视频内容分析：现状与六大趋势

姜育刚

Fudan Vision and Learning Lab

https://fvl.fudan.edu.cn/

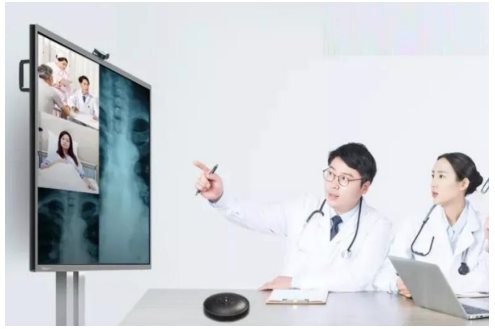
@MLA2022　2022.11.05

视频大数据

- 视频正逐步成为信息化传播的主流方式

视频监控

视频购物



视频医疗

视频教育

海量的视频数据

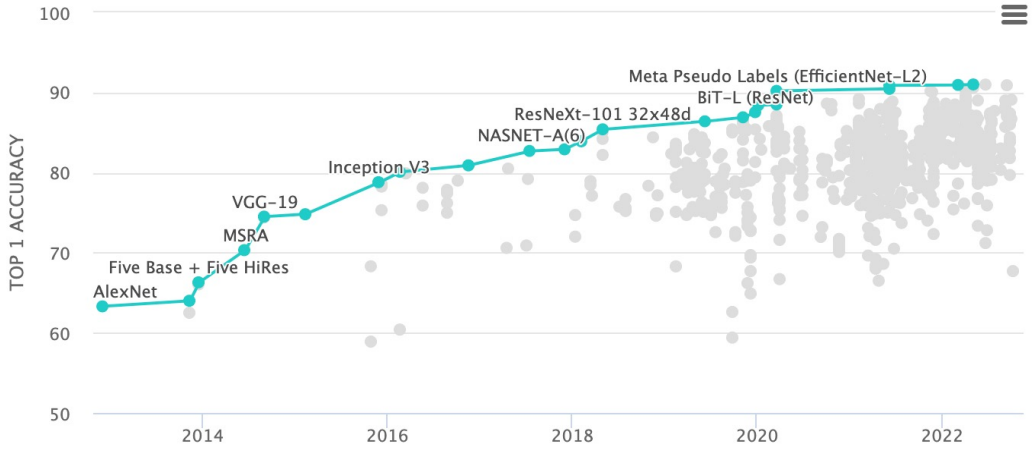
《思科公司2022报告》

- 视频数据超过了**82%互联网数据**
- **1秒**流经网络的视频内容有**100万分钟**

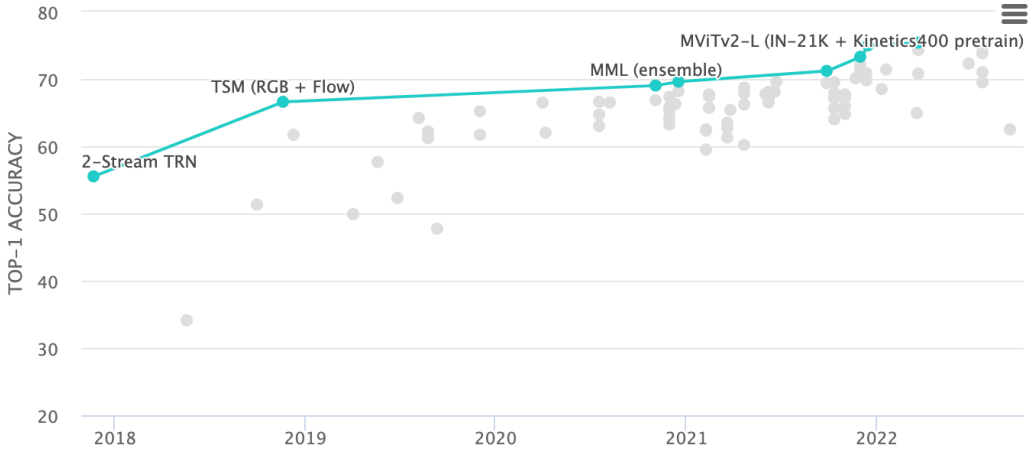
亟需发展高效的视频内容分析技术

视频大数据

- 相比于图像内容分析，视频内容分析更具有挑战性

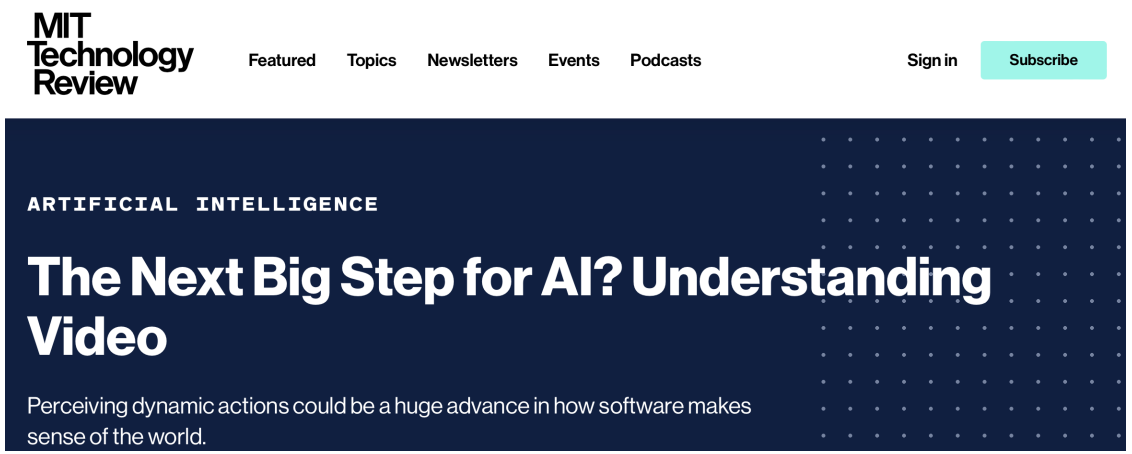


ImageNet



Something-Something V2

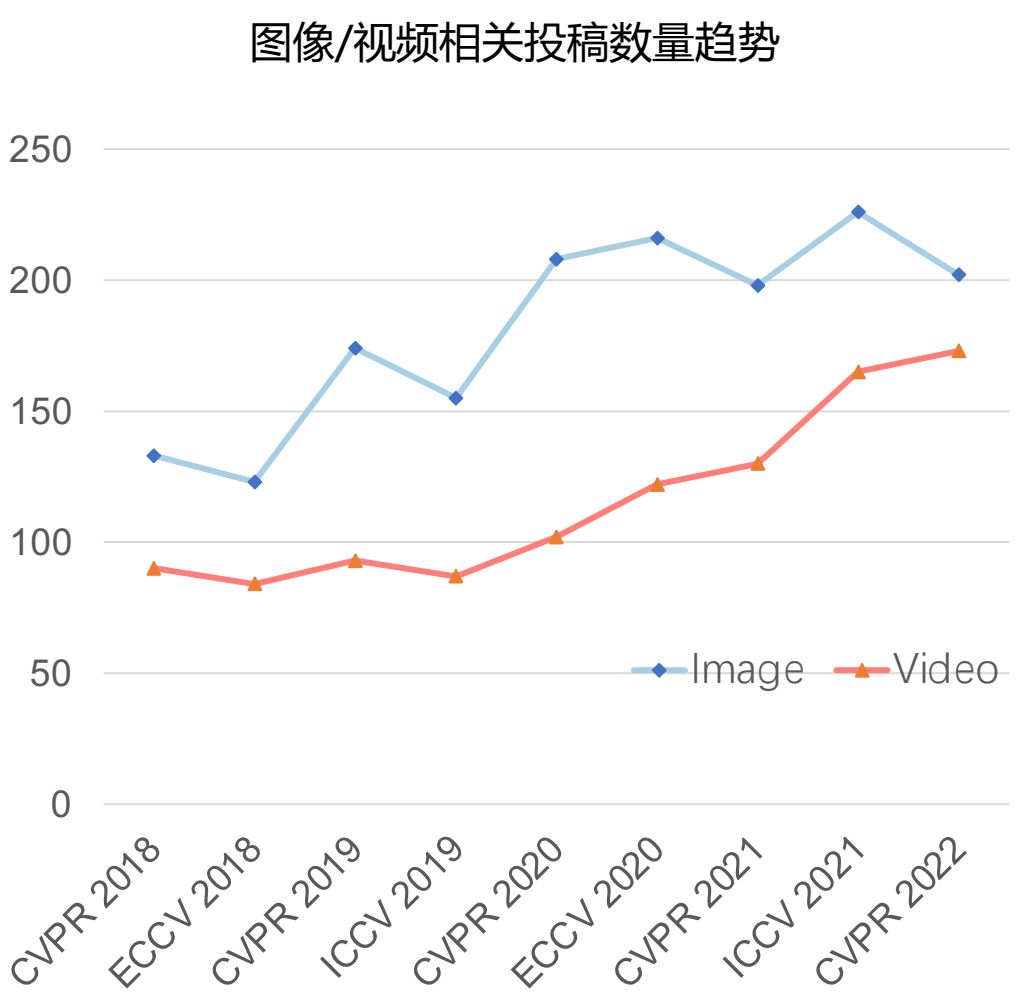
- 人工智能的下一个重要进展是视频理解

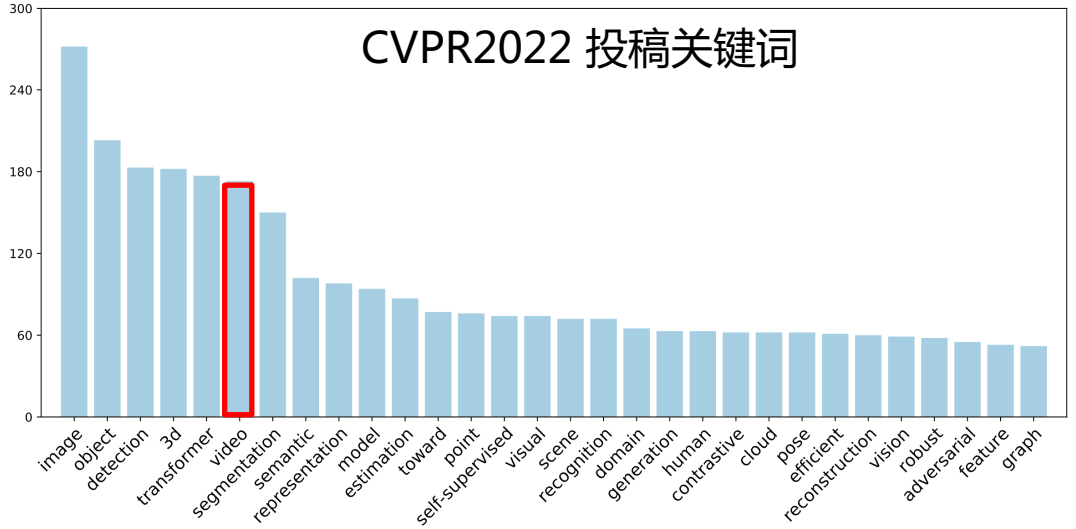


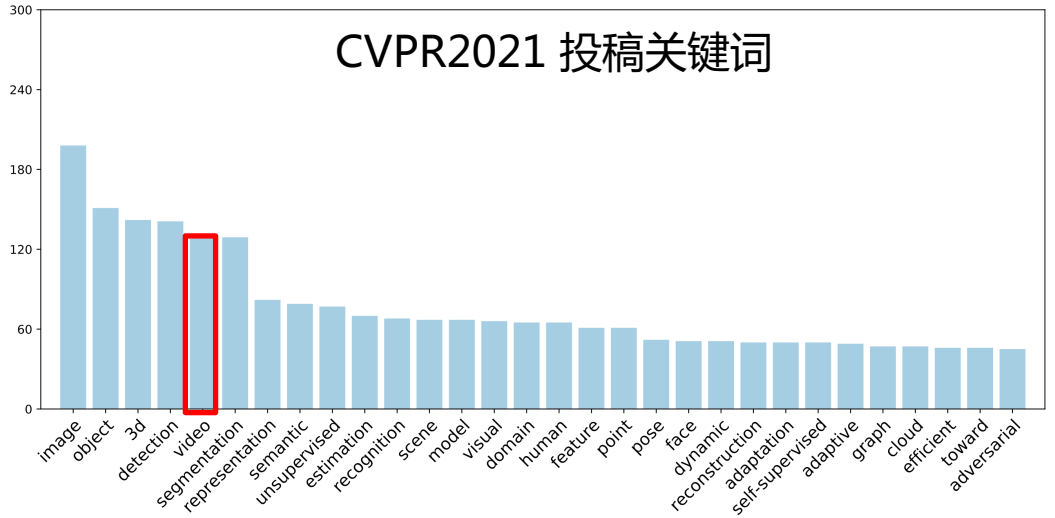
The next challenge may be teaching machines to understand not just what a video contains, but what's happening in the footage as well. That could have some practical benefits, perhaps leading to powerful new ways of searching, annotating, and mining video footage. It also figures to give robots or self-driving cars a better understanding of how the world around them is unfolding.

视频内容分析是计算机视觉前沿问题

- 研究热度持续攀升







提纲

技术路线回顾

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

趋势四：知识融入从外部到联合

趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

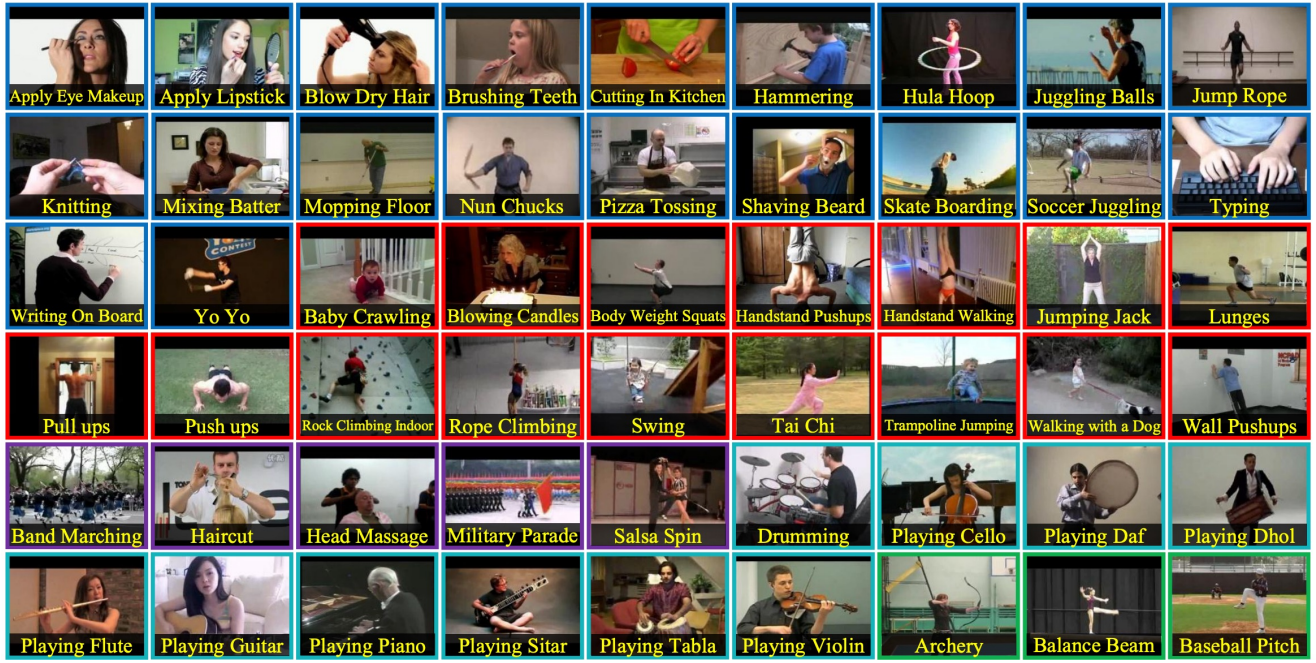
技术路线回顾

技术路线回顾

2012

UCF-101 dataset

(13300个视频, 101个动作类别)

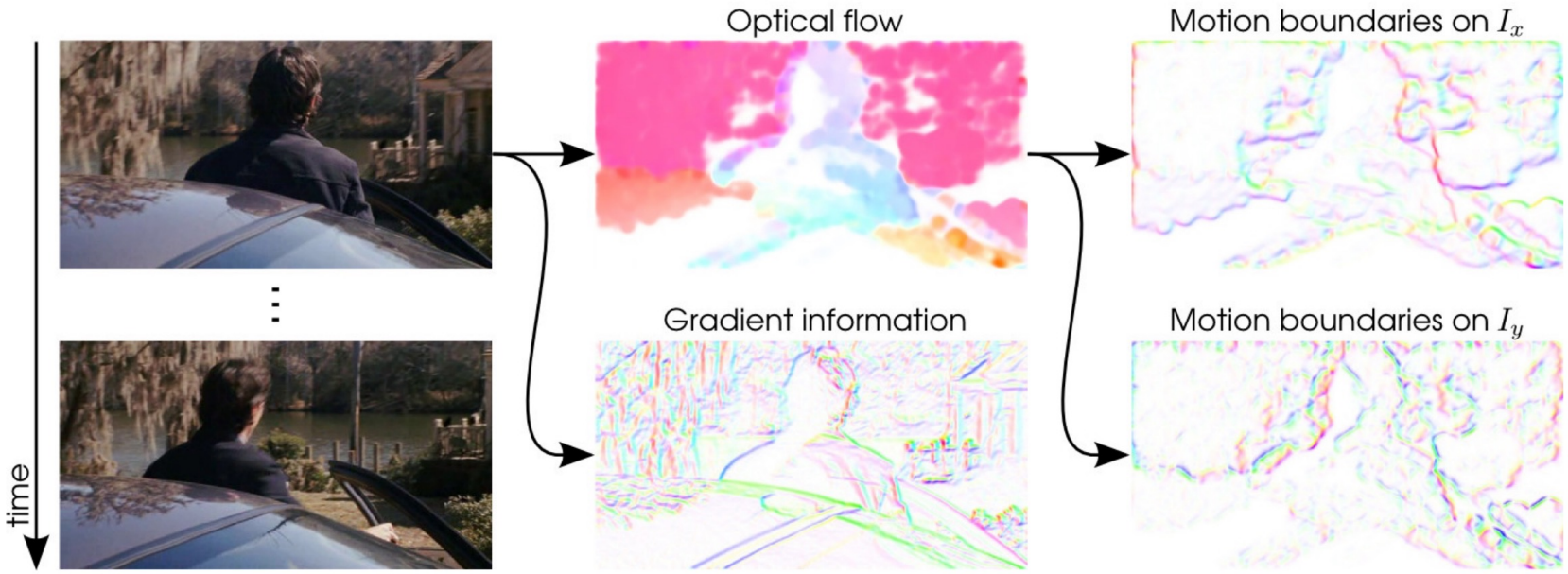


Khurram Soomro et al, “UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild”, 2012

技术路线回顾

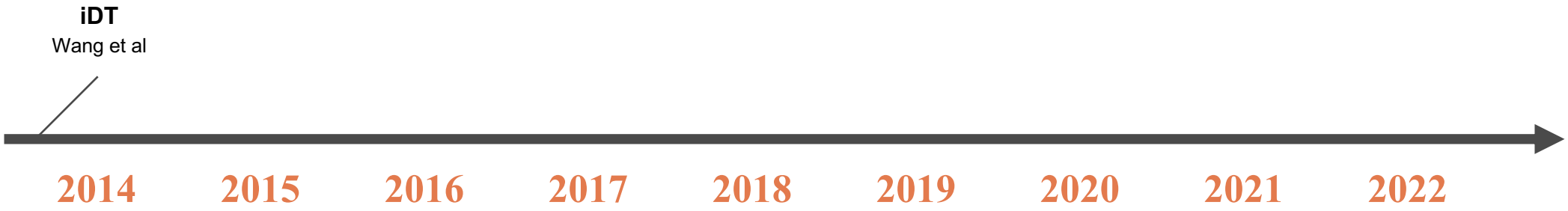
2013

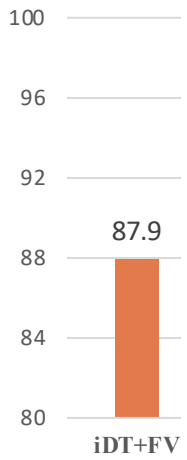
Improved Dense Trajectory



Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." CVPR 2013.

技术路线回顾





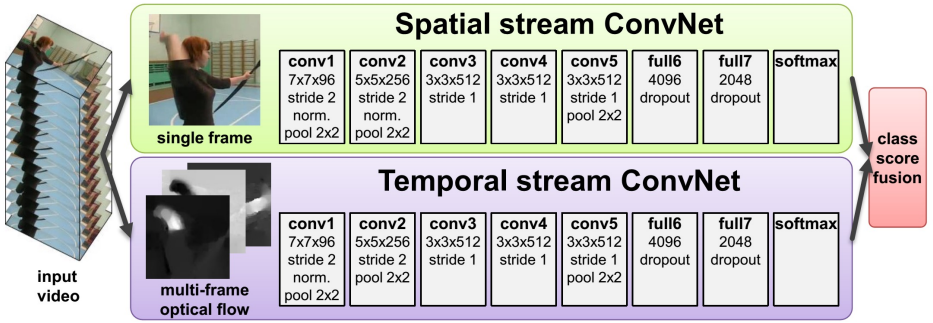
UCF-101

(13.3K videos, 101 action classes)

技术路线回顾

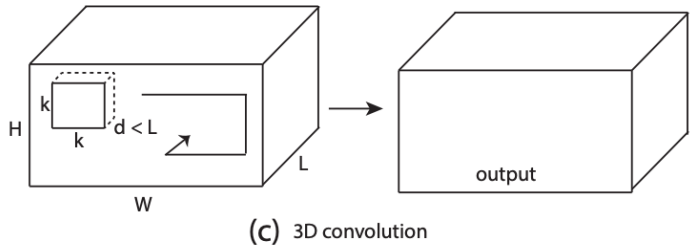
2014~2015

Two-stream networks



Karen Simonyan. et al. “Two-stream convolutional networks for action recognition in videos.” NeurIPS 2014.

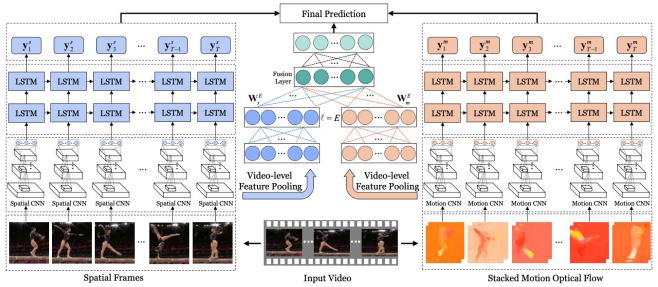
C3D



(c) 3D convolution

Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." CVPR 2015.

