Understanding Human Activity from Visual Data

Mehrsan Javan

CTO

Sportlogiq

SPORT LOGiQ

Outline

- Introduction and definitions
- Industry applications – why sports
- Technology evolution & core concepts
  - Classical feature-based approaches combined with structured output models
  - Deep learning, transformer-based architectures & large models
  - Vision-Language Models (VLMs)
- Computational & deployment challenges
- Conclusion & future directions

Fine-Grained Understanding Is the First Building Block

- Activity Detection: Identifies where and when an activity occurs
- Activity Recognition: Labels without spatio-temporal localization
- Action Grounding: Maps specific actions to visual cues in a video in response to a textual query
- Video Captioning: Generates natural language descriptions for video content, often capturing a sequence of activities and their relationships.

Actions are complex. There is ambiguity in defining and labelling complex actions which are a set of related consecutive atomic actions. We don't want a label such as "person-throws-cat-into-trash-bin-after-petting"

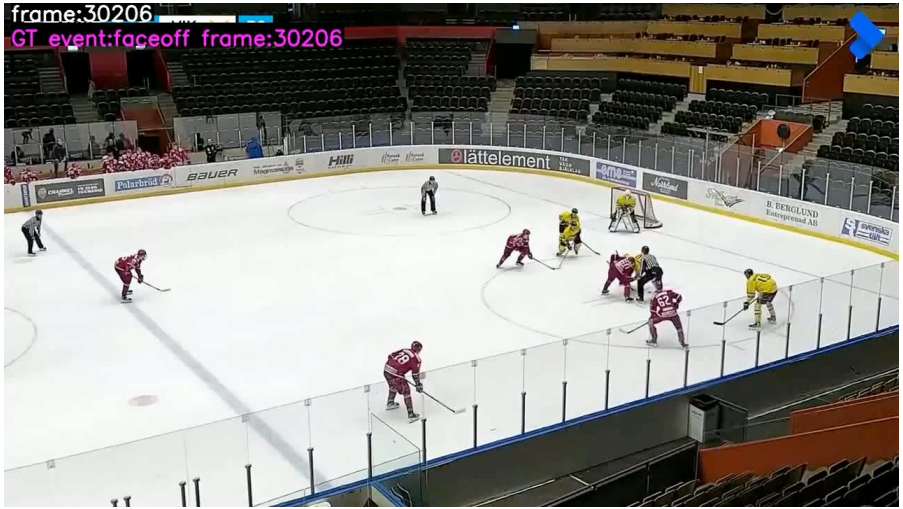
Why It Is Difficult

- Many applications need spatio-temporal localization
- Challenges
  - Large variation in appearances and viewpoint & occlusions, non-rigid motion, temporal inconsistencies
  - Prohibitive manual collection of training samples & rare occurrences
  - Complex actions and not well-defined action vocabularies
- Existing datasets are still small

Perrett et al. HD-EPIC Dataset, 2025

Industry Applications

- Surveillance, autonomous cars, robotics, retail, etc.
- Sport analytics an ultimate testbed
  - Structured environment, rich annotated datasets
  - Multi-agent interactions and strategic complexity
  - Visual similarities of multiple actions
  - Need for precise spatio-temporal localization
  - Need to model long temporal context



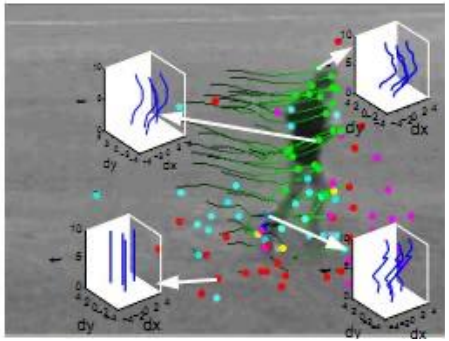
Sports as an Ultimate Testbed

- Sport videos show temporally dense, fine-grained, multi-person interactions
- Sports analytics demand exact timing and positioning of actions
- Understanding context is critical for correct interpretation of actions

