视频动作分析与识别

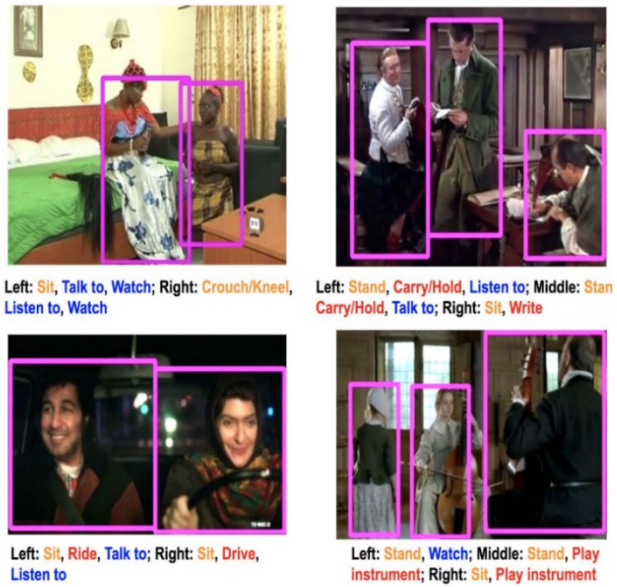
王利民
南京大学 计算机科学与技术系

ARP 报告，VALSE 2023, 无锡

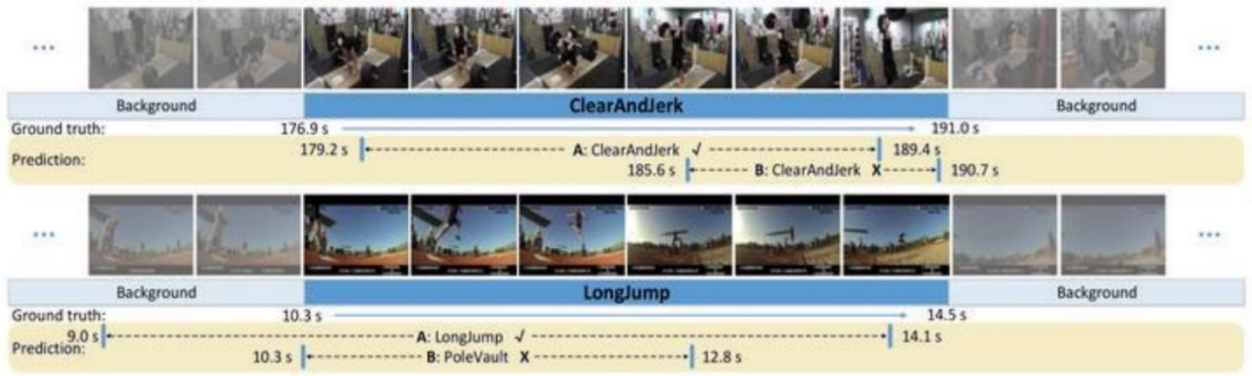
MCG MULTIMEDIA COMPUTING GROUP 媒体计算研究组

视频动作分析任务

视频识别任务



视频时空检测任务



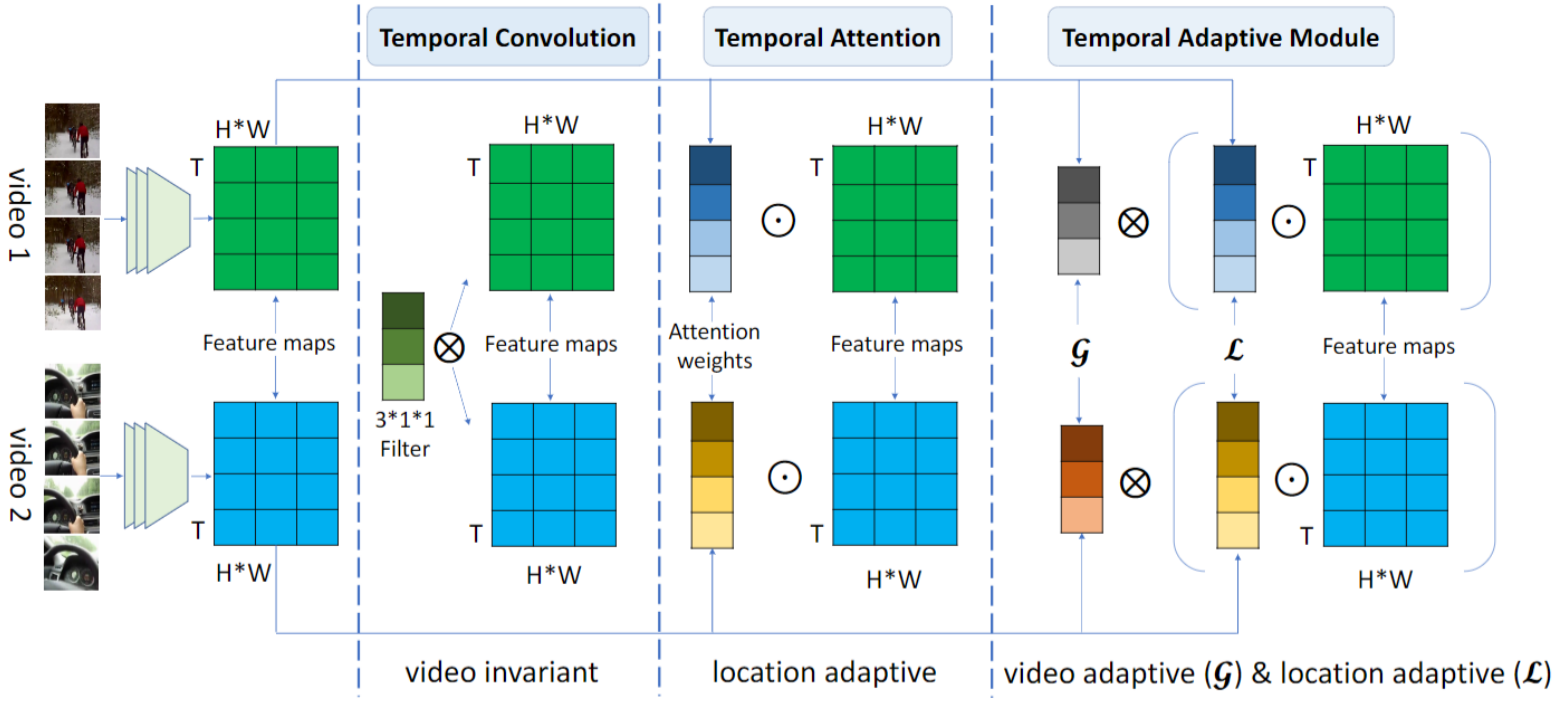
视频时序检测任务

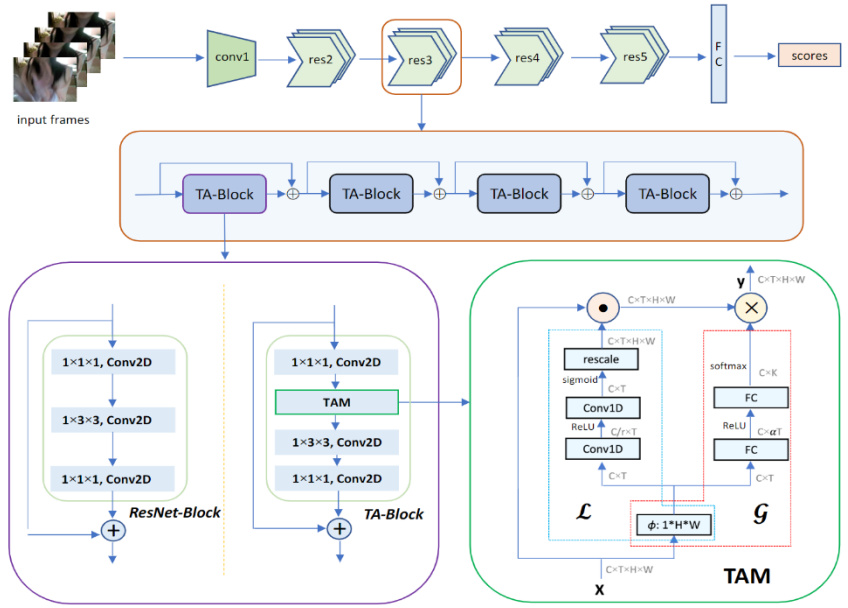
- **视频动作识别** [UCF,HMDB,Kinetics]
  - 类比于图像分类
  - 为下游任务提供基础模型
- **视频动作时空检测** [JHMDB,AVA,MultiSports]
  - 针对多人运动场景
  - Frame-level detection
  - Tube-level detection
- **视频动作时序检测**[THUMOS,ActivityNet,FineAction]
  - 针对长视频场景
  - 检测感兴趣的时间片段
- **其他任务**

提纲

- **视频基础模型：结构设计与预训练策略**
- 视频动作检测方法：时序与时空检测
- 视频分析数据集: FineAction 和 MultiSports

时序自适应模块（TAM)





两层次的时序自适应建模方法：局部注意力增强 和 时序动态融合模块

$$Y = \mathcal{G}(\hat{X}) \otimes (\mathcal{L}(\hat{X}) \odot X)$$

Zhaoyang Liu et al. , **TAM: Temporal Adaptive Module for Video Recognition** in ICCV 2021.

时序自适应模块（TAM)

Baseline

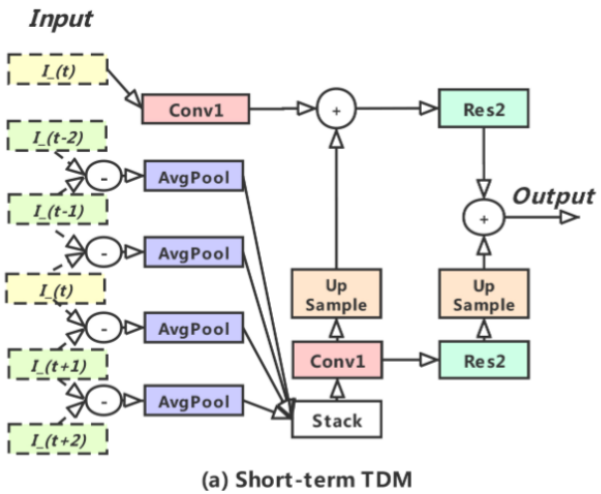
Ours

Models	FLOPs (of single view)	Params	Top-1	Top-5
C2D	42.95G	24.33M	70.2%	88.9%
C2D-Pool	42.95G	24.33M	73.1%	90.6%
C2D-TConv	53.02G	28.10M	73.3%	90.7%
C2D-TIM [24]	43.06G	24.37M	74.7%	91.7%
$I3D_{3\times1\times1}$	62.55G	32.99M	74.3%	91.6%
TSM* [23]	42.95G	24.33M	74.1%	91.2%
TEINet* [24]	43.01G	25.11M	74.9%	91.8%
NL C2D [41]	64.49G	31.69M	74.4%	91.5%
Global branch	43.00G	24.33M	75.6%	91.9%
Local branch	43.07G	25.59M	73.3%	90.7%
Global branch + SE [13]	43.02G	24.65M	75.9%	92.1%
TANet-R	43.02G	25.59M	76.0%	92.2%
TANet	43.02G	25.59M	**76.3%**	**92.6%**

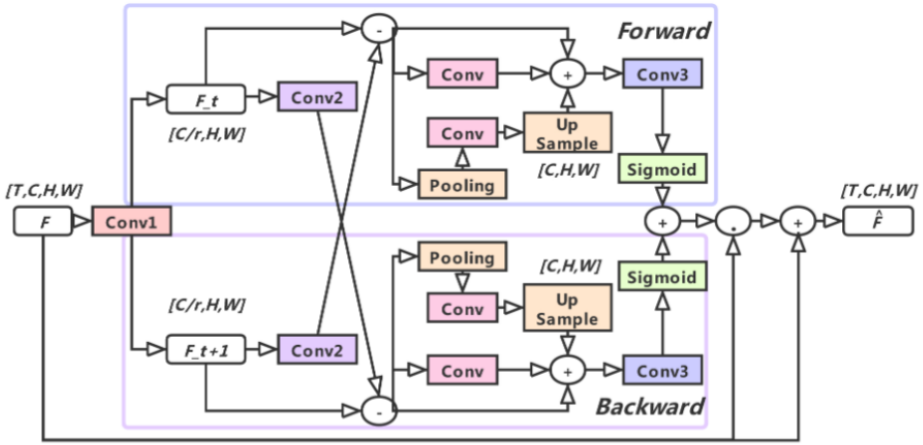
Methods	Backbones	Training Input	GFLOPs	Top-1	Top-5
TSN [40]	InceptionV3	3×224×224	3×250	72.5%	90.2%
ARTNet [38]	ResNet18	16 ×112×112	24×250	70.7%	89.3%
I3D [1]	InceptionV1	64×224×224	108×N/A	72.1%	90.3%
R(2+1)D [36]	ResNet34	32×112×112	152×10	74.3%	91.4%
NL I3D [41]	ResNet50	128×224×224	282×30	76.5%	92.6%
ip-CSN [35]	ResNet50	8×224×224	1.2×10	70.8%	-
TSM [23]	ResNet50	16×224×224	65×30	74.7%	91.4%
TEINet [24]	ResNet50	16×224×224	86×30	76.2%	92.5%
bLVNet-TAM [6]	bLResNet50	48×224×224	93×9	73.5%	91.2%
SlowOnly [8]	ResNet50	8×224×224	42×30	74.8%	91.6%
$SlowFast_{4\times16}$ [8]	ResNet50	(4+32)×224×224	36×30	75.6%	92.1%
$SlowFast_{8\times8}$ [8]	ResNet50	(8+32)×224×224	66×30	77.0%	92.6%
I3D* [2]	ResNet50	32 ×224×224	335 ×30	76.6%	-
TANet-50	ResNet50	8×224×224	43×30	76.3%	92.6%
TANet-50	ResNet50	16×224×224	86×12	76.9%	92.9%
X3D-XL [7]	-	16×312×312	48×30	79.1%	93.9%
CorrNet [37]	ResNet101	32×10×3	224×30	79.2%	-
ip-CSN [35]	ResNet152	32 ×224×224	83×30	79.2%	93.8%
$SlowFast_{16\times8}$ [8]	ResNet101	(16+64)×224×224	213×30	78.9%	93.5%
TANet-101	ResNet101	8×224×224	82×30	77.1%	93.1%
TANet-101	ResNet101	16×224×224	164×12	78.4%	93.5%
TANet-152	ResNet152	16×224×224	242×12	**79.3%**	**94.1%**

Zhaoyang Liu et al. , **TAM: Temporal Adaptive Module for Video Recognition** in ICCV, 2021.