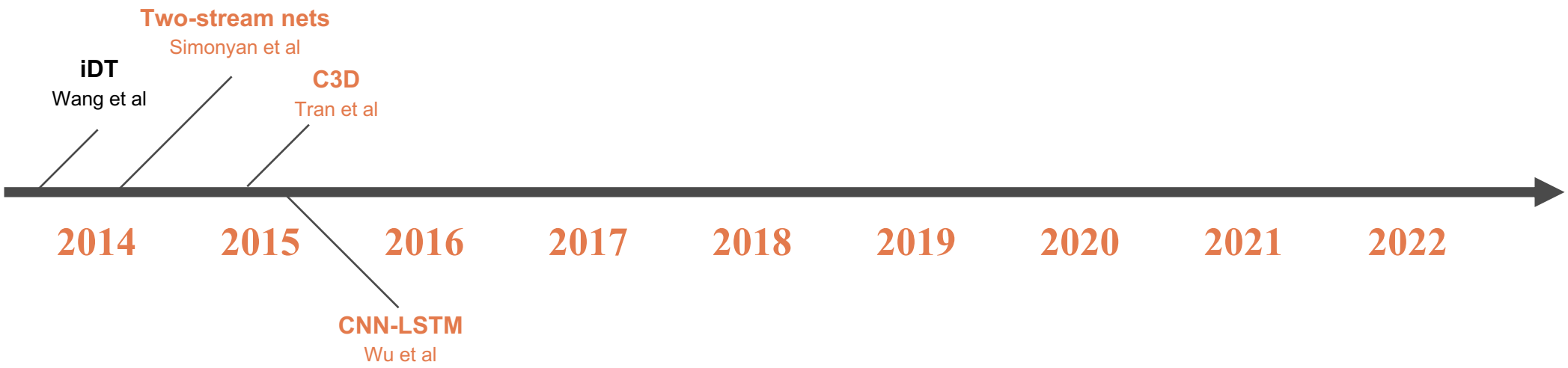
CNN-LSTM framework

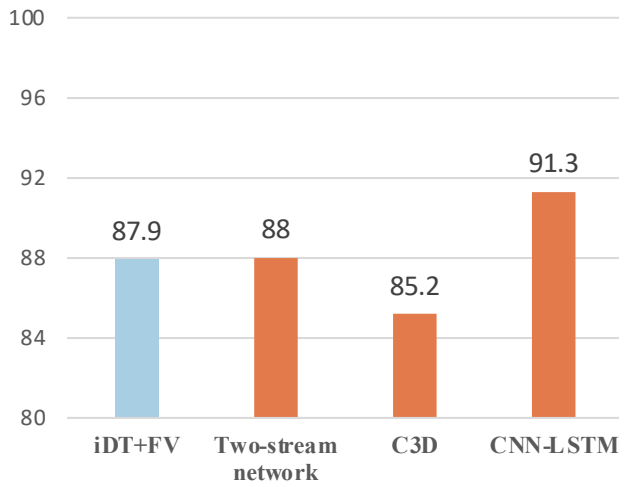


Zuxuan, Wu, et al. “Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification.” MM 2015.

(13.3K videos, 101 action classes)

技术路线回顾





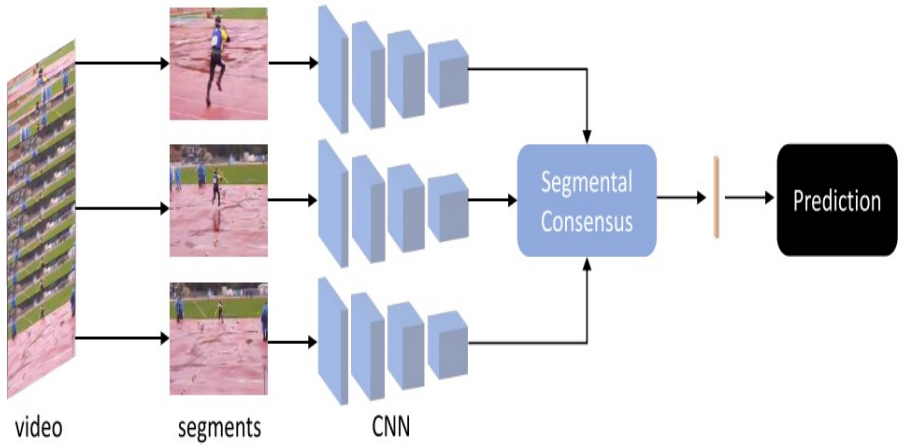
UCF-101

(13.3K videos, 101 action classes)

技术路线回顾

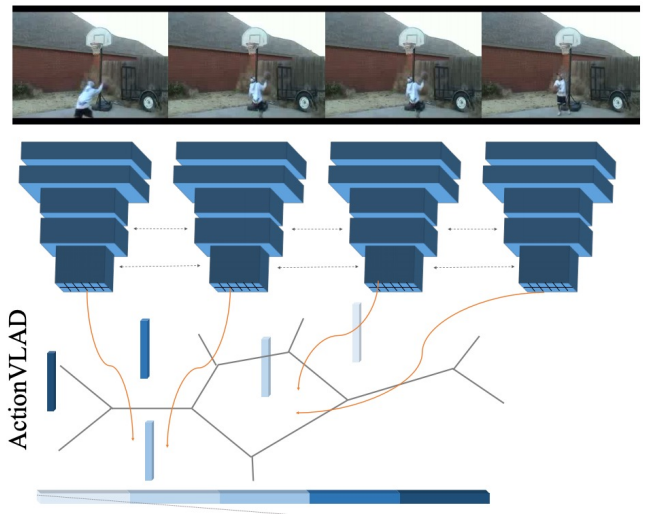
2016

Temporal Segment Network (TSN)



Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." ECCV 2016.

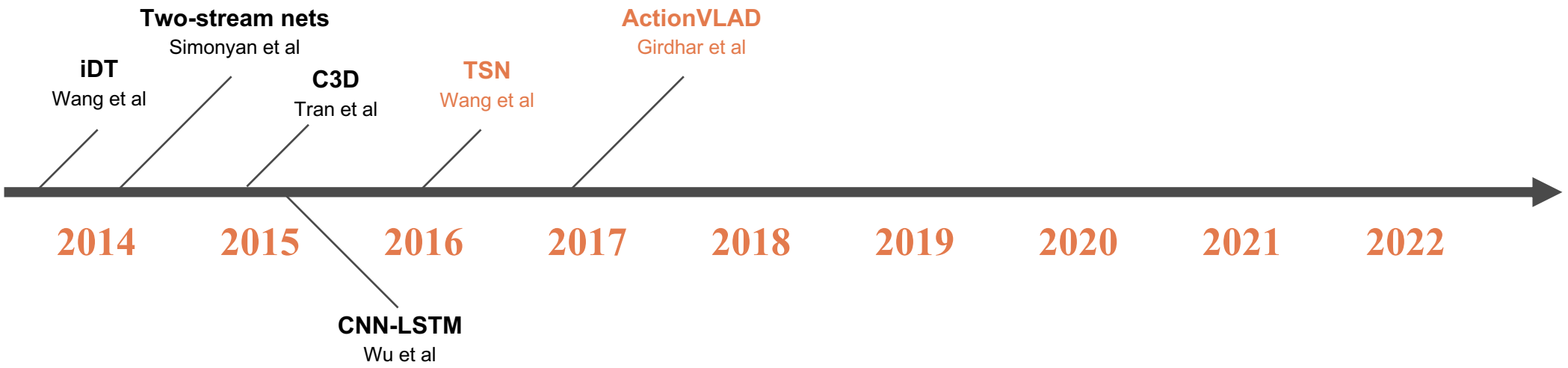
ActionVLAD

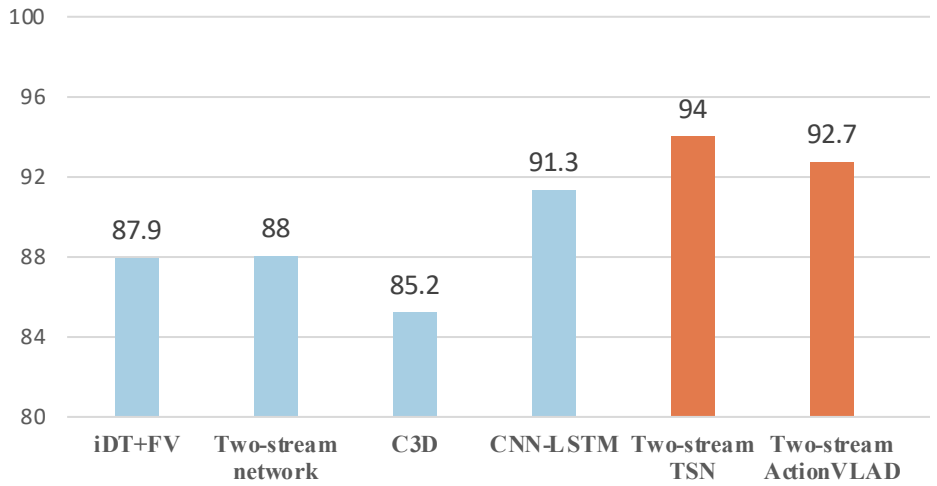


Girdhar, Rohit, et al. "Actionvlad: Learning spatio-temporal aggregation for action classification." CVPR 2017.

(13.3K videos, 101 action classes)

技术路线回顾





UCF-101

(13.3K videos, 101 action classes)

技术路线回顾

Two-stream nets

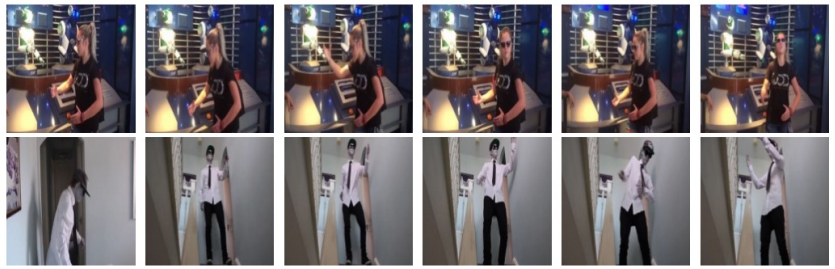
Simonyan et al

ActionVLAD

Girdhar et al

Kinetics-400 dataset

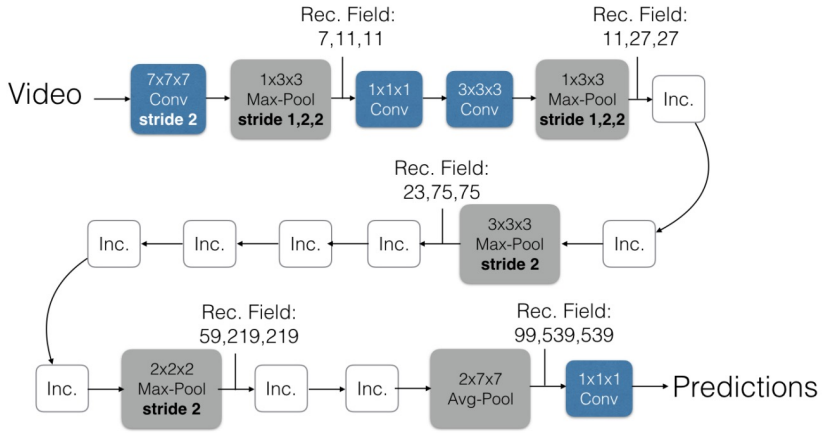
(26万个视频, 400个类别)



(e) robot dancing

(g) riding a bike

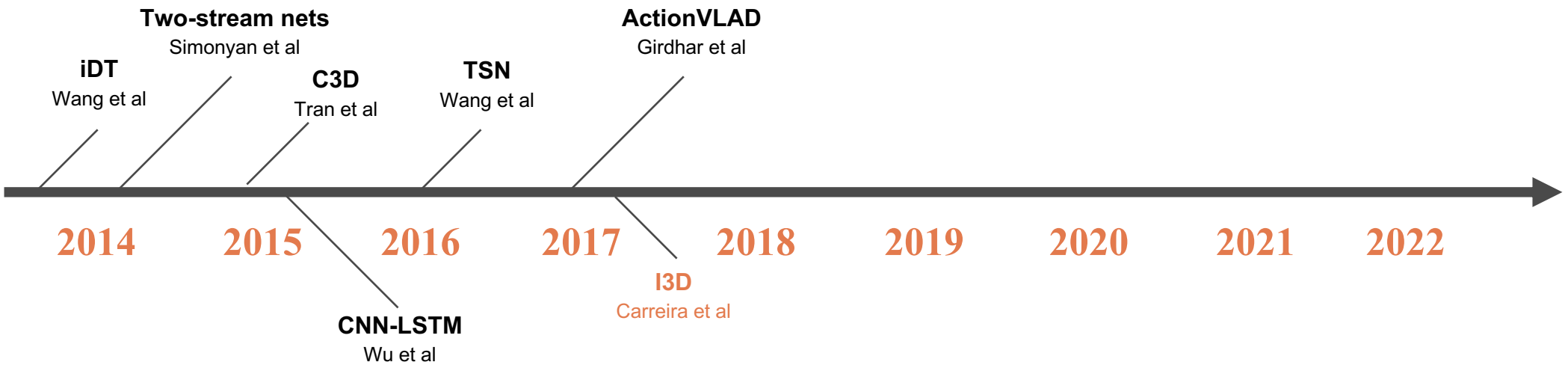
Inflated Inception Network (I3D)

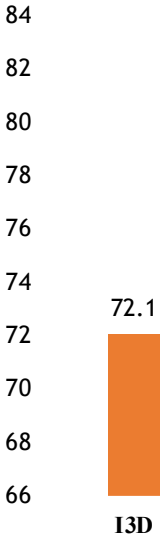


Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." CVPR 2017.

(13.3K videos, 101 action classes)

技术路线回顾





Kinetics-400

(260K videos, 400 action classes)

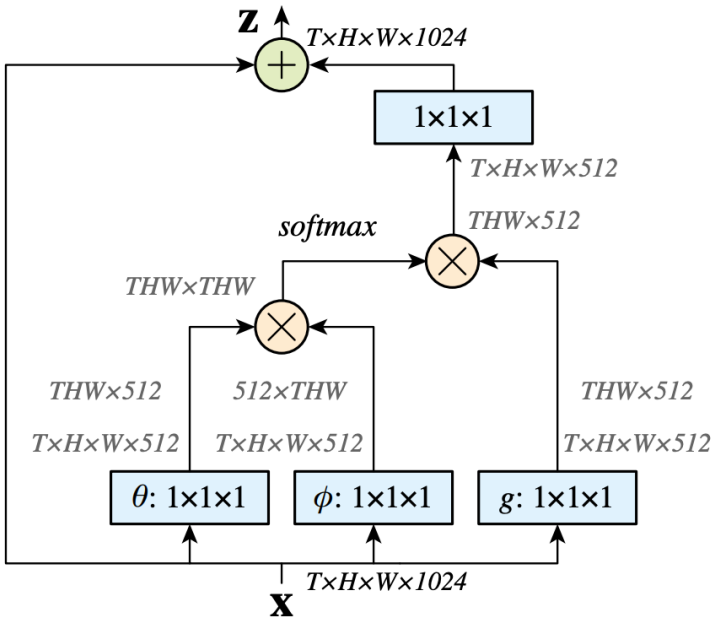
技术路线回顾

Two-stream nets
Simonyan et al.

ActionVLAD
Girdhar et al.

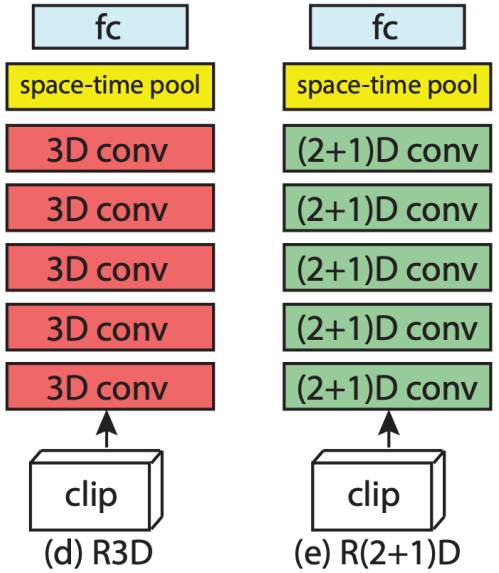
2018

Non-local network



Wang, Xiaolong, et al. "Non-local neural networks." CVPR 2018.

R(2+1)D



Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." CVPR 2018.

(260K videos, 400 action classes)

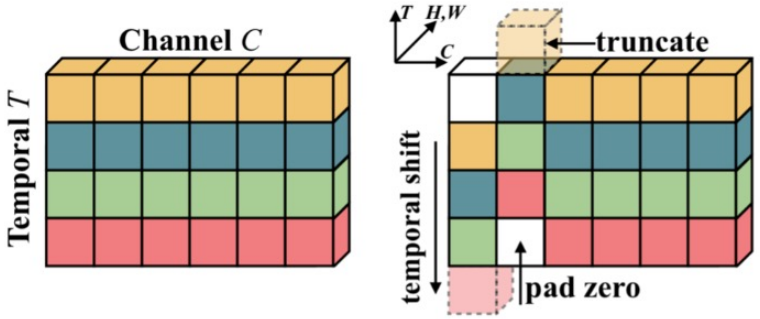
技术路线回顾

Two-stream nets
Simonyan et al.

ActionVLAD
Girdhar et al.

2019

Temporal Shift Module

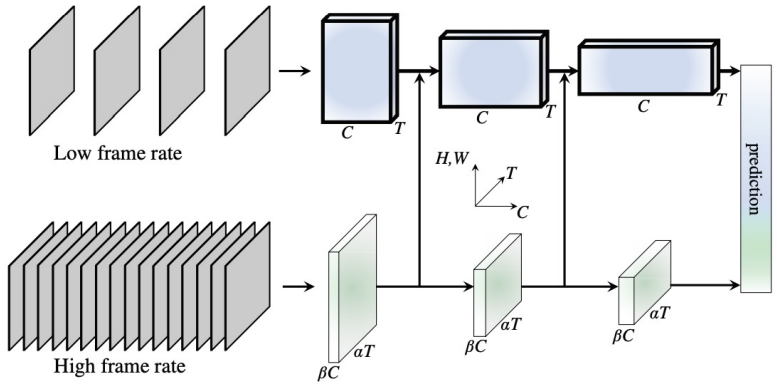


(a) The original tensor without shift.

(b) Offline temporal shift (bi-direction).

Lin, Ji, et al. "Tsm: Temporal shift module for efficient video understanding." ICCV 2019.

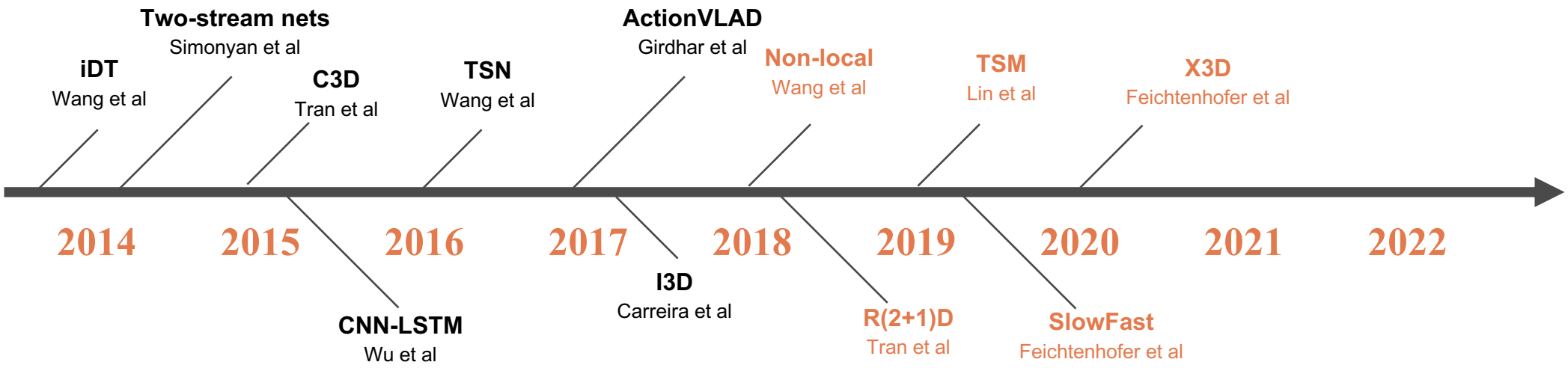
SlowFast Network

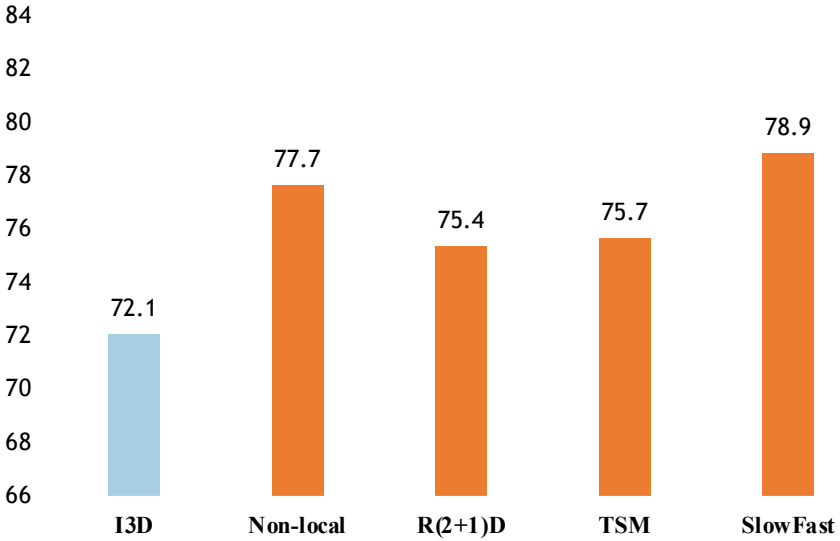


Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." ICCV 2019.

