Action recognition in videos

Cordelia Schmid

Inria

Automatic video understanding

- Huge amount of video is available and growing daily

TV-channels recorded since 60's

30k hours of videos uploaded every hour

770M surveillance cameras world-wide

Automatic video understanding

- Classification of short clips, i.e. answer phone, shake hands

answer phone

hand shake

Hollywood dataset

Automatic video understanding

- Classification of activities, i.e. birthday party, groom an animal

Birthday party

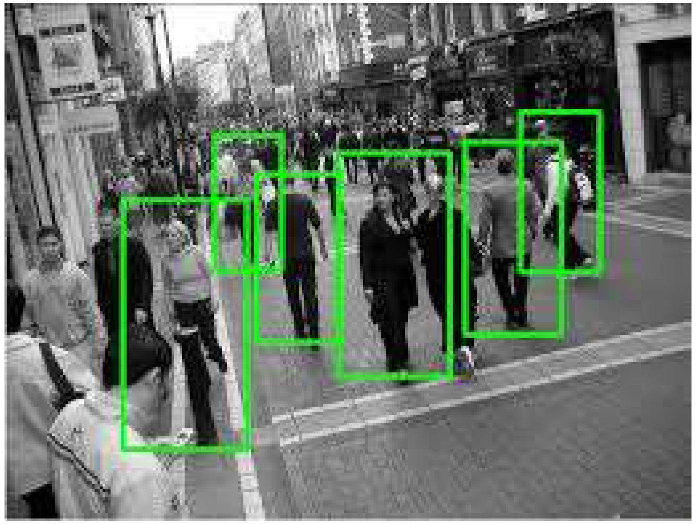
Grooming an animal

TrecVid Multi-media event detection task (MED)

Automatic video understanding

- Car safety & self-driving and video surveillance
  - Detection of humans (pedestrians) and their motion, detection of unusual behavior

Courtesy Volvo



Courtesy Embedded Vision Alliance

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Automatic video understanding

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Action recognition - difficulties

- Large variations in appearance
  - Viewpoint changes
  - Intra-class variation
  - Camera motion

Variation in appearance: viewpoint change

Variation in appearance: intra-class variation

Variation in appearance: camera motion

Action recognition - difficulties

- Large variations in appearance
  - Viewpoint changes
  - Intra-class variation
  - Camera motion
- Manual collection of training data is difficult
  - Many action classes, rare occurrence
  - Pose, object and interaction annotation often a plus
- Action vocabulary is not well defined
  - What is the action granularity?
  - How to represent composite actions?

Action recognition – approaches

- Action recognition from still images
  - Detect human pose + interaction with objects

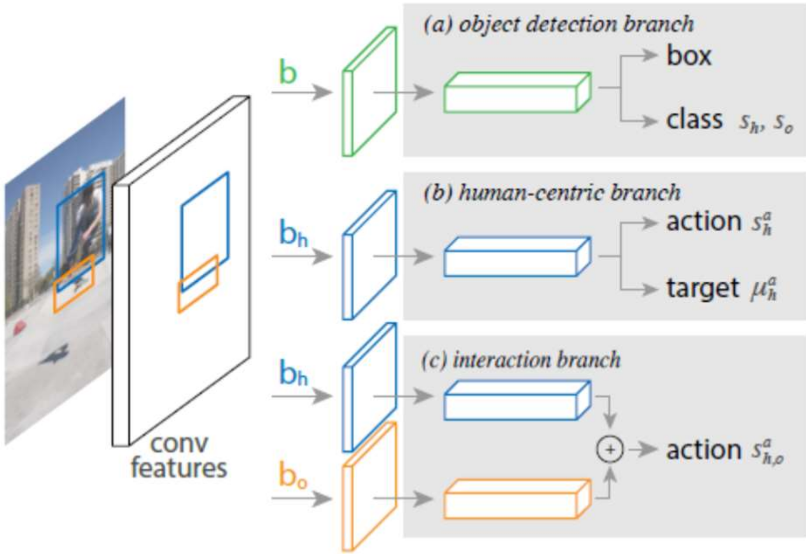
Playing Instrument | Reading | Taking Photo | Riding Horse | Walking

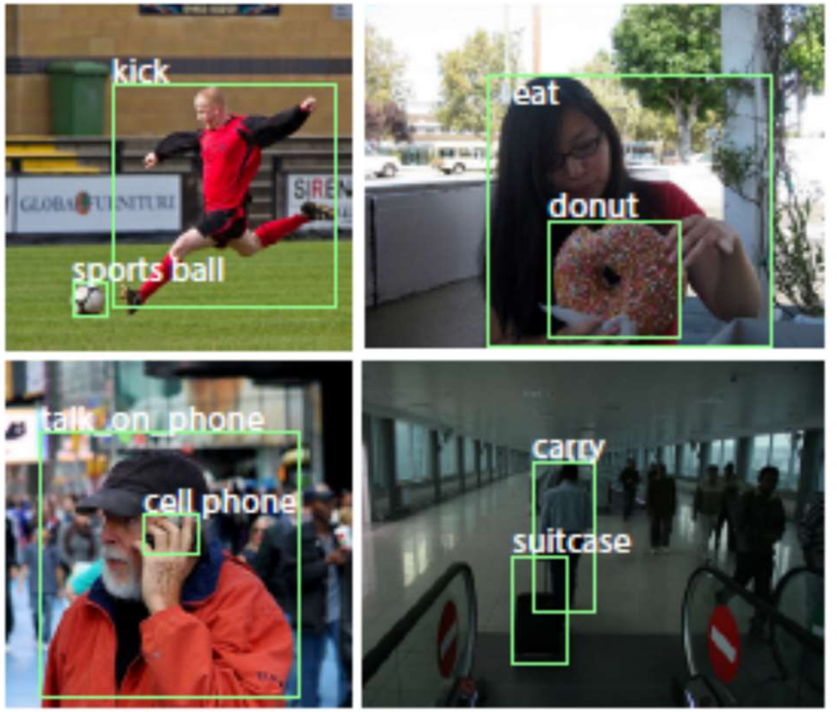
PASCAL VOC Human action classification dataset

[Weakly Supervised Learning of Interactions between Humans and Objects, Prest et al., PAMI 2012]

