processes each frame of video independently in horizontal and vertical pathways. We take the horizontal one (blue chunks in Fig. 4) for example.

Specifically, given one frame of input videos $\mathbf{X} \in \mathbb{R}^{HW \times C}$ that has been projected into a sequence of tokens, we first split $\mathbf{X}$ along horizontal direction. We set chunk length to L and thus obtain $\mathbf{X}_i \in \mathbb{R}^{L \times C}$, where $i \in \{1, ..., HW/L\}$. Furthermore, to reduce computation cost, we also split each $\mathbf{X}_i$ into multiple groups along channel dimension, where each group has D channels. Thus we get split chunks, and each single chunk is $\mathbf{X}_i^k \in \mathbb{R}^{LD}$, where $k \in \{1, ..., C/D\}$. Next, we flatten each chunk into 1D vector and apply a FC weight matrix $\mathbf{W} \in \mathbb{R}^{LD \times LD}$ to transform each chunk, yielding

$$\mathbf{Y}_i^k = \mathbf{X}_i^k \mathbf{W}. \tag{1}$$

After feature transformation, we reshape all chunks $\mathbf{Y}_i^k$ back to the original dimension $\mathbf{Y} \in \mathbb{R}^{H \times W \times C}$. The vertical way (green chunks in Fig. 4) does likewise except splitting the sequence of tokens along vertical direction. To make communication among groups along channel dimension, we also apply a FC layer to process each token individually. Finally, we get the output by element-wise summing horizontal, vertical, and channel features together. The chunks length L hierarchically increases as the network deepens, thereby enabling the FC filter to discover more core semantics progressively from small to big spatial region.

Difference between our MorphFC$_s$ and convolution. (i) Typical convolution utilizes fixed small kernel size (e.g., 3×3), which only aggregates local context. On the contrary, the chunks lengths in `MorphFC`$_s$ hierarchically increase as the network deepens, which can model short-to-long range information progressively. **(ii)** Convolution uses sliding windows to obtain overlapping tokens, which requires cumbersome operations, including unfold, reshape and fold. In contrast, we simply reshape the feature map to obtain our chunks with non-overlapping tokens. **(iii)** As shown in Fig. 5, given a 1×3 input, to get the 1×3

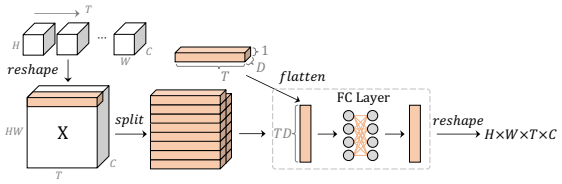

Fig. 6: $\texttt{MorphFC}_\texttt{t}$ on the temporal dimension.

output, the convolution kernel of 1×3 window size needs to slide three times, and each 1×1 output is generated by the shared weight matrix $\mathbf{W}_{conv} \in \mathbb{R}^{3\times1}$. In contrast, FC layer applies weight matrix $\mathbf{W}_{fc} \in \mathbb{R}^{3\times3}$ to the input yielding 1×3 ouput. Each 1×1 output is equivalent to being generated by non-shared weight matrix $\mathbf{W} \in \mathbb{R}^{3\times1}$, which brings more flexible spatial encoding than convolution. **Comparisons with ViP [25].** Our design is related to the well-known ViP designed for image domain, which also leverages the multi-branch features in spatial modeling. Hence, we further discuss the differences. **(i)** The FC filters of whole ViP network have the fixed size and receptive field, thus they only capture global information. On the contrary, our FC filters are morphable, as shown in Fig. 3. In shallow layers, they have small size to model local structure, while in deeper layers, they gradually change to large size to model long-range information. Hence, ours can discover more detailed semantics by progressively operating FC from small to big spatial region. **(ii)**As shown in Fig. 8, at the network level, ours have hierarchical downsampling after each stage but ViP does not. **(iii)** As ViP paper said, ViP is hard to transfer to downstream tasks i.e. segmentation with spatial resolution 2048×512, since its filter size is always equal to the height/weight of features. But it is easy for ours, because the filter size is equal to pre-defined chunk size in the pre-training.