(260K videos, 400 action classes)

技术路线回顾





Kinetics-400

(260K videos, 400 action classes)

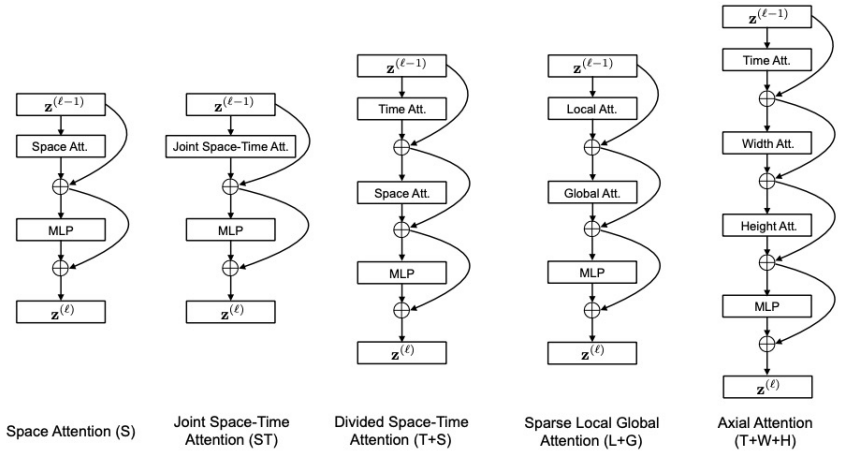
技术路线回顾

Two-stream nets
Simonyan et al

ActionVLAD
Girdhar et al

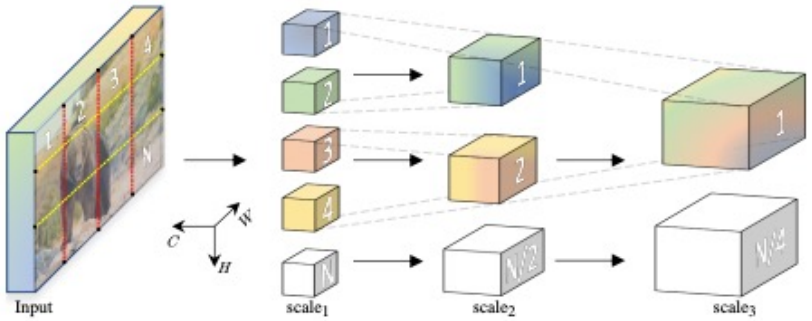
2021

TimeSformer



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

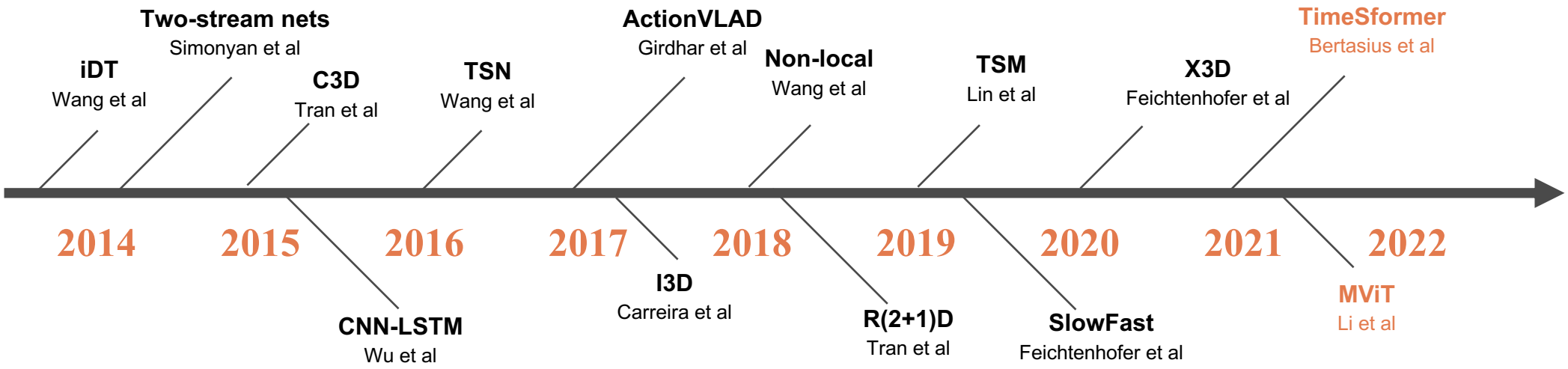
MViT

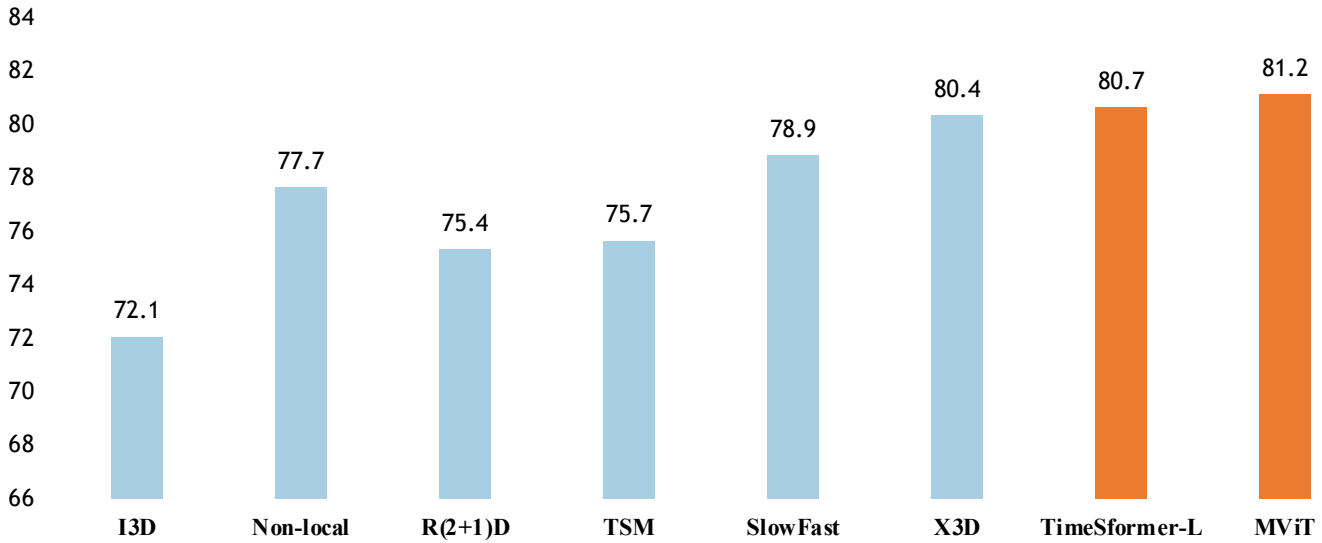


Li, Yanghao, et al. "Multiscale Vision Transformers." ICCV 2021.

(260K videos, 400 action classes)

技术路线回顾





Kinetics-400

(260K videos, 400 action classes)

技术路线回顾

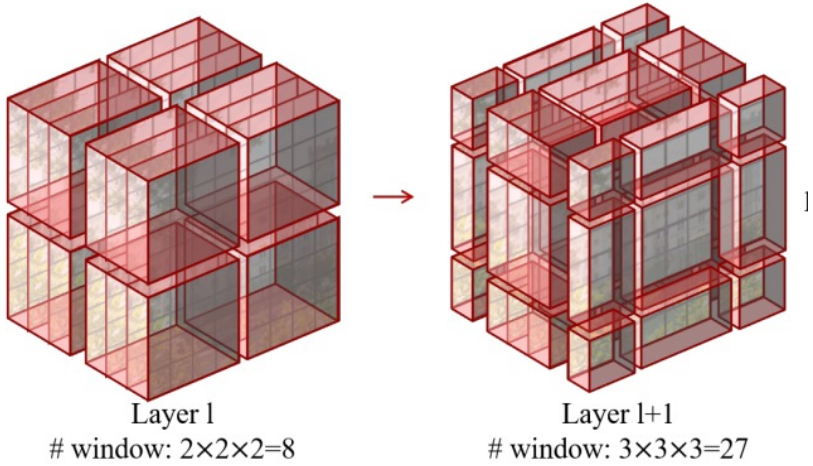
Two-stream nets
Simonyan et al.

ActionVLAD
Girdhar et al.

TimeSformer
Bertasius et al.

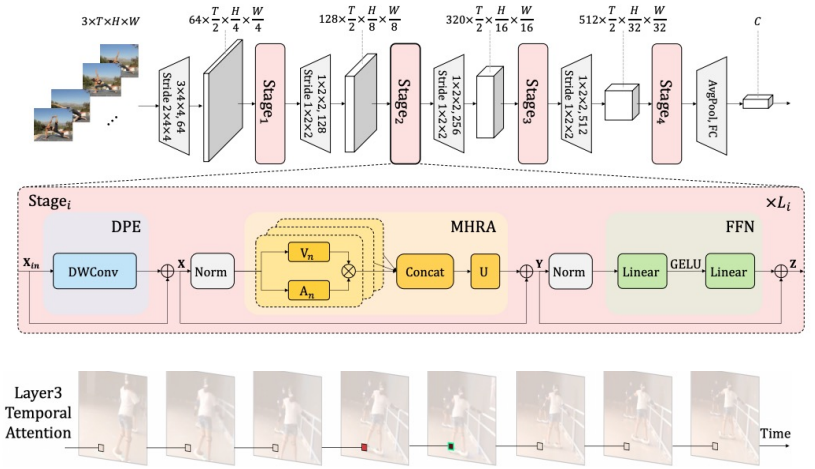
2022

VideoSwin



Liu, Ze, et al. "Video Swin Transformer." CVPR 2022.

Uniformer



Li, Kunchang, et al. "Uniformer: Unified Transformer for Efficient SpatialTemporal Representation Learning." ICLR 2022.

(260K videos, 400 action classes)

技术路线回顾

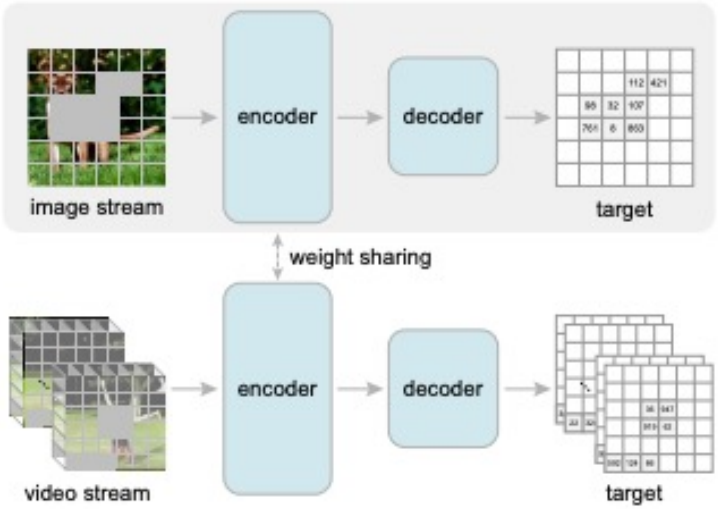
Two-stream nets
Simonyan et al.

ActionVLAD
Girdhar et al.

TimeSformer
Bertasius et al.

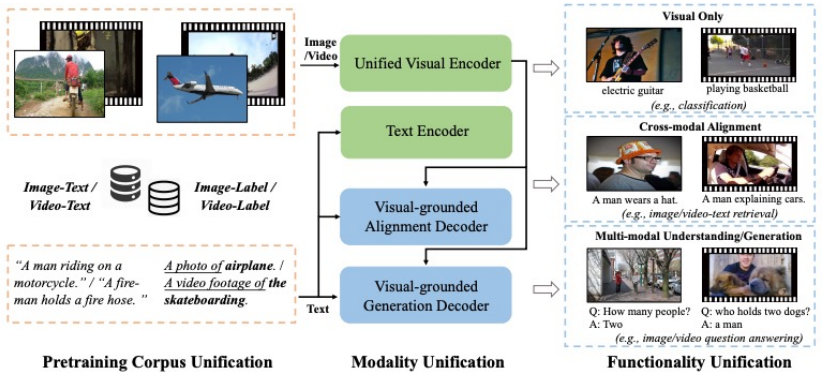
2022

BEVT



Wang, Rui, et al. "BEVT: BERT Pretraining of Video Transformers." CVPR 2022.

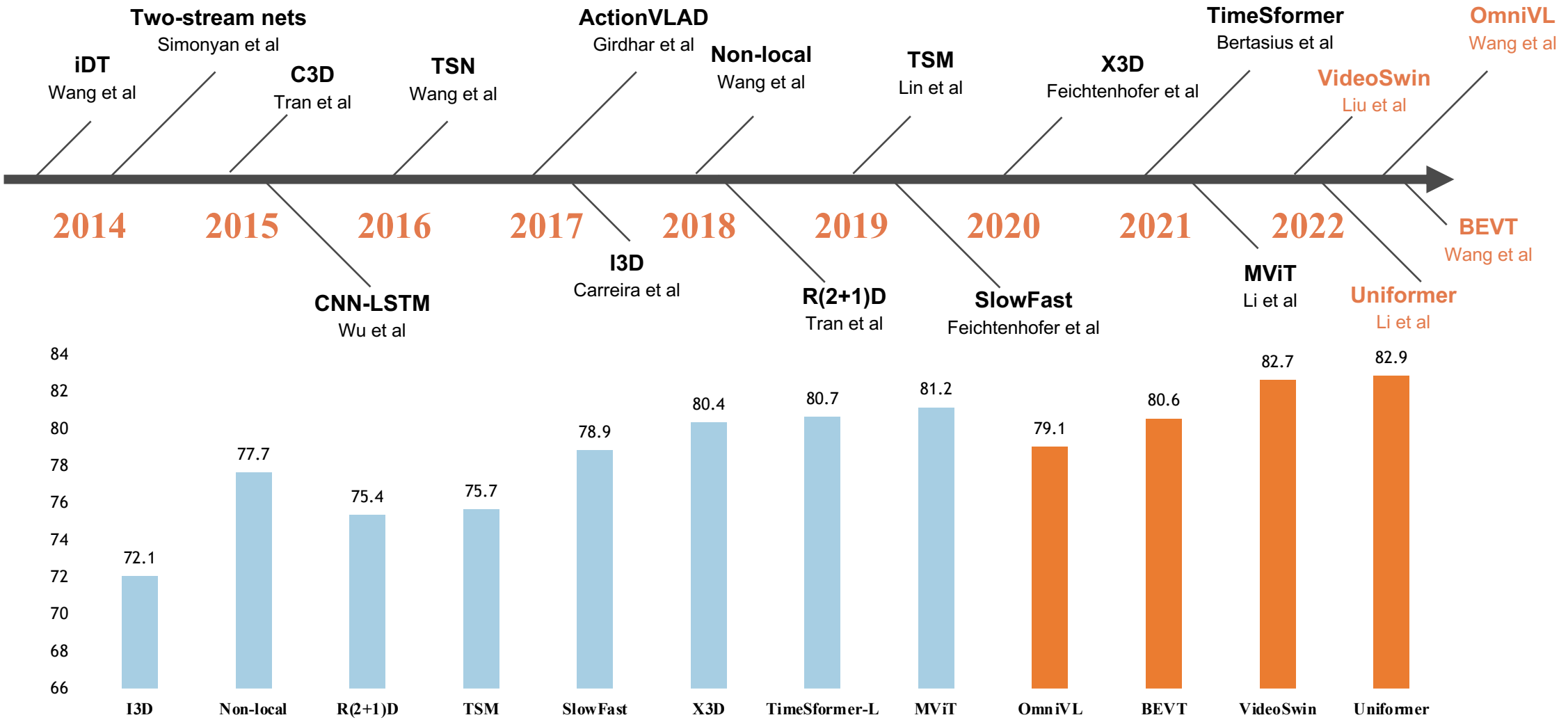
OmniVL



Wang, Junke, et al. "OmniVL: One Foundation Model for Image-Language and Video-Language Tasks." NeurIPS 2022.

(260K videos, 400 action classes)

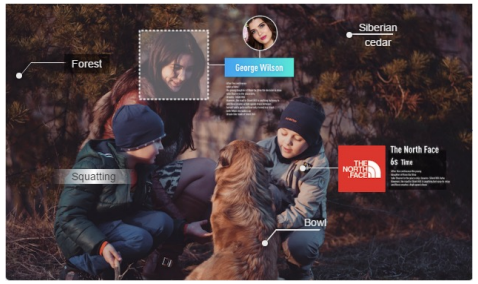
技术路线回顾



Kinetics-400

(260K videos, 400 action classes)

视频内容分析技术得到了广泛应用

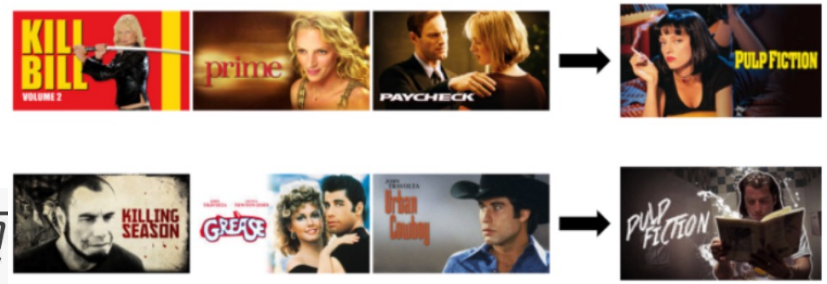


云端智能视频分析理解服务

百度云　腾讯云　阿里云　aws　Google Cloud

视频推荐

NETFLIX　腾讯视频



视频场景营销

抖音　Yi+　快手　周同科技 UNIVER

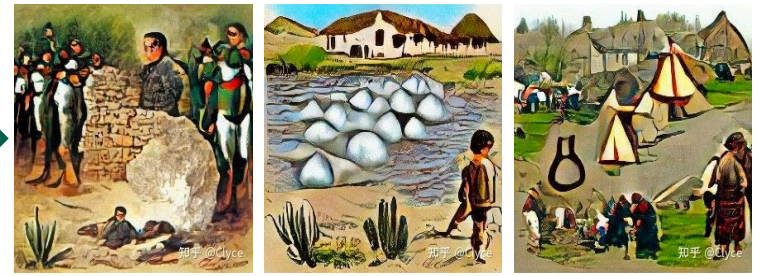
智能视频编辑

视频监控安防

商汤 sensetime　Face++ 旷视　DEEPGLINT 格灵深瞳　……

智能视频生成

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。当时，马孔多是个二十户人家的村庄，一座座土房都盖在河岸上，河水清澈，沿着遍布石头的河床流去，河里的石头光滑、洁白，活象史前的巨蛋。这块天地还是新开辟的，许多东西都叫不出名字，不得不用手指指点点。每年三月，衣衫褴褛的吉卜赛人都要在村边搭起帐篷，在笛鼓的喧嚣声中，向马孔多的居民介绍科学家的最新发明。



提纲

技术路线回顾

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

趋势四：知识融入从外部到联合

趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

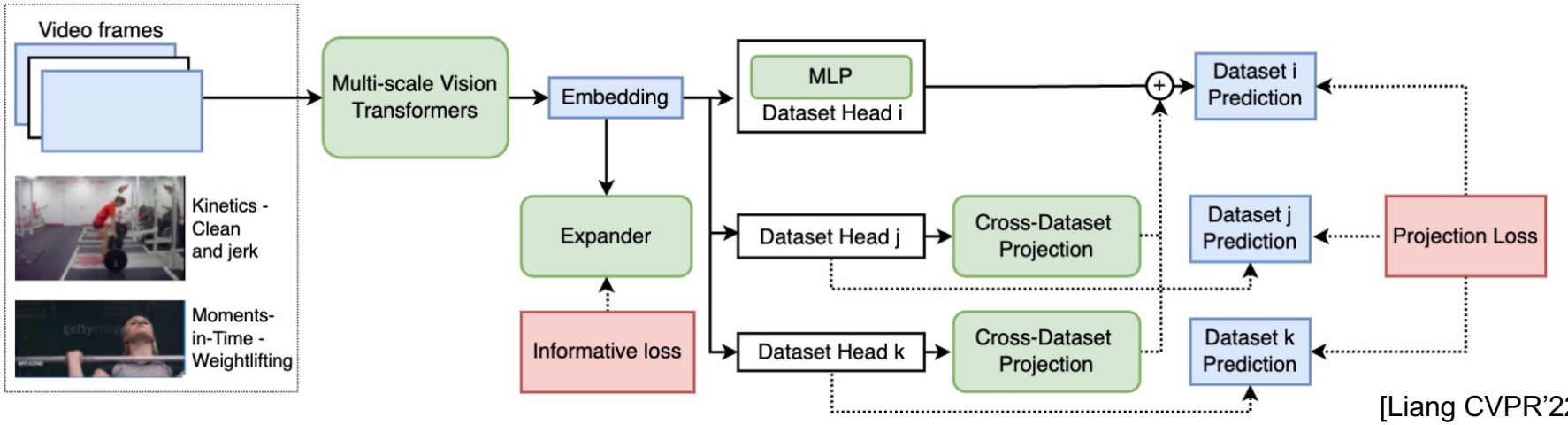
训练数据从单一到多元

- 当前大部分算法在单一数据集训练和测试，未能充分利用所有标注数据

数据集	时间	类别	视频数量	数据来源
HMDB51	2011	51	6849	web video
UCF101	2013	101	13320	web video
Sports-1M	2014	487	1000000	web video
ActivityNet	2015	200	20000	web video
FCVID	2015	239	91223	web video
Charades	2016	157	9848	controlled settings
Youtube8M	2016	1000	237000	web video
AVA	2018	80	57600	movie
Kinetics-400	2017	400	306245	web video
SSV V1	2017	174	108499	controlled settings
SSV V2	2018	174	220847	controlled settings
Kinetics-600	2018	600	495547	web video
Kinetics-700	2019	700	650317	web video
Jester	2019	27	148092	controlled settings
VideoLT	2021	1004	256218	web video

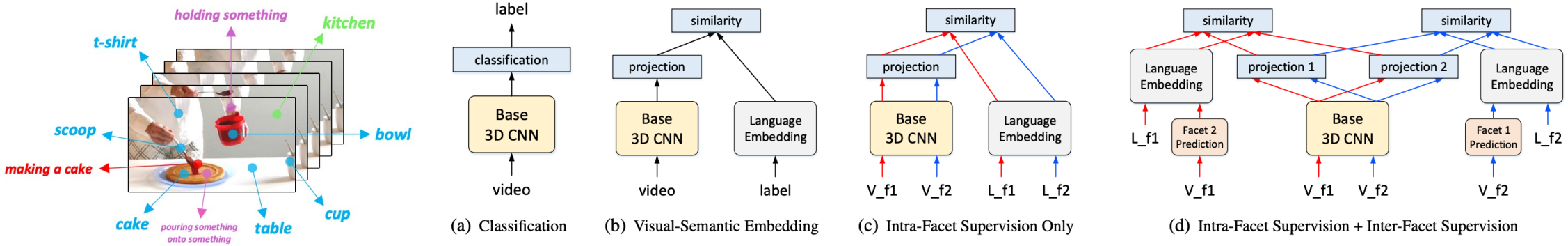
训练数据从单一到多元

- 多个视频分类数据集的联合训练
  - **学习类别关联矩阵，挖掘不同数据集语义关联**



[Liang CVPR'22]

  - **利用标签语义信息，建立数据集关联**

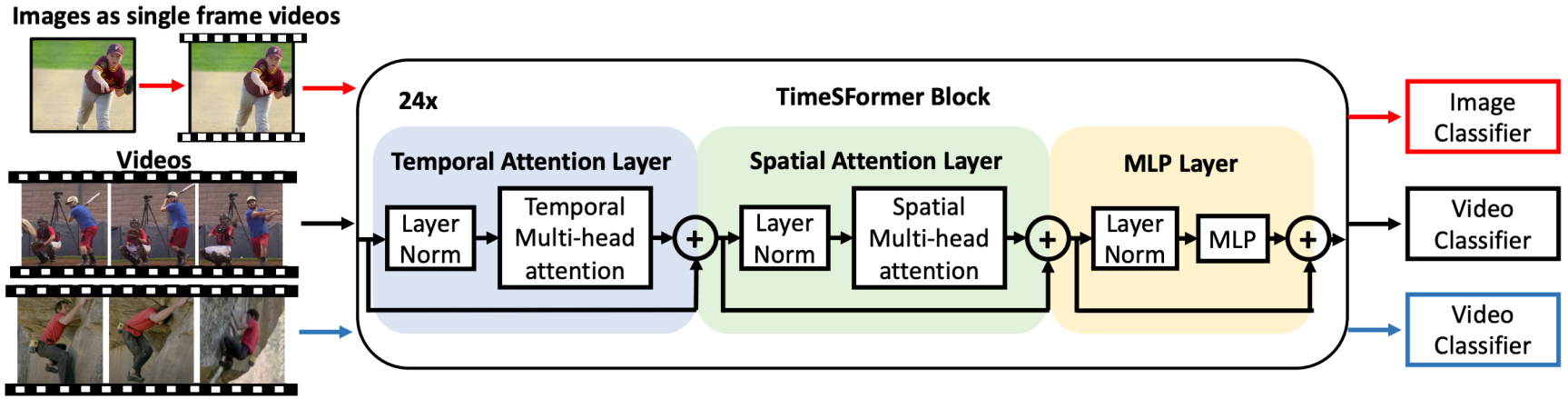


(a) Classification (b) Visual-Semantic Embedding (c) Intra-Facet Supervision Only (d) Intra-Facet Supervision + Inter-Facet Supervision