Action recognition – approaches

- Action recognition from still images
  - Human pose + interaction with objects





[Detecting and Recognizing Human-Object Interactions. G. Gkioxari, R. Girshick, P. Dollar and K. He. CVPR 2018]

Action recognition – approaches

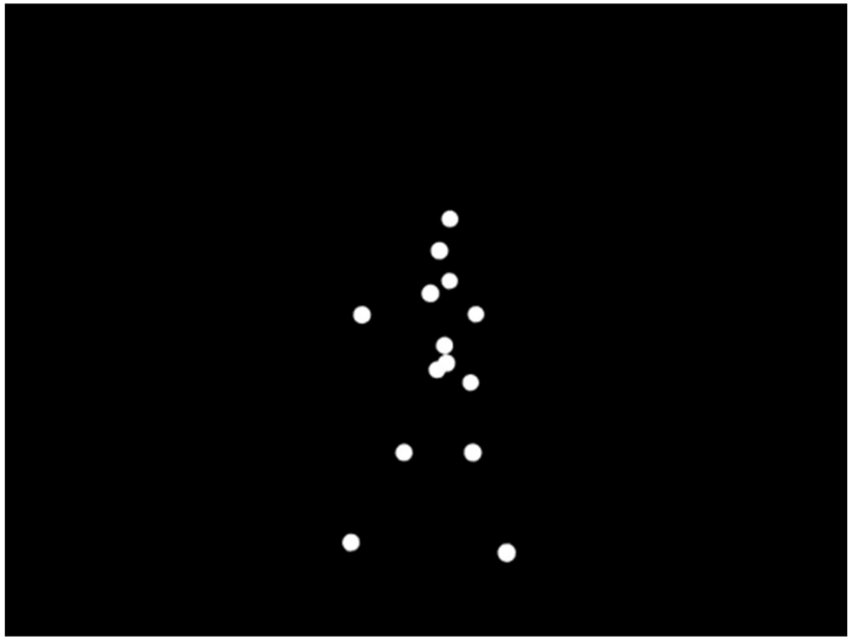
- Motion information necessary to disambiguate actions

Open or close door?

- Motion often sufficient by itself

Motion perception

- Johansson [1973] pioneered studies on sequence based human motion analysis
- Moving light displays enable identification of motion, familiar people and gender



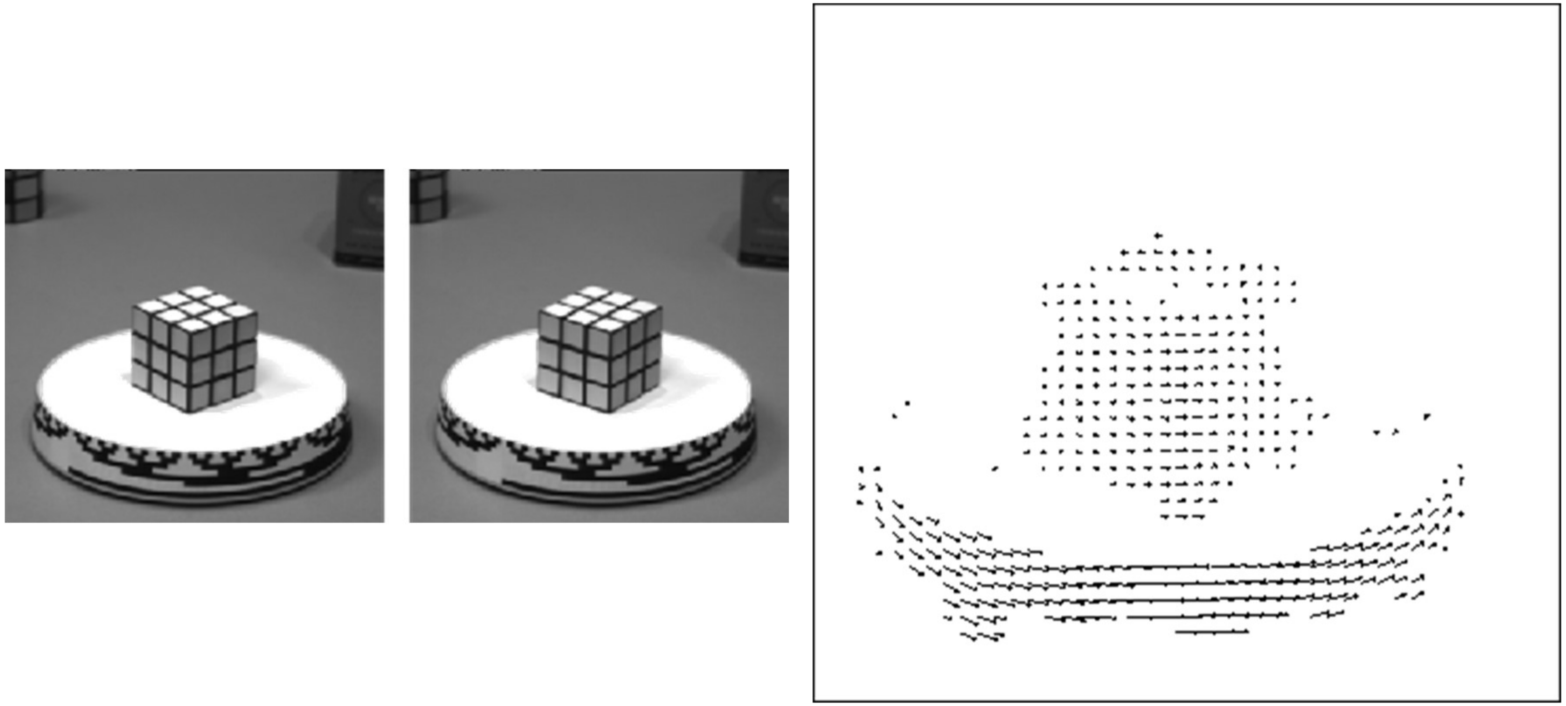
male walker

Overview

- *Optical flow*
- Video classification
- Action localization
- Multi-modal / LLM-based video understanding