Time	Event Name	Player ID	Location (x, y)
0	Faceoff		(68.71, 21.38)
1.96	Loose Puck Recovery	VIK #40	(71.22, 22.38)
5.92	Dump Out	VIK #40	(38.53, -28.42)
9.84	Loose Puck Recovery	MOD #30	(-98.27, -0.76)
11.20	Pass	MOD #30	(-98.27, -1.26)
12.92	Reception	MOD #28	(-98.27, 12.82)
13.40	Pass	MOD #28	(-93.74, 18.86)
14.40	Reception	MOD #55	(-95.25, -3.77)
16.16	Pass	MOD #55	(-93.24, 3.27)
16.76	Reception	MOD #62	(-72.12, 31.94)
19.12	Controlled Exit	MOD #62	(-25.34, 36.46)
19.40	Pass	MOD #62	(-17.8, 34.95)
19.44	Block	VIK #16	(-12.77, 29.92)
…			

Pre-deep Learning Approaches (Before 2015)

- Hand-crafted features
  - Appearance: HOG (Histogram of Oriented Gradients), Extended SURF
  - Motion: HOF (Histogram of Optical Flow), HOG3D
  - Pose-based: Articulated pose estimation
  - Improved Dense Trajectories (IDT) – the dominant approach for motion analysis



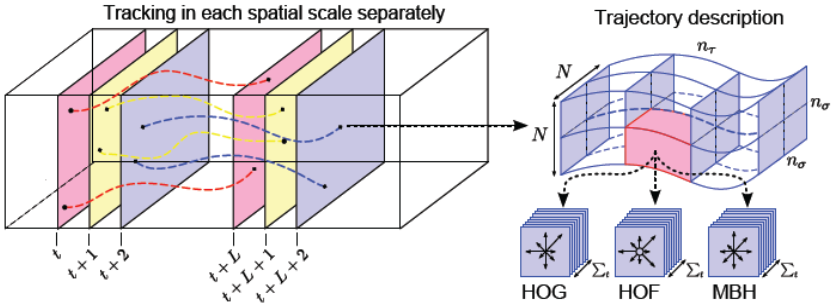


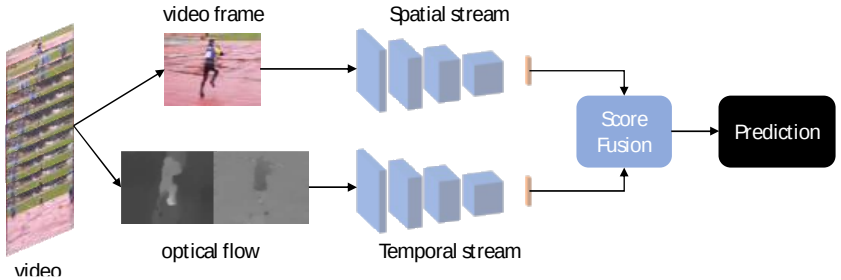
Image credit Matikainen et al. 2009; Wang et al. 2011

Pre-deep Learning Approaches (Before 2015)

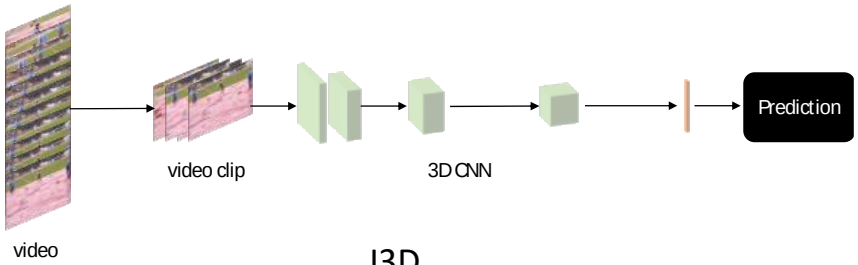
- Inference mechanisms
  - Bag of Features (BoF) and Fisher Vectors
  - Structured SVMs, HMMs, and CRFs
  - Interaction modeling: Probabilistic graphical models (Bayesian Networks, Markov Random Fields)
- Limitations
  - Small datasets with simple activities (e.g., KTH – 6 classes, Weizmann – 10 classes)
  - Limited ability to model group dynamics and complex interactions
  - Lack of generalization and poor performance in unconstrained environments

Deep Learning Era (2014 – present)

- Early CNN-based approaches (2014-2016)
  - Independent 2D CNNs on frames
  - Two-stream networks (2014): CNNs for appearance & motion
  - C3D (2015): First true 3D CNN for spatio-temporal learning
- Advancements in spatio-temporal modeling (2016-2021)
  - TSN (2016), I3D (2017), CSN (2019), SlowFast (2019), X3D (2020), and MoViNets (2021)