[Qiu CVPR'21]

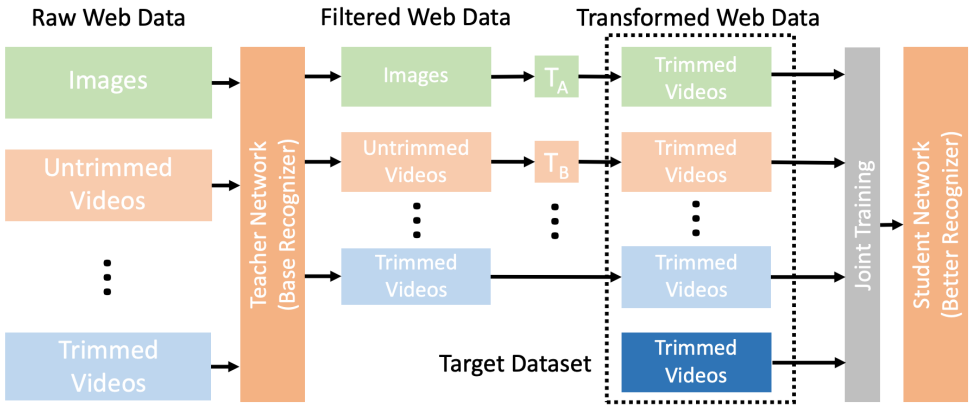
训练数据从单一到多元

- 多个视频分类数据集+图像分类数据集的联合训练
  - 将图像看为单帧的视频参与联合训练



[Zhang arXiv'22]

  - 将图像扩充为视频参与联合训练

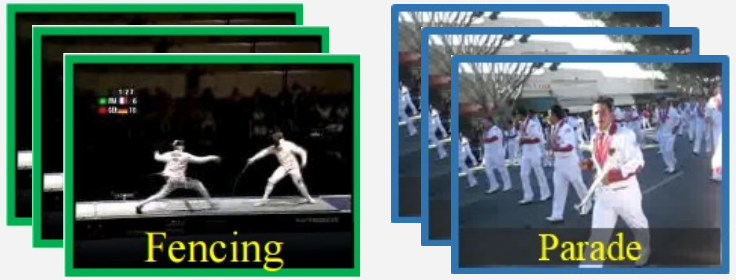


[Duan ECCV'20]

训练数据从单一到多元

- 多个视频分类数据集+视频描述数据集的联合训练

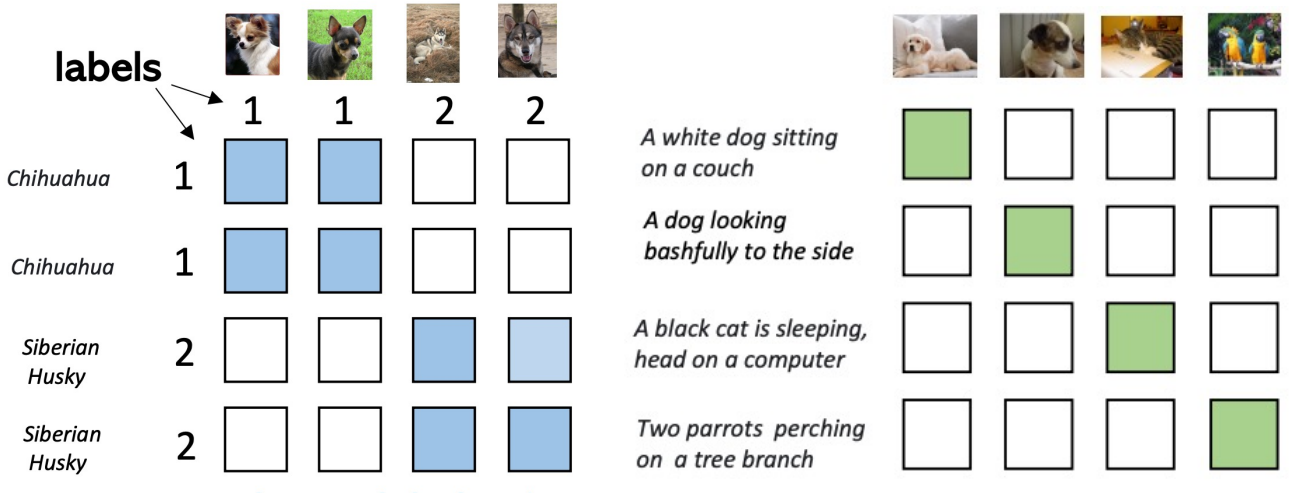
视频分类：语义较少



视频描述：语义丰富

An elderly man is playing the piano in front of a crowd.
A woman walks to the piano and briefly talks to the the elderly man.

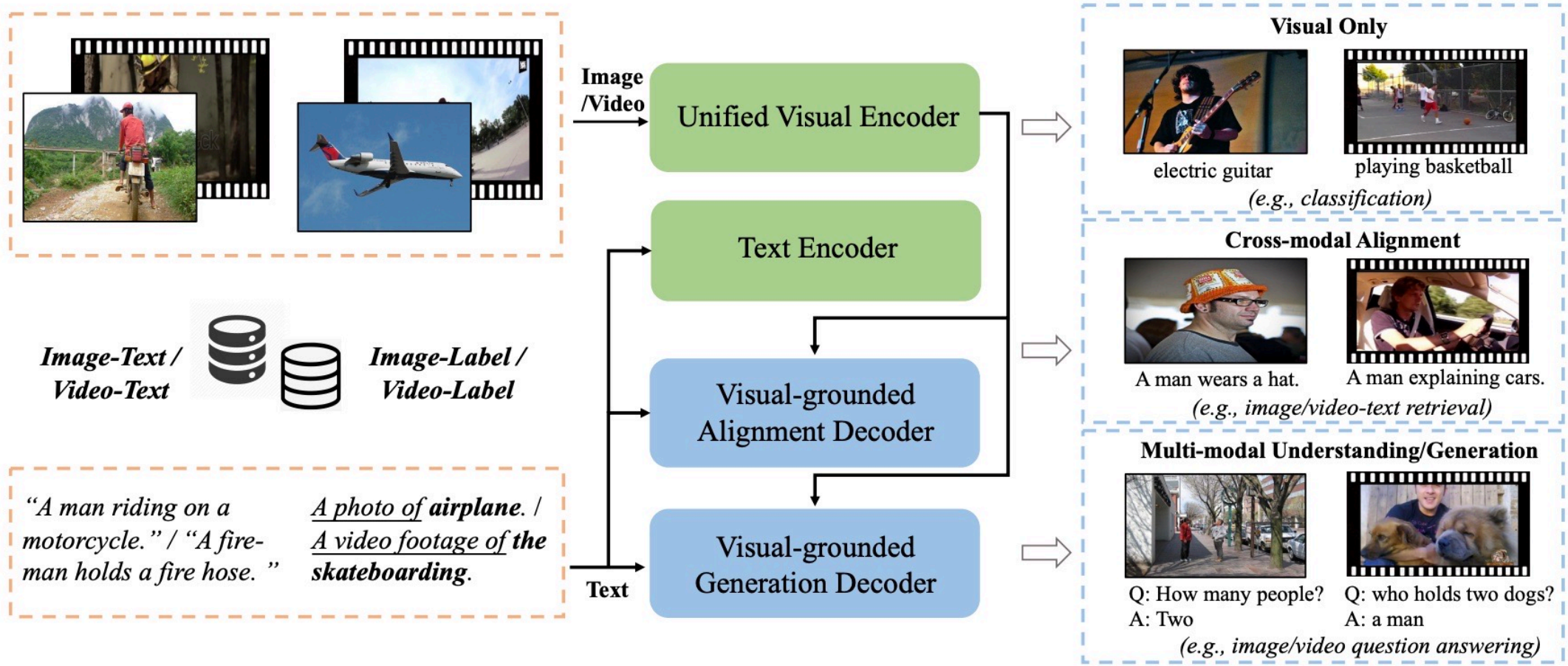
□ 基于对比学习实现“视频—文本—标签”的统一



[Yuan CVPR’22, Yang CVPR’22]

训练数据从单一到多元

- 视频分类+视频描述+图像分类+图像描述数据集的联合训练



[Wang NeurIPS'22]

提纲

技术路线回顾

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

趋势四：知识融入从外部到联合

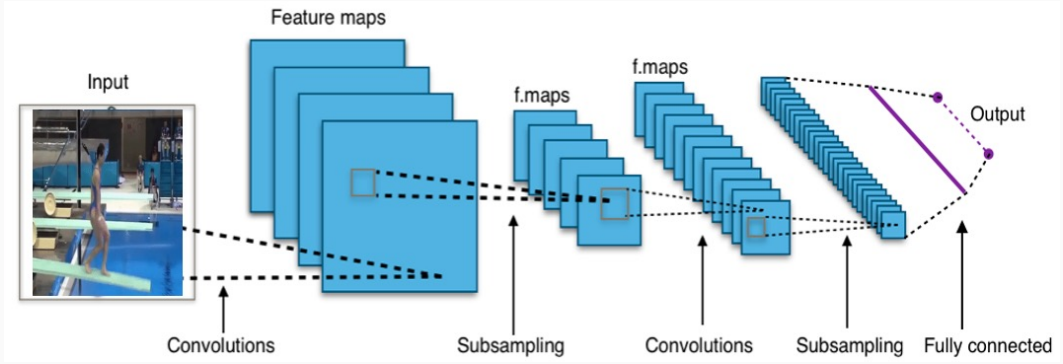
趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

网络结构从分散到统一

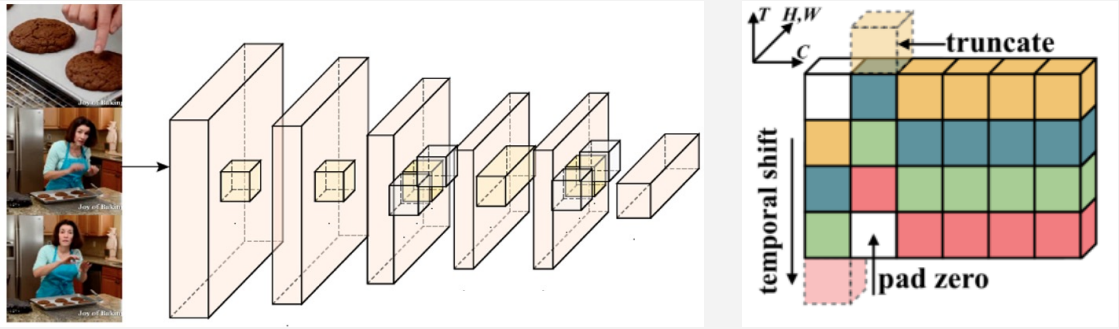
空间信息建模

2D Convolutional NNs (ResNets/InceptionNet)



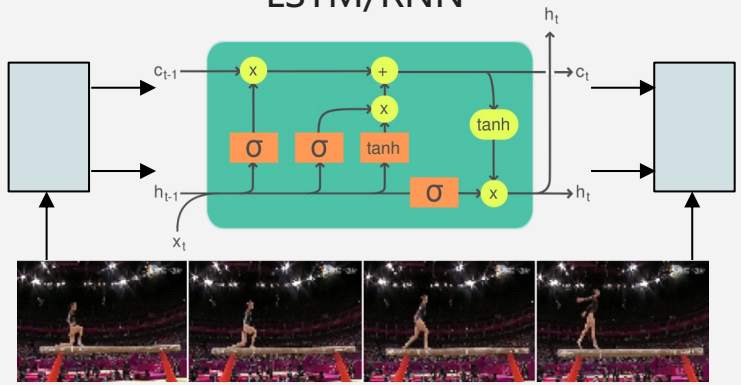
短时运动信息建模

3D Convolutional NNs (I3D/C3D/P3D)



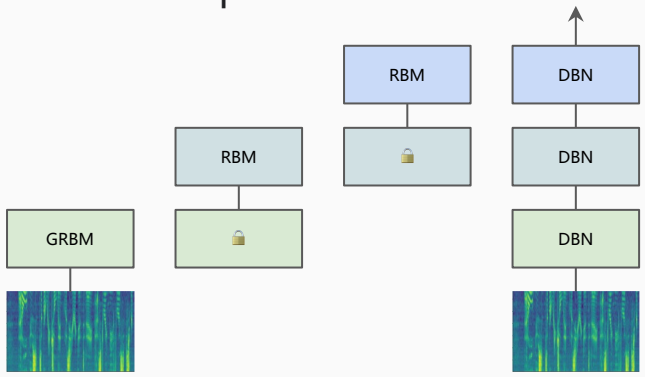
长时时序信息建模

LSTM/RNN



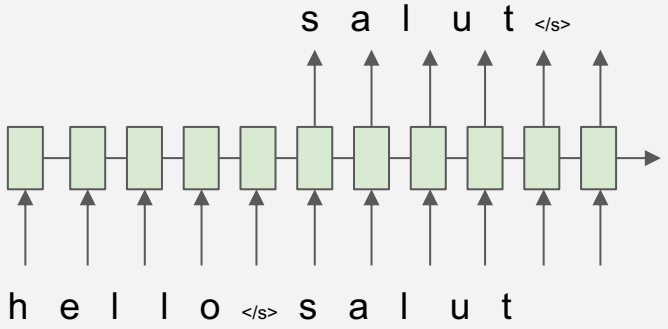
音频信息建模

Deep Belief Network



文本信息建模

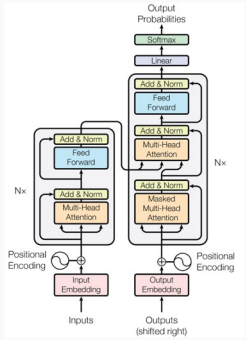
Seq2Seq



网络结构从分散到统一

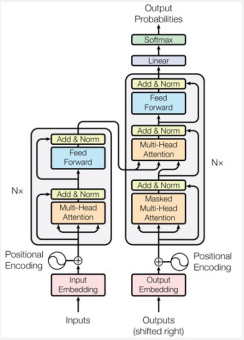
空间信息建模

Transformer



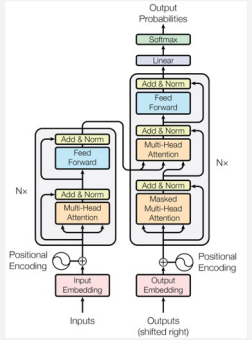
短时运动信息建模

Transformer



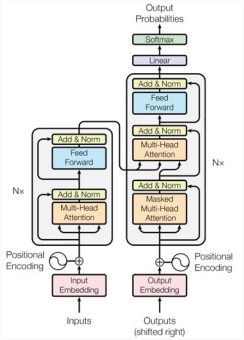
长时时序信息建模

Transformer



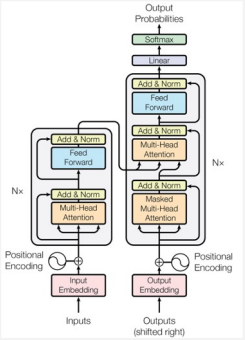
音频信息建模

Transformer



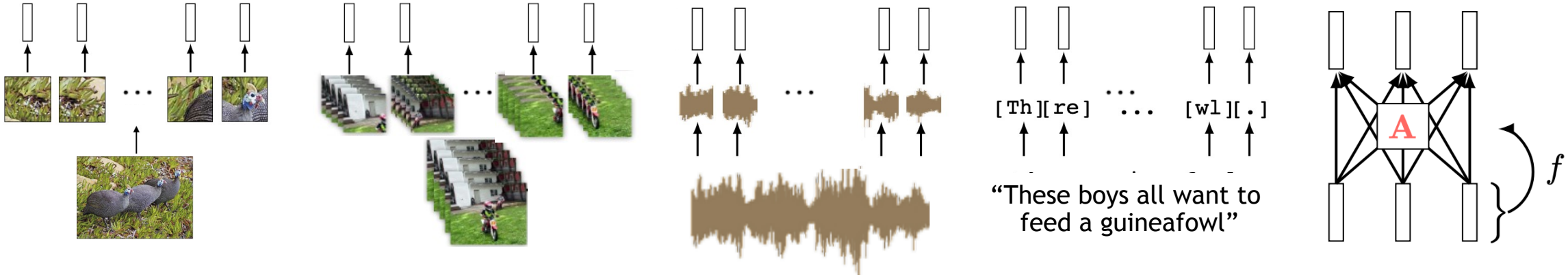
文本信息建模

Transformer



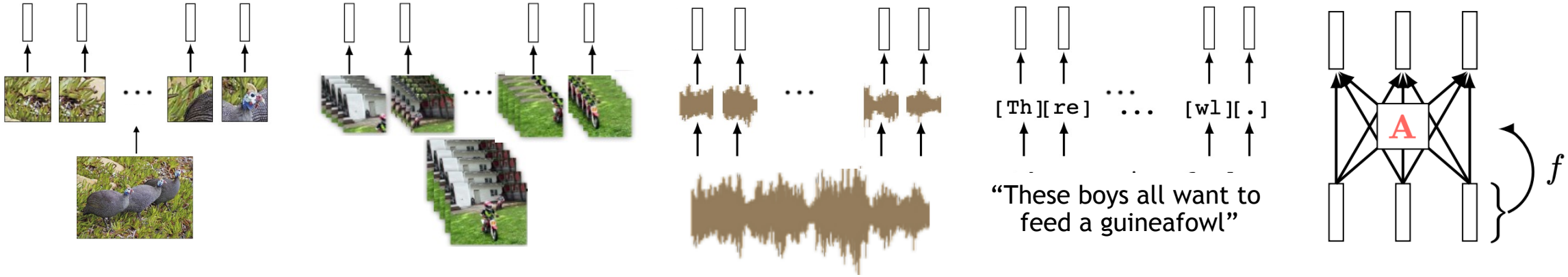
网络结构从分散到统一

- 将输入切成小块，并tokenize，利用Transformer学习token间关系

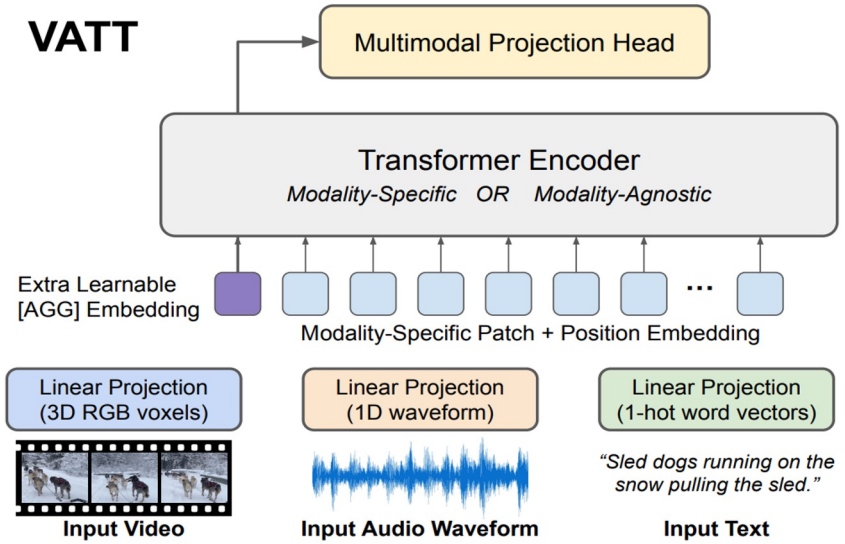


网络结构从分散到统一

- 将输入切成小块，并tokenize，利用Transformer学习token间关系



- 利用**统一网络**学习各模态关系



[Mustafa arXiv'22]

[Akbari CVPR'22]