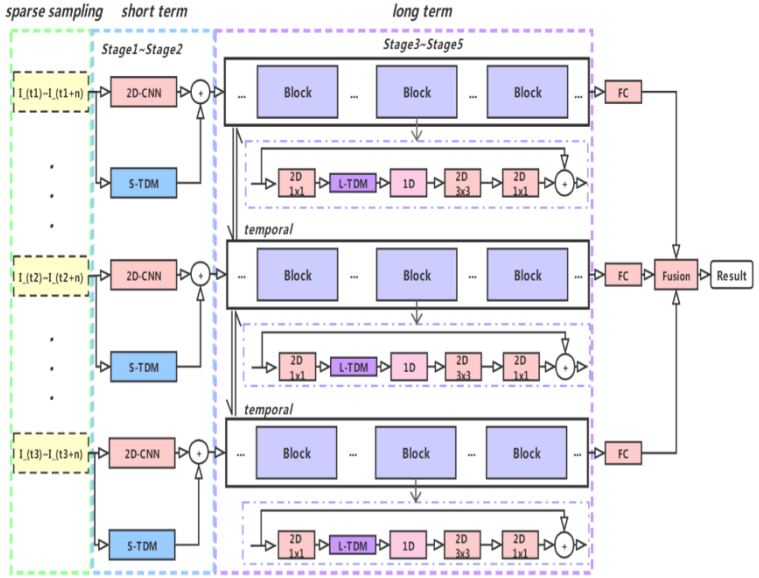
时序差分网络（TDN)



(a) Short-term TDM



(b) Long-term TDM



Short term TDM : $\hat{F}_i = F_i + \mathcal{H}(I_i)$

Long term TDM : $\hat{F}_i = F_i + F_i \odot \mathcal{G}(F_i, F_{i+1})$

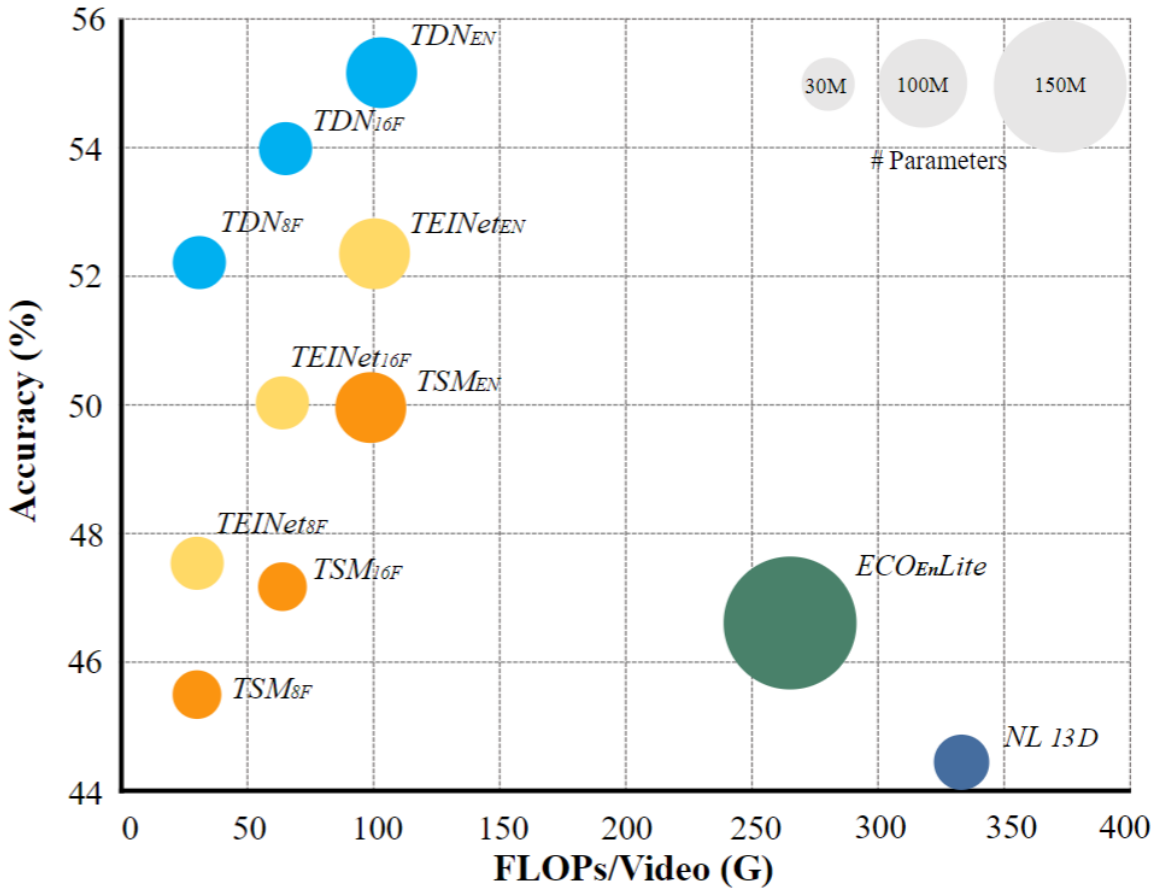
利用时序差分操作显示引导神经网络捕捉视频的高频（运动变化）信息

L. Wang et al. , **TDN: Temporal Difference Networks for Efficient Action Recognition** in CVPR 2021.

时序差分网络（TDN)

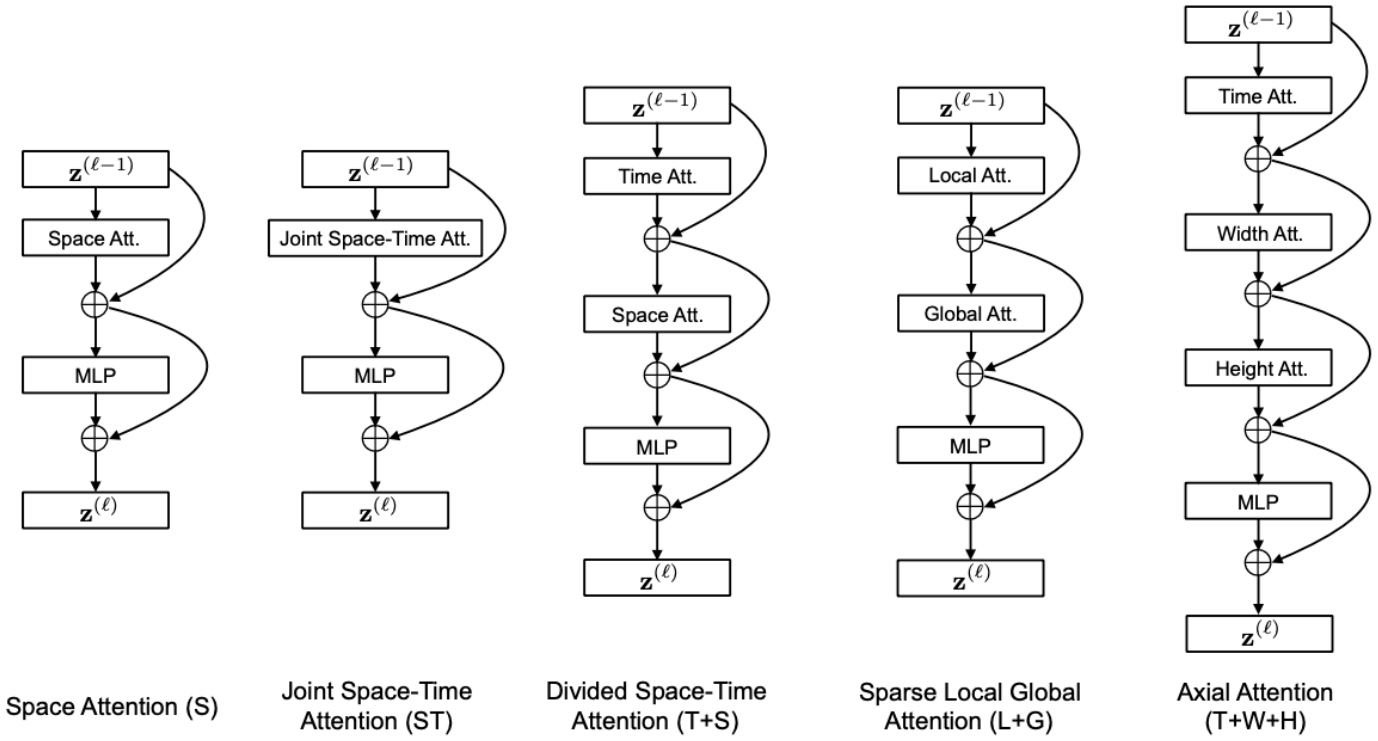
S-TDM	L-TDM	FLOPs	Top1
concat	avg	36.2G	41.5%
concat	diff✓	36.2G	51.4%
diff✓	avg	35.9G	51.6%
diff✓	diff✓	**35.9G**	**52.3%**

Model	FLOPs	Top1	Top5
T-Conv [33]	33G	47.5%	77.5%
T-Conv++ [33]	165G	48.2%	79.1%
TSM [19]	33G	47.1%	76.2%
TSM++ [19]	165G	47.6%	77.9%
TEINet [20]	33G	48.4%	77.2%
TEINet++ [20]	165G	49.0%	79.0%
TDM	36G	**52.3%**	**80.6%**



L. Wang et al. , **TDN: Temporal Difference Networks for Efficient Action Recognition** in CVPR 2021.

Vision Transformer（TimeSformer）



Space Attention (S) | Joint Space-Time Attention (ST) | Divided Space-Time Attention (T+S) | Sparse Local Global Attention (L+G) | Axial Attention (T+W+H)

Attention	Params	K400	SSv2
Space	85.9M	77.6	36.6
Joint Space-Time	85.9M	78.1	58.5
Divided Space-Time	121.4M	**78.5**	**59.5**
Sparse Local Global	121.4M	76.8	56.3
Axial	156.8M	74.6	56.2

Method	Top-1	Top-5	TFLOPs
ARTNet (Wang et al., 2018a)	69.2	88.3	6.0
I3D (Carreira & Zisserman, 2017)	71.1	89.3	N/A
R(2+1)D (Tran et al., 2018)	72.0	90.0	17.5
MFNet (Chen et al., 2018b)	72.8	90.4	N/A
Inception-ResNet (Bian et al., 2017)	73.0	90.9	N/A
bLVNet (Fan et al., 2019)	73.5	91.2	0.84
A^2-Net (Chen et al., 2018c)	74.6	91.5	N/A
TSM (Lin et al., 2019)	74.7	N/A	N/A
S3D-G (Xie et al., 2018)	74.7	93.4	N/A
Oct-I3D+NL (Chen et al., 2019a)	75.7	N/A	0.84
D3D (Stroud et al., 2020)	75.9	N/A	N/A
GloRe (Chen et al., 2019b)	76.1	N/A	N/A
I3D+NL (Wang et al., 2018b)	77.7	93.3	10.8
ip-CSN-152 (Tran et al., 2019)	77.8	92.8	3.2
CorrNet (Wang et al., 2020a)	79.2	N/A	6.7
LGD-3D-101 (Qiu et al., 2019)	79.4	94.4	N/A
SlowFast (Feichtenhofer et al., 2019b)	79.8	93.9	7.0
X3D-XXL (Feichtenhofer, 2020)	80.4	94.6	5.8
TimeSformer	78.0	93.7	**0.59**
TimeSformer-HR	79.7	94.4	5.11
TimeSformer-L	**80.7**	**94.7**	7.14

Table 2. Video-level accuracy on Kinetics-400. TimeSformer-L achieves the best reported accuracy.

Gedas Bertasius et al., **Is Space-Time Attention All You Need for Video Understanding?,** in ICML 2021

Vision Transformer（ViViT）

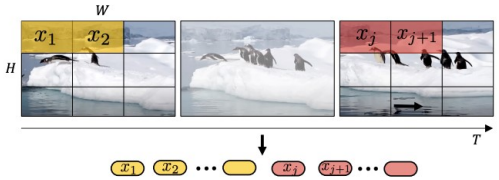


Figure 2: Uniform frame sampling: We simply sample n_t frames, and embed each 2D frame independently following ViT [15].

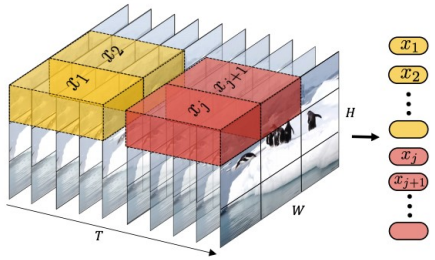
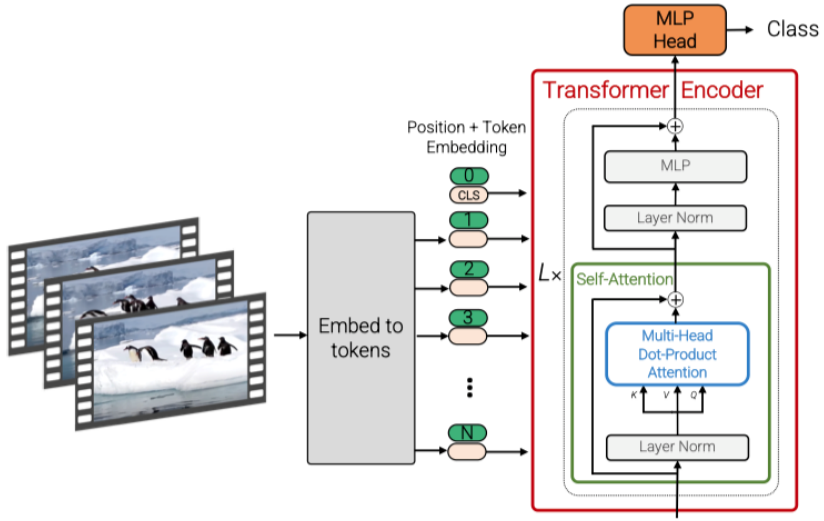


Figure 3: Tubelet embedding. We extract and linearly embed non-overlapping tubelets that span the spatio-temporal input volume.



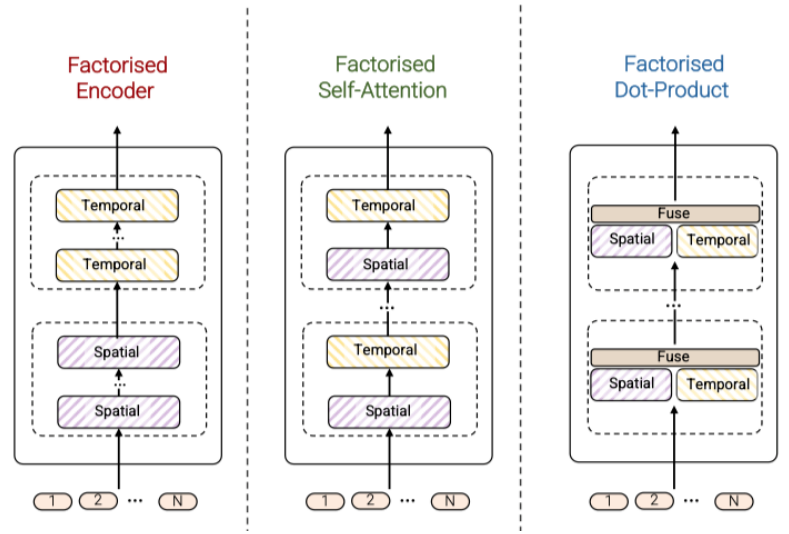


Table 1: Comparison of input encoding methods using ViViT-B and spatio-temporal attention on Kinetics. Further details in text.

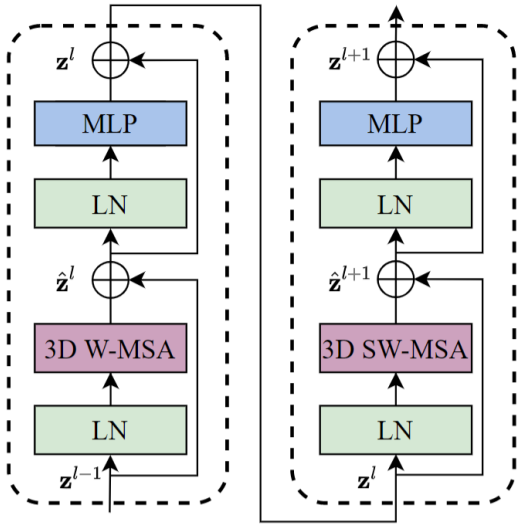
	Top-1 accuracy
Uniform frame sampling	78.5
Tubelet embedding	
Random initialisation [24]	73.2
Filter inflation [8]	77.6
Central frame	79.2

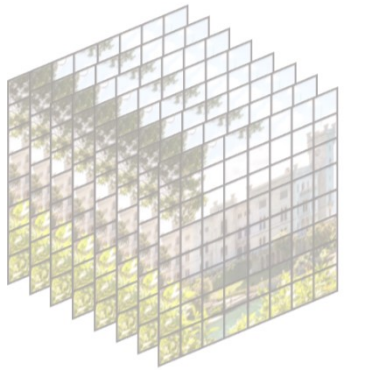
Table 2: Comparison of model architectures using ViViT-B as the backbone, and tubelet size of 16×2. We report Top-1 accuracy on Kinetics 400 (K400) and action accuracy on Epic Kitchens (EK). Runtime is during inference on a TPU-v3.