Two Stream Networks

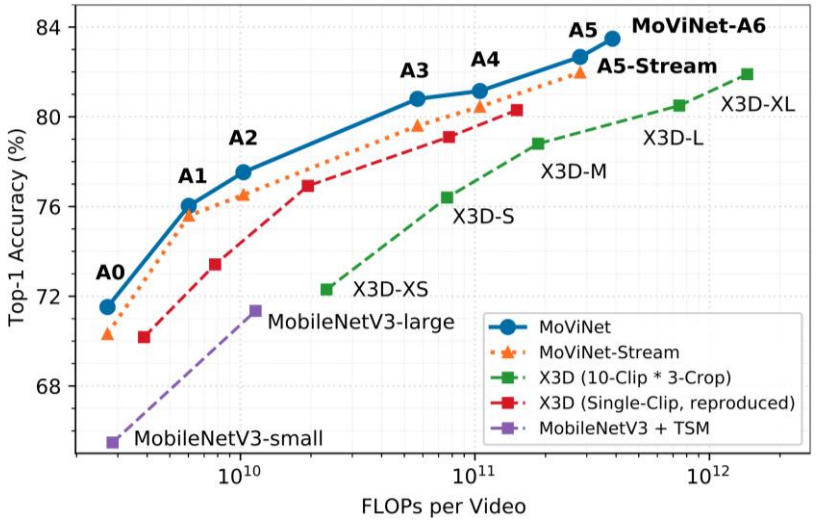


I3D

Image credit Zhu et al. 2020

Deep Learning Era (2014 – present)

- 3D CNN models issues
  - Short temporal attention span up to a few seconds
  - Longer video understanding needs attention-based layers
  - Don’t scale well with more data
  - Difficulties in scaling to action detection and distinguishing actions with subtle differences
- They still stand a chance compared to transformers for action recognition with small training sets

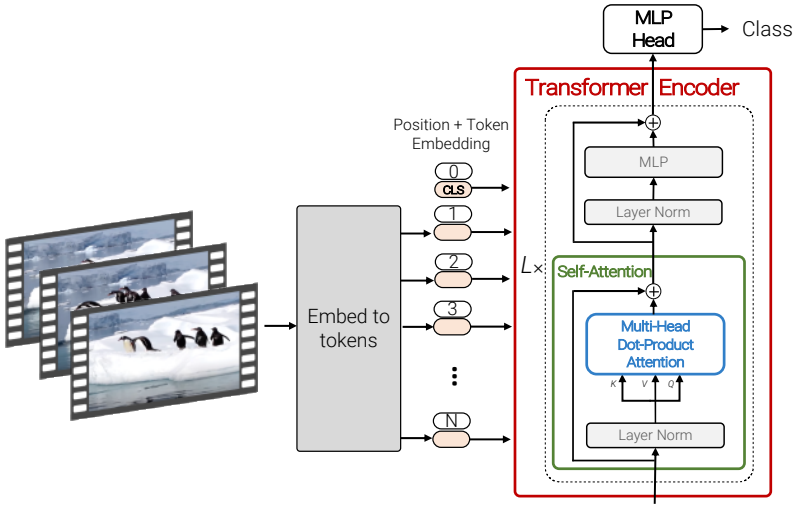


Accuracy vs. FLOPs on Kinetics 600.
MoViNets are more accurate than 2D networks and more efficient than 3D networks.

Image credit Kondratyuk et al. 2021

Vision Transformer Era (2020 – present)

- ViT (Vision Transformer): Self-attention for spatial feature learning
- Extension to video transformers (TimeSformer, VideoMAE, ViViT, MViT, UniFormer)
  - Strengths: Handles long-range dependencies, better scene understanding
  - Challenges: High computational cost, data efficiency issues
  - More favorable to larger datasets

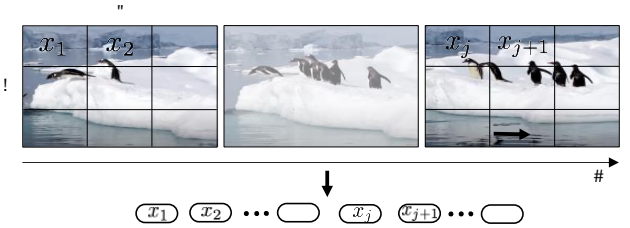


ViViT. Joint Spatio-Temporal Attention Space/Time Factorizations

Image credit Arnab et al. 2021

Transformers for Video Understanding (2020-Present)