提纲

技术路线回顾

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

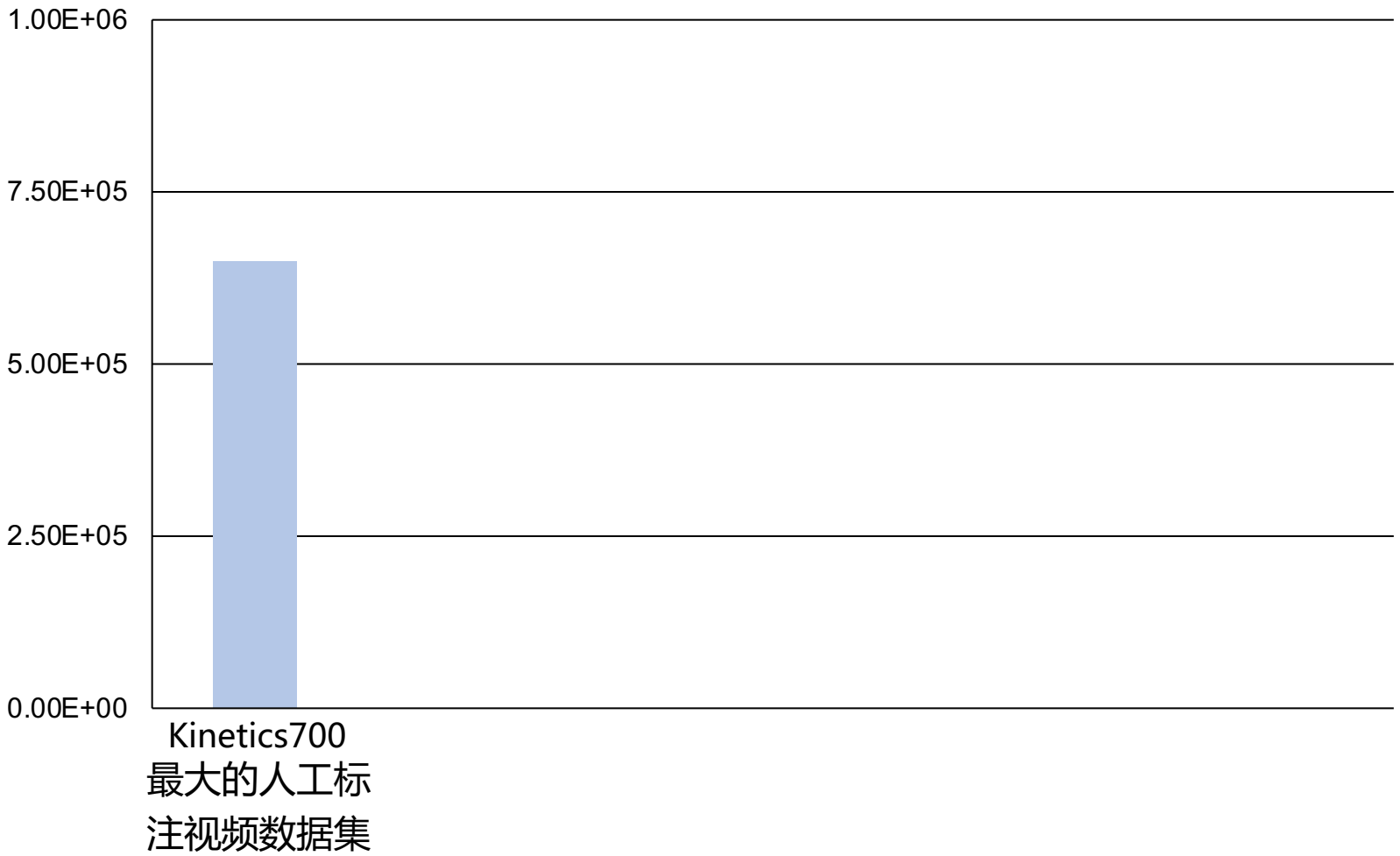
趋势四：知识融入从外部到联合

趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

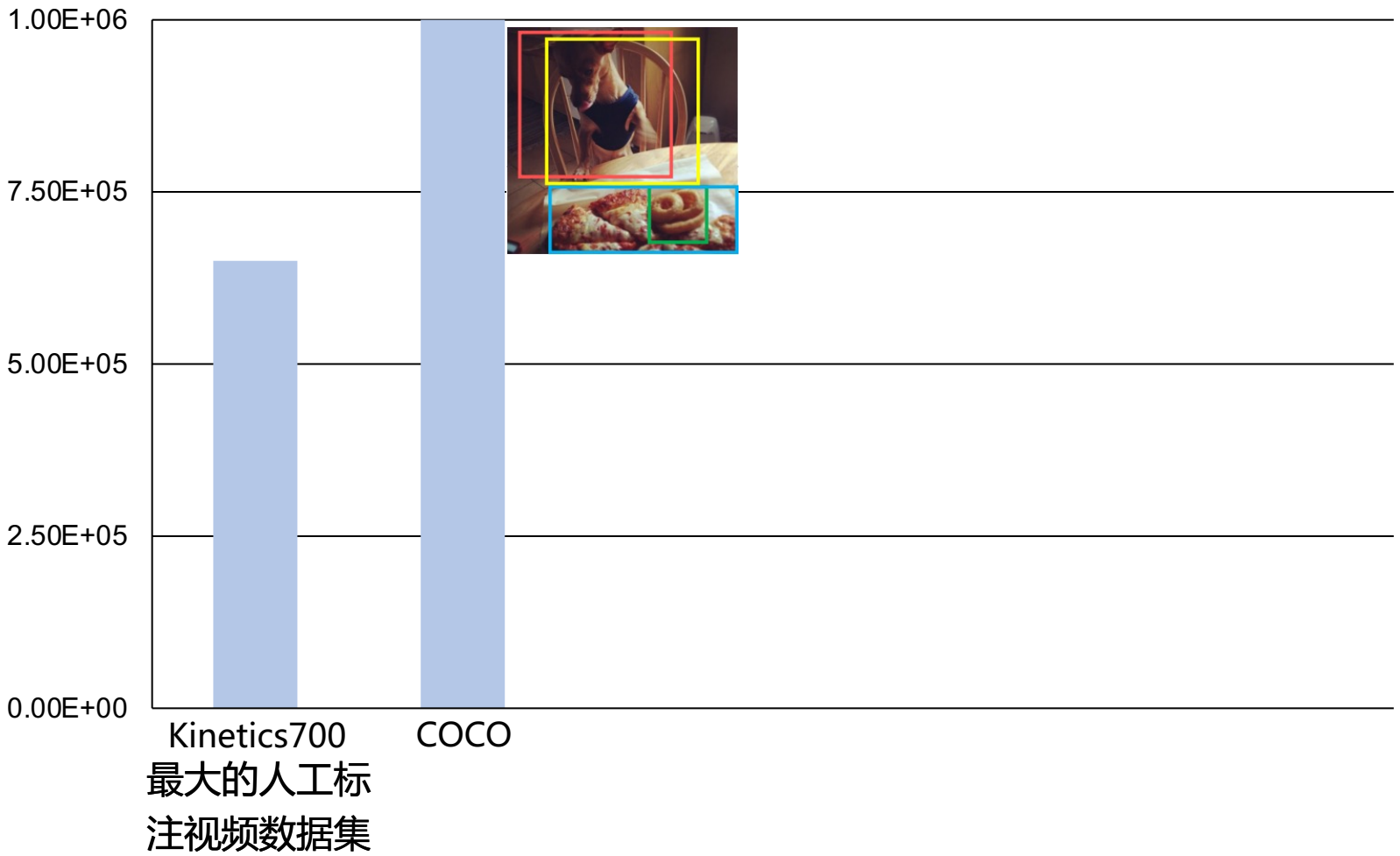
监督信号从人标到自学

- 当前深度学习的成功依赖于大规模人工标注的数据集，人工标注耗时耗力



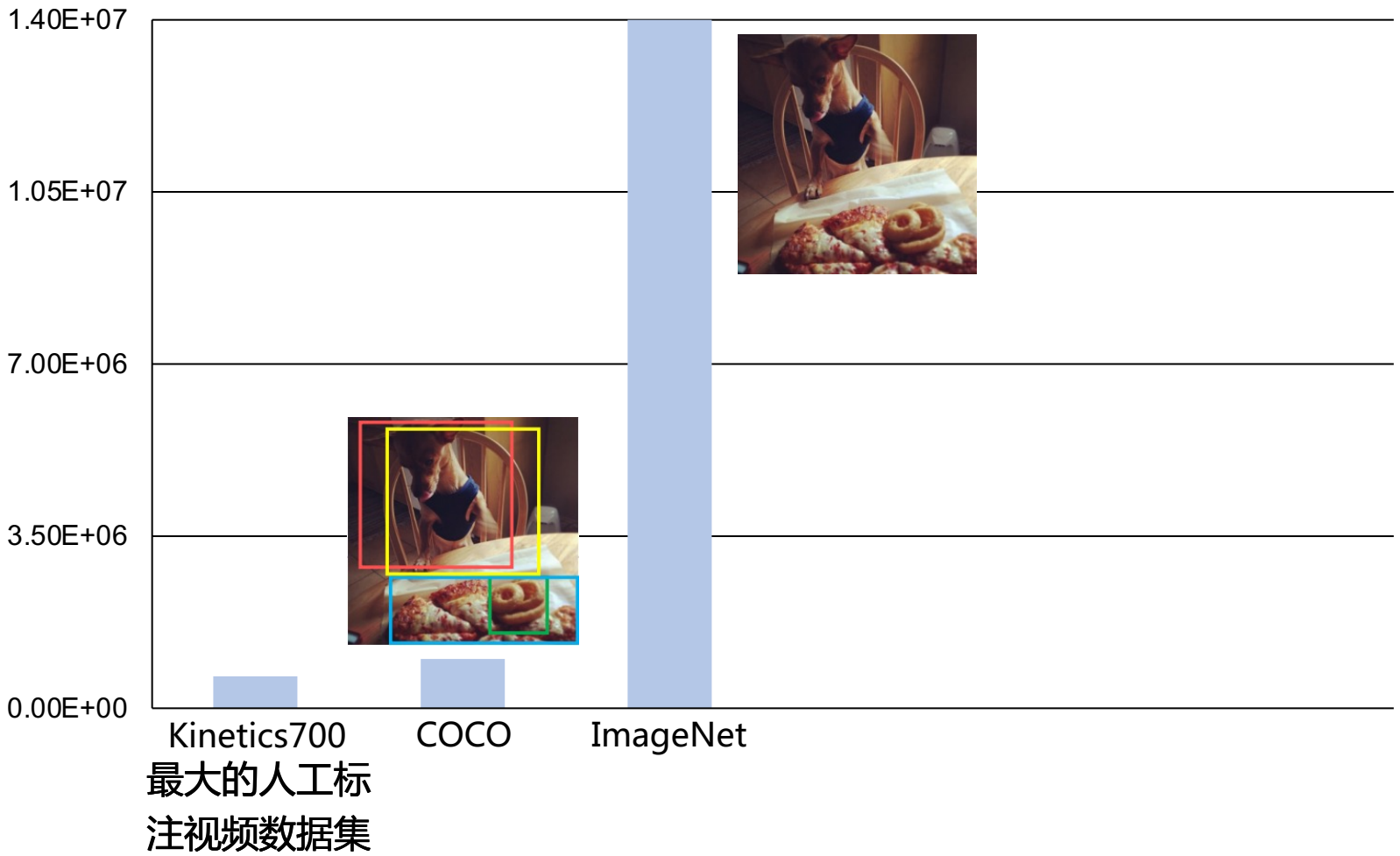
监督信号从人标到自学

- 当前深度学习的成功依赖于大规模人工标注的数据集，人工标注耗时耗力



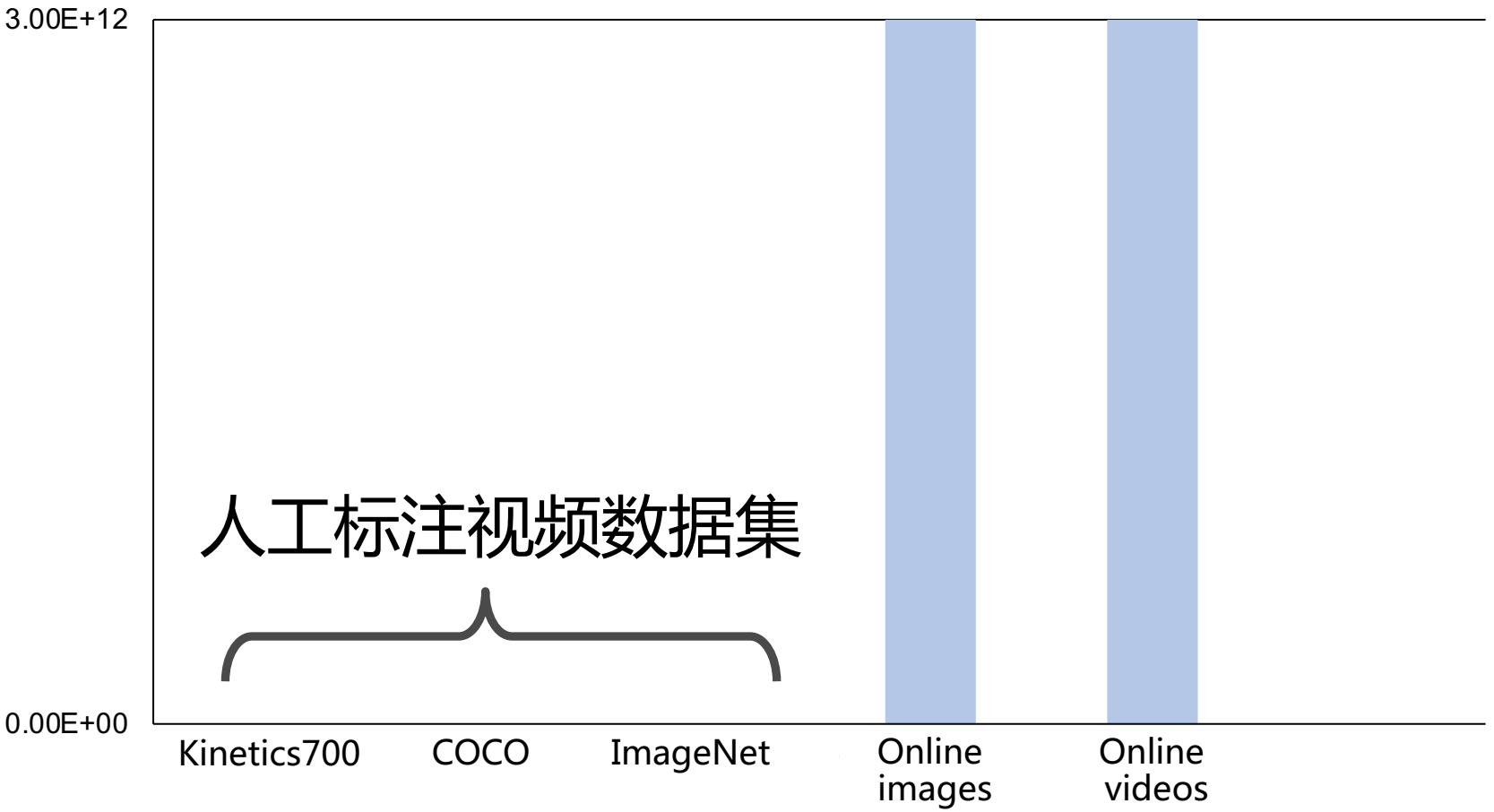
监督信号从人标到自学

- 当前深度学习的成功依赖于大规模人工标注的数据集，人工标注耗时耗力



监督信号从人标到自学

- 当前深度学习的成功依赖于大规模人工标注的数据集，人工标注耗时耗力



监督信号从人标到自学

- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

监督信号从人标到自学

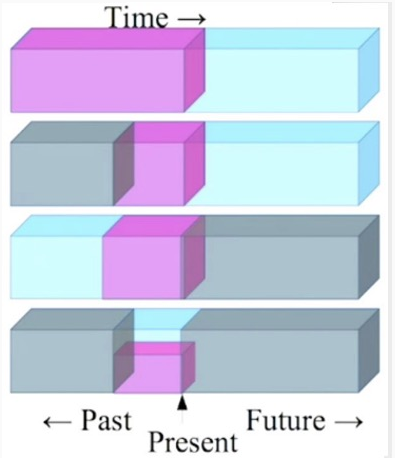
- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

自我预测（Self-Prediction）

图灵奖得主Lecun：从输入数据一部分预测另一部分

从过去预测未来　　从未来预测过去

从刚刚预测未来　　从下半部分预测上半部分



监督信号从人标到自学

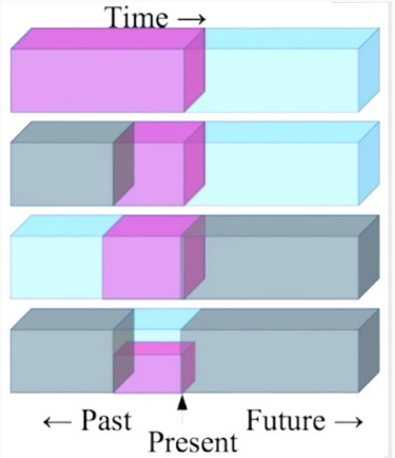
- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

自我预测（Self-Prediction）

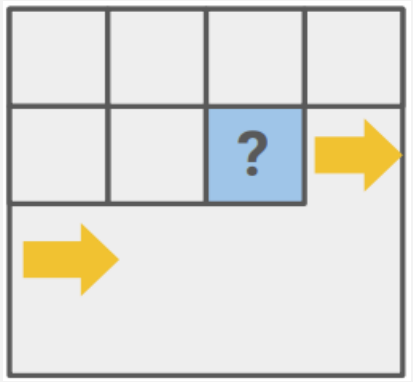
图灵奖得主Lecun：从输入数据一部分预测另一部分

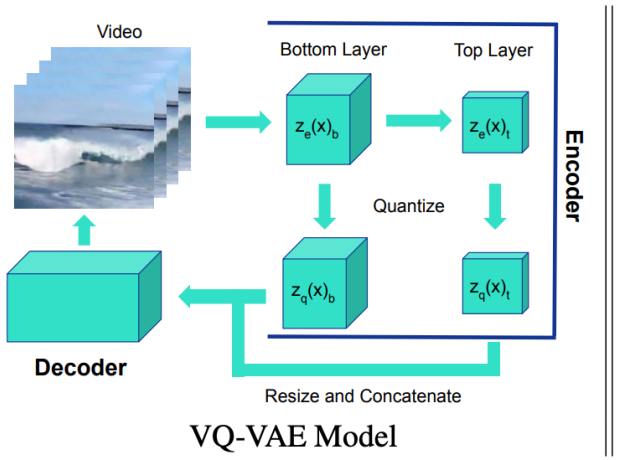
从过去预测未来　　从未来预测过去

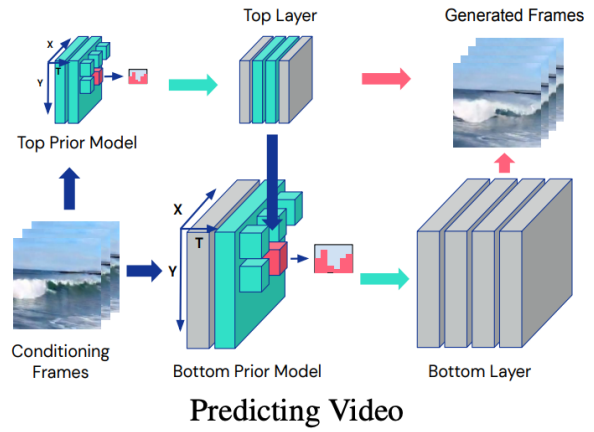
从刚刚预测未来　　从下半部分预测上半部分



自回归生成







[Walker arXiv'21]

监督信号从人标到自学

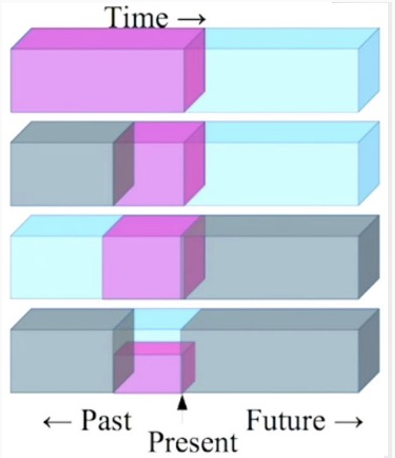
- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

自我预测（Self-Prediction）

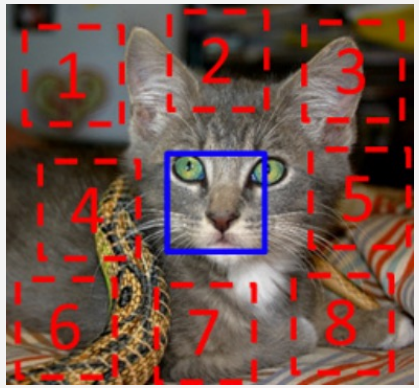
图灵奖得主Lecun：从输入数据一部分预测另一部分

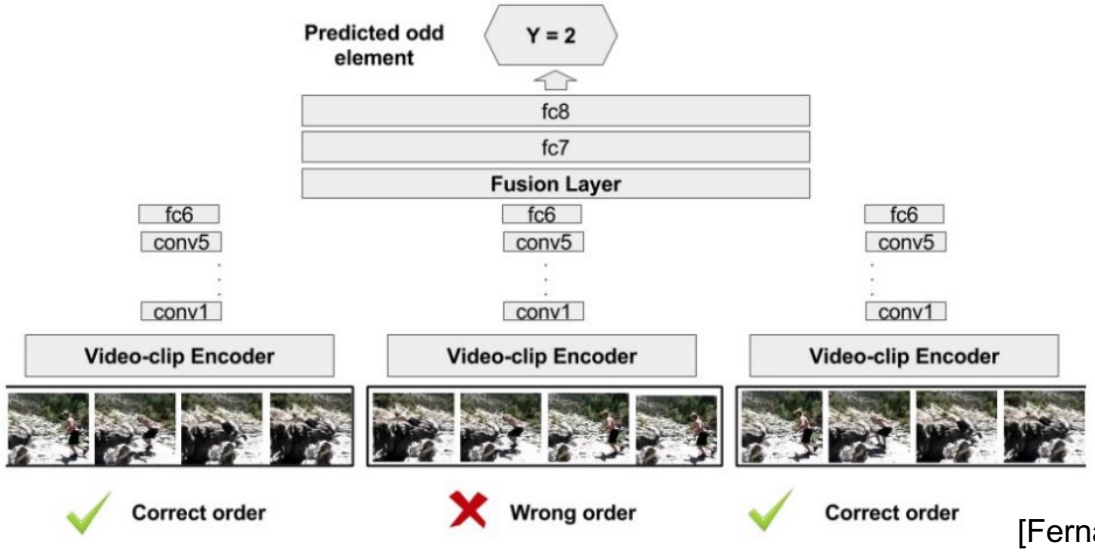
从过去预测未来　　从未来预测过去

从刚刚预测未来　　从下半部分预测上半部分



数据性质预测





[Fernando CVPR'17]

监督信号从人标到自学

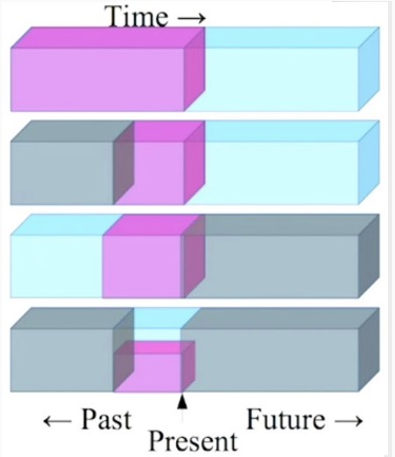
- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

自我预测（Self-Prediction）

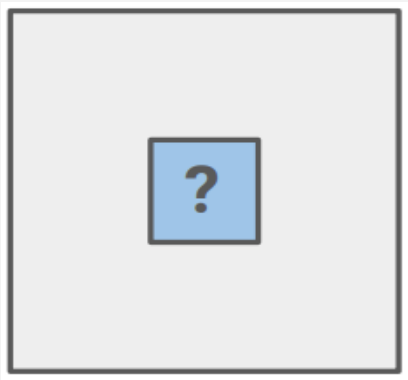
图灵奖得主Lecun：从输入数据一部分预测另一部分

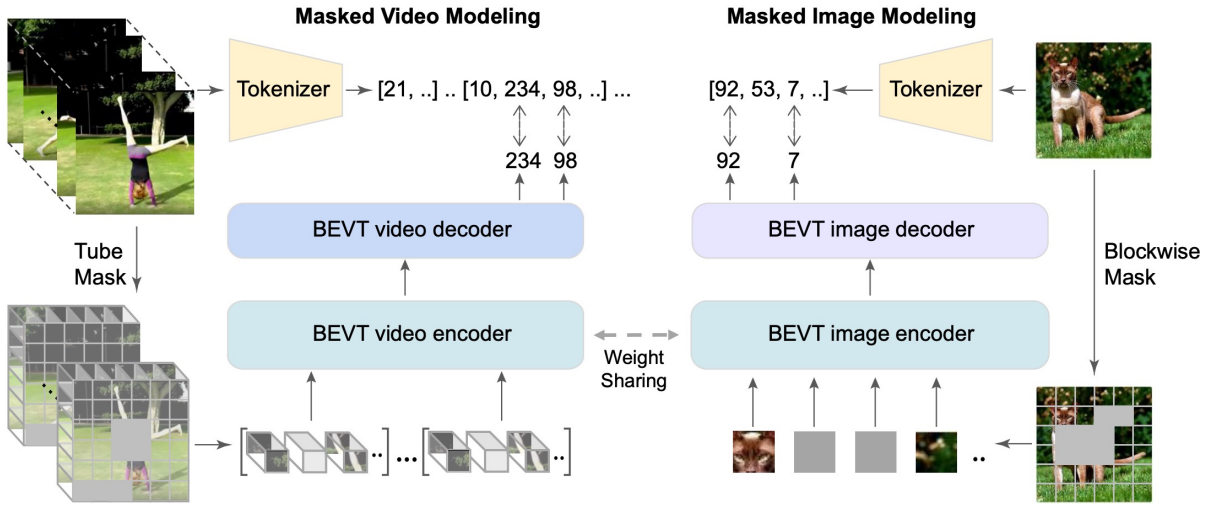
从过去预测未来　　从未来预测过去

从刚刚预测未来　　从下半部分预测上半部分



掩码预测





[Wang CVPR'22]