Motion field

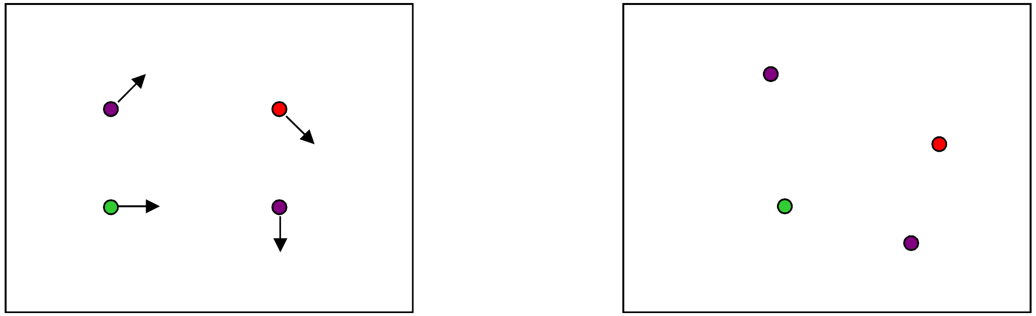
- The motion field is the projection of the 3D scene motion into the image



Optical flow

- Definition:
  - optical flow is the *apparent* motion of brightness patterns in the image
- Ideally, optical flow would be the same as the motion field
  - However, apparent motion can be caused by lighting changes without any actual motion
  - For example: a uniform rotating sphere under fixed lighting
    vs. a stationary sphere under moving illumination

Estimating optical flow



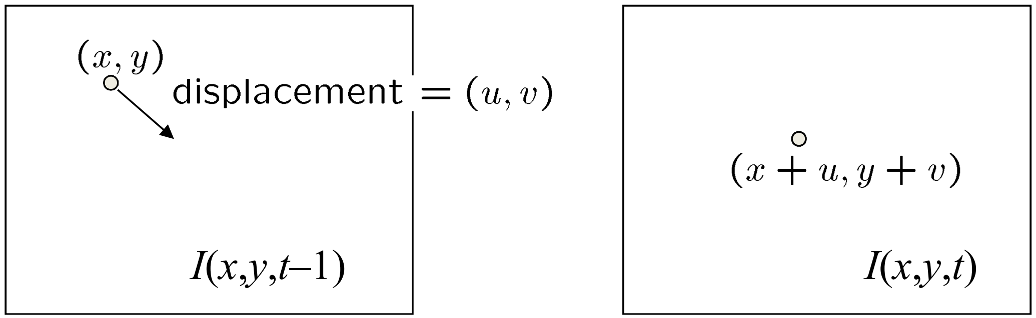
$I(x,y,t-1)$ $I(x,y,t)$

Given two subsequent frames, estimate the apparent motion field $u(x,y)$ and $v(x,y)$ between them

Key assumptions for the flow estimation in “classical” approaches

- Brightness constancy: projection of the same point looks the same in every frame
- Small motion: points do not move very far
- Spatial coherence: points move like their neighbors

The brightness constancy constraint



Brightness Constancy Equation:

$$I(x, y, t-1) = I(x + u(x, y), y + v(x, y), t)$$

Linearizing the right side using Taylor expansion (small motion):

$$I(x, y, t-1) \approx I(x, y, t) + I_x\, u(x, y) + I_y\, v(x, y)$$

Hence, $\quad I_x u + I_y v + I_t \approx 0$

The brightness constancy constraint

$$I_x u + I_y v + I_t = 0$$

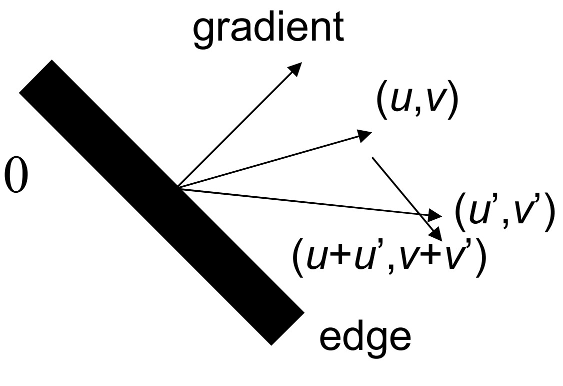
- How many equations and unknowns per pixel?
  - One equation, two unknowns

- What does this constraint mean?

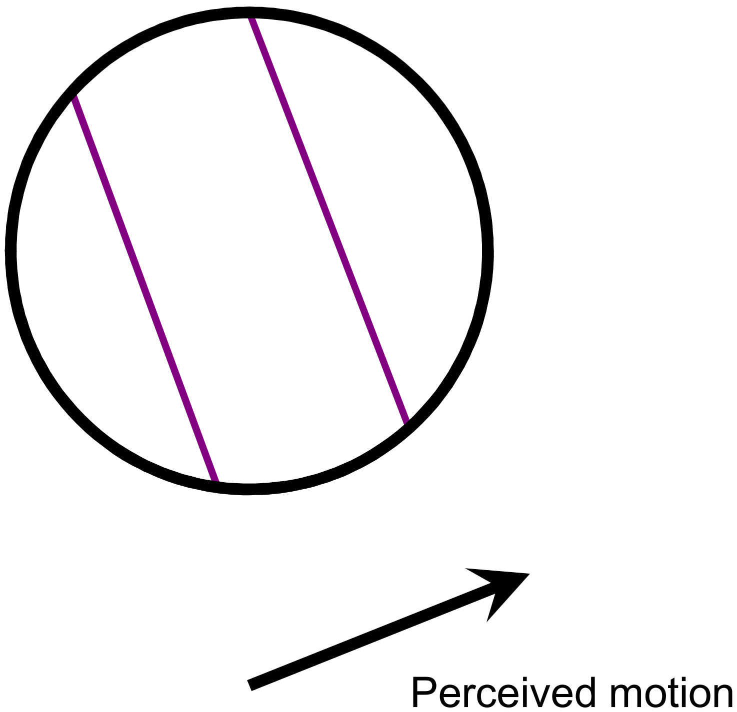
$$\nabla I \cdot (u, v) + I_t = 0$$

- The component of the flow perpendicular to the gradient (i.e., parallel to the edge) is unknown