3.2 MorphFC$_t$ on Temporal Modeling

In addition to the horizontal and vertical pathways in $\texttt{MorphFC}_\texttt{s}$, we introduce another temporal pathway $\texttt{MorphFC}_\texttt{t}$. It aims at capturing long-term temporal information using the simple FC layer with low computation cost. Specifically, as shown in Fig. 6, given an input video clip tokens $\mathbf{X} \in \mathbb{R}^{H\times W\times T\times C}$, we first split X into a couple of groups along channel dimension (D channels in each group) to reduce computation cost and get $\mathbf{X}^k \in \mathbb{R}^{H\times W\times T\times D}$, where $k \in \{1, ..., C/D\}$. For each spatial position s, we concatenate features across all frames into a chunk $\mathbf{X}_s^k \in \mathbb{R}^{TD}$, where $s \in \{1, ..., HW\}$. Then we apply a FC matrix $\mathbf{W} \in \mathbb{R}^{TD\times TD}$, to transform temporal features and get

$$\mathbf{Y}_s^k = \mathbf{X}_s^k \mathbf{W}. \tag{2}$$

Finally, we reshape all chunks $\mathbf{Y}_s^k \in \mathbb{R}^{TD}$ back to original tokens dimension and output $\mathbf{Y} \in \mathbb{R}^{H\times W\times T\times C}$. In this way, the FC filter can simply aggregate token relations along time dimension in the chunk to model temporal dependencies.

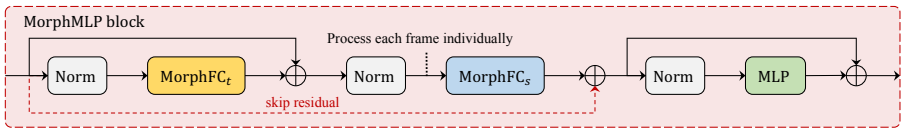

Fig. 7: Spatial-Temporal MorphMLP Block.

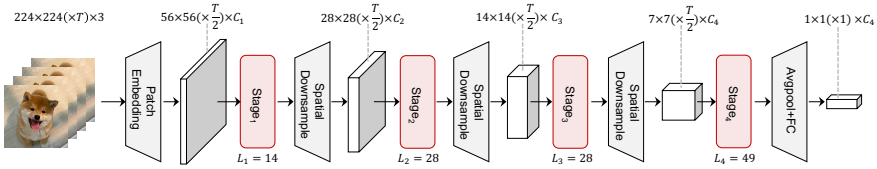

Fig. 8: Architecture of our MorphMLP L means chuck length.

Spatial-Temporal MorphMLP block. Based on the `MorphFC`$_s$ and `MorphFC`$_t$, we propose a factorized spatial-temporal MorphMLP block in the video domain for efficient video representation learning. As shown in Fig. 7, our MorphMLP block contains `MorphFC`$_t$, `MorphFC`$_s$ and MLP [62] modules in a sequential order. On one hand, it is difficult for joint spatial-temporal optimization [3]. On the other hand, factorizing spatial and temporal modeling is able to reduce the computation cost significantly. Therefore, we place temporal and spatial `MorphFC`$_s$ layers in the sequential style. The LN [2] layer is applied before each module, and the standard residual connections are used after `MorphFC`$_t$ and MLP module. Instead of applying a standard residual connection [23] after `MorphFC`$_s$, we add a skip residual connection (red line) between the original input and output features from `MorphFC`$_s$ layer. We found that such a connection can make training more stable.

3.3 Network architecture

For video recognition, as shown in Fig. 8, we hierarchically stack spatial-temporal MorphMLP blocks to build up our network. Given an video sequence $\mathbf{X} \in \mathbb{R}^{H\times W\times T\times 3}$, taking H=W=224 for example, our MorphMLP backbone first performs patch embedding on the video clip and gets a sequence of tokens with dimension $56\times56\times T/2 \times C_1$. Then, we have four sequential stages and each of them contains a couple of MorphMLP blocks. The feature size remains unchanged as passing through layers inside the same stage. At the end of each stage excluding the last one, we expand the channel dimension and downsample the spatial resolution of features by ratio 2.

Note that we set chunk lengths of `MorphFC`$_s$ to be 14, 28, 28, 49 for stage 1-4, respectively. Horizontal/vertical chunks with lengths 14, 28, 28, 49 of stage 1-4 can cover quarter, one, two, all rows/columns of feature maps of stage 1-4, re-

spectively. In shallow layers, our network can learn detailed representation from the local spatial context in small chunk length, e.g., length 14 for $56\times56\times C_1$ feature map. In deep layers, our network can capture long-range information from the global semantic context in considerable chunk length, e.g., length 49 for $7\times7\times C_4$ feature map. With downsampling the spatial resolution and expanding chunk length as the network goes deeper, our MorphMLP is capable of discovering more core semantics progressively by operating the FC filter from small to big spatial regions.

We provide two model variants for the video recognition depending on the number of MorphMLP blocks in four stages: $\{3, 4, 9, 3\}$ for MorphMLP-Small(S) and $\{4, 6, 15, 4\}$ for MorphMLP-Base(B). The numbers of channels of four stages are $\{112, 224, 392, 784\}$. Additionally, for image-domain architecture, we simply exclude the temporal dimension and drop $\texttt{MorphFC}_\texttt{t}$ in the MorphMLP block. In addition to small and base settings, we provide two extra model variants for image domain, depending on the number of MorphMLP blocks in four stages, i.e., $\{3, 4, 7, 3\}$ for MorphMLP-Tiny(T) and $\{4, 8, 18, 6\}$ for MorphMLP-Large(L).