	K400	EK	FLOPs ($\times 10^9$)	Params ($\times 10^6$)	Runtime (ms)
Model 1: Spatio-temporal	80.0	43.1	455.2	88.9	58.9
Model 2: Fact. encoder	78.8	43.7	284.4	100.7	17.4
Model 3: Fact. self-attention	77.4	39.1	372.3	117.3	31.7
Model 4: Fact. dot product	76.3	39.5	277.1	88.9	22.9
Model 2: Ave. pool baseline	75.8	38.8	283.9	86.7	17.3

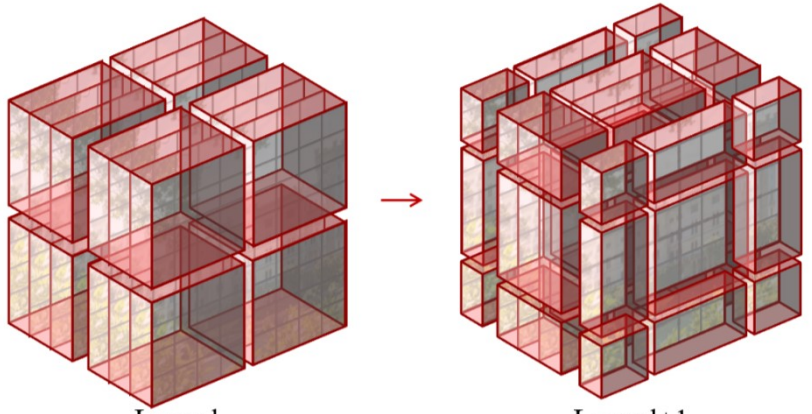
Anurag Arnab et al., **ViViT: A Video Vision Transformer**, in ICCV 2021

Vision Transformer（Video Swin）





3D tokens: T'×H'×W' = 8×8×8
Window size: P×M×M = 4×4×4

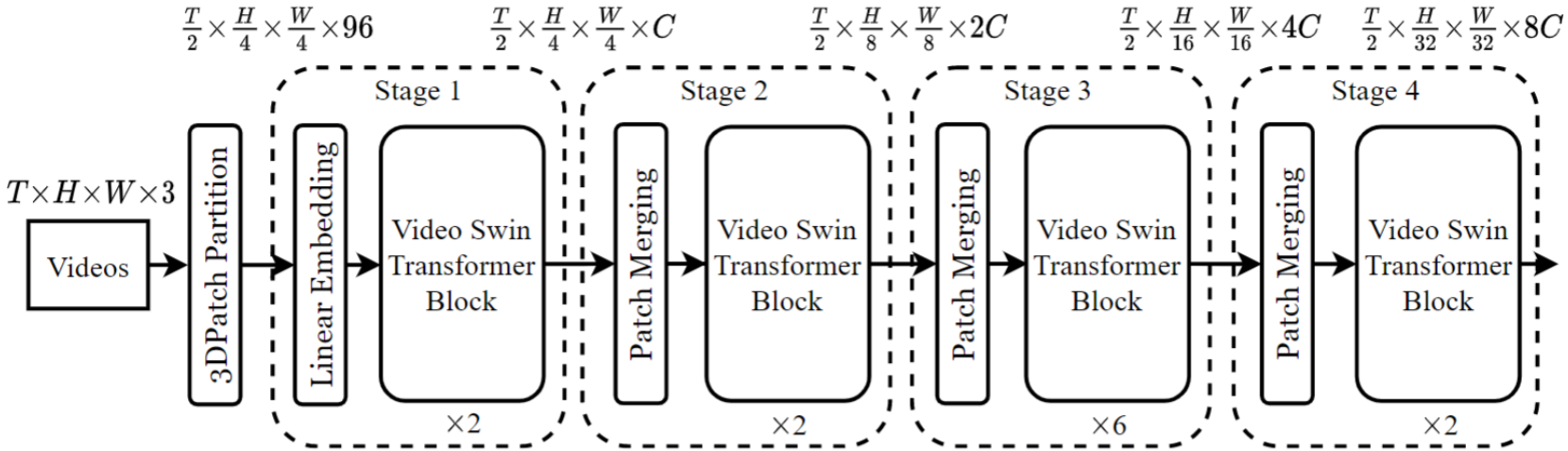


Layer l
window: 2×2×2=8

Layer l+1
window: 3×3×3=27

3D local window to perform self-attention

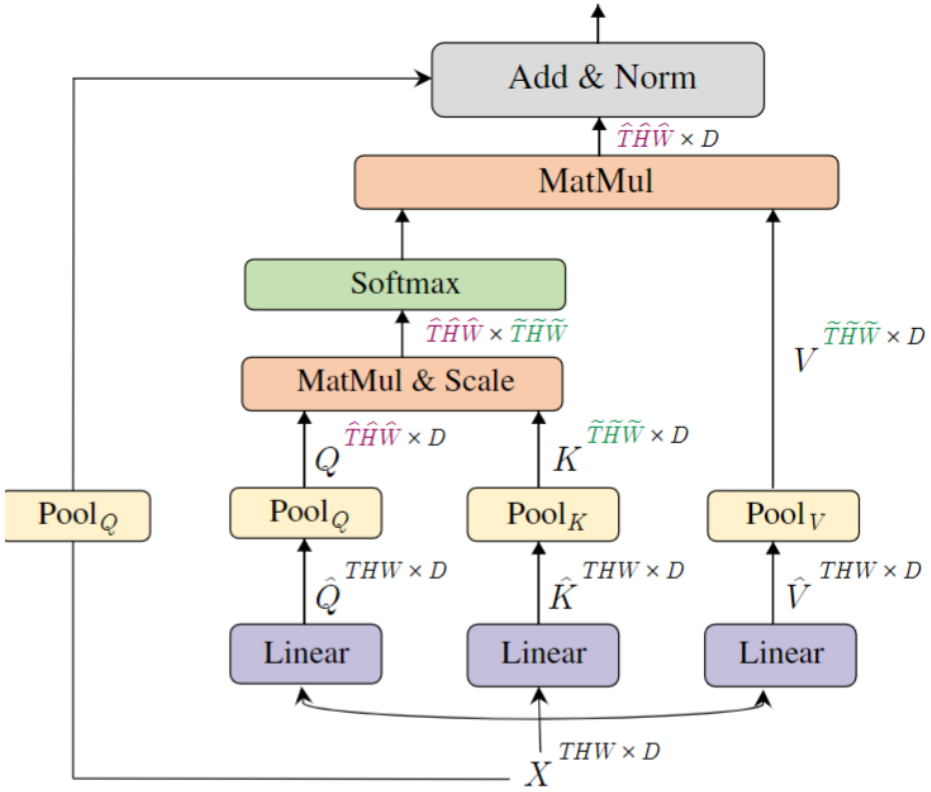
A token



	Top-1	Top-5	FLOPs	Param
joint	78.8	93.6	88	28.2
split	76.4	92.1	83	42.0
factorized	78.5	93.5	95	36.5

Ze Liu et al., **Video Swin Transformer**, in CVPR 2022

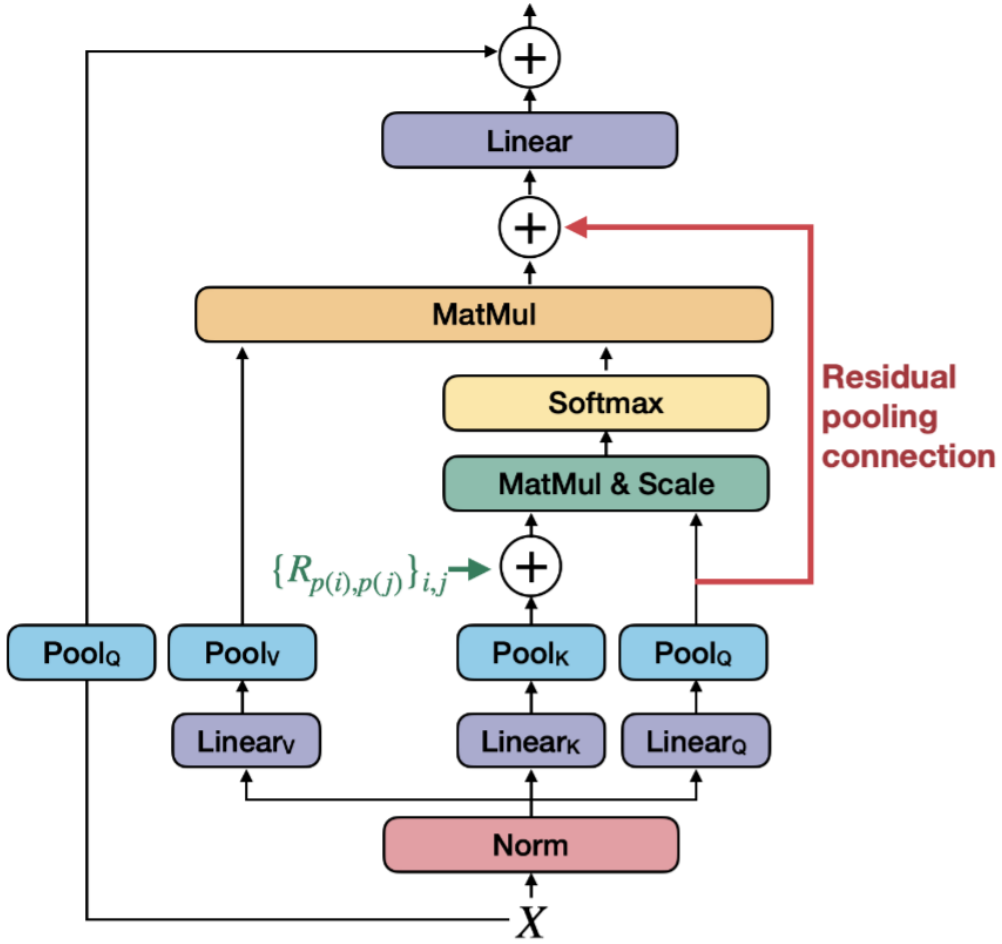
Vision Transformer（MVIT v1）



model	pre-train	top-1	top-5	FLOPs×views	Param
Two-Stream I3D [14]	-	71.6	90.0	216 × NA	25.0
ip-CSN-152 [102]	-	77.8	92.8	109×3×10	32.8
SlowFast 8×8 +NL [34]	-	78.7	93.5	116×3×10	59.9
SlowFast 16×8 +NL [34]	-	79.8	93.9	234×3×10	59.9
X3D-M [33]	-	76.0	92.3	6.2×3×10	3.8
X3D-XL [33]	-	79.1	93.9	48.4×3×10	11.0
ViT-B-VTN [84]	ImageNet-1K	75.6	92.4	4218×1×1	114.0
ViT-B-VTN [84]	ImageNet-**21K**	78.6	93.7	4218×1×1	114.0
ViT-B-TimeSformer [8]	ImageNet-**21K**	80.7	94.7	2380×3×1	121.4
ViT-L-ViViT [1]	ImageNet-**21K**	81.3	94.7	3992×3×4	310.8
ViT-B (our baseline)	ImageNet-**21K**	79.3	93.9	180×1×5	87.2
ViT-B (our baseline)	-	68.5	86.9	180×1×5	87.2
MViT-S	-	76.0	92.1	32.9×1×5	26.1
MViT-B, 16×4	-	78.4	93.5	70.5×1×5	36.6
MViT-B, 32×3	-	80.2	94.4	170×1×5	36.6
MViT-B, 64×3	-	**81.2**	**95.1**	455×3×3	36.6

Haoqi Fan et al., **Multiscale Vision Transformers**, in ICCV 2021

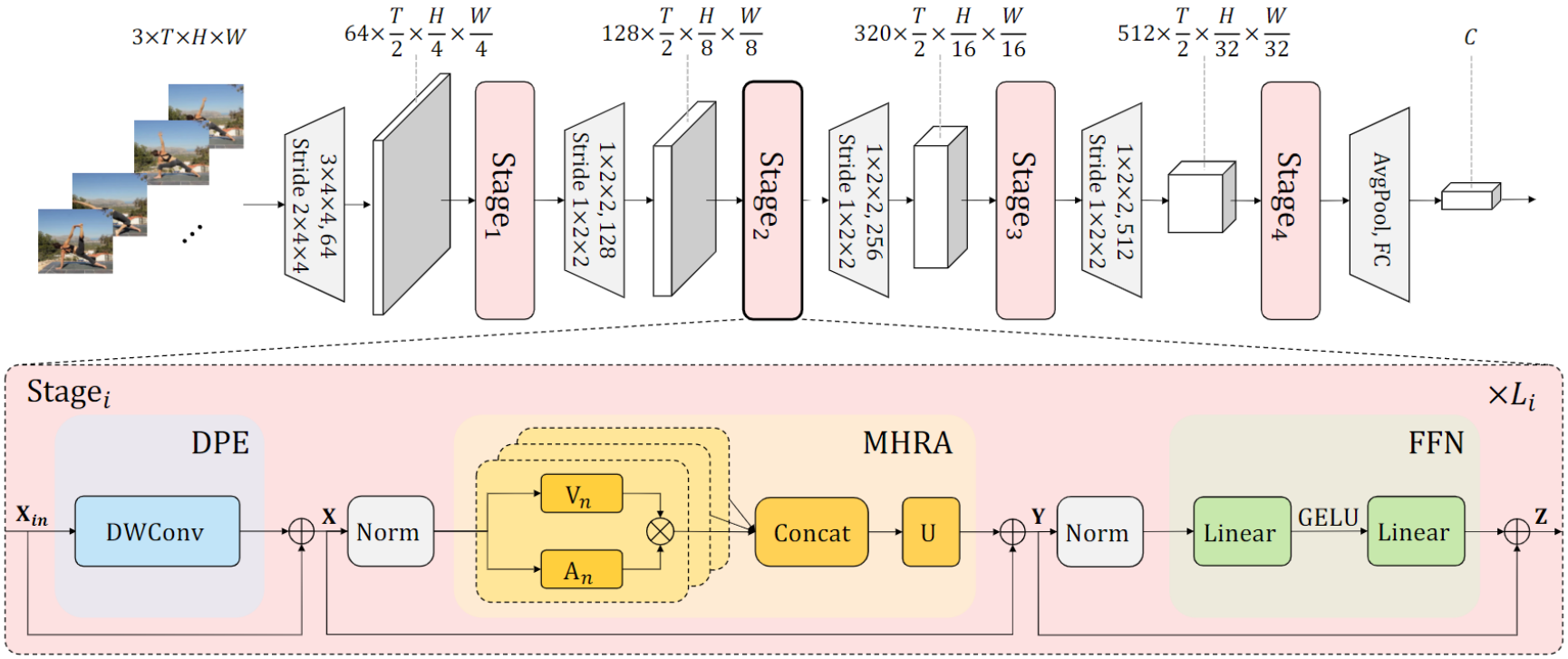
Vision Transformer（MVIT v2）



model	pre-train	top-1	top-5	FLOPs×views	Param
SlowFast 16×8 +NL [23]	-	79.8	93.9	234×3×10	59.9
X3D-XL [22]	-	79.1	93.9	48.4×3×10	11.0
MoViNet-A6 [45]	-	81.5	95.3	386×1×1	31.4
MViTv1, 16×4 [21]	-	78.4	93.5	70.3×1×5	36.6
MViTv1, 32×3 [21]	-	80.2	94.4	170×1×5	36.6
MViTv2-S, 16×4	-	81.0	94.6	64×1×5	34.5
MViTv2-B, 32×3	-	**82.9**	**95.7**	225×1×5	51.2
ViT-B-VTN [59]	IN-21K	78.6	93.7	4218×1×1	114.0
ViT-B-TimeSformer [3]		80.7	94.7	2380×3×1	121.4
ViT-L-ViViT [1]		81.3	94.7	3992×3×4	310.8
Swin-L↑ 384^2 [56]		84.9	96.7	2107×5×10	200.0
MViTv2-L↑ 312^2, 40×3		**86.1**	**97.0**	2828×3×5	217.6

Yanghao Li et al., **MViTv2: Improved Multiscale Vision Transformers for Classification and Detection**, in CVPR 2022

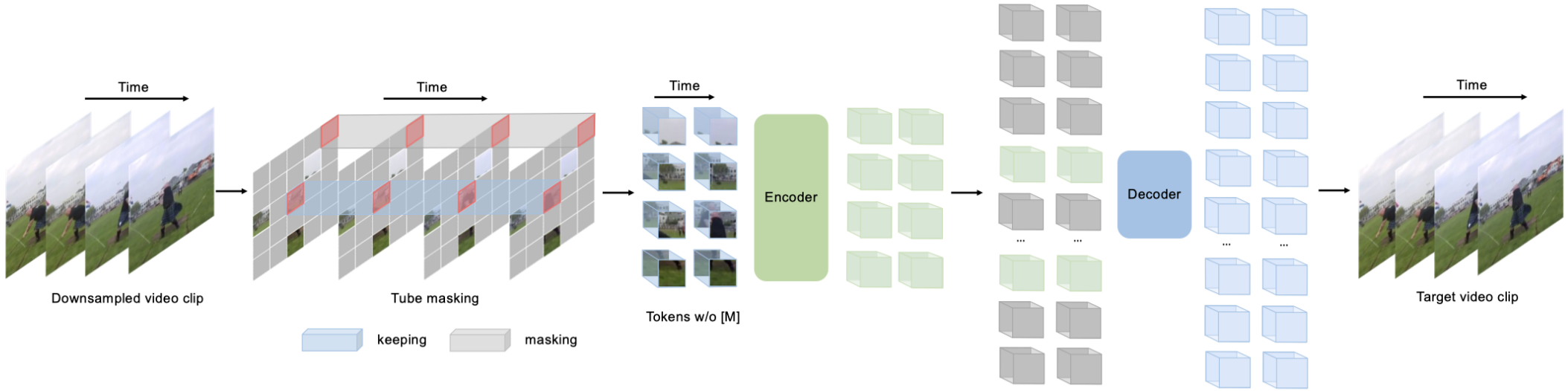
UniFormer



Method	Basic Operation	Tackle Local Redundancy	Capture Global Dependency	Efficiency GFLOPs	Efficiency Top-1
X3D (Feichtenhofer, 2020)	PWConv-DWConv-PWConv	✔	✗	5823	80.4
TimeSformer (Bertasius et al., 2021)	Divided MHSA	✗	✔	7140	80.7
Our UniFormer	Joint MHRA	✔	✔	**168**	**80.8**

K. Li et al., UniFormer: unified transformer for efficient spatiotemporal representation learning, in ICLR 2022

Video Transformer (VideoMAE v1)



case	ratio	SSV2	K400
tube	75	68.0	79.8
tube	90	**69.6**	**80.0**
random	90	68.3	79.5
frame	87.5*	61.5	76.5

Mask方法

case	SSV2	K400
from scratch	32.6	68.8
ImageNet-21k sup.	61.8	78.9
IN-21k+K400 sup.	65.2	-
VideoMAE	**69.6**	**80.0**