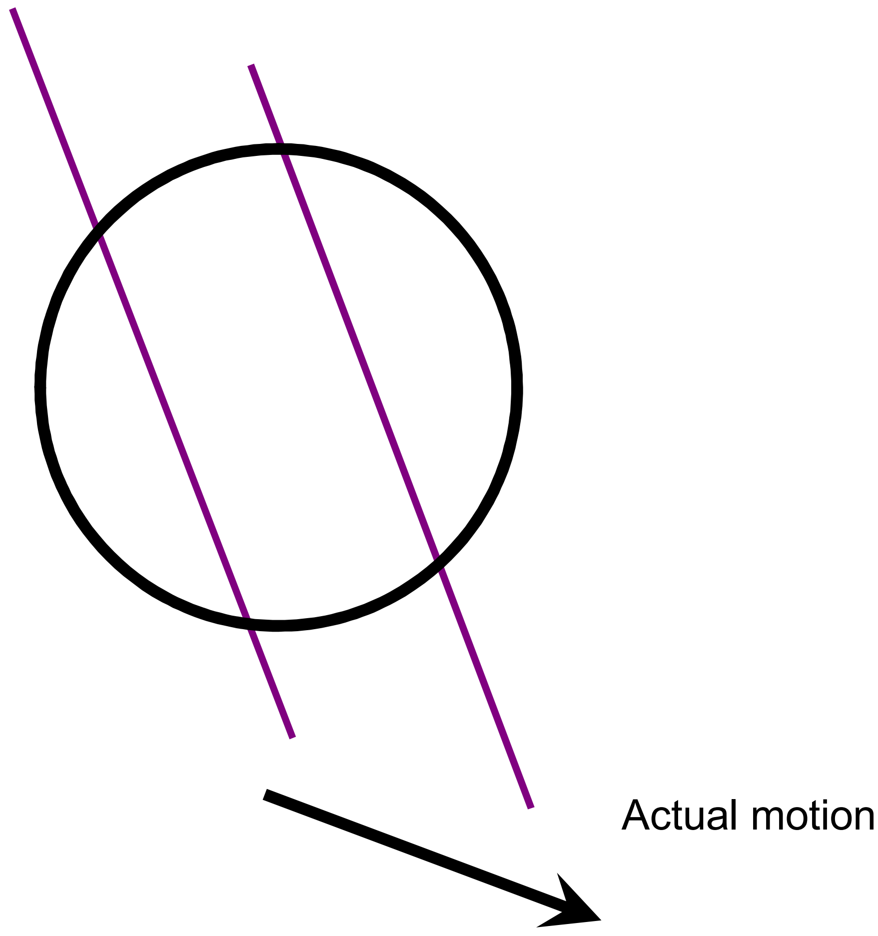
If (u, v) satisfies the equation,
so does $(u+u', v+v')$ if $\nabla I \cdot (u', v') = 0$



The aperture problem



The aperture problem



Solving the aperture problem

- How to get more equations for a pixel?
- **Spatial coherence constraint:** pretend the pixel's neighbors have the same (u,v)
  - E.g., if we use a 5x5 window, that gives us 25 equations per pixel

$$\begin{bmatrix} I_x(\mathbf{x}_1) & I_y(\mathbf{x}_1) \\ I_x(\mathbf{x}_2) & I_y(\mathbf{x}_2) \\ \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_y(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(\mathbf{x}_1) \\ I_t(\mathbf{x}_2) \\ \vdots \\ I_t(\mathbf{x}_n) \end{bmatrix}$$

B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*,1981.

Lucas-Kanade flow

- Linear least squares problem

$$\begin{bmatrix} I_x(\mathbf{x}_1) & I_y(\mathbf{x}_1) \\ I_x(\mathbf{x}_2) & I_y(\mathbf{x}_2) \\ \vdots & \vdots \\ I_x(\mathbf{x}_n) & I_y(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} I_t(\mathbf{x}_1) \\ I_t(\mathbf{x}_2) \\ \vdots \\ I_t(\mathbf{x}_n) \end{bmatrix}$$

$$\underset{n\times 2}{\mathbf{A}} \ \underset{2\times 1}{\mathbf{d}} = \underset{n\times 1}{\mathbf{b}}$$

Solution given by $(\mathbf{A}^T\mathbf{A})\mathbf{d} = \mathbf{A}^T\mathbf{b}$

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

The summations are over all pixels in the window

Lucas-Kanade flow

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

- Recall the Harris corner detector: M = $A^T A$ is the *second moment matrix*
- When is the system solvable?
  - By looking at the eigenvalues of the second moment matrix
  - The eigenvectors and eigenvalues of M relate to edge direction and magnitude
  - The eigenvector associated with the larger eigenvalue points in the direction of fastest intensity change, and the other eigenvector is orthogonal to it

Uniform region

– gradients have small magnitude
– small λ_1, small λ_2
– system is ill-conditioned