4 Experiment

In this section, we first examine the performance of MorphMLP and evaluate its spatiotemporal effectiveness on Kinetics-400[6], and Something-Something V1&V2 [21] datasets. For fair comparisons and due to GPU resources limitation, we only report MorphMLP-S and B for video classification. Then we verify the effectiveness of its adaption to image domain, including ImageNet-1K [12] image classification and ADE20K[76] semantic segmentation.

4.1 Video Classification on Kinetics-400

Settings. Kinetics-400[6] is a large-scale scene-related video benchmark. It contains around 240K training videos and about 20K validation videos in 400 classes. Our code heavily relies on `PySlowFast` [15] repository and the training recipe mainly follows MViT[16]. We directly load the parameters of $\texttt{MorphFC}_\texttt{s}$ pre-trained on ImageNet and randomly initialize the parameters of $\texttt{MorphFC}_\texttt{t}$ in the video domain. We adopt a dense sampling strategy [66] and AdamW optimizer to train the whole network. The warm-up epoch, total epoch, batch size, base learning rate, and weight decay are 10, 60, 64, 2e-4, and 0.05 respectively. We utilize the stochastic depth rates 0.1 and 0.3 for MorphMLP-S and B.

Results. As shown in Table 1, our method achieves outstanding performance with fewer computation costs. Compared with CNN models such as SlowFast[19], our MorphMLP requires 8× fewer GFLOPS but achieves 1.9% accuracy improvement (80.8% *vs.* 78.9%). With only ImageNet-1K pre-training, our method surpasses most of the self-attention based Transformer backbones with larger dataset pre-training. For example, compared with ViViT-L[1] pre-trained on ImageNet-21K, our MorphMLP obtains better performance with 20× fewer computations. When our model is scaled larger, the accuracy increases as well. Since

Table 1: Comparisons with the state-of-the-art on Kinetics-400[6]. Our MorphMLP achieves outstanding results with much fewer computation costs. For example, compared with VideoSwin-T, our MorphMLP-S only requires 2× fewer GFLOPs but gets 0.9% accuracy improvement (79.7% *vs.* 78.8%).

Method	Pretrain	#Frame	GFLOPs	K400	
				Top-1	Top-5
Self-Attention Free – CNN					
SlowFast R101[19]	-	(16+64)×3×10	6390	78.9	93.5
CorrNet-101[63]	-	32× 3×10	6720	79.2	-
ip-CSN[60]	Sports1M	32× 3×10	3264	79.2	93.8
X3D-XL[18]	-	16×3×10	1452	79.1	93.9
SmallBig$_{EN}$[35]	IN-1K	(8+32)×3×4	5700	78.7	93.7
TDN$_{EN}$[64]	IN-1K	(8+16)×3×10	5940	79.4	94.4
CT-Net$_{EN}$[32]	IN-1K	(16+16)×3×4	2641	79.8	94.2
Self-Attention Based – Transformer					
Timesformer-L[3]	IN-21K	96×3×1	7140	80.7	94.7
VidTr-L[75]	IN-21K	32×3×10	11760	79.1	93.9
ViViT-L[1]	IN-21K	16×3×4	17357	80.6	94.7
X-ViT[4]	IN-21K	16×3×1	850	80.2	94.7
Mformer[46]	IN-21K	32×3×10	11085	80.2	94.8
Mformer-L[46]	IN-21K	32×3×10	35550	80.2	94.8
MViT-B,16×4 [16]	-	16×1×5	355	78.4	93.5
MViT-B,32×3 [16]	-	32×1×5	850	80.2	94.4
VideoSwin-T[41]	IN-1K	32×3×4	1056	78.8	93.6
VideoSwin-B[41]	IN-1K	32×3×4	3384	80.6	94.6
Self-Attention Free – MLP-Like					
MorphMLP-S	IN-1K	16×1×4	268	78.7	93.8
MorphMLP-S	IN-1K	32×1×4	532	79.7	94.2
MorphMLP-B	IN-1K	16×1×4	392	79.5	94.4
MorphMLP-B	IN-1K	32×1×4	788	**80.8**	**94.9**

the computation cost is relatively low, our method still has great potential for better performance. It demonstrates that our MorphMLP is a strong MLP-Like backbone for video recognition.

4.2 Video Classification on Something-Something

Settings. Something-Something [21] is another large-scale dataset, in which the temporal relationship modeling is critical for action understanding. It includes two versions, i.e., V1 and V2, both of which contain plentiful videos over 174 categories. We adopt the same training setting as used for Kinetics-400, except that a random horizontal flip is not applied. We utilize the sparse sampling strategy. The warm-up epoch, total epoch, batch size, base learning rate, and weight decay are 5, 50, 64, 4e-4, and 0.05, respectively. We set the stochastic depth rates to be 0.3 and 0.6 for Morph-S and B respectively.