预训练策略

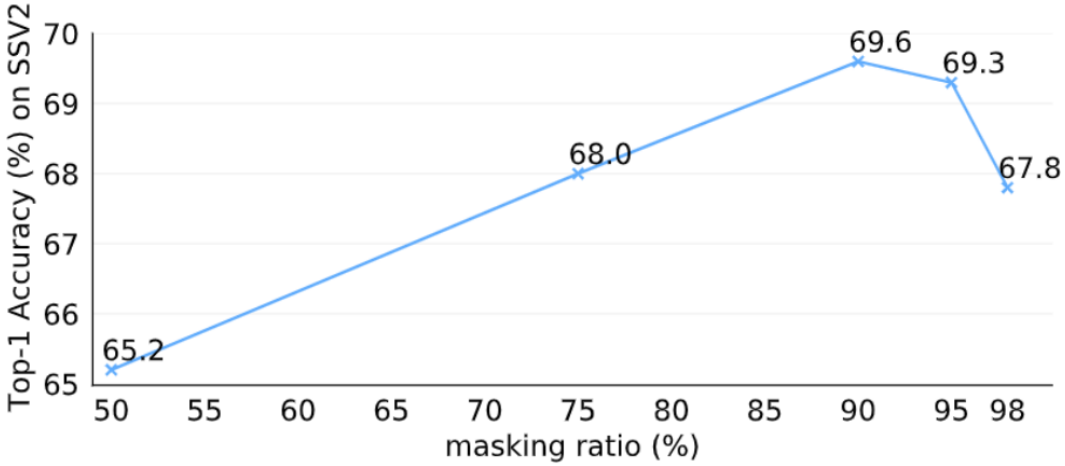
dataset	method	SSV2	K400
IN-1K	ImageMAE	64.8	78.7
K400	VideoMAE	68.5	**80.0**
SSV2	VideoMAE	**69.6**	79.6

预训练数据

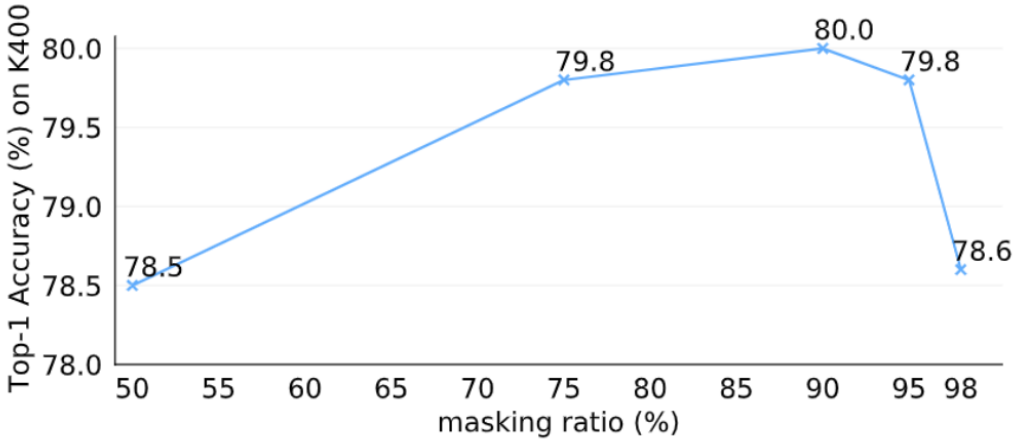
Z. Tong et al., **VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training**, in NeurIPS 2022

Video Transformer (VideoMAE v1)

视频数据冗余性



(a) Performance on SSV2

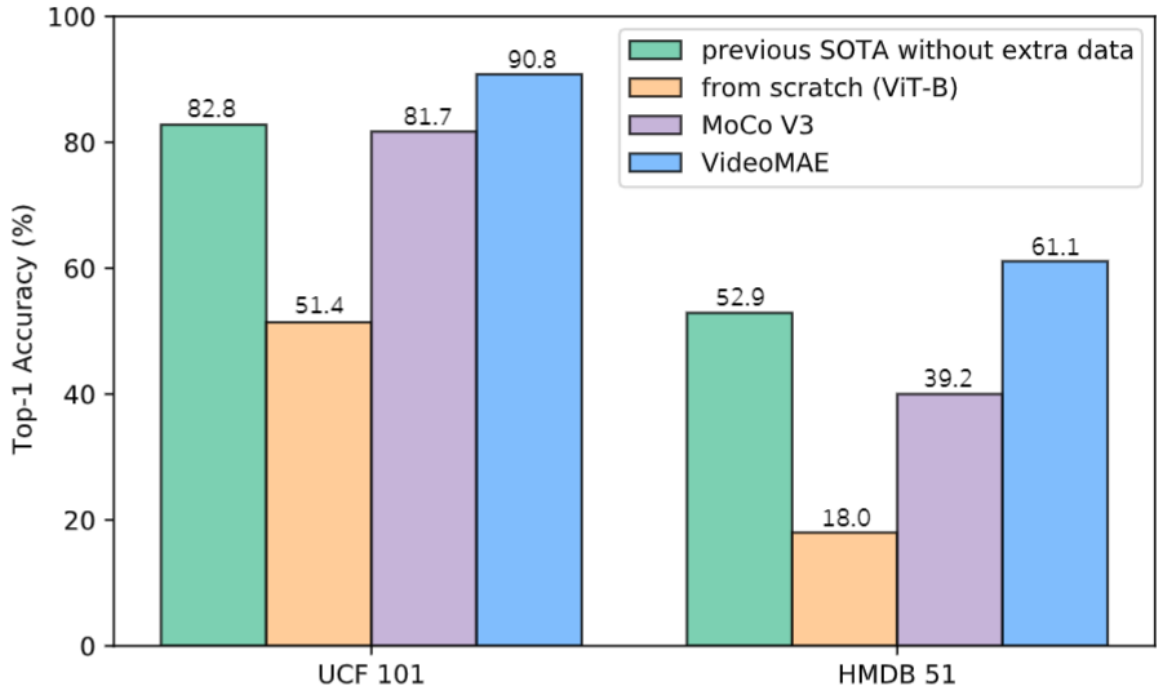


(b) Performance on Kinetics-400

Z. Tong et al., **VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training**, in NeurIPS 2022

Video Transformer (VideoMAE v1)

VideoMAE数据高效性



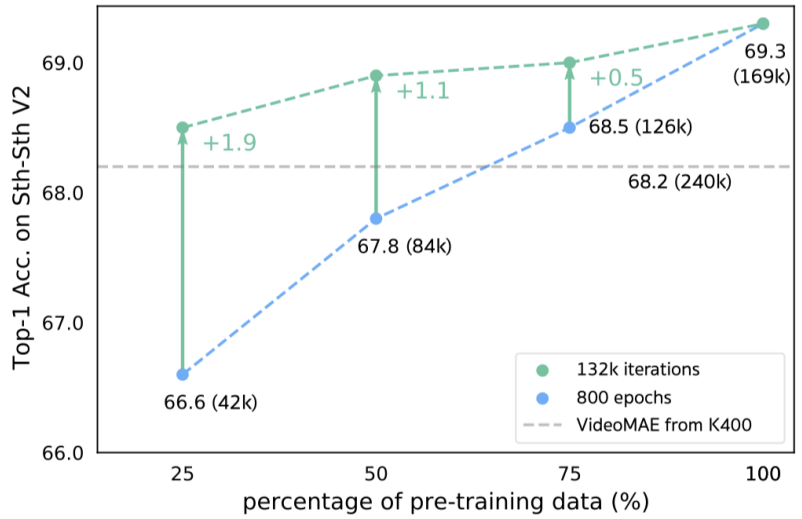
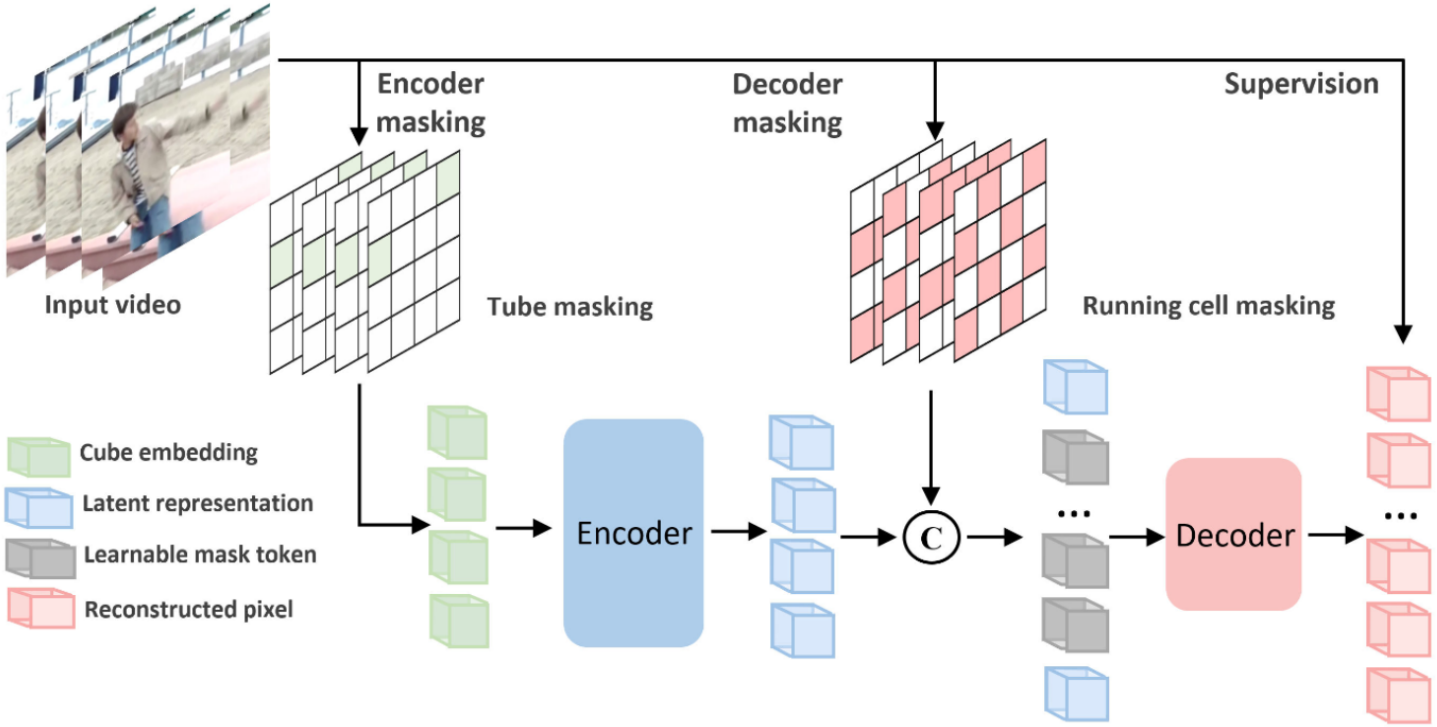


Figure 6. **Data efficiency** of VideoMAE representations. Our default backbone is 16-frame ViT-B described in Table 1. ● denotes that all models are trained for the **same** 132k iterations, and ● denotes that all models are trained for the **same** 800 epochs. Note that it takes 132k iterations to pre-train the model for 800 epochs on the full training set.

Z. Tong et al., **VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training**, in NeurIPS 2022

Video Transformer (VideoMAE v2)



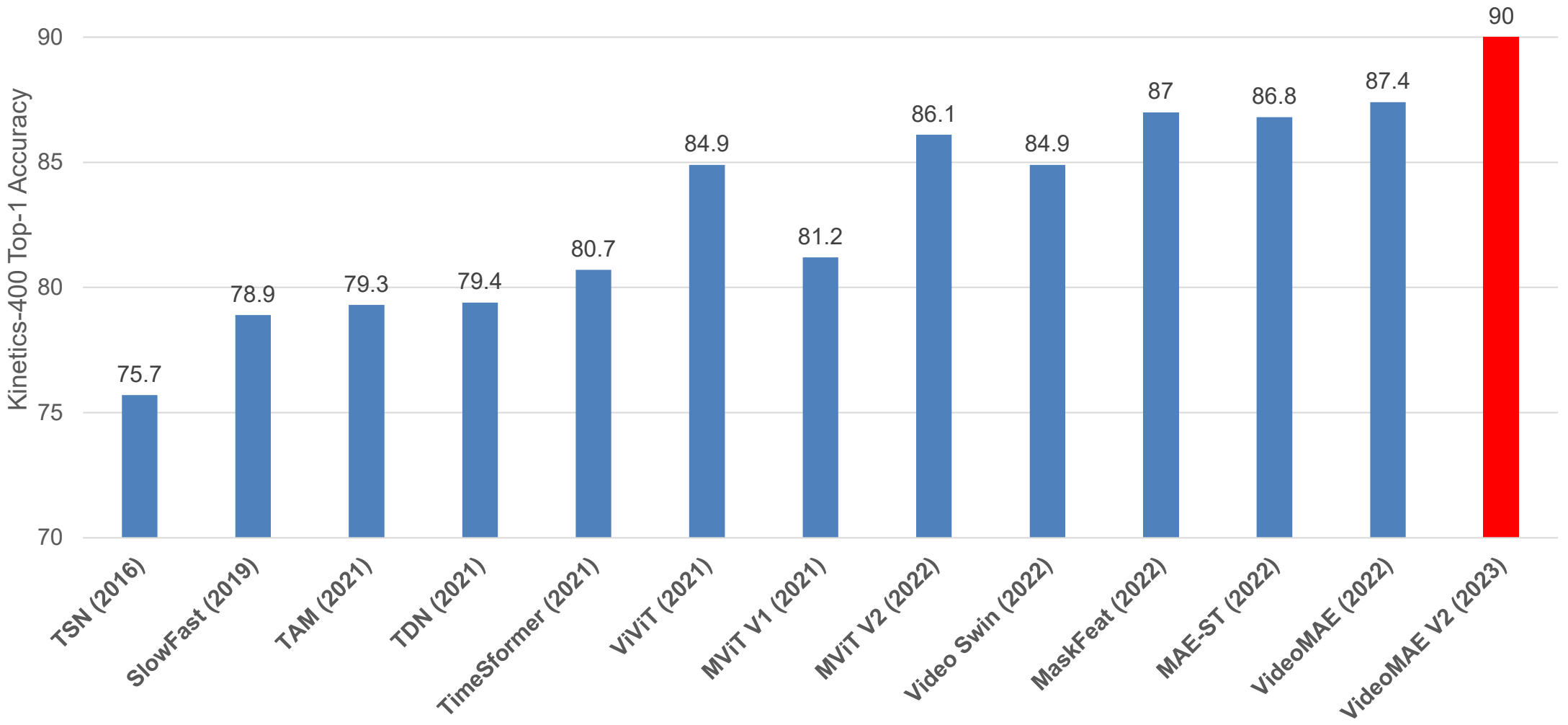
Study on dual masking and masking ratio

Decoder Masking	ρ^d	Top-1	FLOPs
None	0%	**70.28**	35.48G
Frame	50%	69.76	25.87G
Random	50%	64.87	25.87G
Running cell [1]	50%	66.74	25.87G
Running cell [2]	25%	70.22	31.63G
Running cell [2]	50%	70.15	25.87G
Running cell [2]	75%	70.01	21.06G