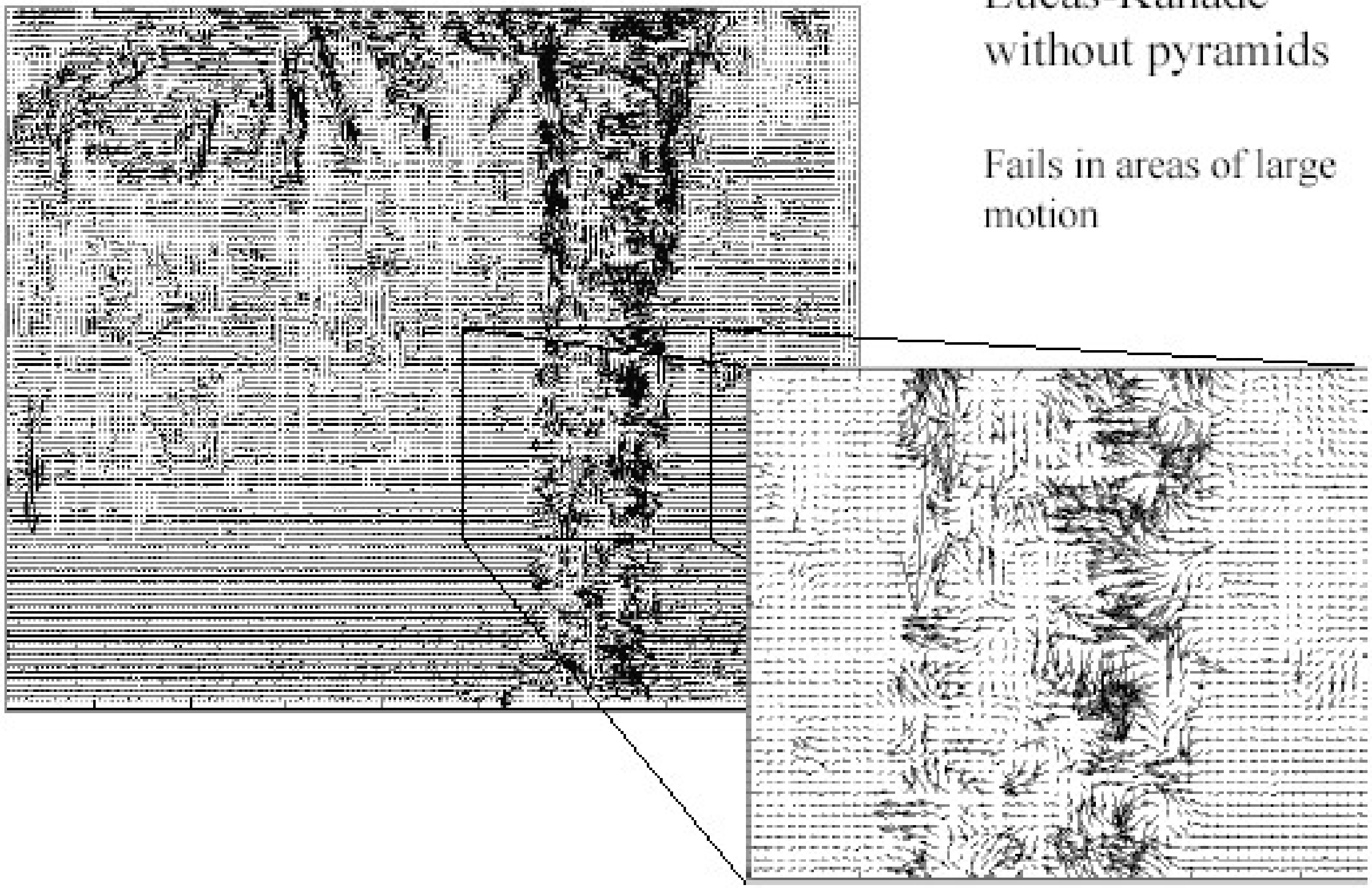
Edge

– gradients have one dominant direction
– large λ_1, small λ_2
– system is ill-conditioned

High-texture or corner region

– gradients have different directions, large magnitudes
– large λ_1, large λ_2
– system is well-conditioned