监督信号从人标到自学

- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

对比学习（Contrastive Learning）

相近的样本在特征空间中更近、相远的样本在特征空间中更远

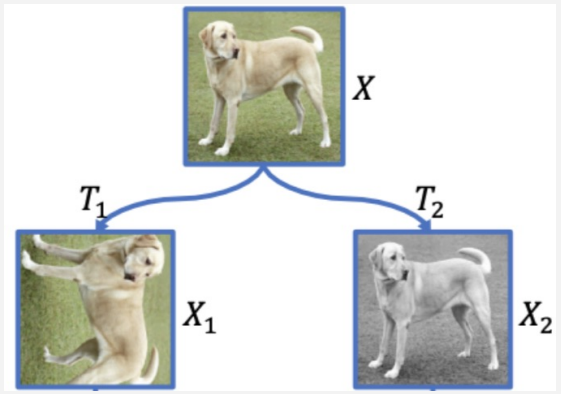
监督信号从人标到自学

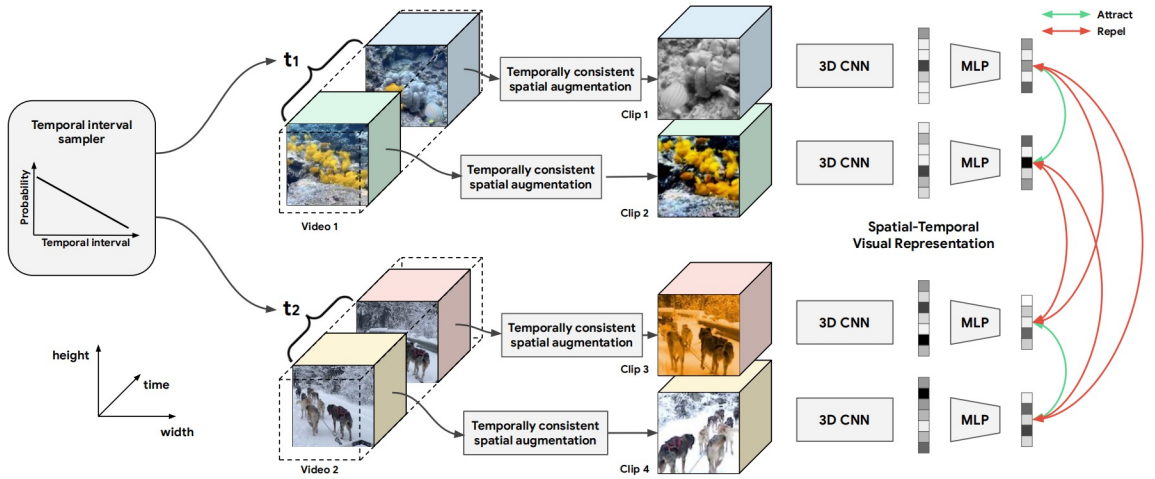
- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

对比学习（Contrastive Learning）

相近的样本在特征空间中更近、相远的样本在特征空间中更远

基于数据增强的对比





[Qian CVPR'21]

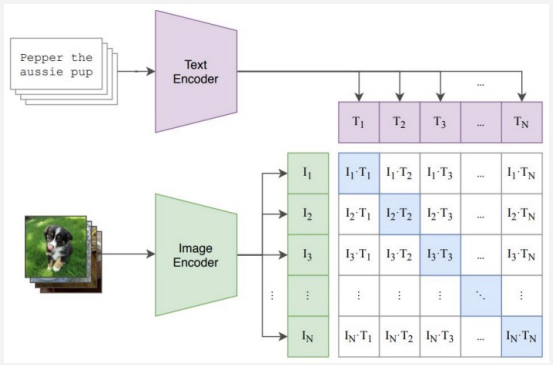
监督信号从人标到自学

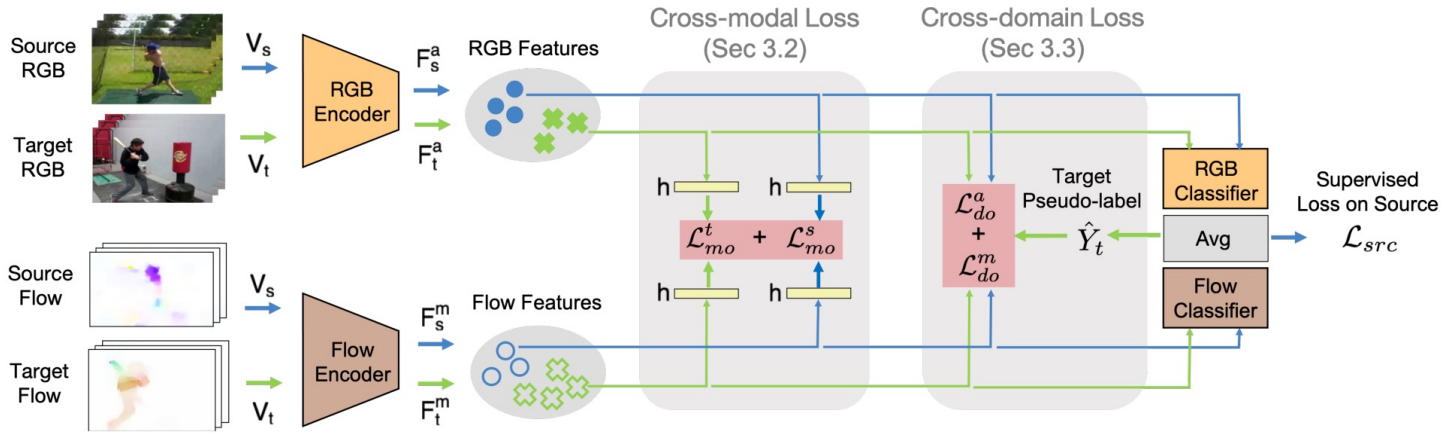
- 自监督学习通过设计前置任务（pretext tasks）**从数据中自动挖掘监督信息**

对比学习（Contrastive Learning）

相近的样本在特征空间中更近、相远的样本在特征空间中更远

模态间对比





[Kim ICCV'21]

提纲

技术路线回顾

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

趋势四：知识融入从外部到联合

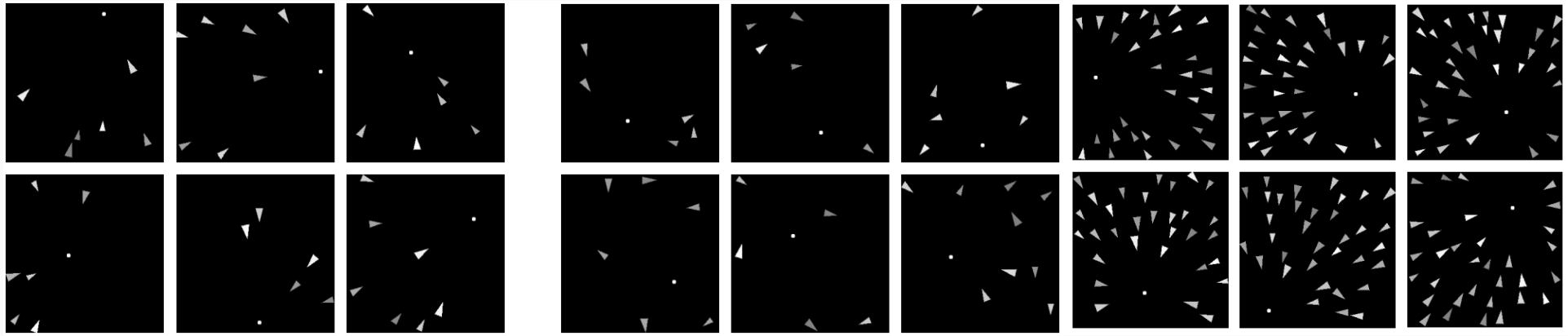
趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

知识融入从外部到联合

- 当前深度学习的成功主要由数据驱动，忽略了对外部知识的有效利用

Common Fate 任务 [Yan arXiv17]

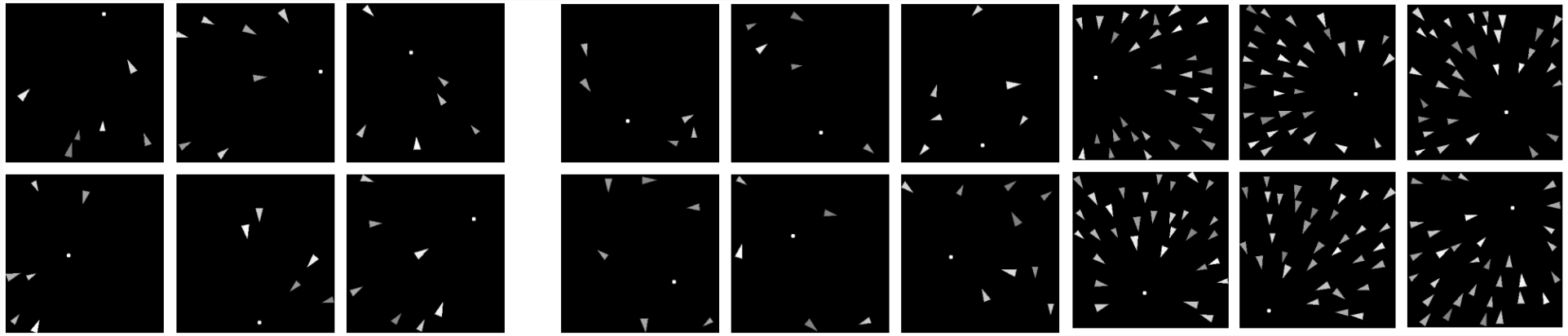


正样本　　负样本

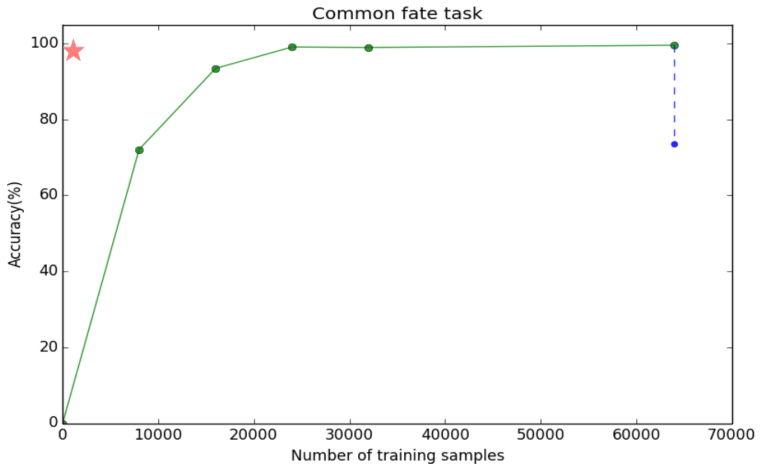
知识融入从外部到联合

- 当前深度学习的成功主要由数据驱动，忽略了对外部知识的有效利用

Common Fate 任务 [Yan arXiv17]



正样本　　负样本



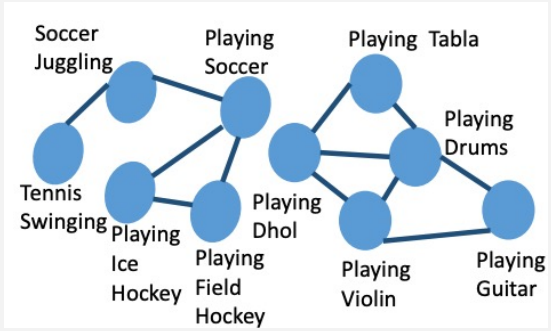
- **两万个**训练样本才能提升准确率至90%以上
- 人只需要**20个样本**就可得到100%的准确率
- 三角变大后准确率大幅下降
- **缺乏知识导致学习难、泛化弱**

潘云鹤院士称“**人工智能正走向数据与知识的双轮驱动**”

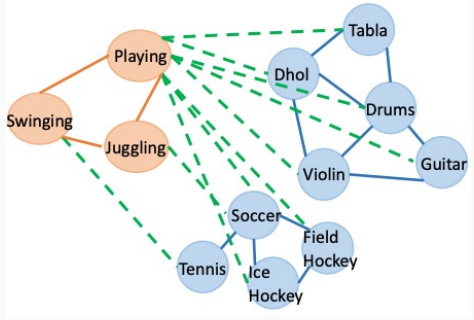
知识融入从外部到联合

- 当前大部分知识图谱独立于视频分析网络，未能与视觉特征深度交互

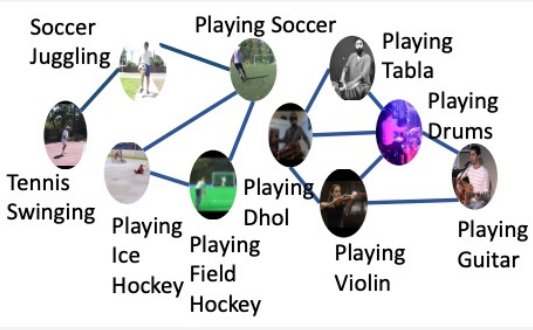
动作图谱



动作—物体图谱



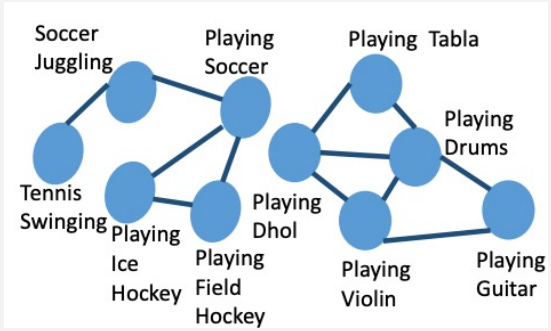
视觉图谱



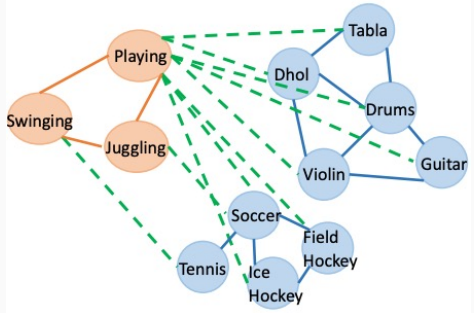
知识融入从外部到联合

- 当前大部分知识图谱独立于视频分析网络，未能与视觉特征深度交互

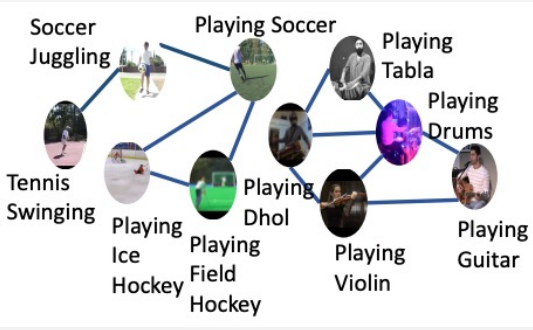
动作图谱

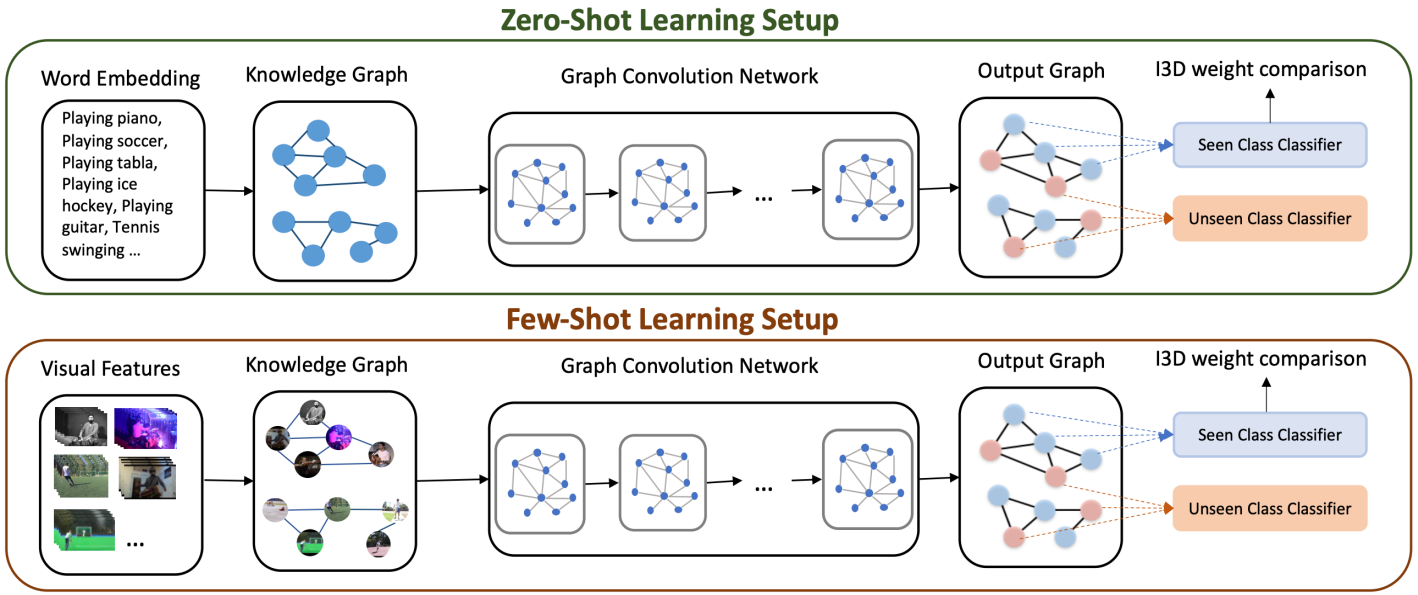


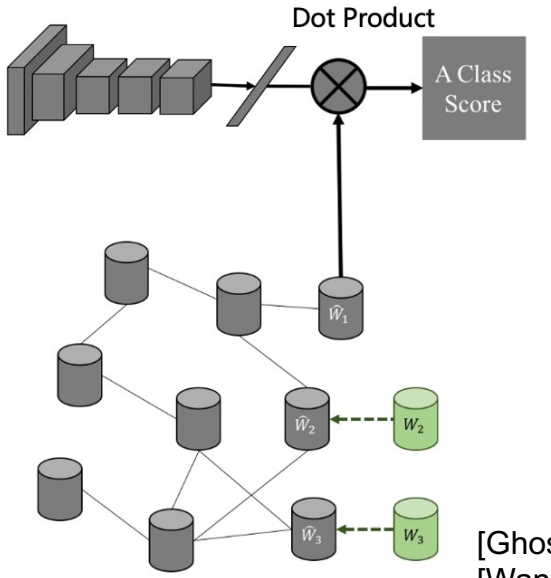
动作—名词图谱



视觉图谱



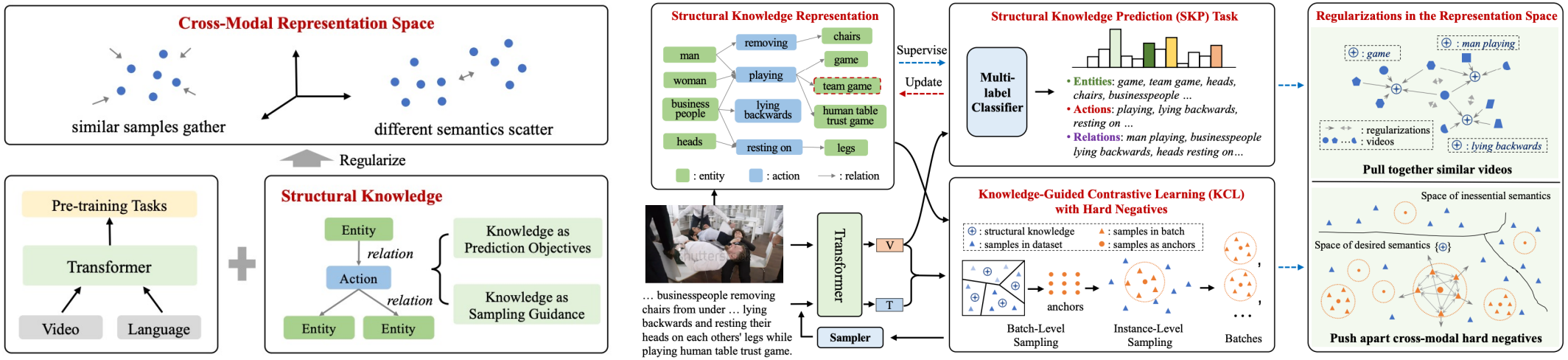




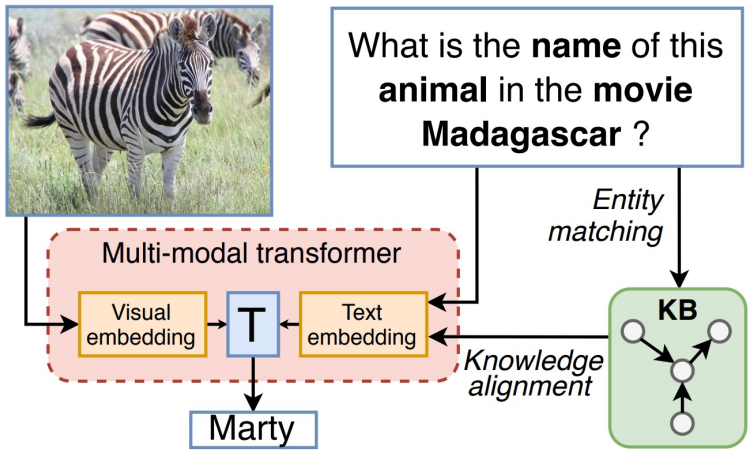
[Ghosh arXiv'20]
[Wang CVPR'18]

知识融入从外部到联合

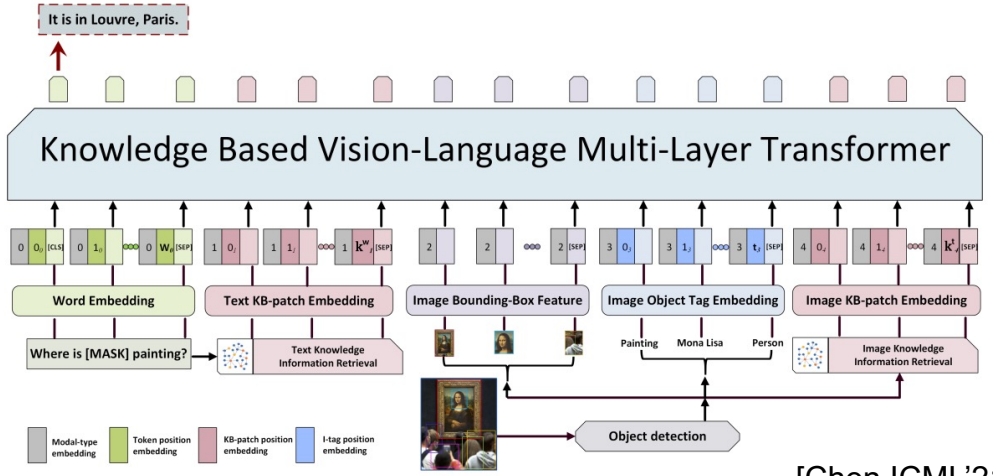
- 知识图谱与视觉特征联合优化，约束视觉特征学习



[Li MM'22]



[Shevchenko ACL'21]



[Chen ICML'21]