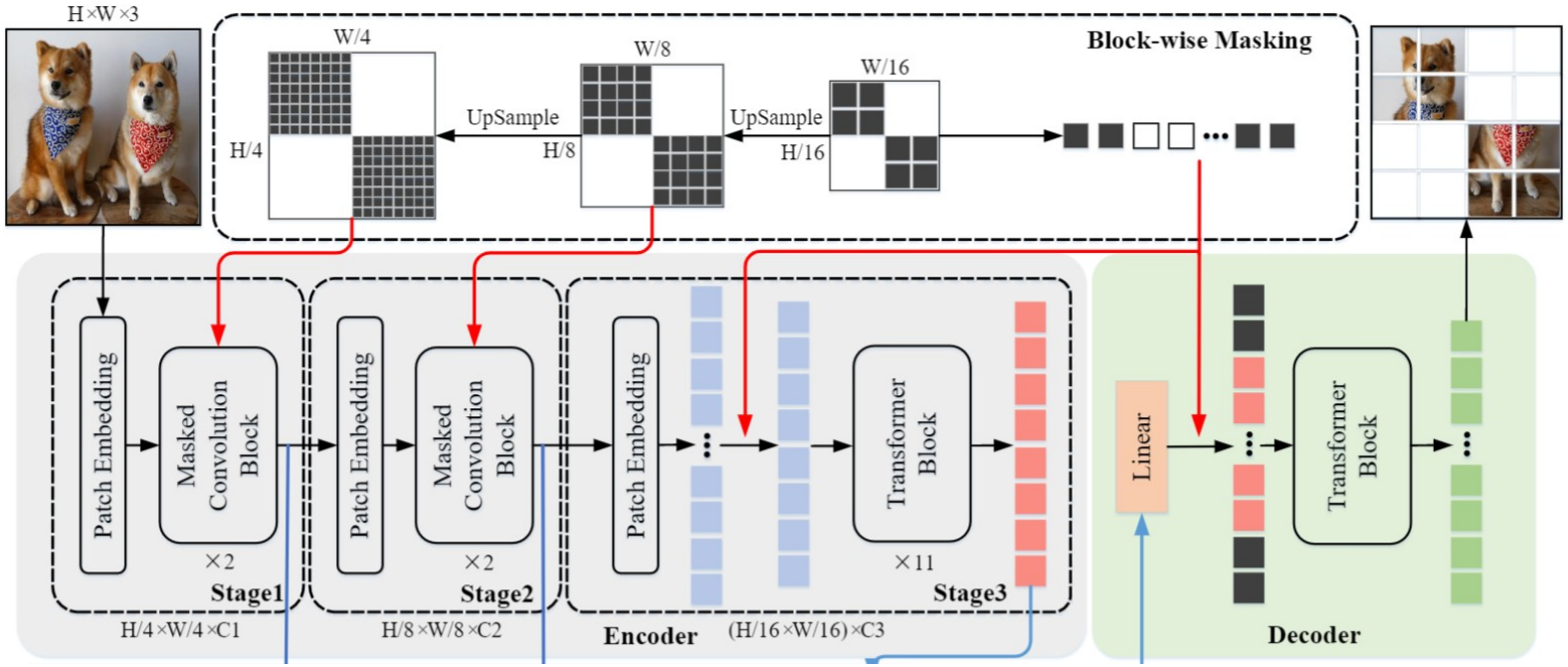
提出了编码器-解码器双掩码策略，提升预训练效率，更好支撑大规模预训练

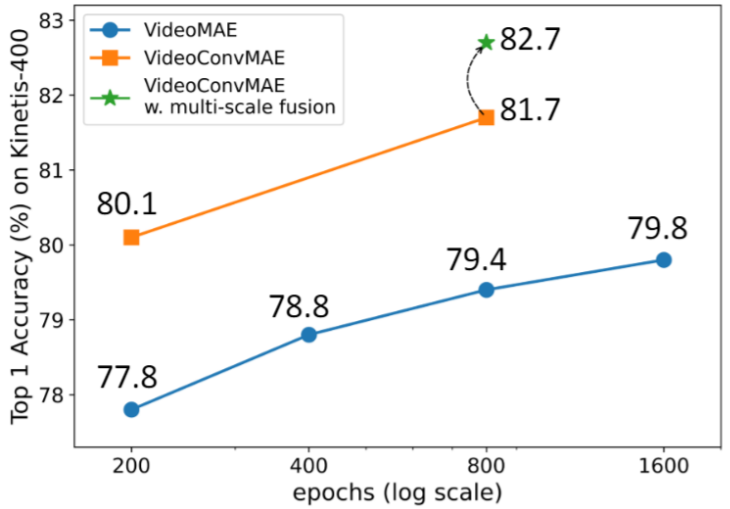
L. Wang et al., **VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking,** in CVPR 2023

Performance on Kinetics-400



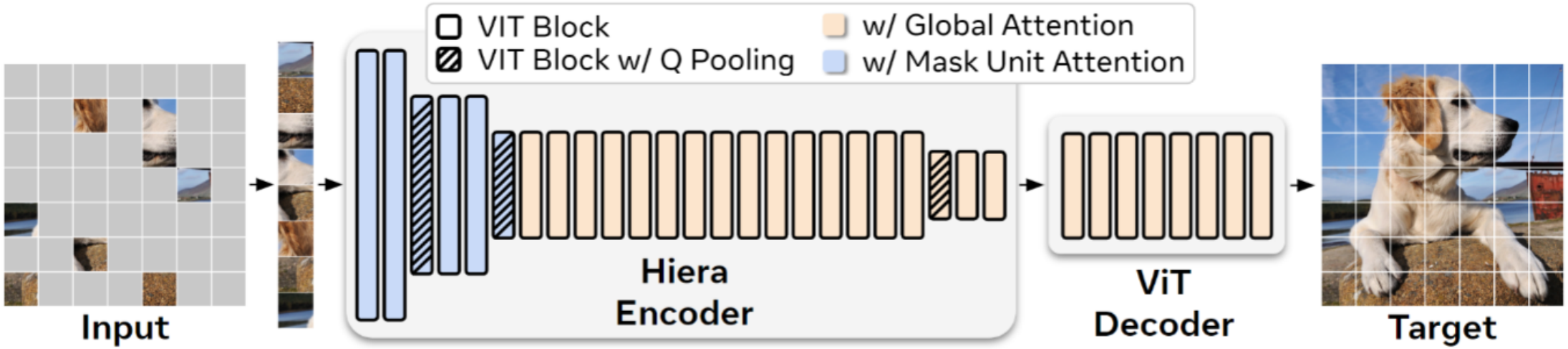
ConvMAE





P Gao et al., ConvMAE: Masked Convolution Meets MaskedAutoencoders, in NeurIPS 2022

Hiera

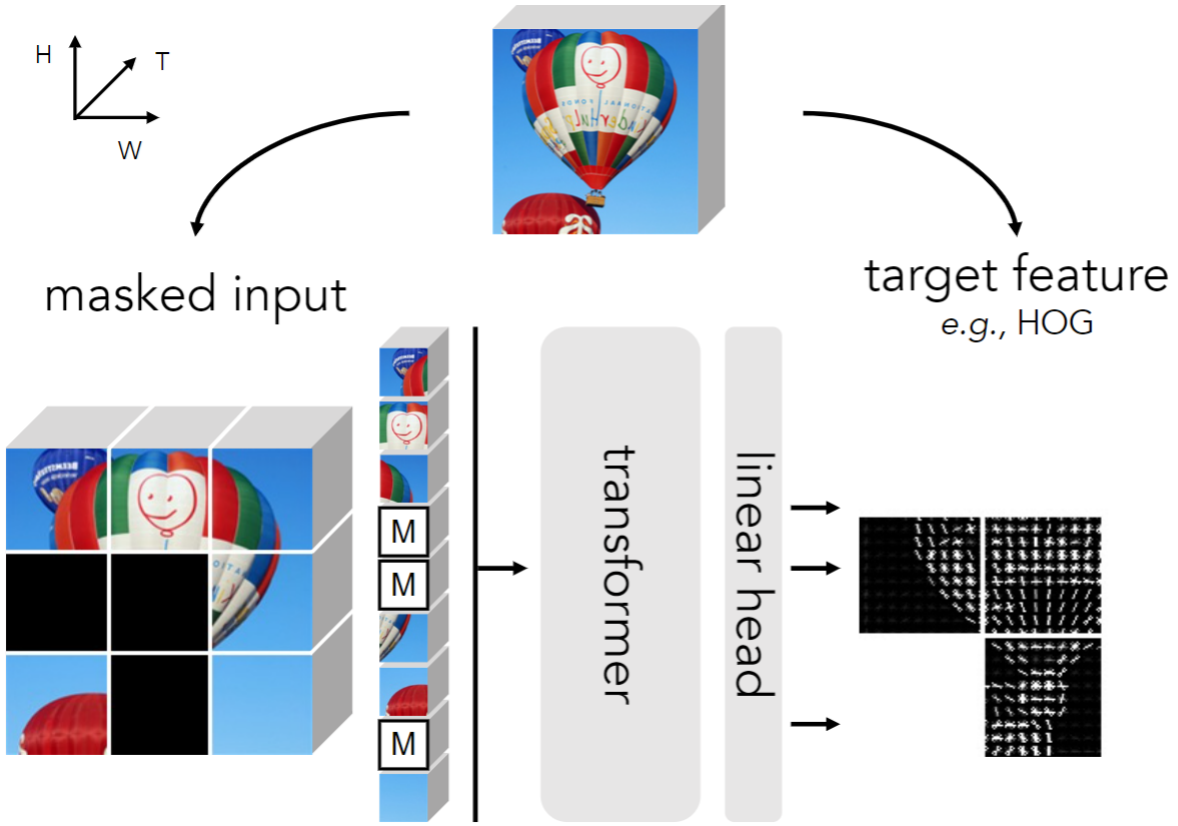


Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	85.6	253.3	85.3	20.7
b. replace convs with maxpools *	84.4	99.9†	84.1	10.4†
c. delete stride=1 maxpools *	85.4	309.2	84.3	26.2
d. set kernel size equal to stride	**85.7**	369.8	**85.5**	29.4
e. delete q attention residuals	85.6	374.3	**85.5**	29.8
f. replace kv pooling with MU attn	85.6	**531.4**	**85.5**	**40.8**

backbone	pretrain	acc.	FLOPs (G)	Param
ViT-B	MAE	81.5	180×3×5	87M
Hiera-B	MAE	84.0	**102**×3×5	**51M**
Hiera-B+	MAE	**85.0**	133×3×5	69M
MViTv2-L	-	80.5	**377**×1×10	218M
MViTv2-L	MaskFeat	84.3	**377**×1×10	218M
ViT-L	MAE	85.2	597×3×5	305M
Hiera-L	MAE	**87.3**	413×3×5	**213M**
ViT-H	MAE	86.6	1192×3×5	**633M**
Hiera-H	MAE	**87.8**	**1159**×3×5	672M

C. Ryali et al., **Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles**, in ICML 2023

MaskFeat

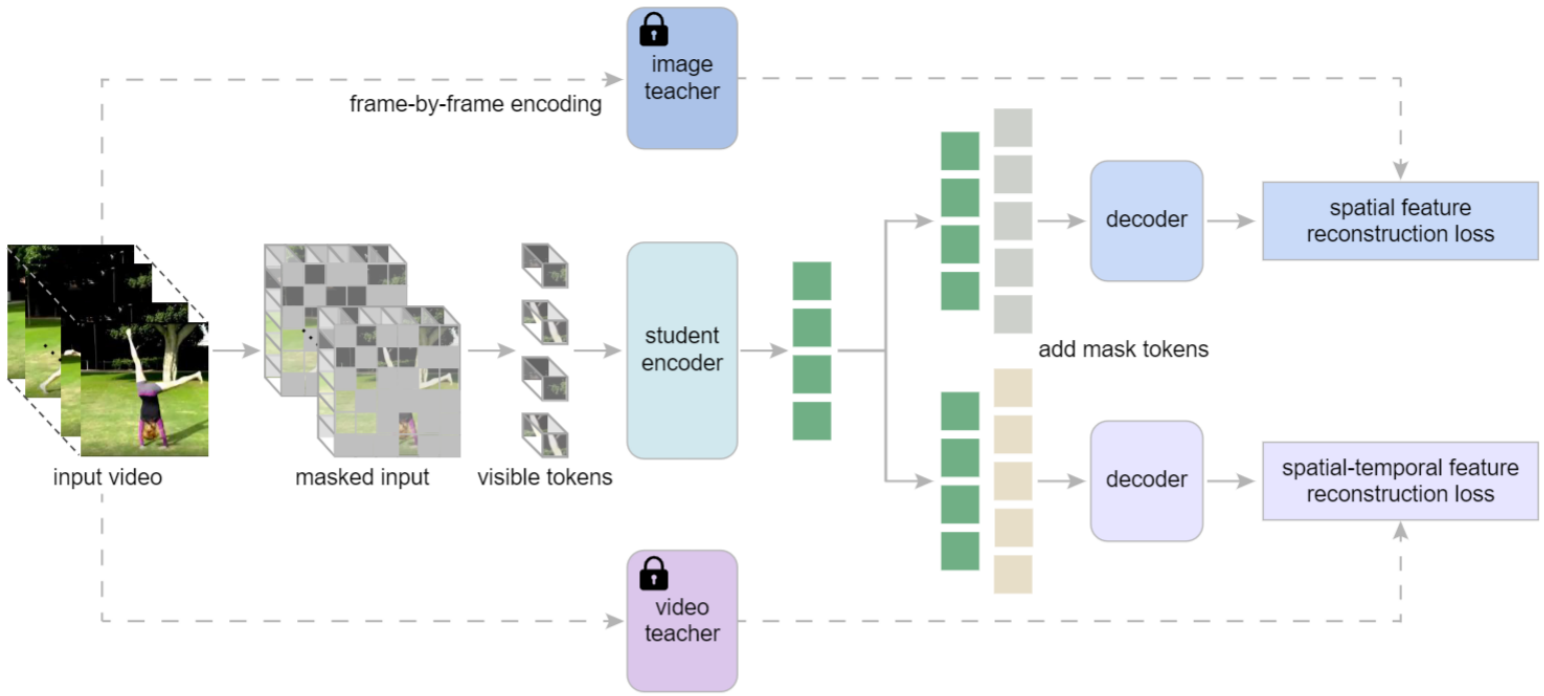


feature type	one-stage	variant	top-1
scratch	-	MViT-S [56]	81.1
pixel	✓	RGB	80.7
image descriptor	✓	HOG [22]	**82.2**
dVAE	✗	DALL-E [73]	81.7
unsupervised feature	✗	DINO [9], ViT-B	**82.5**
supervised feature	✗	MViT-B [31]	81.9

Table 1. **Comparing target features for MaskFeat (*video*).** All variants are pre-trained with MaskFeat for 300 epochs on MViT-S, 16×4. We report fine-tuning accuracy on K400. Default is gray.

C. Wei et al., **Masked Feature Prediction for Self-Supervised Visual Pre-Training**, in CVPR 2022

MVD



student	teachers		K400 top-1	SSv2 top-1
	image	video		
ViT-S	✓	✗	80.4	69.4
	✗	✓	80.1	70.0
	✓	✓	**80.6**	**70.7**
ViT-B	✓	✗	82.3	71.4
	✗	✓	82.1	71.8
	✓	✓	**82.7**	**72.5**

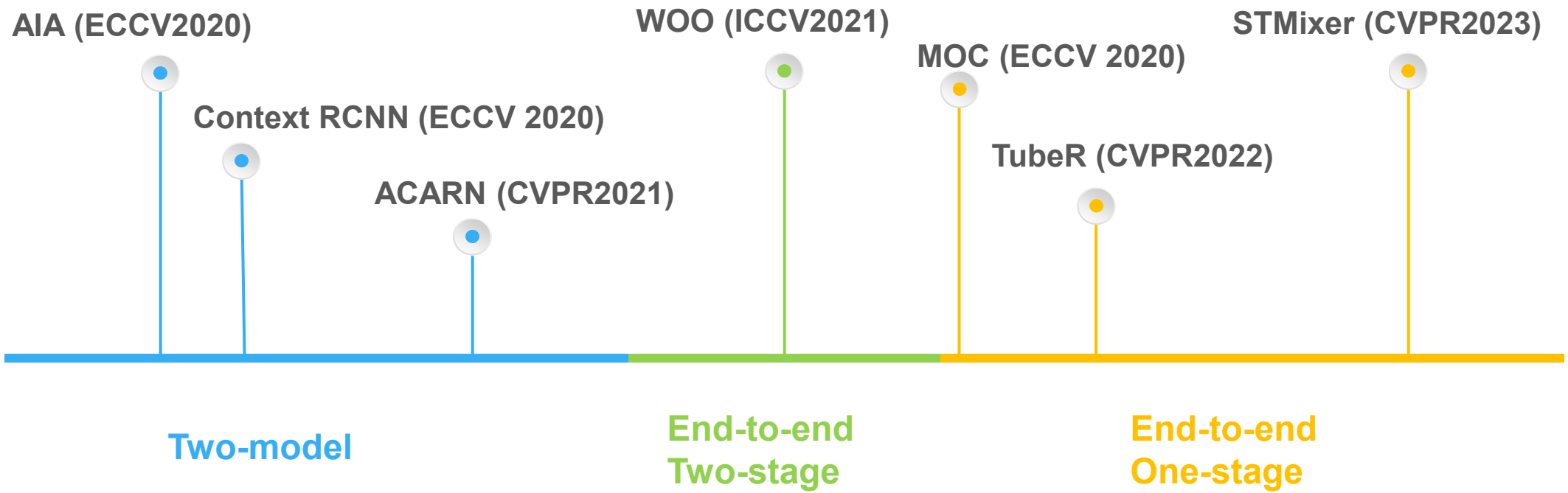
student	teacher	K400 top-1		SSv2 top-1	
		ViMAE	MVD	ViMAE	MVD
ViT-S	ViT-B	79.0	80.6 ↑**1.6**	66.4	70.7 ↑**4.3**
ViT-S	ViT-L	79.0	81.0 ↑**2.0**	66.4	70.9 ↑**4.5**
ViT-B	ViT-B	81.5	82.7 ↑**1.2**	69.7	72.5 ↑**2.8**
ViT-B	ViT-L	81.5	83.4 ↑**1.9**	69.7	73.7 ↑**4.0**
ViT-L	ViT-L	85.2	86.0 ↑**0.8**	74.0	76.1 ↑**2.1**