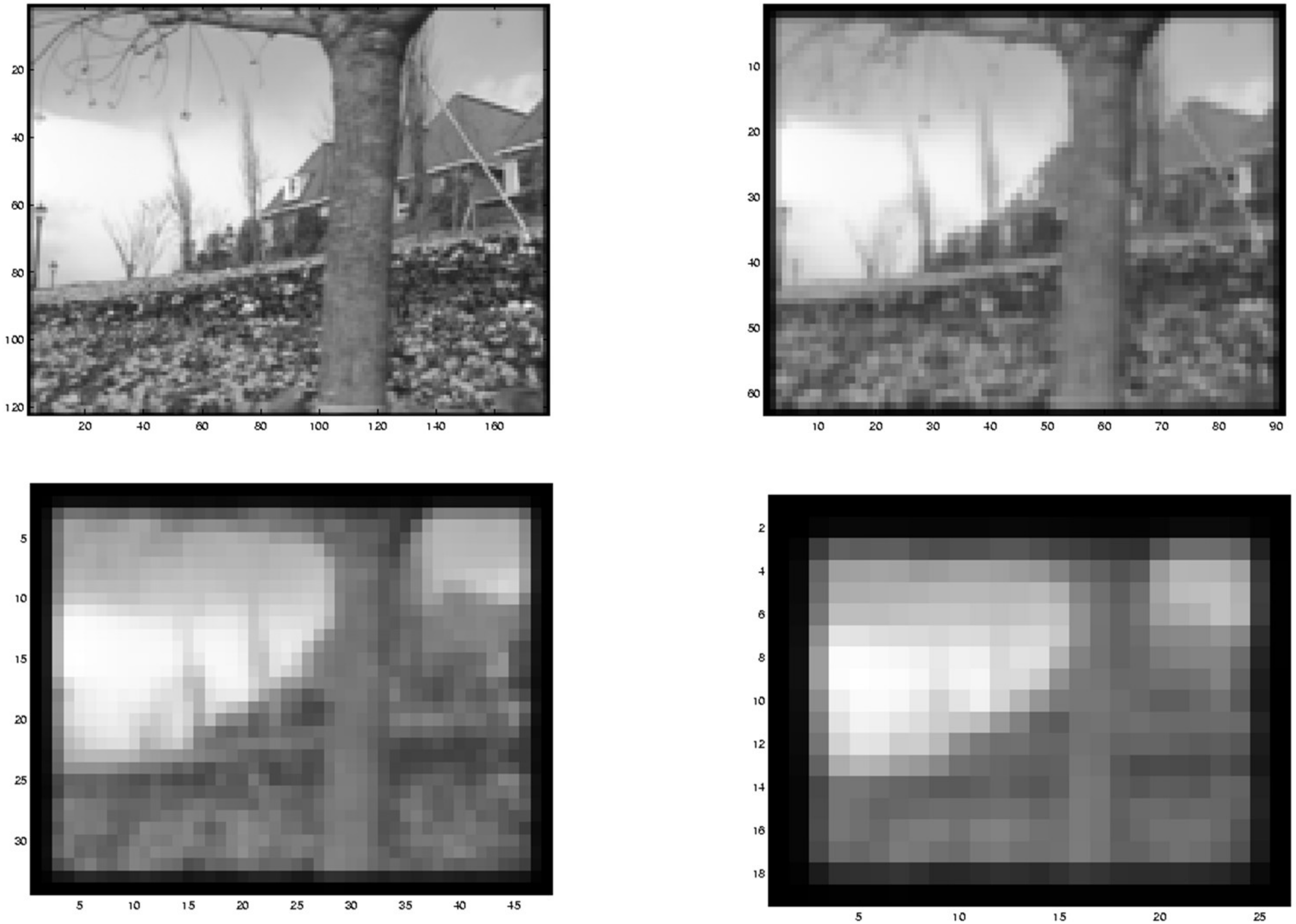
Optical Flow Results



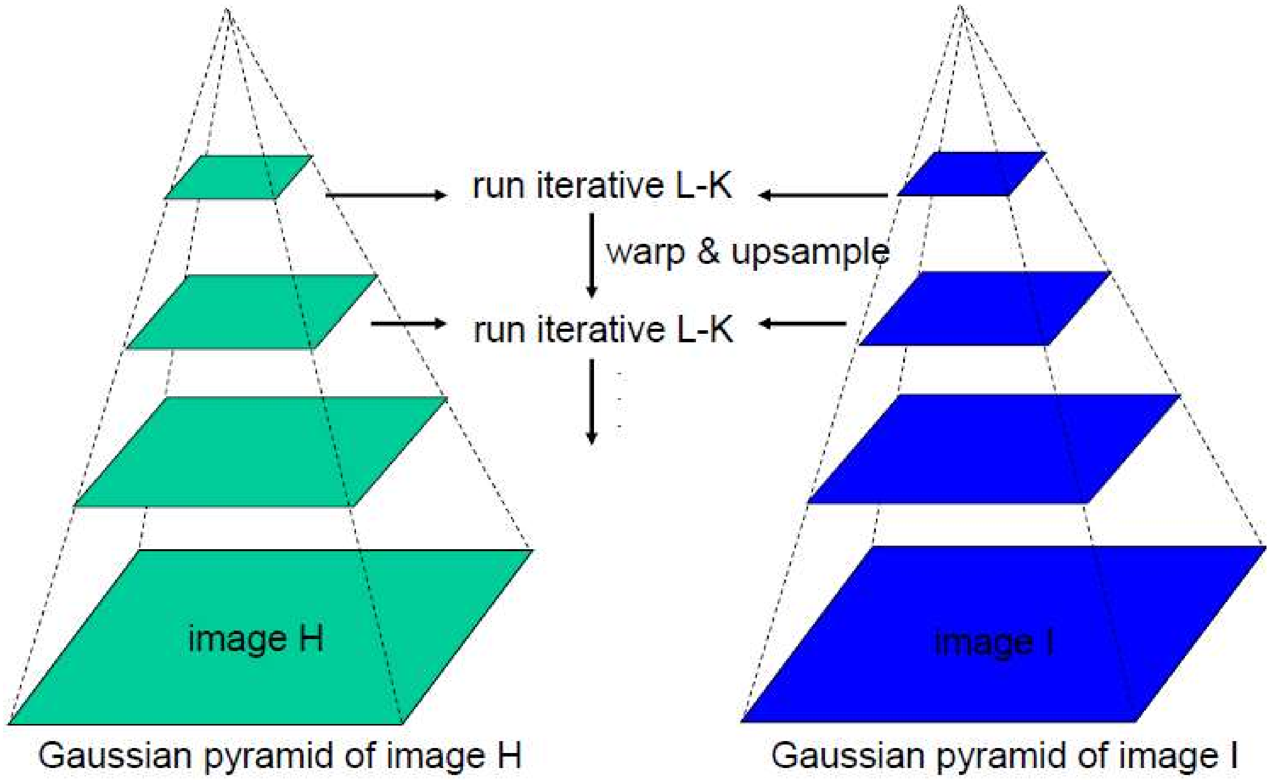
Lucas-Kanade
without pyramids

Fails in areas of large
motion

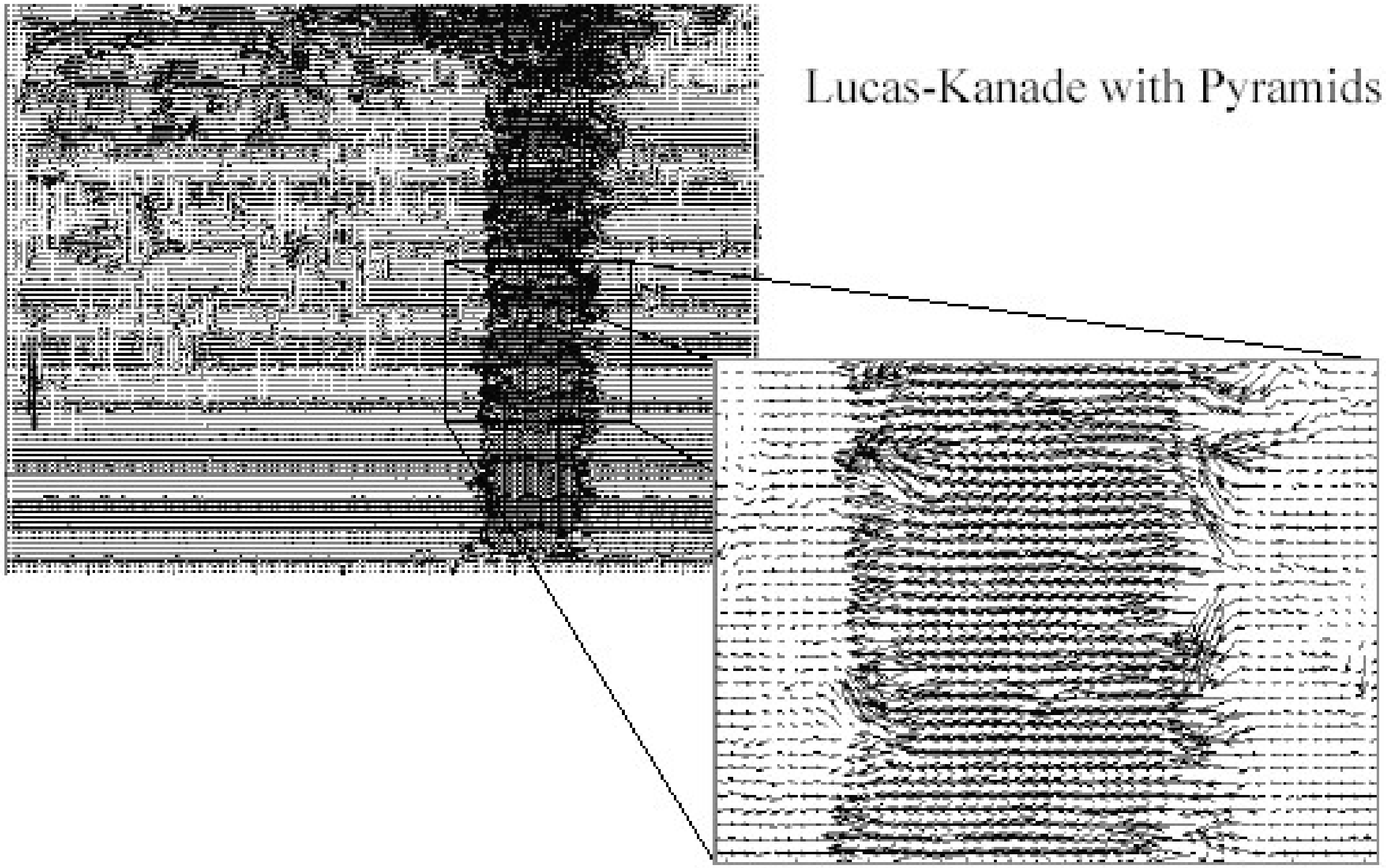
Multi-resolution registration



Coarse to fine optical flow estimation



Optical Flow Results



Lucas-Kanade with Pyramids

Horn & Schunck algorithm

Additional smoothness constraint :