Results. The comparison results on Something V2&V1 are shown in Table 2 and Table 3 respectively. For SSV2, CNN architectures perform worse than Transformer architectures since they are limited to capturing local spatial and temporal information and struggle to model long-term dependencies. Transformer architectures can achieve better results, but they heavily rely on large-scale dataset pre-training which requires high computation. Compared with CT-Net[32], our

Table 2: Comparisons with the SOTA on SSV2 [21]. Our MorphMLP outperforms previous sota Transformers and CNNs with IN-1K pretraining only.

Method	Pretrain	#Frame	GFLOPs	SSV2	
				Top-1	Top-5
Self-Attention Free – CNN					
SlowFast R50[19]	K400	(8+32)×3×1	197	61.7	46.6
TSM[38]	K400	16×3×2	374	63.4	88.5
STM[28]	IN-1K	16×3×10	1995	64.2	89.8
bLVNet[17]	IN-1K	32×3×10	3870	65.2	90.3
TEA[36]	IN-1K	16×3×10	2100	65.1	-
CT-Net[32]	IN-1K	16×3×2	450	65.9	90.1
Self-Attention Based – Transformer					
Timesformer[3]	IN-21K	16×3×1	5109	62.5	-
VidTr-L[75]	IN-21K+K400	32×3×10	10530	60.2	-
ViViT-L[1]	IN-21K+K400	16×3×4	11892	65.4	89.8
X-ViT[4]	IN-21K	32×3×1	1269	65.4	90.7
Mformer[46]	IN-21K+K400	16×3×1	1110	66.5	90.1
Mformer-L[46]	IN-21K+K400	32×3×1	3555	68.1	91.2
MViT-B,16×4[16]	K400	16×3×1	510	67.1	90.8
MViT-B,32×3[16]	K400	32×3×1	1365	67.7	90.9
MViT-B-24,32×3[16]	K600	32×3×1	708	68.7	91.5
Self-Attention Free – MLP-like					
MorphMLP-S	IN-1K	16×3×1	201	67.1	90.9
MorphMLP-S	IN-1K	32×3×1	405	68.3	91.3
MorphMLP-B	IN-1K	16×3×1	294	67.6	91.3
MorphMLP-B	IN-1K	32×3×1	591	**70.1**	**92.8**

Table 3: Comparisons with the state-of-the-art on Something-Something V1 [21].

Method	Pretrain	#Frame	GFLOPs	SSV1	
				Top-1	Top-5
I3D[67]	IN-1K+K400	32×3×2	918	41.6	72.2
NLI3D[67]	IN-1K+K400	32×3×2	1008	44.4	76.0
NLI3D+GCN[67]	IN-1K+K400	32×3×2	1818	46.1	76.8
TSM[38]	IN-1K+K400	16×1×1	65	47.2	77.1
SmallBig[35]	IN-1K	16×1×1	105	49.3	79.5
TEINet[42]	IN-1K	16×3×10	1980	51.0	-
TEA[36]	IN-1K	16×3×10	2100	52.3	81.9
CT-NET[32]	IN-1K	16×3×2	447	53.4	81.7
MorphMLP-S	IN-1K	16×1×1	67	50.6	78.0
MorphMLP-S	IN-1K	16×3×1	201	53.9	81.3
MorphMLP-B	IN-1K	16×3×1	294	55.5	82.4
MorphMLP-B	IN-1K	32×3×1	591	**57.4**	**84.5**

MorphMLP can reduce 2.5× computation but achieves 1.2% accuracy gain. Compared with the-state-of-art method MViT[16], which is pre-trained on large video dataset Kinetics-600, our MorphMLP only pre-trained on ImageNet-1K can obtain better performance (70.1% *vs.* 68.7%) with smaller GFLOPS (591G *vs.* 708G). For SSV1, our MorphMLP also achieves outstanding results.

The superior results of our method on this dataset can be attributed to our unique progressively core semantics discovering manner and efficient spatial-temporal block design in MorpMLP. Table 6 and 7 can also demonstrate our point. Note that even if we do not add any complicated and unique temporal attention operation, our simple method can achieve such great performance. This indicates that our model can serve as a strong backbone for further improvement.

Table 4: ImageNet-1K results. As shown in (a), our method achieves the best performance among SOTA MLP-Like architectures. From (b), we can see that our MorphMLP also achieves the comparable results with SOTA self-attention based and hybrid models even with small computation.

(a) Comparisons with MLP-Like models.