R. Wang et al., **Masked Video Distillation: Rethinking Masked Feature Modeling for Self-supervised Video Representation Learning**, in CVPR 2023

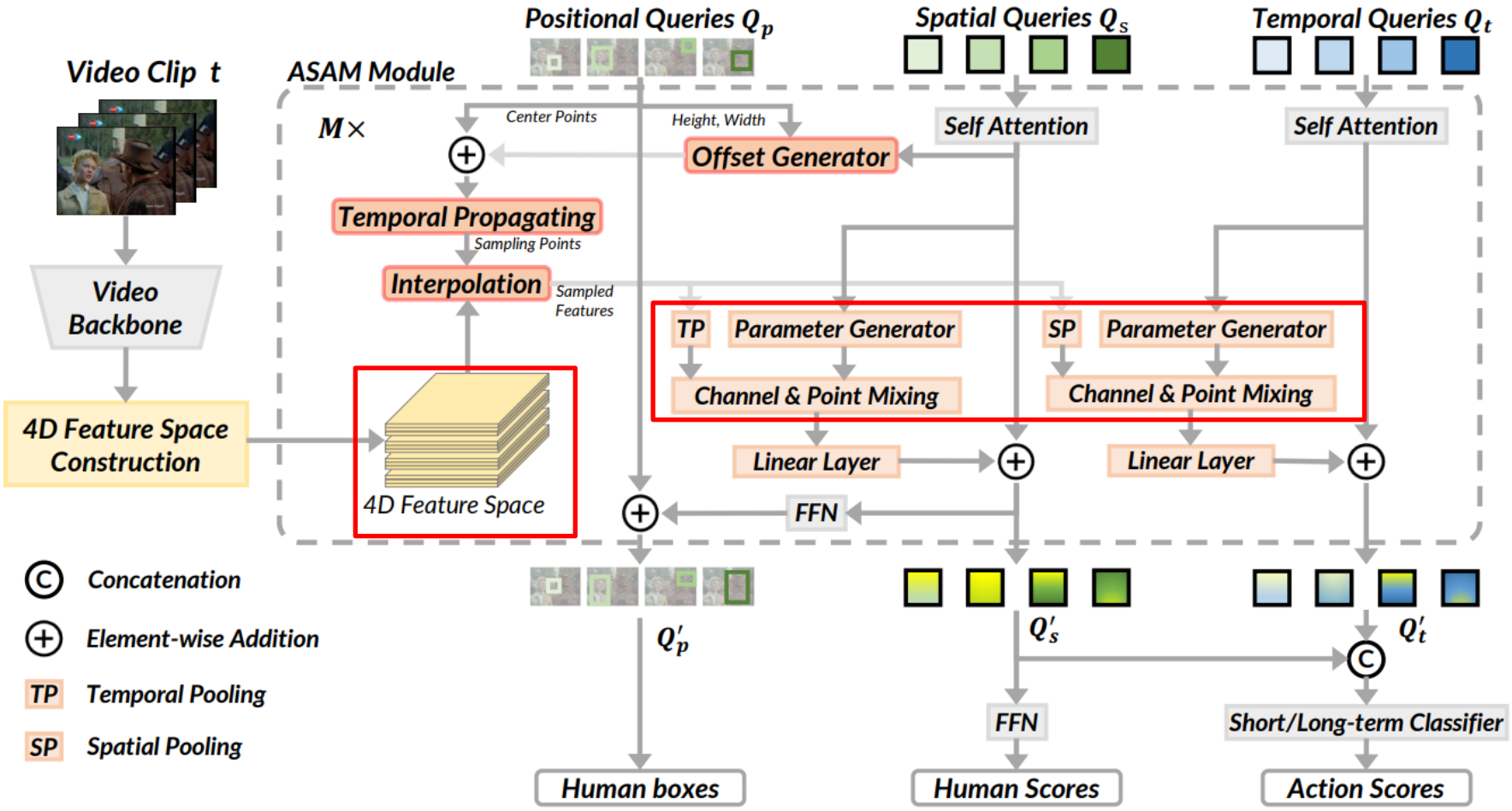
提纲

- 视频基础模型：结构设计与预训练策略
- **视频动作检测方法：时序与时空检测**
- 视频分析数据集: FineAction 和 MultiSports

时空动作检测方法归类

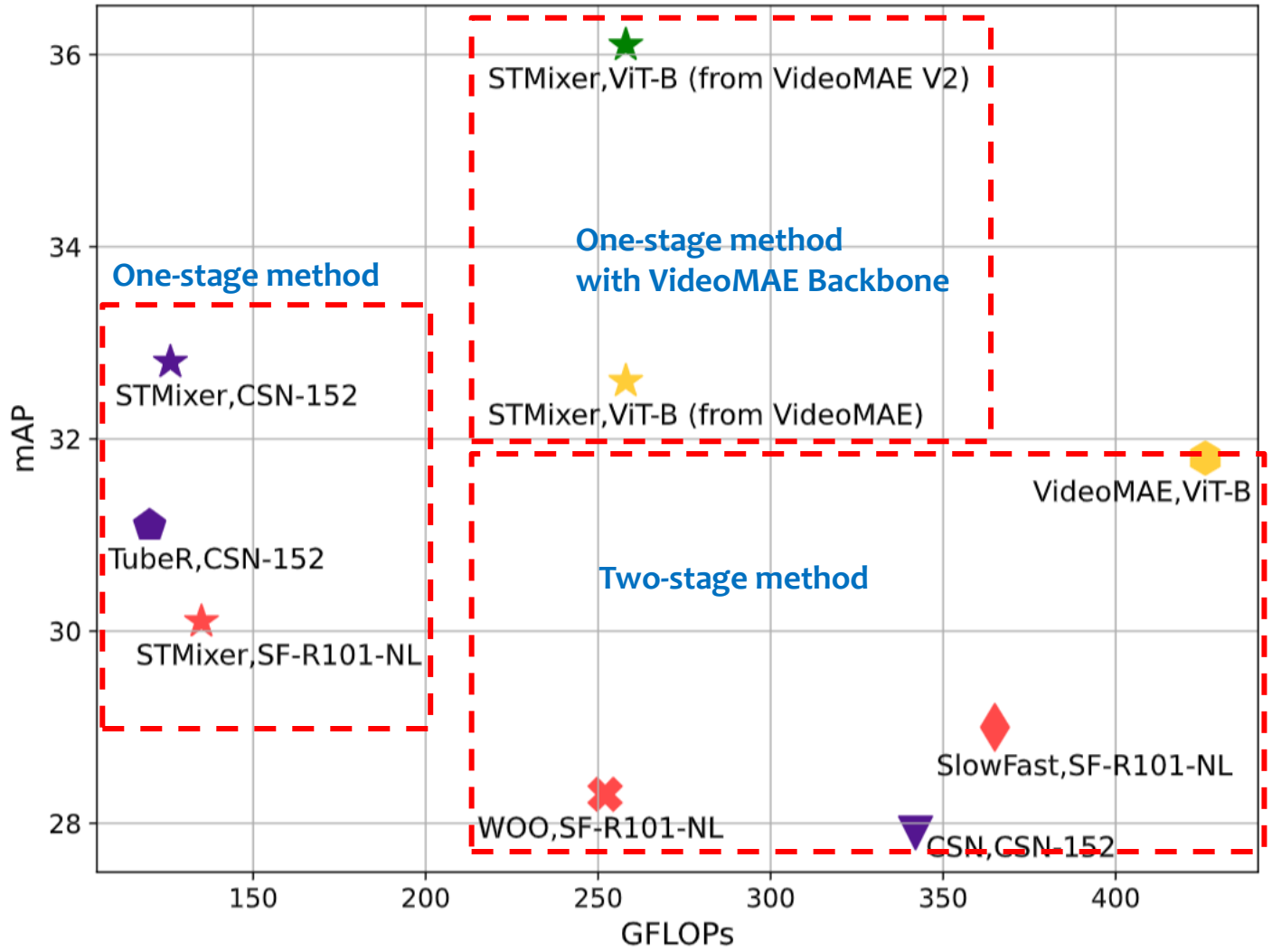

AIA (ECCV2020)
Context RCNN (ECCV 2020)
ACARN (CVPR2021)
WOO (ICCV2021)
MOC (ECCV 2020)
TubeR (CVPR2022)
STMixer (CVPR2023)
Two-model
End-to-end
Two-stage
End-to-end
One-stage

STMixer



T. Wu et.al, **STMixer: One-stage Sparse Action Detector**, in CVPR 2023

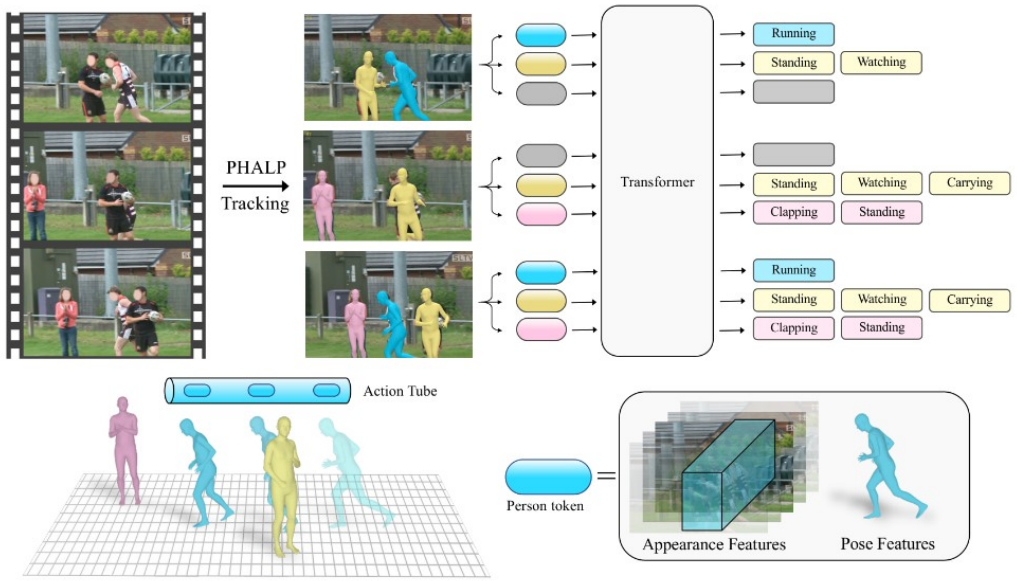
STMixer: Accuracy vs. Complexity



T. Wu et.al, **STMixer: One-stage Sparse Action Detector,** in CVPR 2023

最新工作1

- Introducing extra modal: LART (CVPR2023)
  - Use an offline 3D pose tracker to generate 3D pose tracklets
  - The representation of each person in each frames is a combination of pose embedding and appearance feature

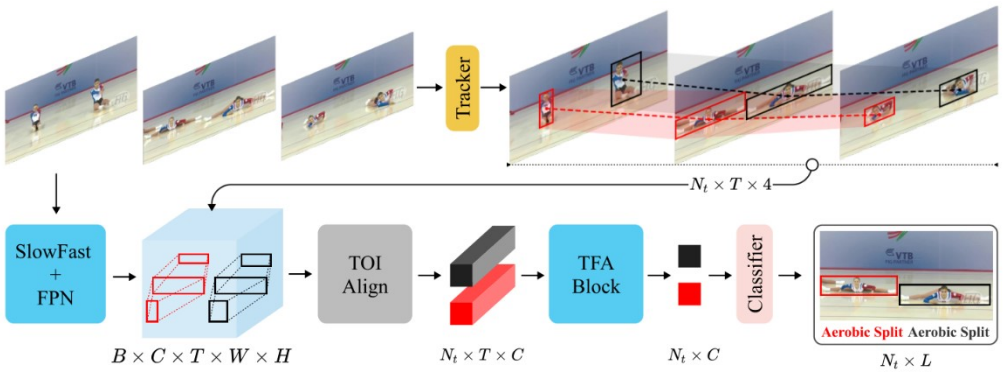


Model	Pretrain	mAP
SlowFast R101, 8×8 [18]	K400	23.8
MViTv1-B, 64×3 [14]		27.3
SlowFast 16×8 +NL [18]	K600	27.5
X3D-XL [16]		27.4
MViTv1-B-24, 32×3 [14]		28.7
Object Transformer [67]		31.0
ACAR R101, 8×8 +NL [44]		31.4
ACAR R101, 8×8 +NL [44]	K700	33.3
MViT-L↑312, 40×3 [38],	IN-21K+K400	31.6
MaskFeat [65]	K400	37.5
MaskFeat [65]	K600	38.8
Video MAE [17,55]	K600	39.3
Video MAE [17,55]	K400	39.5
LART	K400	42.3 (+2.8)

Jathushan Rajasegaran et al. On the Benefits of 3D Pose and Tracking for Human Action Recognition. CVPR 2023

最新工作2

- Introducing tracking information: TAAD (WACV2023)
  - 3D RoIAlign may fail to capture meaningful spatio-temporal features if the position or shape of the actor shows large motion and variability through the frames
  - Use an offline tracker to generated actor tracks
  - Given the actor tracks, RoIAlign can be performed in each frame (TOI-Align) and the features are then aggregated (TFA)



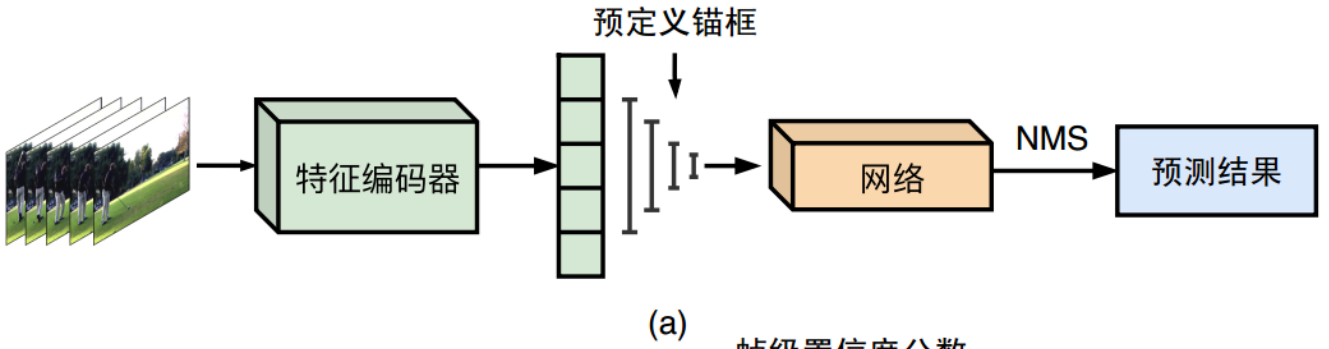
Method	f-mAP 0.5	v-mAP 0.2	v-mAP 0.5	v-mAP .1:.9
YOWO [20, 21]	25.2	12.9	9.7	–
MOC [20, 21]	25.2	12.9	9.7	–
SlowFast-R50 [12, 20]	27.7	24.2	9.7	–
SlowFast-R101 [27]	29.5	28.1	8.4	12.3
SlowFast-R101+PCCA [27]	42.2	41.0	20.0	20.9
Baseline (ours)	49.6	54.1	31.3	28.9
Baseline + tracks (ours) †	50.6	56.3	33.0	30.9
TAAD + MaxPool (ours)	53.9	58.6	34.8	32.4
TAAD + ASPP (ours)	54.4	59.2	36.0	33.0
TAAD + TCN (ours)	**55.3**	**60.6**	**37.0**	**33.7**

Setting new SOTA on Multisports dataset with large motion

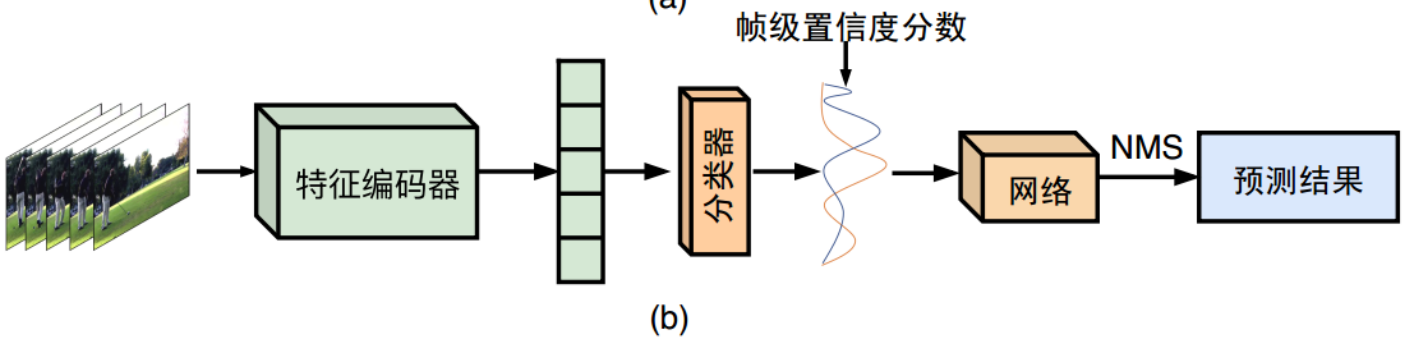
Gurkirt Singh et al. Spatio-Temporal Action Detection Under Large Motion. WACV 2023