- nearby point have similar optical flow
- additional constraint $||\nabla u||^2, \ ||\nabla v||^2$ small

$$e_s = \iint ((u_x^2 + u_y^2) + (v_x^2 + v_y^2))dxdy,$$

In addition to OF constraint equation term

$$e_c = \iint (I_x u + I_y v + I_t)^2 dxdy,$$

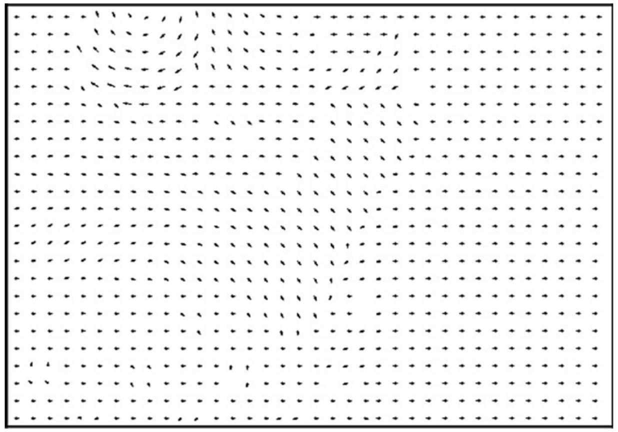
$\boxed{\text{minimize } e_s + \lambda e_c}$ λ regularization parameter

Coupled PDEs solved with iterative methods + finite differences

B.K.P. Horn and B.G. Schunck, "Determining optical flow." *Artificial Intelligence*,1981

Horn & Schunck

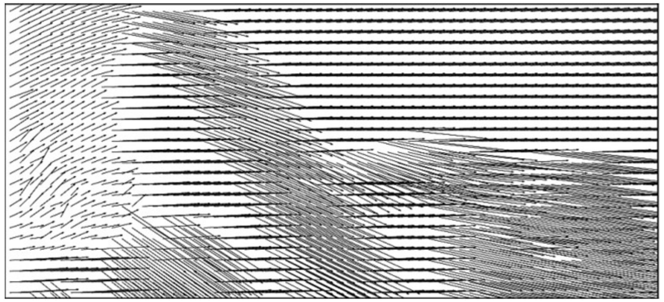
- Works well for small displacements
  - For example Middlebury sequence



Large displacement estimation in optical flow

Large displacement is difficult for optical flow estimation due to:

- locality and smoothness constraints



MPI Sintel dataset

Large displacement optical flow

- Classical optical flow [Horn and Schunck 1981]
  - energy: $E(\mathbf{w}) = \iint \underbrace{E_{data}}_{\text{color/gradient constancy}} + \alpha \underbrace{E_{smooth}}_{\text{smoothness constraint}} \mathbf{dx}$
  - minimization using a coarse-to-fine scheme

- Large displacement approaches:
  - LDOF [Brox and Malik 2011]
    a matching term, penalizing the difference between flow and HOG matches
    $$E(\mathbf{w}) = \iint E_{data} + \alpha E_{smooth} + \beta E_{match} \mathbf{dx}$$
  - MDP-Flow2 [Xu *et al.* 2012]
    expensive fusion of matches (SIFT + PatchMatch) and estimated flow at each level
  - DeepFlow [Weinzaepfel *et al.* 2013]
    deep matching + flow refinement with variational approach

Experimental results: datasets

- MPI-Sintel [Butler *et al.* 2012]
  - ► sequences from a realistic animated movie
  - ► large displacements (>20px for 17.5% of pixels)
  - ► atmospheric effects and motion blur

Experimental results: datasets

- KITTI [Geiger *et al.* 2013]
  - ► sequences captured from a driving platform
  - ► large displacements (>20px for 16% of pixels)
  - ► real-world: lightings, surfaces, materials

Experimental results: sample results

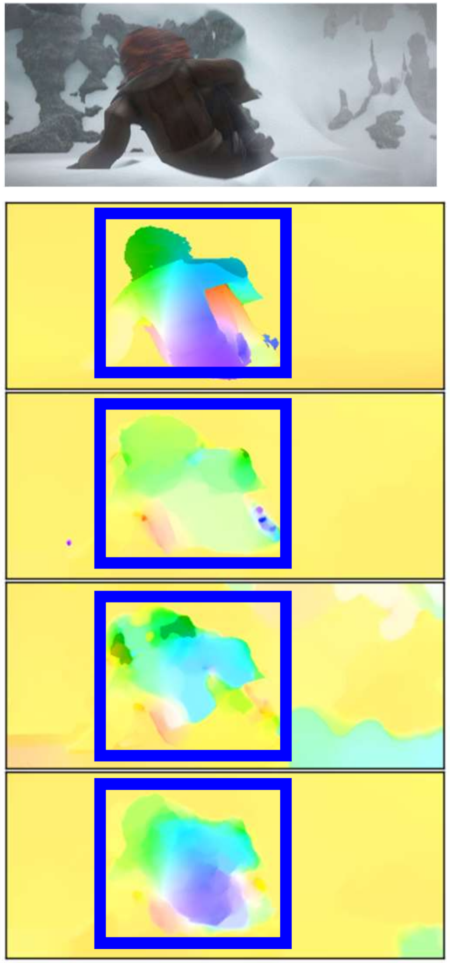
Ground-truth

LDOF [Brox & Malik 2011]

MDP-Flow2 [Xu *et al.* 2012]

DeepFlow [Weinzaepfel *et al.* 2013]

Experimental results: sample results



Ground-truth

LDOF [Brox & Malik 2011]