- Transformer characteristics:
  - Scale with larger datasets
  - Can naturally handle any input which can get “tokenized”
  - Inherent attention mechanism for spatio-temporal information encoding
  - Handle long-range dependencies, better scene understanding
  - Can accommodate multiple modalities



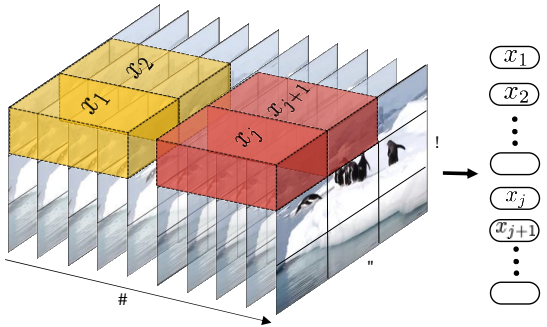
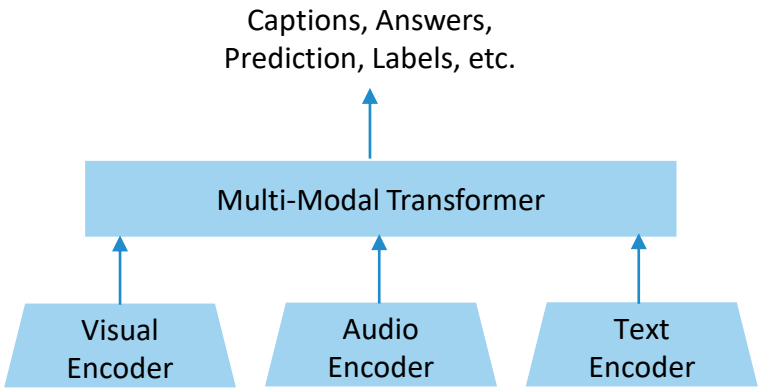


Image credit Arnab et al. 2021

Foundational Vision-Language Models (VLMs)

- Visual encoding with transformers
- Independent modality encoding: Other modalities like audio, text (e.g., captions, language queries
- Cross-modal fusion with transformers
- Encoding and decoding: Creating a unified representation or generating outputs like captions, answers, or predictions



Foundational Vision-Language Models (VLMs)

- Input tokenization
  - Video: Pre-trained visual encoders, e.g., TimeSformer, ViViT, VideoMAE
  - Audio: Embeddings with pre-trained models, e.g., wav2vec 2.0 or HuBERT
  - Text: Queries, captions, and metadata tokenized using BPE (in LLaMA, T5, or GPT)
- Transformer-based processing
  - Cross-modal fusion through shared attention layers
- Common architectures
  - InternVideo2, Video-LLaMA, OmniVL

Foundational Vision-Language Models (VLMs)

- Architectures tailored for activity understanding
  - Activity-focused encoders process spatio-temporal features, e.g., motion patterns, actor interactions, e.g., InternVideo, TimeSformer
  - InternVideo2: Supports temporal video QA, event localization, trained on activity datasets
  - Video-LLaMA / OmniVL: Aligns video frames with fine-grained descriptions of actions

Security Video Analysis

- Gemini2.5 summary:
  - Security camera footage of a woman throwing a kitten into a wheelie bin on a street.
- Detailed description:
  - The video is security camera footage capturing an incident on a street. A middle-aged woman interacts seemingly affectionately with a cat near a green trash bin. After glancing around, she abruptly picks up the cat, drops it into the bin, secures the lid, and leaves the scene.

Event Breakdown with Timestamps

- Timestamps match the actual times

  0:00 - 0:01: … a static view from a security camera overlooking a residential street. A green wheelie bin is prominent in the foreground left. A black car is parked further down on the left, and a teal car is parked on the right. The timestamp indicates 21/08/2010, 19:54:13.
  0:01 - 0:04: A woman with grey hair, wearing a white short-sleeved top and dark trousers/skirt, carrying a dark shoulder bag, walks into the frame from the right side, heading left along the pavement.
  0:05 - 0:08: The woman stops near the green bin as a small, dark-colored cat emerges from the bushes/garden area to her left and approaches her. She bends slightly and begins stroking the cat with her right hand.
  0:09 - 0:12: A white car drives past on the road (left to right). The woman continues to pet the cat.
  0:13 - 0:15: The woman looks up and glances over her left shoulder, scanning the street briefly in both directions (up towards the right, then down towards the left).
  0:16 - 0:17: She bends down further and quickly scoops up the cat with both hands.
  0:18 - 0:20: Turning towards the bin, she uses her right hand to lift the lid of the green wheelie bin. A dark-colored car drives past in the background (left to right).
  0:21 - 0:22: She drops the cat directly into the open bin.
  0:23: She immediately pushes the lid down firmly with her right hand, closing the bin.
  0:24 - 0:25: The woman turns away from the bin to her left and promptly walks away down the pavement, exiting the frame to the left. The bin remains closed.