时序动作检测方法归类

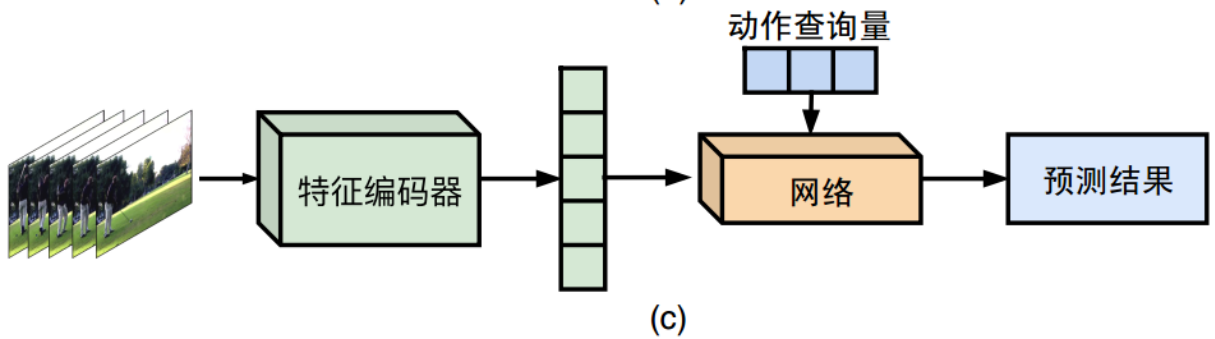
自顶向下



自底向上

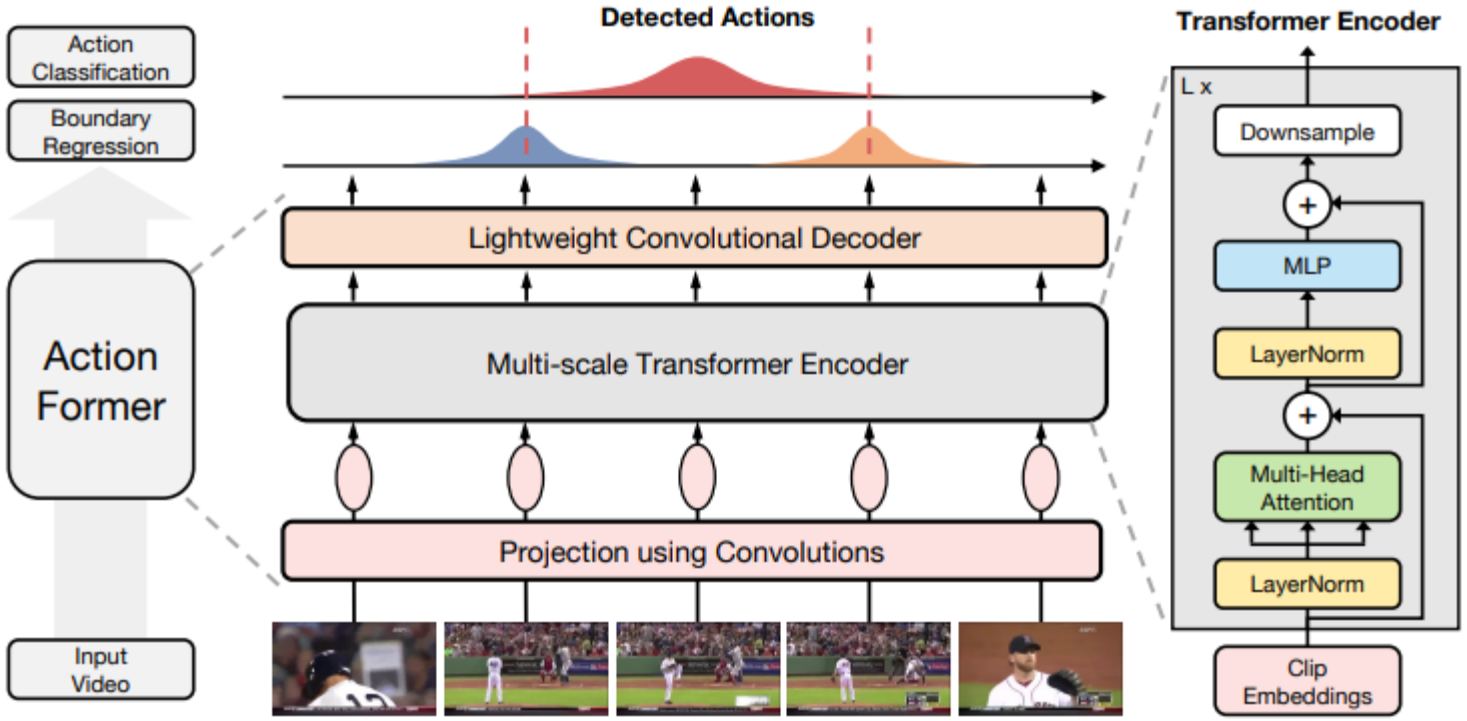


稀疏查询



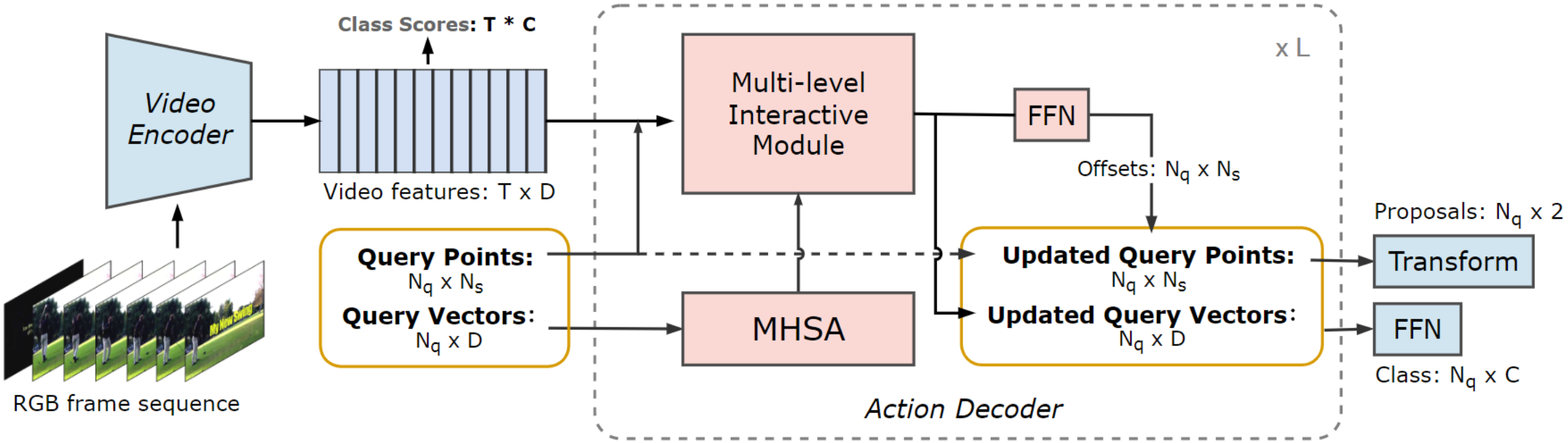
自顶向下：ActionFormer

- ActionFormer: Localizing Moments of Actions with Transformers (ECCV2022)



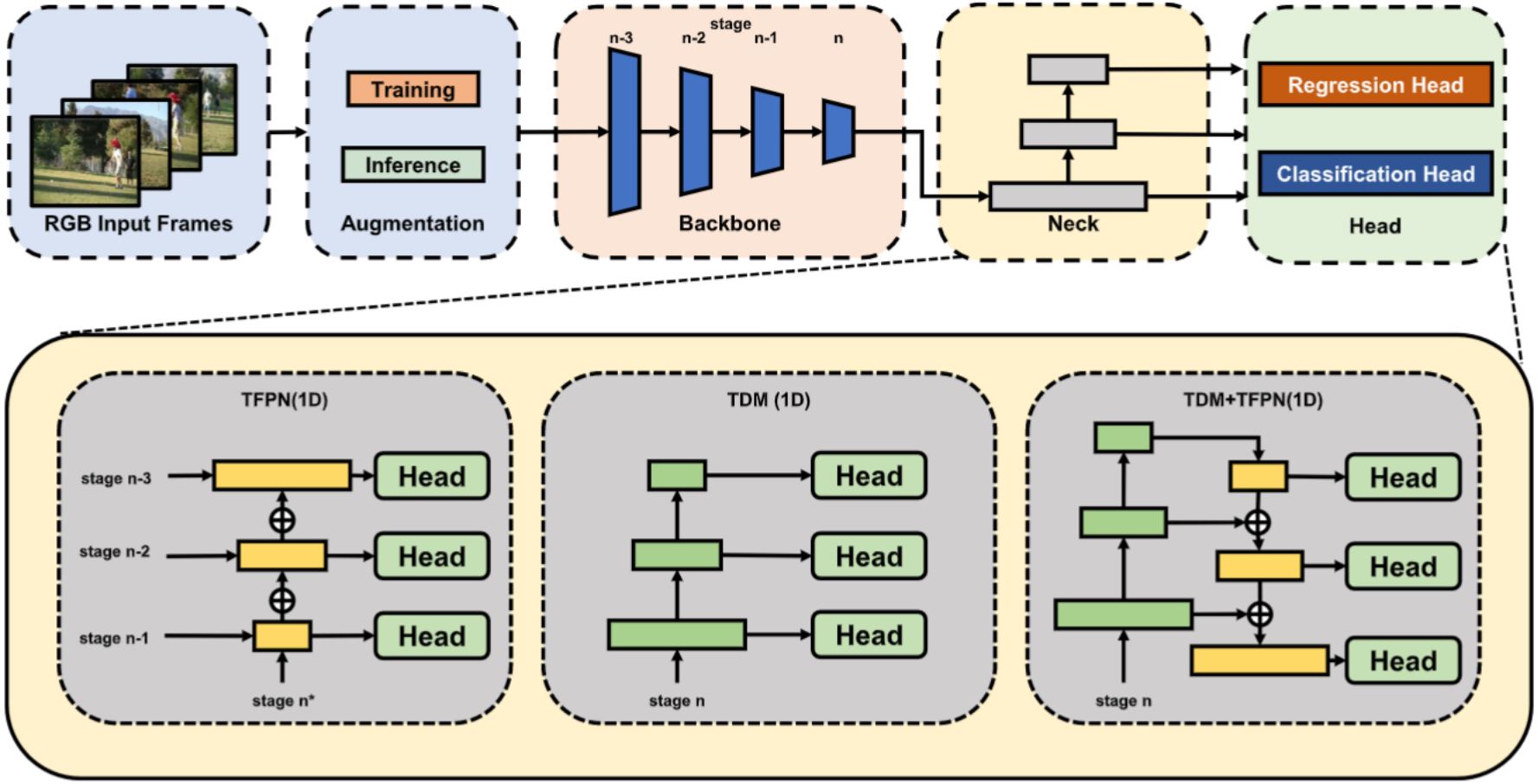
稀疏查询：PointTAD

- PointTAD: Multi-Label Temporal Action Detectionwith Learnable Query Points (NeurIPS 2022)



端到端方法：BasicTAD

- **BasicTAD: An astounding RGB-Only baseline for temporal action detection, in CVIU 2023.**



端到端方法：BasicTAD

Table 4
Study on the effectiveness of end-to-end training based on the anchor-based BasicTAD with different backbones on THUMOS14 (Jiang et al., 2014). "e2e" is short for end-to-end training. We freeze all layers in the backbone to construct a non-end-to-end training manner.

Backbone	e2e	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7	Avg
C3D	✔	**54.4**	**50.8**	**45.7**	**37.4**	**26.1**	**42.9**
	✗	32.0	27.2	22.0	14.4	7.6	20.6
I3D	✔	**59.5**	**56.0**	**51.4**	**41.8**	**28.3**	**47.4**
	✗	35.2	29.7	23.1	15.9	8.1	22.4
R50-I3D	✔	**62.8**	**59.5**	**53.8**	**43.6**	**30.1**	**50.0**
	✗	36.5	31.8	26.7	19.6	12.3	25.4
SlowOnly (×8)	✔	**63.1**	**59.5**	**54.3**	**43.6**	**30.5**	**50.2**
	✗	37.3	32.4	26.6	19.8	13.0	25.8

M. Yang et al., BasicTAD: An astounding RGB-Only baseline for temporal action detection, in CVIU 2023

端到端方法：BasicTAD

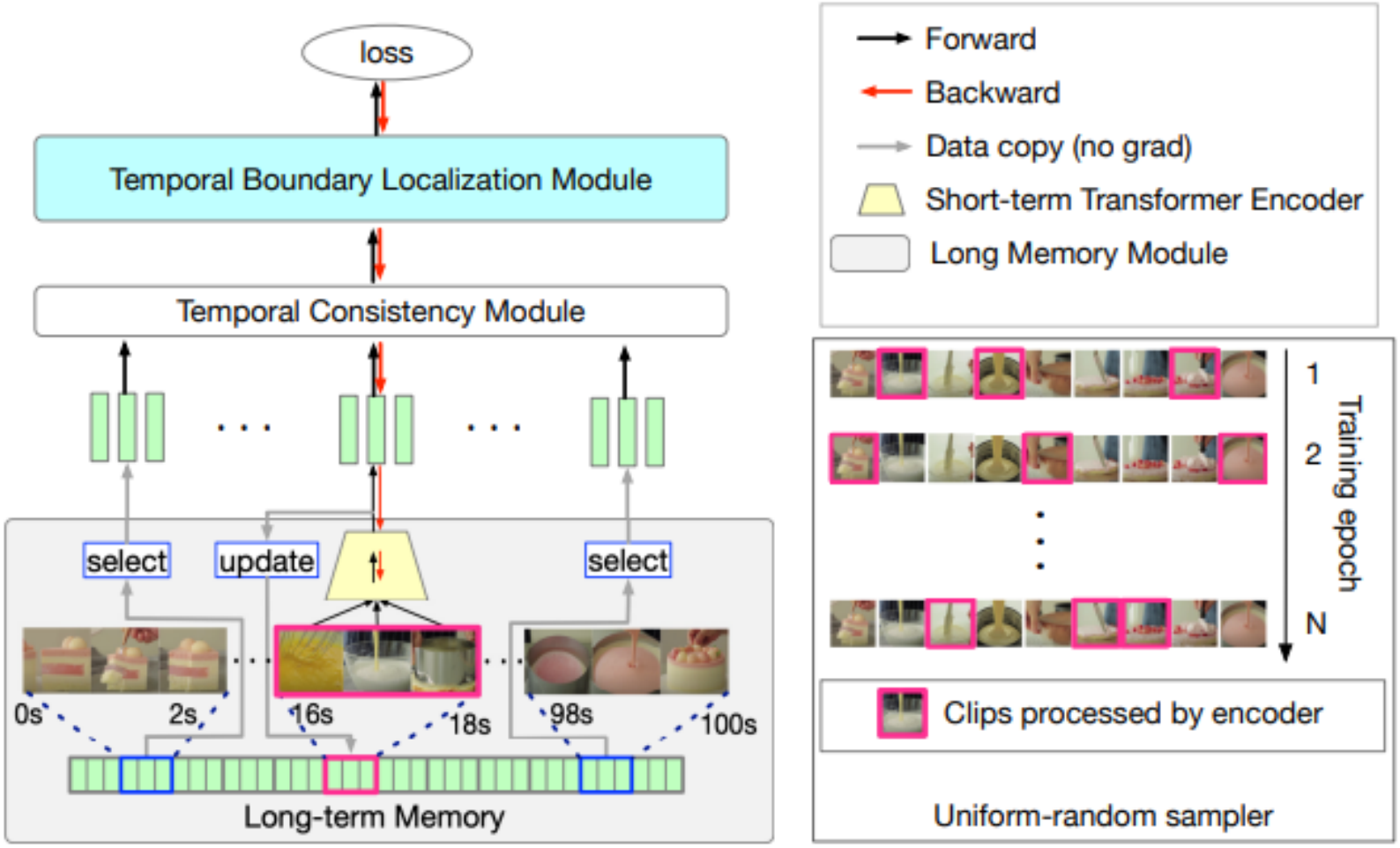
Table 13
Comparison with state of the art on the THUMOS14. "RGB-Only" means whether to use other input modalities besides RGB input.