提纲

技术路线回顾

趋势一: 训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

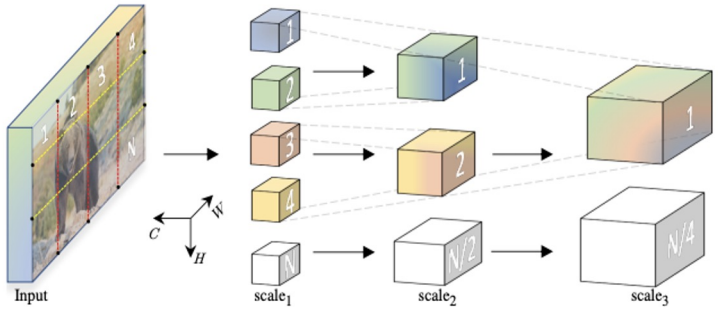
趋势四：知识融入从外部到联合

趋势五：部署推理从普适到专用

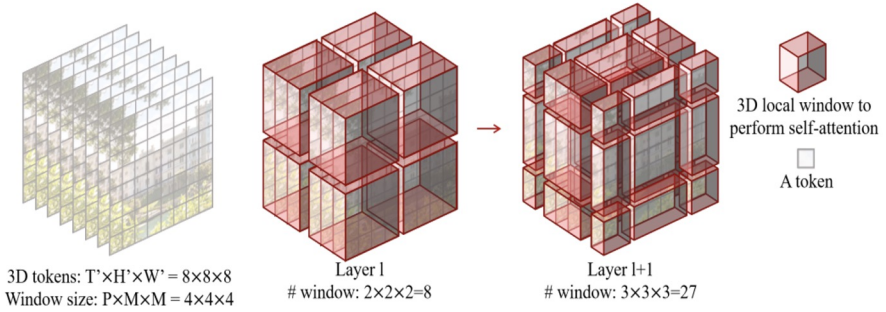
趋势六：可信学习从探索到应用

部署推理从普适到专用

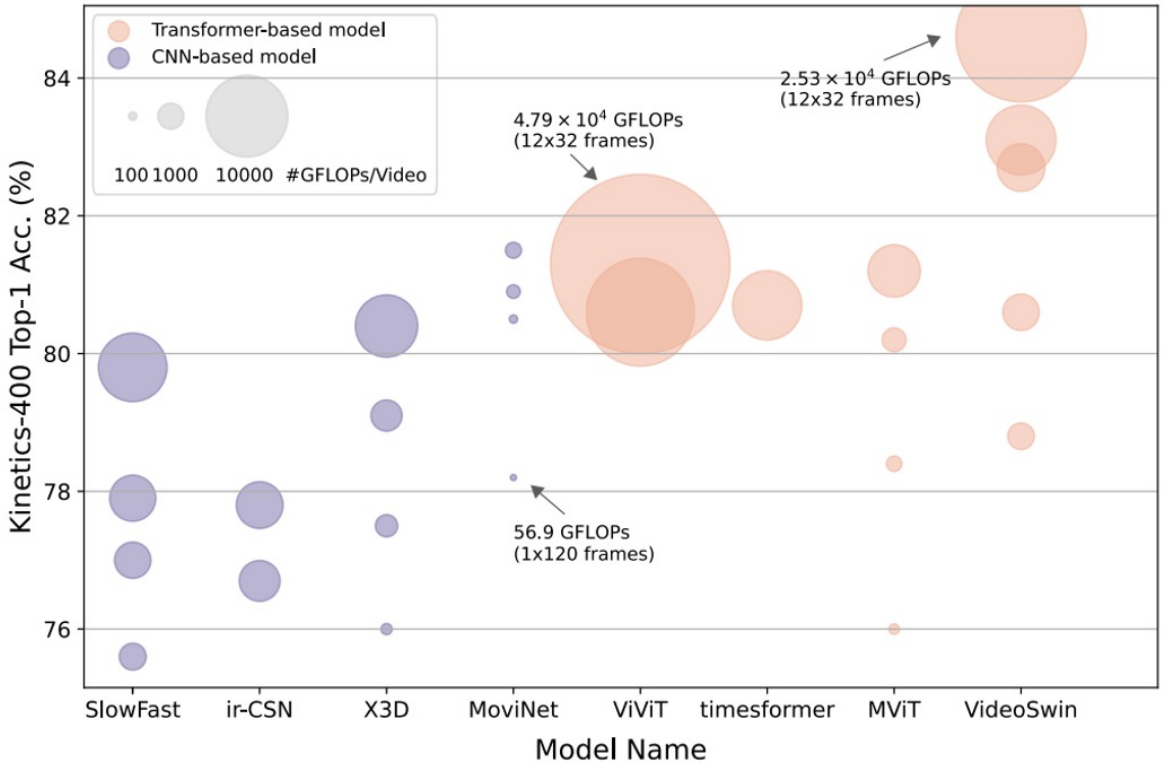
- 深度神经网络性能强，但参数量大、功耗高，难以部署在计算资源有限的场景中



[Fan ICCV'21]



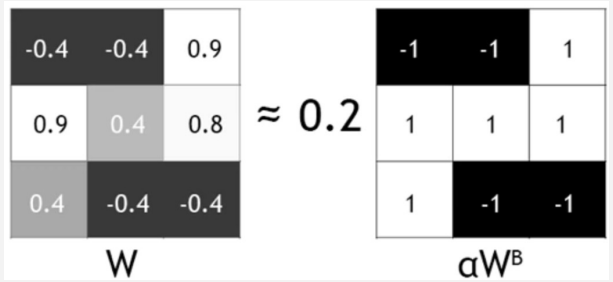
[Liu CVPR]



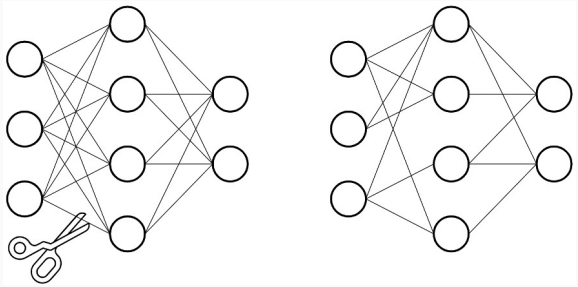
部署推理从普适到专用

- 传统的模型压缩方法对所有样本采用相同的计算资源

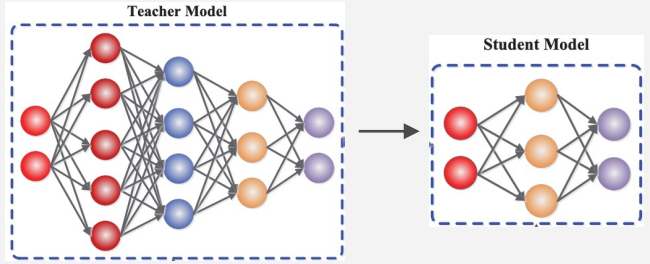
二值化神经网络



模型剪枝



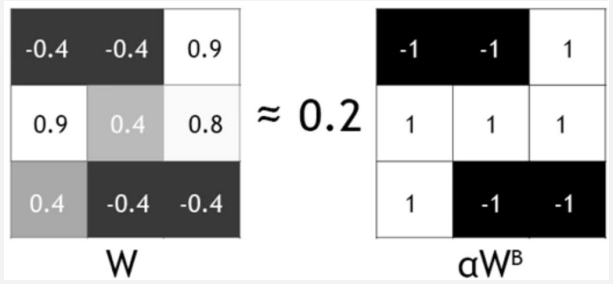
知识蒸馏



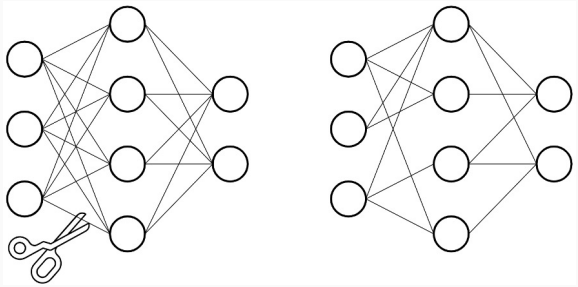
部署推理从普适到专用

- 传统的模型压缩方法对所有样本采用相同的计算资源

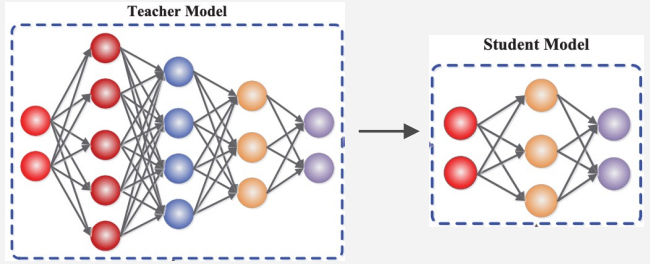
二值化神经网络



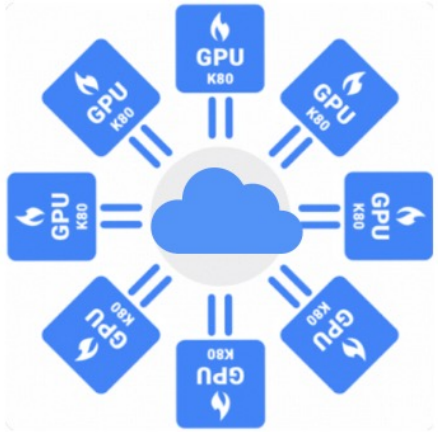
模型剪枝

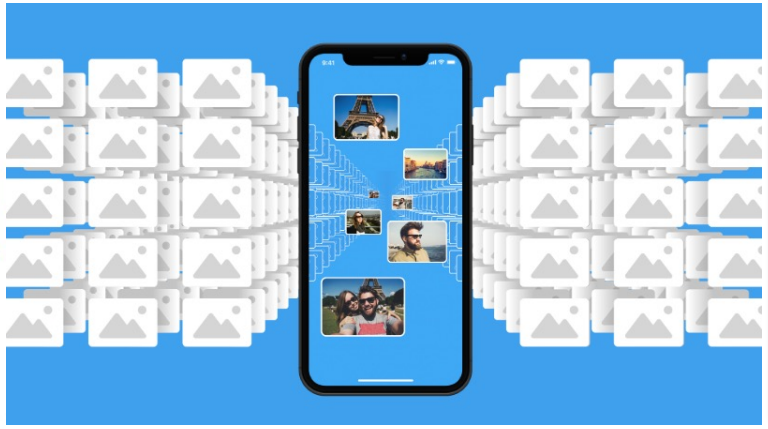


知识蒸馏



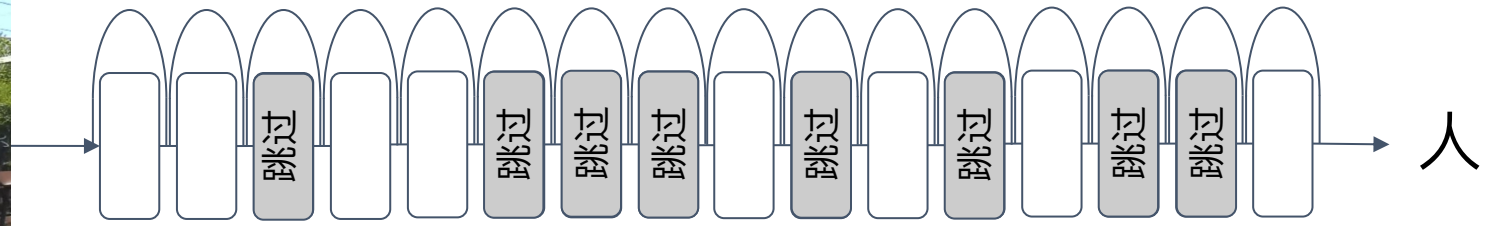
- 缺乏针对不同应用场景，动态调整计算资源的能力



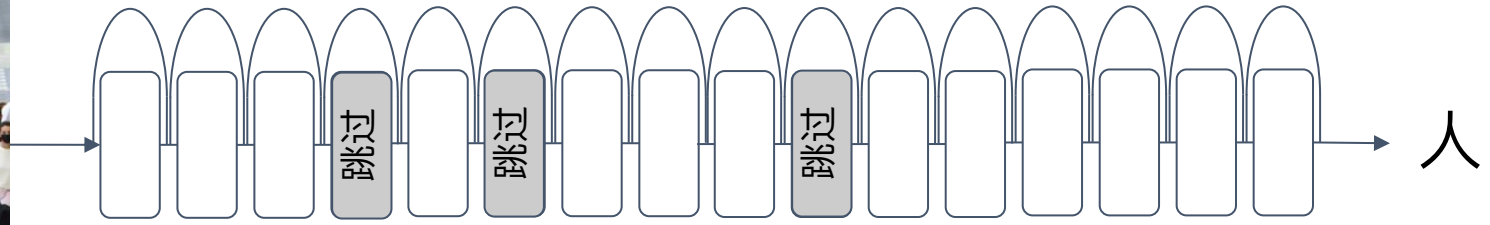


部署推理从普适到专用

- 训练得到网络后，**针对不同场景动态调整计算资源**



简单场景？使用8层



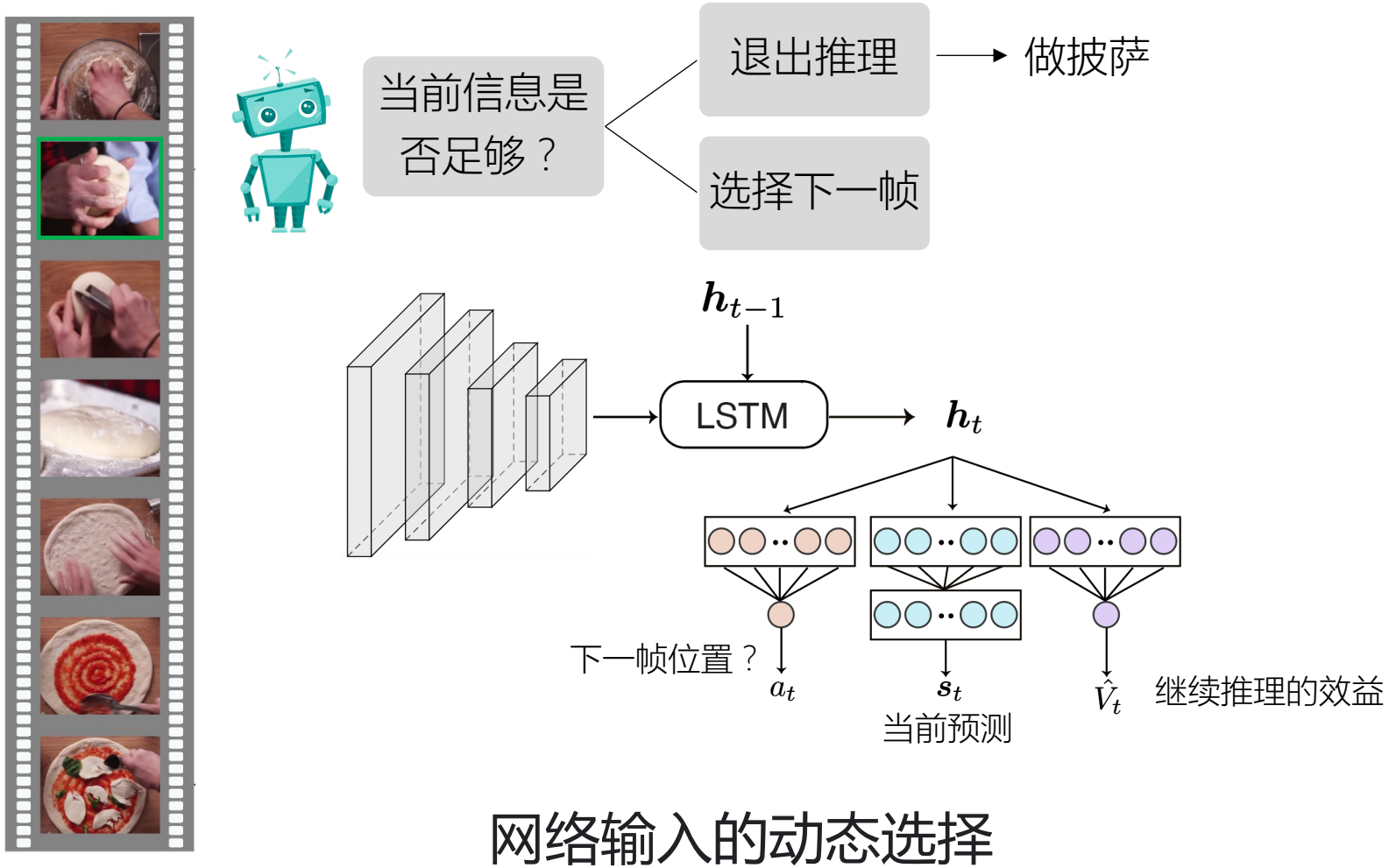
复杂场景？ 使用13层！

网络结构的动态选择

[Wu CVPR'18]

部署推理从普适到专用

- 训练得到网络后，**针对不同场景动态调整计算资源**

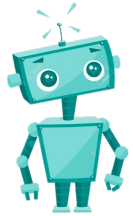


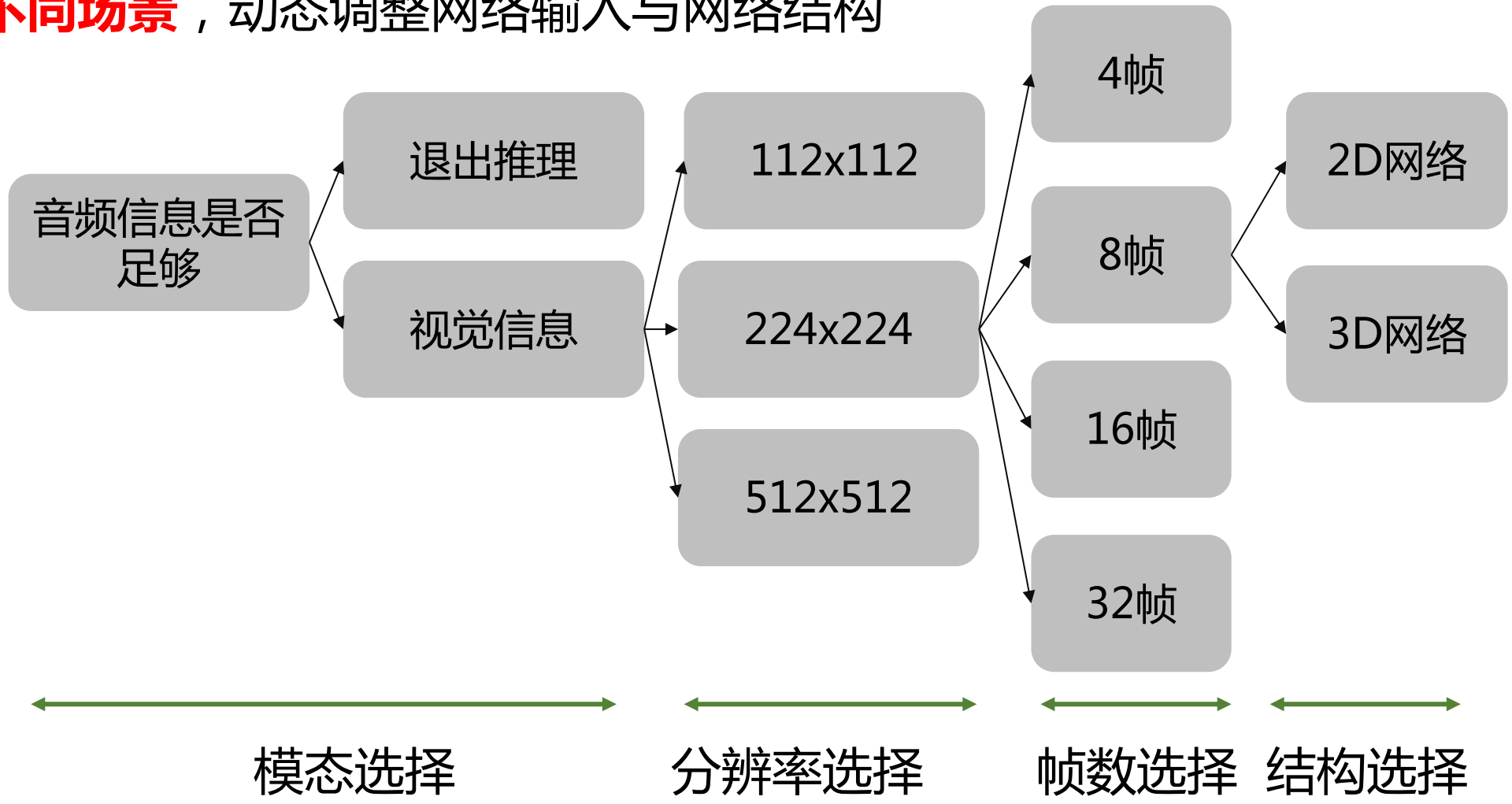
网络输入的动态选择

[Wu TPAMI'22]

部署推理从普适到专用

- **针对不同场景**，动态调整网络输入与网络结构





提纲

技术路线回顾

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

趋势四：知识融入从外部到联合

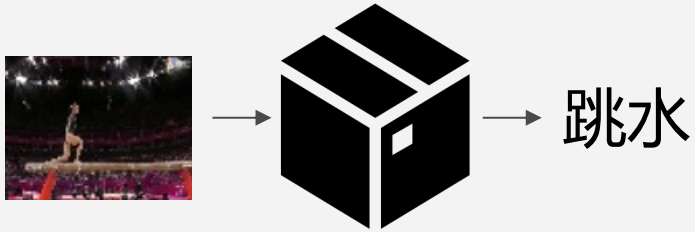
趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