Complex Action Localization In Sports

- Hockey segment – Gemini2.5
  - From approximately 0:05 onwards, the play is characterized by back-and-forth action primarily within the neutral zone. Both teams struggle to gain clean entries into the offensive zone, resulting in frequent changes of possession near the blue lines and center ice due to stick checks, intercepted passes, and defensive pressure. Neither team establishes sustained offensive zone time or generates a clear scoring chance during this period.

Time	Event Name	Player ID	Location (x, y)
0:01	Loose Puck Recovery	VIK #15	(105, -5)
0:03	Dump Out	VIK #15	(95, -15)
0:04	Loose Puck Recovery	MOD #?	(-90, -20)
0:05	Pass	MOD #?	(-85, -35)
0:06	Reception	MOD #?	(-60, -38)
0:06	Controlled Exit	MOD #?	(-25, -38)
0:07	Loose Puck Recovery	VIK #?	(10, 5)
0:09	Loose Puck Recovery	MOD #28	(-5, 5)
0:14	Loose Puck Recovery	VIK #15	(70, 10)
0:16	Pass	VIK #15	(55, 0)
0:16	Reception	VIK #?	(45, -5)
0:17	Loose Puck Recovery	MOD #?	(5, -10)
0:19	Pass	MOD #?	(15, 5)
…			

Complex Actions Localization In Sports

- At the first glance everything seems to make sense
- The text seems to be reasonable
- The output structure follows NHL play-by-play data

Let’s rewatch the game from 0:05 sec

Complex Actions Localization In Sports

- The sequence of events and locations are wrong.
- Inability to maintain continuous trajectories of all people resulted in wrong player ID.
- The context of puck possession is not understood by the model.
- Output is dominated by presence of “Pass”, “Reception”, and “Loose Puck Recovery” as they are repeated frequently in the NHL play-by-play data.

Complex Actions Localization In Sports

Ground Truth			
Time	**Event Name**	**Player ID**	**Location (x, y)**
...			
11.20	Pass	MOD #30	(-98.27 , -1.26)
13.40	Pass	MOD #28	(-93.74 , 18.86)
16.16	Pass	MOD #55	(-93.24 , 3.27)
19.12	Controlled Exit	MOD #62	(-25.34 , 36.46)
19.40	Pass	MOD #62	(-17.8 , 34.95)
22.16	Controlled Entry	MOD #55	(25.45 , -37.47)
25.80	Pass	MOD #55	(95.86 , 2.26)
33.16	Pass	MOD #62	(93.35 , 27.41)
34.44	Pass	MOD #41	(58.65 , 38.47)
36.96	Shot	MOD #55	(63.17 , -14.34)
37.28	Goal	MOD #55	(63.17 , -14.34)

Gemini 2.5 Detection			
Time	**Event Name**	**Player ID**	**Location (x, y)**
...			
0:14	Loose Puck Recovery	VIK #15	(70, 10)
0:16	Pass	VIK #15	(55, 0)
0:17	Loose Puck Recovery	MOD #?	(5, -10)
0:19	Pass	MOD #?	(15, 5)
0:20	Loose Puck Recovery	VIK #?	(30, -5)
0:22	Loose Puck Recovery	MOD #28	(-10, 0)
0:24	Pass	MOD #28	(20, -10)
0:25	Loose Puck Recovery	VIK #?	(40, -15)
0:27	Loose Puck Recovery	MOD #14	(15, -10)
0:31	Pass	MOD #14	(-10, 20)
0:32	Loose Puck Recovery	VIK #?	(-15, 25)

Limitations of VLMs for Activity Understanding

- Most reasoning and understanding in VLMs for action detection is handled by large-scale language models, while visual encoding has seen limited improvements
- Challenges in visual encoding
  - Compression on motion information and losing temporal granularity
  - Frame sampling and keyframe-based processing resulting in motion discontinuity
  - Multiple human interactions and group dynamics are left to be learnt implicitly
  - Lack of hierarchical action modeling in current VLMs