Model	Param	FLOPs	Top-1
Mixer-B/16[55]	59M	12.7G	76.4
Mixer-B/16†[55]	59M	12.7G	77.3
ResMLP-S12[56]	15M	3.0G	76.6
ResMLP-S24[56]	30M	6.0G	79.4
ResMLP-B24 [56]	116M	23.0G	81.0
gMLP-Ti[39]	6M	1.4G	72.3
gMLP-S [39]	20M	4.5G	79.6
gMLP-B [39]	73M	15.8G	81.6
S^2-MLP-wide [72]	71M	14.0G	80.0
S^2-MLP-deep [72]	51M	10.5G	80.7
ViP-Small/7 [24]	25M	6.9G	81.5
ViP-Medium/7[24]	55M	16.3G	82.7
ViP-Large/7 [24]	88M	24.4G	83.2
AS-MLP-T [37]	28M	4.4G	81.3
AS-MLP-S [37]	50M	8.5G	83.1
AS-MLP-B [37]	88M	15.2G	83.3
CycleMLP-B2[8]	27M	3.9G	81.6
CycleMLP–B3[8]	38M	6.9G	82.6
CycleMLP–B4[8]	52M	10.1G	83.0
CycleMLP–B5[8]	76M	12.3G	83.1
MorphMLP-T	23M	3.9G	81.6
MorphMLP-S	38M	6.9G	82.6
MorphMLP-B	58M	10.2G	83.2
MorphMLP-L	76M	12.5G	**83.4**

(b) Comparisons with SOTA models.

Model	Family	Scale	Param	FLOPs	Top-1
ResNet50 [23]	CNN	224^2	26M	4.1G	79.2
DeiT-S [57]	Trans	224^2	22M	4.6G	79.8
ResNest50[74]	CNN	224	28M	4.3G	80.6
T2T-ViT-14 [73]	Trans	224^2	22M	4.8G	81.5
PVT-S [65]	Trans	224^2	25M	3.8G	79.8
Swin-T [40]	Trans	224^2	29M	4.5G	81.3
GFNet-H-S [48]	FFT	224^2	32M	4.5G	81.5
BoT-S1-50 [51]	Hybrid	224^2	21M	4.3G	79.1
CoAtNet-0[11]	Hybrid	224^2	23M	4.2G	81.6
MorphMLP-T	MLP	224^2	23M	3.9G	**81.6**
ResNet101 [23]	CNN	224^2	45M	7.9G	79.8
ResNest101[74]	CNN	224^2	48M	8.0G	82.0
RegNetY-8G [47]	CNN	224^2	39M	8.0G	81.7
T2T-ViT-19 [73]	Tran	224^2	39M	8.5G	81.9
PVT-M [65]	Trans	224^2	44M	6.7G	81.2
BoT-S1-59 [51]	Hybrid	224^2	34M	7.3G	81.7
CoAtNet-1[11]	Hybrid	224^2	42M	8.4G	**83.3**
MorphMLP-S	MLP	224^2	38M	6.9G	82.6
ViT-B/16 [57]	Trans	384^2	86M	55.4G	77.9
DeiT-B [57]	Trans	224^2	86M	17.5G	81.8
DeiT-B [57]	Trans	384^2	86M	55.4G	83.1
T2T-ViT-24 [73]	Tran	224^2	64M	13.8G	82.3
Swin-B [40]	Trans	224^2	88M	15.4G	83.4
CaiT-S36 [58]	Trans	224^2	68M	13.9G	83.3
MorphMLP-L	MLP	224^2	76M	12.5G	**83.4**

4.3 Image Classification on ImageNet-1K

Settings. We train our models from scratch on the ImageNet-1K dataset [12], which consists of 1.2M training images and 50K validation images from 1,000 categories. Our code is implemented based on `DeiT`[57] repository, and we follow the same training strategy proposed in DeiT[57], including strong data augmentation and regularization. Stochastic depth rates are set to be 0.1, 0.1, 0.2, 0.3 for our 4 model variants. We adopt AdamW [43] optimizer with cosine learning rate schedule [44] for 300 epochs, while the first 20 epochs are used for linear warm-up[20]. The total batch size, weight decay, and initial learning rate are set to 1024, 5×10^{-2} and 0.01 respectively.

Results. As shown in Table 4a, our MorphMLP outperforms the state-of-the-art MLP-Like architectures. Compared with ViP-S[24], our method can get much higher accuracy (82.6% *vs.* 81.5%) with similar GFLOPS (7.0G *vs.* 6.9G). This demonstrates the effectiveness of our progressively short-to-long range pattern. In Table 4b, our MorphMLP can achieve competitive results with popular self-attention based models. Compared with other tiny models, e.g., Swin-T[40], our method can achieve better results (81.6% *vs.* 81.3%) with fewer parameters and

Table 5: Semantic segmentation with Semantic FPN [30] on ADE20K [76] val.