可信学习从探索到应用

- 视频内容分析取得了巨大成功，但也存在诸多问题

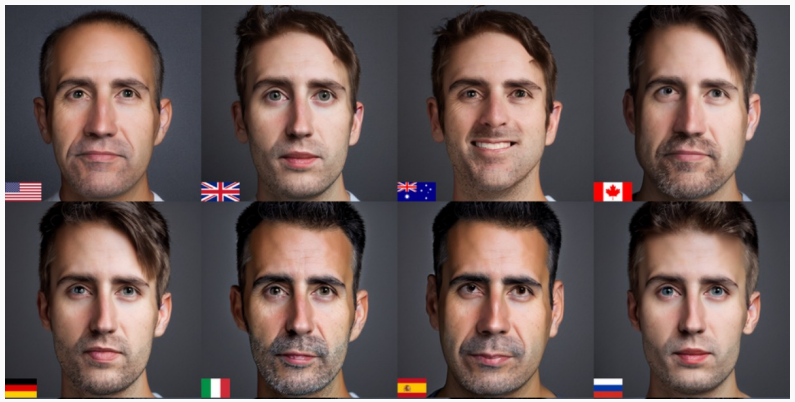
不可解释



不公平

不鲁棒

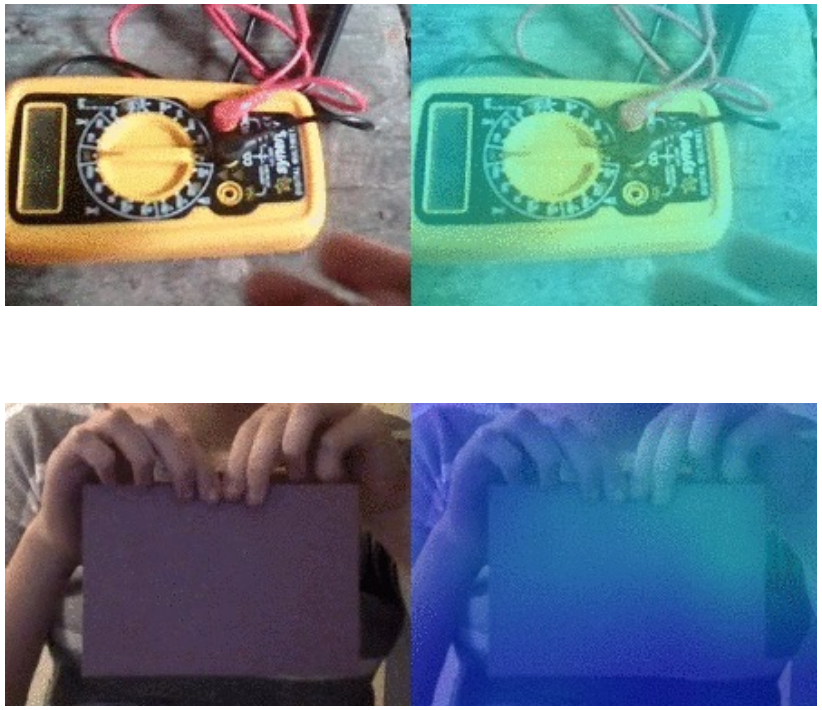
算法滥用



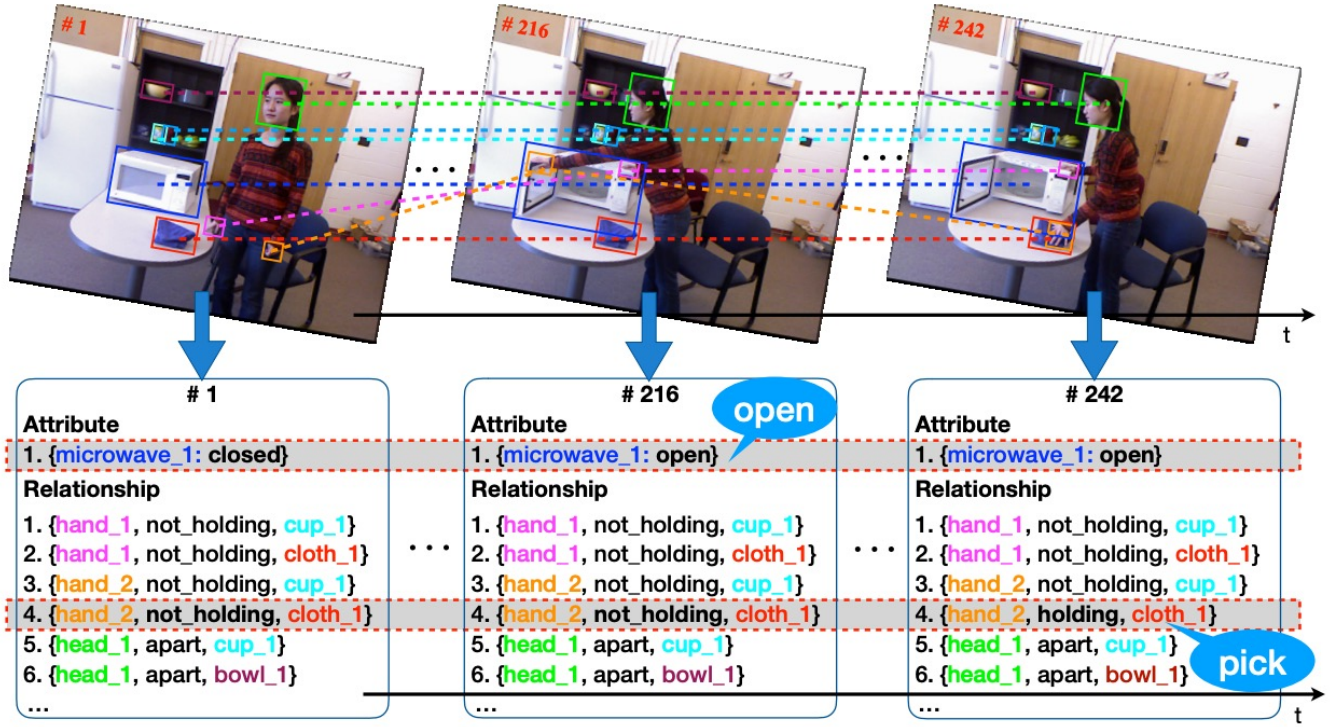
生成技术
被滥用

可信学习从探索到应用

- 可解释性：可视化激活图、梯度、决策依据以及所学知识，提高模型可解释性



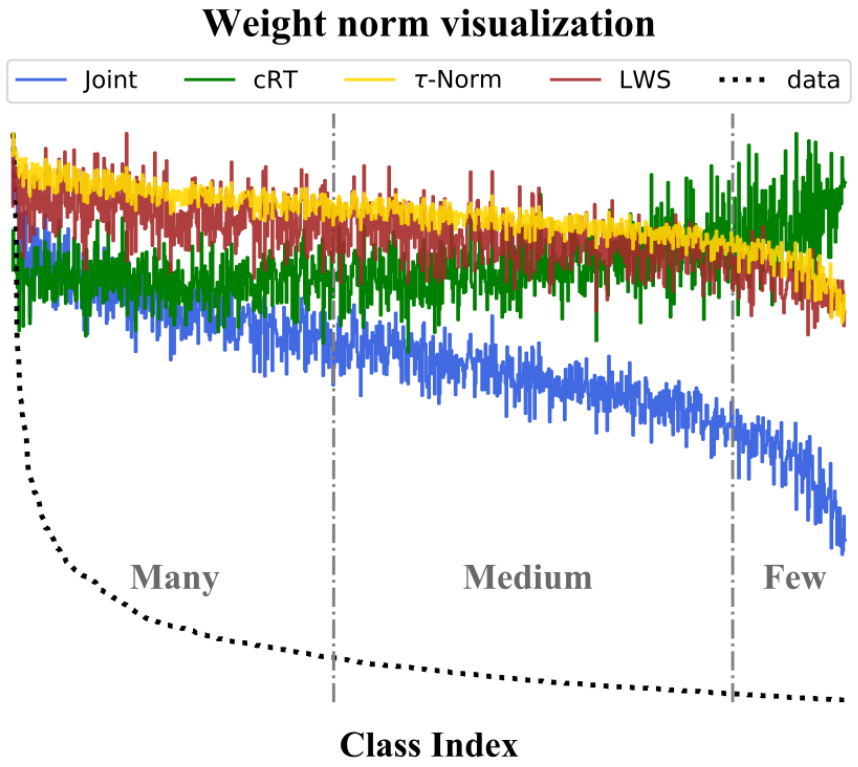
基于类别激活图的可视化 [Zhou CVPR'19]



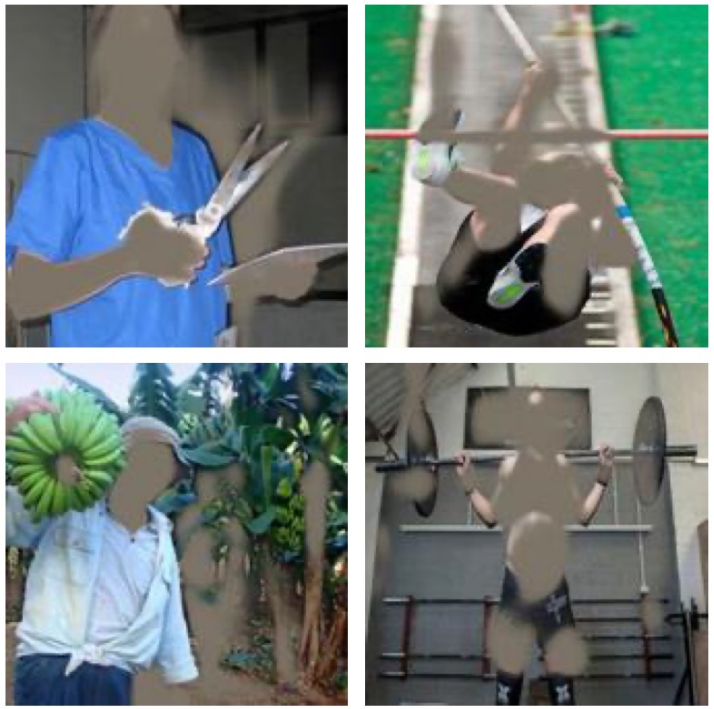
基于属性转移的动作识别 [Zhuo MM'19]

可信学习从探索到应用

- 公平性：通过重采样、移除性别/年龄属性等缓解训练数据不均衡导致的偏见



数据重采样 [Kang ICLR'20]



性别/年龄属性消除 [Wang ICCV'19]

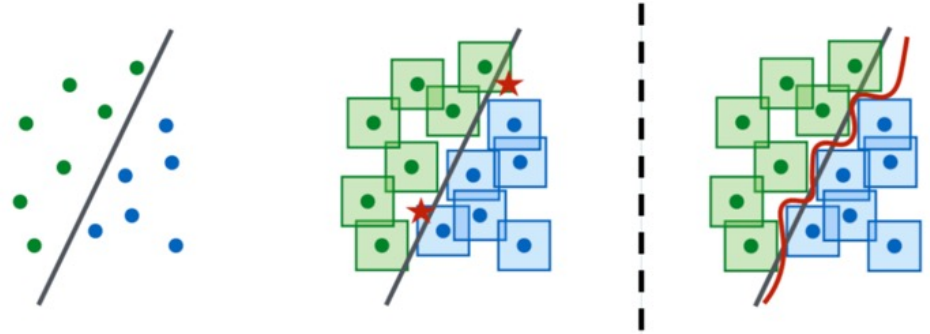
可信学习从探索到应用

- 鲁棒性：对抗训练、数据增强等方式提升视频内容分析的鲁棒性

◻ 提升对损坏输入的鲁棒性

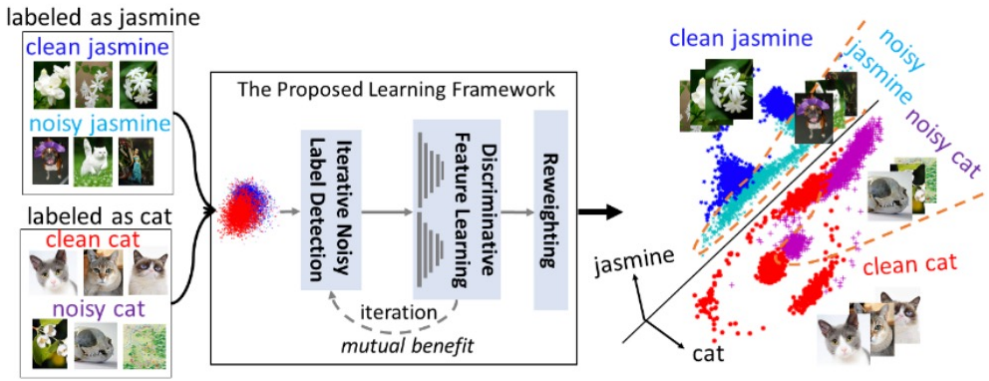
[Schiappa arXiv'22]

◻ 提升对抗鲁棒性



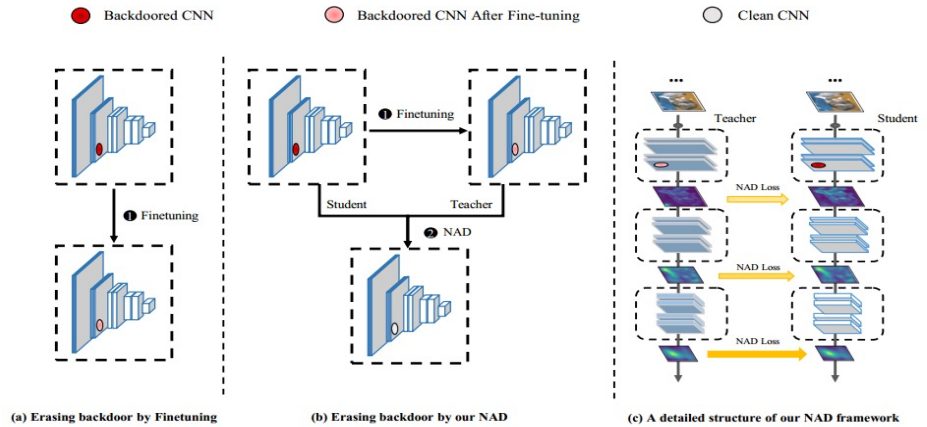
[Madry ICLR'18]

◻ 提升对噪声标注的鲁棒性



[Wang CVPR'18]

◻ 降低被后门攻击的风险

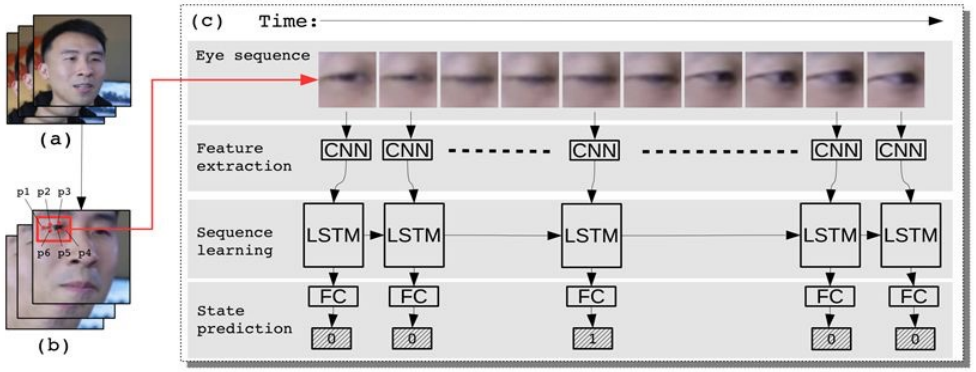


[Li ICLR'21]

可信学习从探索到应用

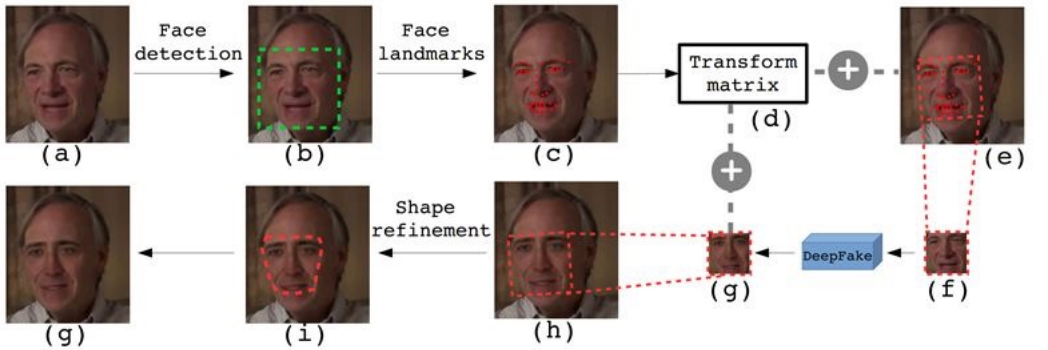
- 伪造视频的鉴别

□ 基于眨眼频率估计的换脸视频鉴别



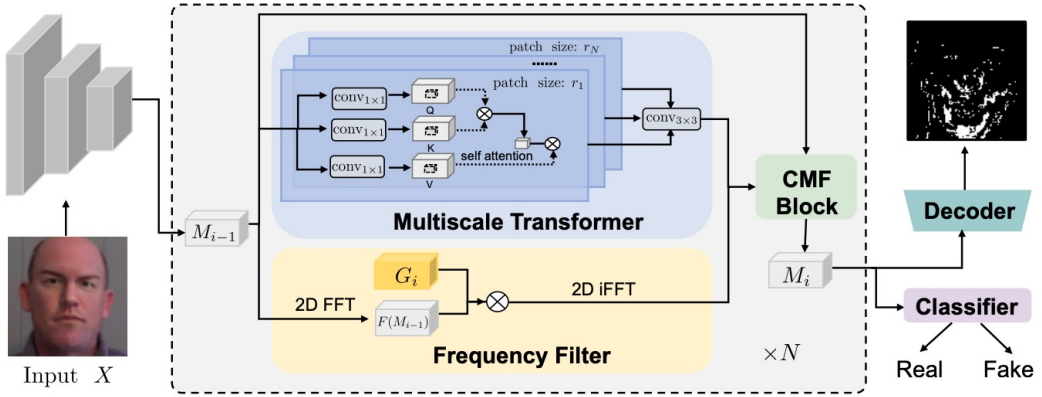
[Güera AVSS'18]

□ 基于头部姿势估计的换脸视频鉴别



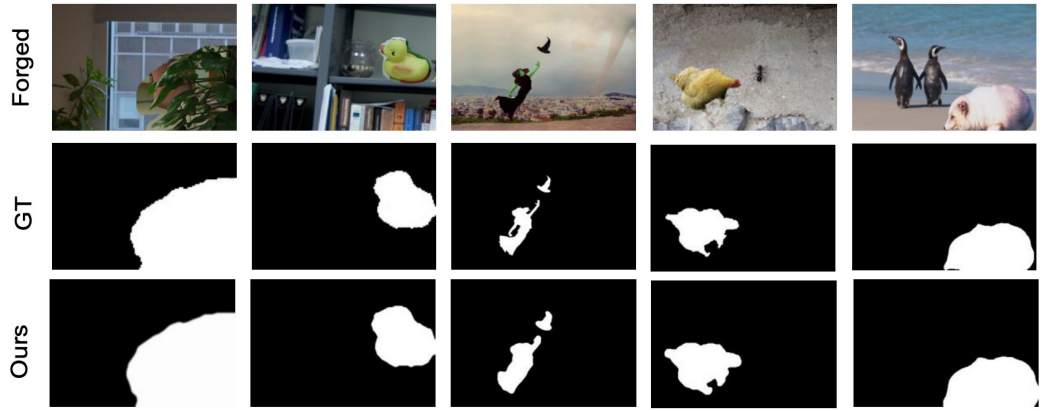
[Li CVPRW'19]

□ 基于多尺度特征的换脸检测



[Wang ICMR'22]

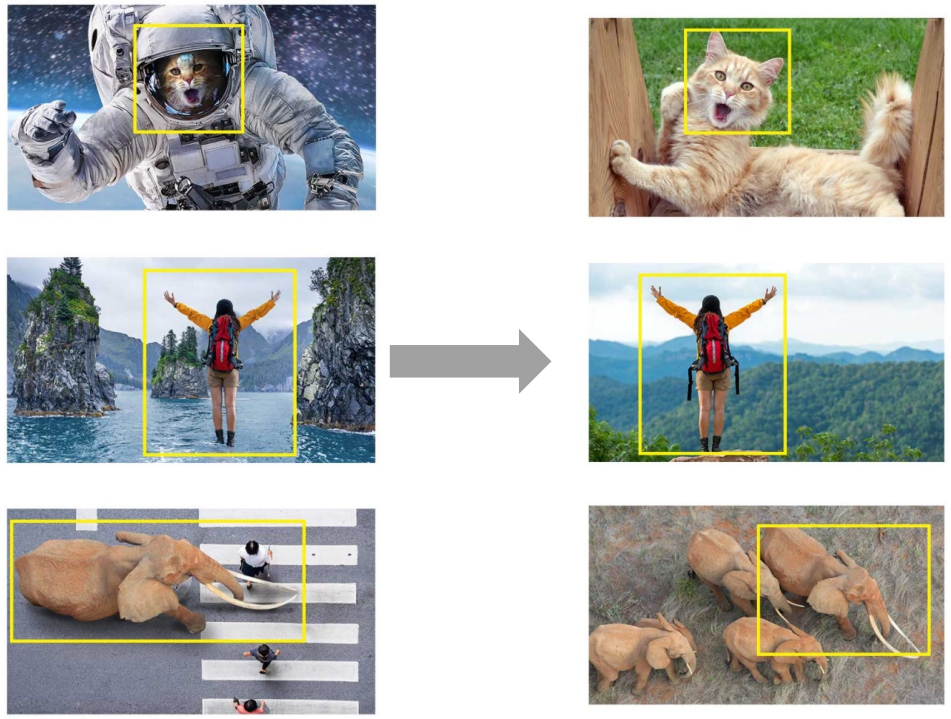
□ 基于多域特征的ObjectFormer检测模型

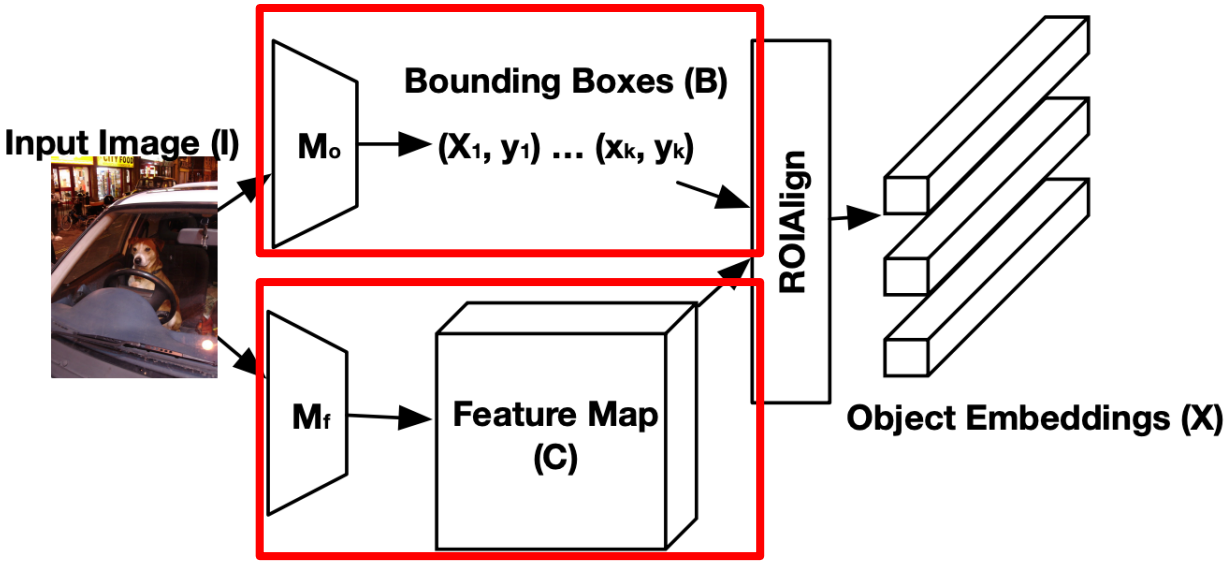


[Wang CVPR'22]

可信学习从探索到应用

- 伪造视频的溯源





FACEBOOK AI Research Publications

ML APPLICATIONS | INTEGRITY

Here's how we're using AI to help detect misinformation

November 19, 2020

基于物体级别的表征，找出用于篡改的原始数据

[Chen CVPR'21]

总结

- 六大趋势

趋势一：训练数据从单一到多元

趋势二：网络架构从分散到统一

趋势三：监督信号从人标到自学

趋势四：知识融入从外部到联合

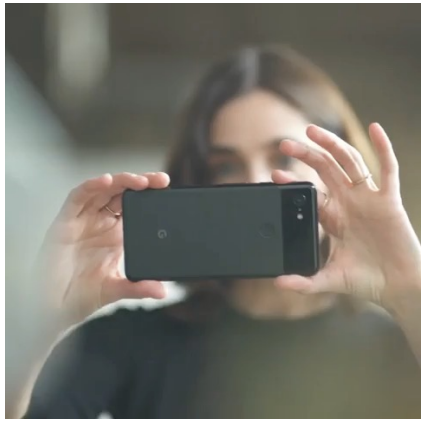
趋势五：部署推理从普适到专用

趋势六：可信学习从探索到应用

- 更好的视频内容分析驱动未来世界的智能感知

4K/8K视频理解

智能机器人



AR/VR

…

谢谢！

Fudan Vision and Learning Lab

https://fvl.fudan.edu.cn/