Type	Method	Backbone	RGB-Only	mAP @0.3	mAP @0.4	mAP @0.5	mAP @0.6	mAP @0.7	mAP @Avg	FLOPs
Multi-stage	BSN (Lin et al., 2018)	TSN	✗	53.5	45.0	36.9	28.4	20.0	36.8	–
	MGG (Liu et al., 2019)	TSN	✗	53.9	46.8	37.4	29.5	21.3	37.8	–
	BMN (Lin et al., 2019b)	TSN	✗	56.0	47.4	38.8	29.7	20.5	38.5	–
	DBG (Lin et al., 2019b)	TSN	✗	57.8	49.4	39.8	30.2	21.7	39.8	–
	RTD-Net (Tan et al., 2021)	I3D	✗	58.5	53.1	45.1	36.4	25.0	43.6	–
	TCANet (Qing et al., 2021)	TSN	✗	60.6	53.2	44.6	36.8	26.7	44.4	–
	G-TAD (Xu et al., 2020)	TSN	✗	66.4	60.4	51.6	37.6	22.9	47.8	–
	AFSD (Lin et al., 2021)	I3D	✗	67.3	62.4	55.5	43.7	31.1	52.0	2780.0G
	DCAN (Chen et al., 2022)	TSN	✗	68.2	62.7	54.1	43.9	32.6	52.3	–
	SP-TAD (Wu et al., 2021)	I3D	✗	69.2	63.3	55.9	45.7	33.4	53.5	–
	TadTR (Liu et al., 2022a)	R50-SlowFast	✓	69.4	64.3	56.0	46.4	34.9	54.2	475.0G
One-stage	SSAD (Lin et al., 2017c)	TSN	✗	43.0	35.0	24.6	–	–	–	–
	DBS (Gao et al., 2019)	TSN	✗	50.6	43.1	34.3	24.4	14.7	33.4	–
	A2Net (Yang et al., 2020)	I3D	✗	58.6	54.1	45.5	32.5	17.2	41.6	–
	PBRNet (Liu and Wang, 2020)	I3D	✗	58.5	54.6	51.3	41.8	29.5	47.1	–
	R-C3D (Xu et al., 2017)	C3D	✓	44.8	35.6	28.9	–	–	–	1360.0G
	GTAN (Long et al., 2019)	P3D	✓	57.8	47.2	38.8	–	–	–	–
	DaoTAD (Wang et al., 2021a)	R50-I3D	✓	62.8	59.5	53.8	43.6	30.1	50.0	206.7G
	$\text{DaoTAD}^{112}_{3,96}$ (Wang et al., 2021a)	R50-SlowOnly	✓	63.2	59.7	54.4	45.6	32.2	51.0	133.3G
	$\text{PlusTAD}^{112}_{3,96}$ (Anchor-based)	R50-SlowOnly	✓	**68.4**	**65.0**	**58.6**	**49.2**	**33.5**	**54.9**	136.4G
	$\text{PlusTAD}^{112}_{3,96}$ (Anchor-free)	R50-SlowOnly	✓	**70.4**	**65.5**	**57.6**	**46.0**	**33.2**	**54.5**	151.5G
	$\text{PlusTAD}^{160}_{6,192}$ (Anchor-based)	R50-SlowOnly	✓	**72.3**	**68.4**	**62.0**	**52.4**	**37.0**	**58.4**	519.3G
	$\text{PlusTAD}^{160}_{6,192}$ (Anchor-free)	R50-SlowOnly	✓	**75.5**	**70.8**	**63.5**	**50.9**	**37.4**	**59.6**	533.1G

M. Yang et al., BasicTAD: An astounding RGB-Only baseline for temporal action detection, in CVIU 2023

端到端方法：TallFormer

- TallFormer: Temporal Action Localization with a Long-memory Transformer (ECCV 2022)



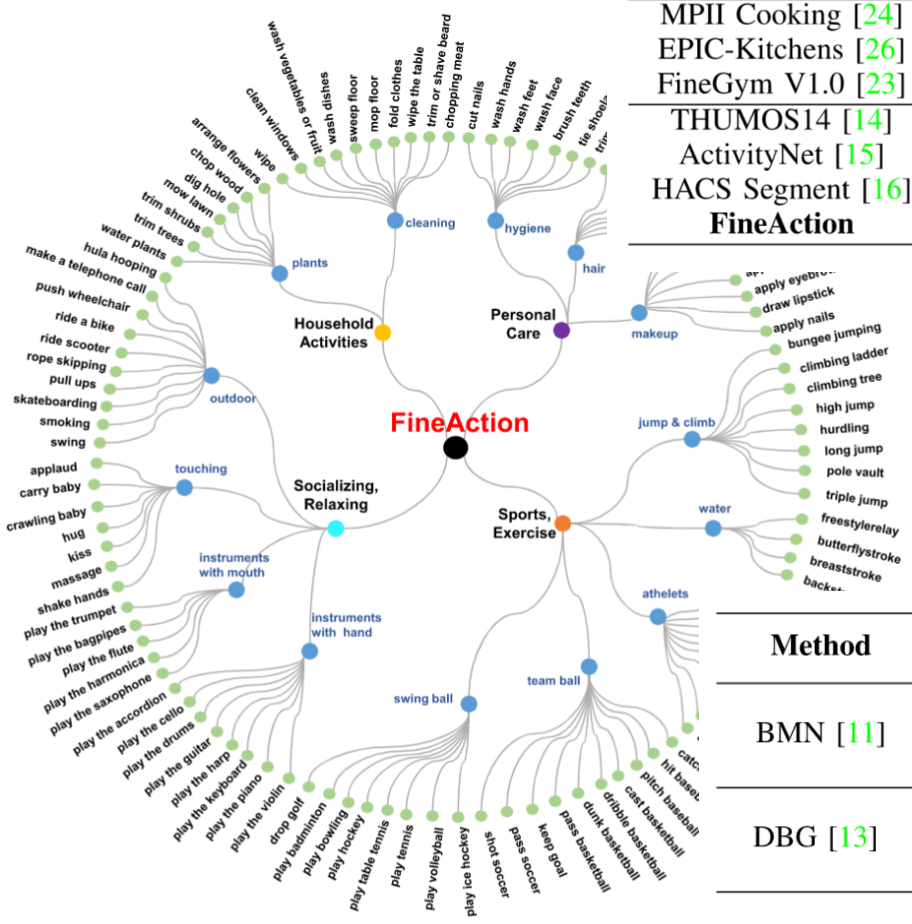
提纲

- 视频基础模型：结构设计与预训练策略
- 视频动作检测方法：时序与时空检测
- **视频分析数据集**: FineAction 和 MultiSports

FineAction: A Fine-Grained Video Dataset for Temporal Action Localization

Yi Liu, Limin Wang, *Member, IEEE*, Yali Wang, Xiao Ma, and Yu Qiao, *Senior Member, IEEE*

Statistics & Performance



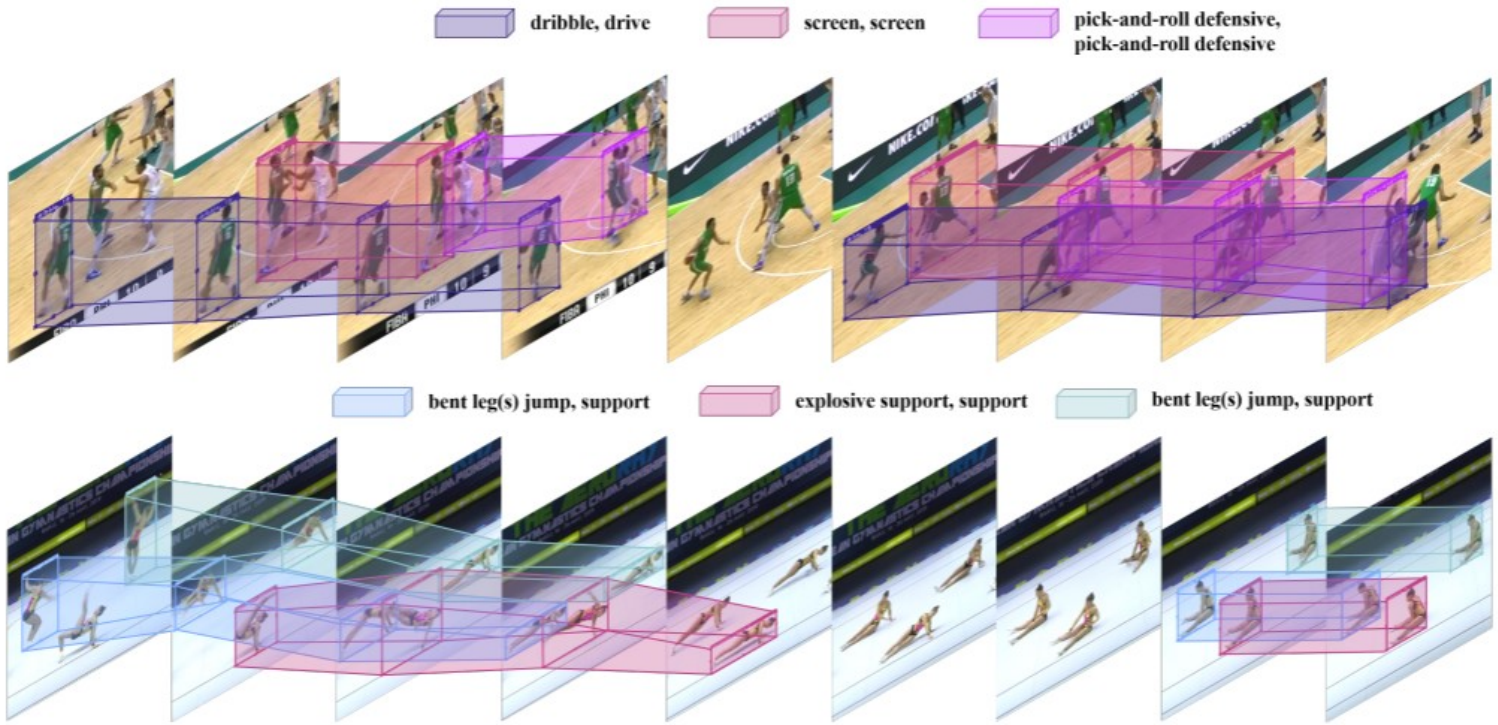
Database	Category	M-L	Video	Instance	Overlap	Duration	Action type	Main task
MPII Cooking [24]	65	✓	45	5,609	0.1%	11.1 m	kitchens	Action Classification
EPIC-Kitchens [26]	4,025	✓	700	89,979	28.1%	3.1 s	kitchens	Action Classification
FineGym V1.0 [23]	530	✓	303	32,697	0.0%	1.7 s	sports	Action Classification
THUMOS14 [14]	20	×	413	6,316	17.5%	4.3 s	sports	Temporal Action Localization
ActivityNet [15]	200	×	19,994	23,064	0.0%	49.2 s	daily events	Temporal Action Localization
HACS Segment [16]	200	×	49,485	122,304	0.0%	33.2 s	daily events	Temporal Action Localization
FineAction	106	✓	16,732	103,324	11.5%	7.1 s	daily events	Temporal Action Localization

Database	0-2 s	2-6 s	6-15 s	>15 s	Ins / Vid
THUMOS14 [14]	2,029	2,753	1,437	99	15.29
ActivityNet [15]	900	3,253	4,426	14,485	1.15
HACS Segment [16]	8,874	29,644	31,982	51,804	2.47
FineAction	**66,890**	15,253	10,523	10,586	**6.17**

Method	Modality	Action Proposal Generation				Temporal Action Localization			
		AR@5	AR@10	AR@100	AUC	mAP@0.50	mAP@0.75	mAP@0.95	Avg.mAP
BMN [11]	RGB	8.62	11.20	22.74	17.49	12.56	7.49	2.62	7.86
	Flow	9.85	12.72	24.18	18.94	14.49	8.92	3.19	9.23
	RGB+Flow	9.99	12.84	24.34	19.19	14.44	8.92	3.12	9.25
DBG [13]	RGB	6.82	9.01	21.26	15.48	8.57	5.01	1.93	5.31
	Flow	8.27	10.90	23.37	17.70	11.03	6.95	2.70	7.20
	RGB+Flow	7.82	10.45	23.07	17.24	10.65	6.43	2.50	6.75
G-TAD [12]	RGB	7.96	10.45	20.86	16.06	10.88	6.52	2.19	6.87
	Flow	8.87	11.60	22.01	17.09	12.58	8.18	2.56	8.26
	RGB+Flow	9.02	11.83	23.17	17.65	13.74	8.83	3.06	9.06
Ours	RGB+Flow	11.47	15.10	29.61	23.07	22.01	12.09	3.88	13.17

MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions

Yixuan Li Lei Chen Runyu He Zhenzhi Wang Gangshan Wu Limin Wang*
State Key Laboratory for Novel Software Technology, Nanjing University, China



Benchmark

Method	Res	MultiSports			UCF101-24			JHMDB			AVA
		F@0.5	V@0.2	V@0.5	F@0.5	V@0.2	V@0.5	F@0.5	V@0.2	V@0.5	F-mAP@0.5
ROAD [43]	300 × 300	3.90	0.00	0.00	70.7	69.8	40.9	-	60.8	59.7	-
YOWO [22]	224 × 224	9.28	10.78	0.87	71.10	72.97	46.42	74.51	88.05	82.57	-
MOC [26] (K=7)	288 × 288	22.51	12.13	0.77	78.0	82.8	53.8	70.8	77.3	77.2	-
MOC [26] (K=11)	288 × 288	25.22	12.88	0.62	-	-	-	-	-	-	-
SlowOnly Det., 4 × 16 [10]	short side 256	16.70	15.71	5.50	-	-	-	-	-	-	20.02
SlowFast Det., 4 × 16 [10]	short side 256	27.72	24.18	9.65	-	-	-	-	-	-	24.56

Table 3. Comparison to state-of-the-art methods on *MultiSports*, UCF101-24, JHMDB and AVA

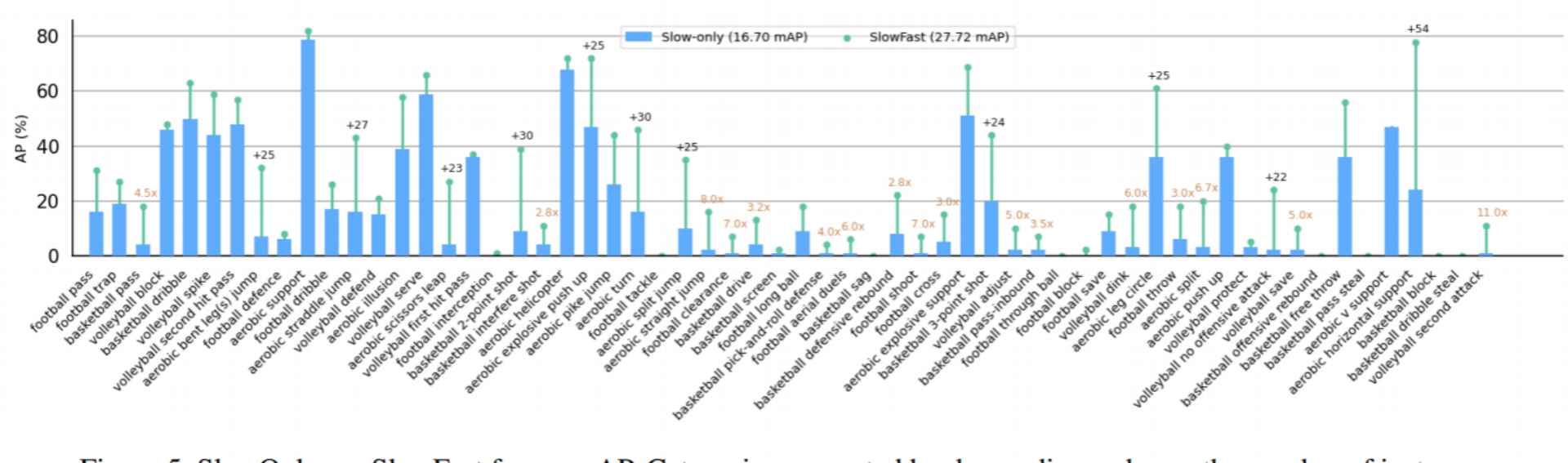


Figure 5. SlowOnly vs. SlowFast frame-mAP. Categories are sorted by descending order on the number of instances.

总结

- 视频表示基础模型
  - Transformer模型结构成为主流
  - 自监督预训练方法成为主流
  - **如何构建更加简洁、通用的视频基础模型?**

- 视频动作检测方法
  - 端到端的检测方案越来越重要（容易拥抱强大基础模型）
  - 基于稀疏查询的检测范式越来越流行
  - **如何构建更加统一、灵活的视频检测范式?**

- 视频动作数据集：细粒度动作分析数据集

Code & Model

- VideoMAE code & model
  - https://github.com/MCG-NJU/VideoMAE
  - https://github.com/OpenGVLab/VideoMAEv2
  - https://github.com/OpenGVLab/InternVideo
  - https://github.com/OpenGVLab/Ask-Anything
- Temporal and spatial action detection code
  - https://github.com/MCG-NJU/AdaMixer
  - https://github.com/MCG-NJU/STMixer
  - https://github.com/MCG-NJU/PointTAD
  - https://github.com/MCG-NJU/BasicTAD

https://github.com/MCG-NJU

https://github.com/OpenGVLab