Method	Arch	#Param.(M)	mIoU
ResNet50[23]	CNN	28.5	36.7
PVT-S[65]	Trans	28.2	39.8
Swin-T[40]	Trans	31.9	41.5
GFNet-H-T[48]	FFN	26.6	41.0
CycleMLP-B2 [8]	MLP-Like	30.6	42.4
MorphMLP-T	MLP-Like	26.4	**43.0**
ResNet101[23]	CNN	47.5	38.8
ResNeXt101-32×4d[70]	CNN	47.1	39.7
PVT-M[65]	Trans	48.0	41.6
GFNet-H-S[48]	FFN	47.5	42.5
CycleMLP-B3[8]	MLP-Like	42.1	44.5
MorphMLP-S	MLP-Like	41.0	**44.7**
PVT-L[65]	Trans	65.1	42.1
Swin-S[40]	Trans	53.2	45.2
CycleMLP-B4 [8]	MLP-Like	55.6	45.1
MorphMLP-B	MLP-Like	59.3	**45.9**

GFLOPS (23M *vs.* 29M, 3.9G *vs.* 4.5G). As for larger settings, our method can achieve comparable result to Swin-B [40] with fewer GFLOPS.

4.4 Semantic Segmentation on ADE20K

Settings. We conduct semantic segmentation experiments on ADE20K[76], which consists of 20K training images and 2K validation images over 150 semantic categories. Our code is based on `mmsegmentation` [10] and we follow the experiment setting used in PVT[65]. We simply apply Semantic FPN [30] for fair comparisons, and all the backbones are pre-trained on ImageNet-1K. We adopt AdamW [43] optimizer with cosine learning rate schedule [44], while the initial learning rate is 1e-4. The input images are randomly resized and cropped to 512×512 for training, and the shorter sides of images are set to 512 while testing.
Results. The results on ADE20k dataset are shown in Table 5. Our MorphMLP outperforms ResNet[47] and PVT[65] significantly. Compared with Swin-T, our MorphMLP-T can achieve better mIoU with fewer parameters (26.4M *vs.* 31.9M).

4.5 Ablation Study

For Table 6, 8, and 9, we train all the models based on MorphMLP-T for 100 epochs on ImageNet. To explore the variants of our spatial-temporal design, we adopt MorphMLP-S as the backbone on SSV1.
Impact of chunk length. In the MorphMLP, we expand the chunk length gradually. The spatial resolutions of feature maps of Stages 1-4 are 56, 28, 14, 7, respectively. For the horizontal/vertical directions, chunk lengths 14, 28, 28, 49 in Stages1-4 can cover quarter, one, two and all rows/columns of the tokens, respectively, which can discover core semantics progressively by operating the FC filter from small to big spatial region.

As shown in Table 6, there are some alternative ways to set chunk length. The first line represents that $\texttt{MorphFC}_\texttt{s}$ in each stage covers one row of image/video

Table 6: Impact of chunk length.

Stage1	Stage2	Stage3	Stage4	SSV1 Top-1	ImageNet Top-1	ADE20K mIoU
56	28	14	7	48.6	79.1	41.6
3	3	3	3	48.0	78.2	41.0
7	7	7	7	48.4	79.0	41.9
14	14	14	14	48.7	79.0	42.0
14	28	28	7	49.2	79.3	42.0
28	28	28	49	49.1	79.4	42.3
14	28	28	49	**50.6**	**79.6**	**42.6**

Table 7: Detail designs of spatial-temporal MorphMLP block.

Method	Order	Standard Residual	Skip Residual	SSV1 Top-1
Parallel	T‖S	✔		49.2
Sequential	T+S	✔		49.8
	S+T	✔		50.2
	T+S		✔	**50.6**
	S+T		✔	31.7

tokens, which only models global information. The second, third and fourth line utilize the small chunk length, which only captures local structure. The results show that our progressively expanding pattern can perform better than the solely local or global pattern. The reason is that, in the shallow layer, the original texture and shape information of the image/videos is relatively intact. Therefore, it is critical to capture detailed structures in the early stage. The features in the deep layers cover more semantic information, thus long-range relation modeling is significant. **Note** that the improvement brought by expanding chunk lengths on video is larger than image because such pattern is conducive to discovering more fine-grained semantics for many tiny movement actions.

It is also worth noting that since chunk sizes are equal to H, W of features in each stage if input is 224×224, 1st row is no 'Morph' (progressively discovering core semantics), but with hierarchical downsampling only. Last row is our final model (w/ both 'Morph' and same downsampling as the 1st row). Comparisons show benefits are from MorphFC design instead of hierarchical downsampling.