Why VLMs Capabilities Are Limited

- Text dominates reasoning
  - VLMs repurpose frozen vision encoders (ViT, TimeSformer, CLIP) and rely on LLMs to infer actions from text-based descriptions
  - Most VLMs are trained on narration-based datasets (e.g., HowTo100M, YouCook2) rather than detailed action labels.
- Existing motion models are underutilized
  - Traditional models (e.g., I3D, SlowFast, CSN) were designed for detailed motion capture, but VLMs often discard their benefits in favor of high-level features
- Lack of high-quality fine-grained datasets

Conclusion –
Deep Learning Models (CNNs, 3D ConvNets, Transformers)

- Strengths:
  - Precise spatial and temporal localization, when trained on well-annotated datasets
  - Fine-grained motion encoding and modeling short-to-medium range context
  - Highly tunable for domain-specific tasks (e.g., sports, surgery, surveillance)
- Limitations:
  - Require large labeled datasets to generalize well
  - Poor transferability to out-of-distribution scenarios
  - Lack semantic reasoning (e.g., understanding “why” actions happen)

Conclusion – Vision-Language Models (VLMs / Foundation Models)

- Strengths:
  - Zero-shot or few-shot generalization via language prompts
  - Global scene understanding and coarse temporal queries
  - Semantic search, semantic localization, video QA, and descriptive understanding
- Limitations:
  - Weak spatio-temporal grounding, especially for fine-grained or multi-person actions
  - Low temporal resolution and limited motion encoding
  - Relies heavily on text-based reasoning, not designed for frame-accurate detection

Model Selection Guide

Use Case	Recommended Model Type	Spatio-temporal Localization	Data Requirement	Tuning Complexity	Deployment Cost
General activity recognition (with labels)	3D CNN (SlowFast / TSN / MoViNets)	High	Large labeled set	Moderate to High	Moderate
Segment-level recognition (e.g., surveillance)	Transformer family, (SN / ViViT / TimeSFormer)	Medium	Moderate labels	Medium	Moderate
Sports / Fine-grained multi-person actions	Custom 3D CNN / Transformers + trackers	Very High	Dense annotations	High	High
Event retrieval / Zero Shot / Semantic QA	VLMs (e.g., InternVideo, Flamingo, VideoCoCa)	Coarse	Unlabeled	None	High
Generic queries / Video captioning / Summarization	VLMs + prompt engineering	Coarse	Unlabeled or few-shot	None / Low	High

Combining Best of Both Worlds

- Combination of strong visual embedding and long context temporal sequence models
- An example of Sportlogiq's sport video processing output

Future Directions

- Better motion encoding in foundation models
  - Integrate temporal modeling (e.g., from SlowFast, I3D) into VLMs
  - Improve frame sampling strategies and motion tokenization
- Multi-agent interaction modeling
  - Develop graph-based reasoning modules within transformers
  - Capture group dynamics in sports, team-based tasks, and surveillance
- Efficient deployment
  - Lightweight video encoders (e.g., MoViNets, TinyVLMs)
  - Compress and distill foundation models for edge inference

Resources

Pre-deep learning era

Matikainen et al. 2009 Trajectons

Wang et al. 2011, Dense Trajectories

Wang et al. 2013 IDT

Early deep learning

Karpathy et al. 2014 Deepvideo

Simonyan et al. 2014, Two-Stream Networks

Tran et al. 2015 C3D

Wang et al. 2016 TSN

Carreira et al. 2017, I3D

Feichtenhofer et al. 2019 SlowFast

Tran et al. 2019 CSN

Feichtenhofer 2020 X3D

Zhu et al. 2020, A Comprehensive Study of Deep Video Action Recognition

Kondratyuk et al. 2021, MoViNets

Resources

Transformers and VLMs

Radford et al. 2021 CLIP

Dosovitskiy et al. 2021 ViT

Bertasius et al. 2021 TimeSformer

Arnab et al. 2021 ViViT

Li et al. 2021 MViT

Tong et al. 2022 VideoMAE

Li et al. 2022 UniFormer

Wang et al. 2022 OmniVL

Alayrac et al. 2022, Flamingo

Yan et al. 2023 Video CoCa

Wang et al. 2023 VideoMAEv2

Zhang et al. 2023 Video-LLaMA

Wang et al. 2024 InternVideo2

Lu et al. 2024 FACT

Perrett et al. 2025 HD-EPIC