Detail designs of spatial-temporal MorphMLP block. We explore some alternative designs for our spatial-temporal MorphMLP block in the Table 7. To begin with, in addition to applying $\texttt{MorphFC}_\texttt{t}$ and $\texttt{MorphFC}_\texttt{s}$ in a sequential way, we can add the features from $\texttt{MorphFC}_\texttt{t}$ and $\texttt{MorphFC}_\texttt{s}$ in parallel. As shown in Table 7, the parallel way performs worse than the sequential way. We argue that it is more difficult for joint spatial and temporal optimization. Moreover, we explore different spatial-temporal orders and residual connections. Standard residual refers to applying a residual connection after each module in MorphMLP block of Fig. 7. Skip residual means that a connection is applied between input features of MorphMLP block and output features of the $\texttt{MorphFC}_\texttt{s}$ (red line in Figure 7). The results show that sequential temporal and spatial order with skip residual connection is the optimal setting.

Comparisons with convolution. To compared with spatial convolution, we replace the $\texttt{MorphFC}_\texttt{s}$ layer with typical 3×3 and 7×7 convolution on image domain. As shown in Table 8, our $\texttt{MorphFC}_\texttt{s}$ can outperform typical convolution by a large margin. This demonstrates that typical convolution is difficult to capture long-range information, which is crucial to the recognition problem. Furthermore, we adopt two 1D group convolutions along the horizontal and vertical direction, whose kernel sizes are exactly the same as our chunk lengths

Table 8: Spatial design of MorphMLP.

Operation	#Param. (M)	FLOPs (G)	Throughput (images/s)	ImageNet Top-1
3×3 Conv	34.5	6.2	676	77.3
7×7 Conv	113	20.6	532	77.7
Group Conv	23.4	4.0	620	79.0
MorphFC$_s$	23.4	4.0	734	**79.6**

Table 9: Different operations.

Dimension	Style	Weight Sum	SSV1 Top-1	ImageNet Top-1
H+W+C	Transformer	✔	50.6	**79.6**
H+W	Transformer	✔	49.4	78.5
H+W+C	CNN	✔	47.2	77.2
H+W+C	Transformer	✗	50.2	79.3

Table 10: Temporal design.

Operation	#Param.(M)	FLOPs(G)	SSV1
3×1×1 Conv	46.0	62.7	47.9
5×1×1 Conv	52.5	72.9	48.6
MorphFC$_t$	47.0	66.4	**50.6**

Table 11: Training cost.

Video Model	TFLOPs	K400	Training	Cost
SlowFast	1.11	71.0	30 epoch	444h
Timesformer	0.59	75.8	30 epoch	416h
Morph-S	0.27	77.0	30 epoch	408h

in each stage. The results show that our method is much better than group conv in terms of speed and accuracy. This indicates the effectiveness of our MorphFC$_s$.

Moreover, we do comparisons between MorphFC$_t$ and typical temporal convolutions, i.e., 3×1×1 and 5×1×1. As shown in Table 10, our MorphFC$_t$ outperforms typical temporal convolutions greatly. This is because that typical convolutions only focus on local temporal information aggregation. On the contrary, our MorphFC$_t$ is able to capture long-term temporal dependencies.

Importance of different operations. We explore the importance of different operations in Table 9. First, we evaluate the necessity of FC layers from three directions. It shows that each direction plays an important role. Second, we replace the 3×3 convolution with our MorphFC$_s$ layer in ResNet[23]/R(2+1)D[61] and the result shows that Transformer structure is more suitable for our MorphFC than the bottleneck block of CNN. Third, following the ViP[24], we utilize a weighted sum after three directions FC layers. Results show that weighted sum can bring a slight improvement (0.3%).

Training speed. As shown in Table 11, considering speed and accuracy trade-off, our approach is more efficient for training with other SOTA video methods.

5 Conclusion

In this paper, we propose a self-attention free, MLP-Like backbone for video representation learning, named MorphMLP. MorphMLP is capable of progressively discovering core semantics and capturing long-term temporal information. To our best knowledge, we are the first to apply MLP-Like architecture in the video domain. The experiments demonstrate that such self-attention free models can be as strong as and even outperform self-attention based architectures.

Acknowledgements. This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008, and Mike Zheng Shou's Start-Up Grant from NUS. David Junhao Zhang is supported by NUS IDS-ISEP scholarship.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
2. Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. ArXiv **abs/1607.06450** (2016)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (ICML) (2021)
4. Bulat, A., Perez-Rua, J.M., Sudhakaran, S., Martinez, B., Tzimiropoulos, G.: Space-time mixing attention for video transformer. In: Advances in Neural Information Processing Systems (2021)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision (2020)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
8. Chen, S., Xie, E., GE, C., Chen, R., Liang, D., Luo, P.: CycleMLP: A MLP-like architecture for dense prediction. In: International Conference on Learning Representations (2022)
9. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Aˆ 2-nets: Double attention networks. Advances in neural information processing systems (2018)
10. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020)
11. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. In: Advances in Neural Information Processing Systems (2021)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition (2009)
13. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. ArXiv **abs/2107.00652** (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
15. Fan, H., Li, Y., Xiong, B., Lo, W.Y., Feichtenhofer, C.: Pyslowfast. https://github.com/facebookresearch/slowfast (2020)
16. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
17. Fan, Q., Chen, C.F.R., Kuehne, H., Pistoia, M., Cox, D.: More Is Less: Learning Efficient Video Representations by Temporal Aggregation Modules. In: Advances in Neural Information Processing Systems (2019)

18. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
20. Goyal, P., Dollár, P., Girshick, R.B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. ArXiv **abs/1706.02677** (2017)
21. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fründ, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
22. Guo, J., Tang, Y., Han, K., Chen, X., Wu, H., Xu, C., Xu, C., Wang, Y.: Hire-mlp: Vision mlp via hierarchical rearrangement. arXiv preprint arXiv:2108.13341 (2021)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
24. Hou, Q., Jiang, Z., Yuan, L., Cheng, M.M., Yan, S., Feng, J.: Vision permutator: A permutable mlp-like architecture for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
25. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv **abs/1704.04861** (2017)
26. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
27. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: Stm: Spatiotemporal and motion encoding for action recognition. 2019 IEEE International Conference on Computer Vision (ICCV) (2019)
29. Jin, Y., Han, D.K., Ko, H.: Trseg: Transformer for semantic segmentation. Pattern Recognit. Lett. (2021)
30. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
32. Li, K., Li, X., Wang, Y., Wang, J., Qiao, Y.: Ct-net: Channel tensorization network for video classification. In: International Conference on Learning Representations (2021)
33. Li, T., Ke, Q., Rahmani, H., Ho, R.E., Ding, H., Liu, J.: Else-net: Elastic semantic network for continual action recognition from skeleton data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
34. Li, T., Liu, J., Zhang, W., Duan, L.: Hard-net: Hardness-aware discrimination network for 3d early activity prediction. In: European Conference on Computer Vision. pp. 420–436. Springer (2020)

35. Li, X., Wang, Y., Zhou, Z., Qiao, Y.: Smallbignet: Integrating core and contextual views for video classification. 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
36. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
37. Lian, D., Yu, Z., Sun, X., Gao, S.: AS-MLP: An axial shifted MLP architecture for vision. In: International Conference on Learning Representations (2022)
38. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. 2019 IEEE International Conference on Computer Vision (ICCV) (2019)
39. Liu, H., Dai, Z., So, D., Le, Q.: Pay attention to mlps. Advances in Neural Information Processing Systems (2021)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
41. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. ArXiv **abs/2106.13230** (2021)
42. Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T.: Teinet: Towards an efficient architecture for video recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
43. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. ArXiv **abs/1711.05101** (2017)
44. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
45. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
46. Patrick, M., Campbell, D., Asano, Y., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., Henriques, J.F.: Keeping your eye on the ball: Trajectory attention in video transformers. Advances in Neural Information Processing Systems (2021)
47. Radosavovic, I., Kosaraju, R.P., Girshick, R.B., He, K., Dollár, P.: Designing network design spaces. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
48. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in Neural Information Processing Systems (2021)
49. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
50. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
51. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
52. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2015)
53. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning (2019)
54. Tang, C., Zhao, Y., Wang, G., Luo, C., Xie, W., Zeng, W.: Sparse mlp for image recognition: Is self-attention really necessary? arXiv preprint arXiv:2109.05422 (2021)

55. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A.P., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: MLP-mixer: An all-MLP architecture for vision. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
56. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Joulin, A., Synnaeve, G., Verbeek, J., J'egou, H.: Resmlp: Feedforward networks for image classification with data-efficient training. ArXiv **abs/2105.03404** (2021)
57. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J'egou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
58. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., J'egou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
59. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
60. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
61. Tran, D., xiu Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
62. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
63. Wang, H., Tran, D., Torresani, L., Feiszli, M.: Video modeling with correlation networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
64. Wang, L., Tong, Z., Ji, B., Wu, G.: Tdn: Temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
65. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
66. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
67. Wang, X., Gupta, A.: Videos as space-time region graphs. In: European conference on computer vision (2018)
68. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Tomizuka, M., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision. ArXiv **abs/2006.03677** (2020)
69. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems (2021)
70. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
71. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. Advances in Neural Information Processing Systems (2021)

72. Yu, T., Li, X., Cai, Y., Sun, M., Li, P.: S2-mlp: Spatial-shift mlp architecture for vision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2022)
73. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F.E.H., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. ArXiv **abs/2101.11986** (2021)
74. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020)
75. Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., Chen, H., Marsic, I., Tighe, J.: Vidtr: Video transformer without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
76. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision (2019)
77. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)