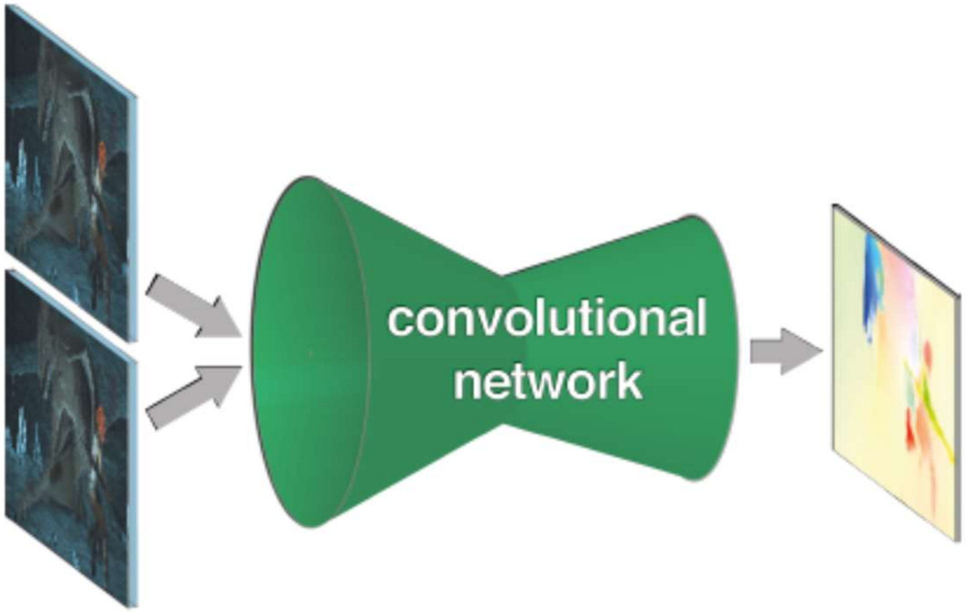
MDP-Flow2 [Xu *et al.* 2012]

DeepFlow [Weinzaepfel *et al.* 2013]

Methods – overview

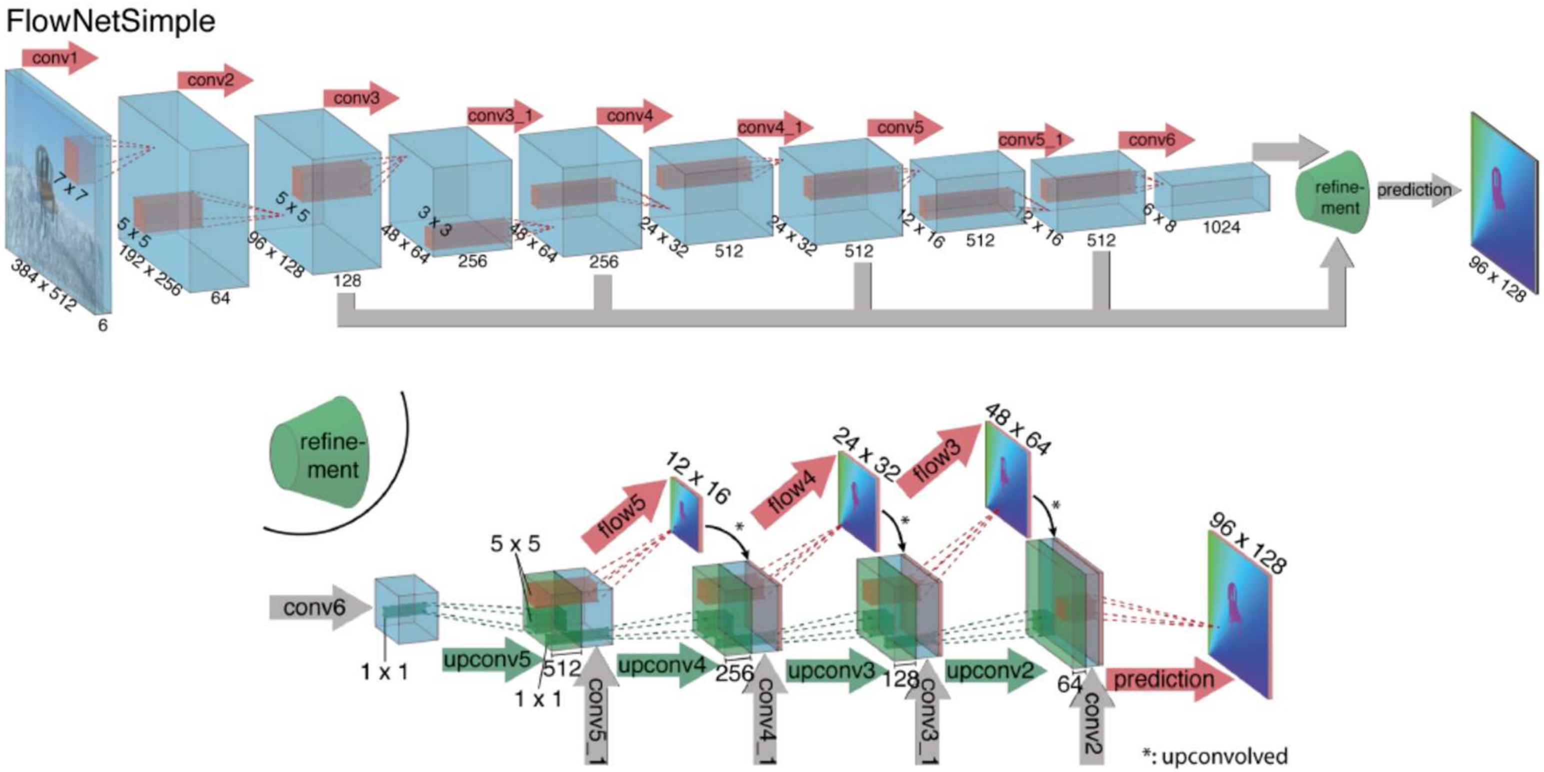
- Brightness constancy assumption
- + spatial coherence constraint: Lucas & Kanade, IJCAI’81
- + smoothness constraint: Horn & Schunk, AI’81
- + addition of matching term: Brox & Malik, PAMI’10
- recently: deep CNN based approaches

CNN to estimate optical flow: FlowNet

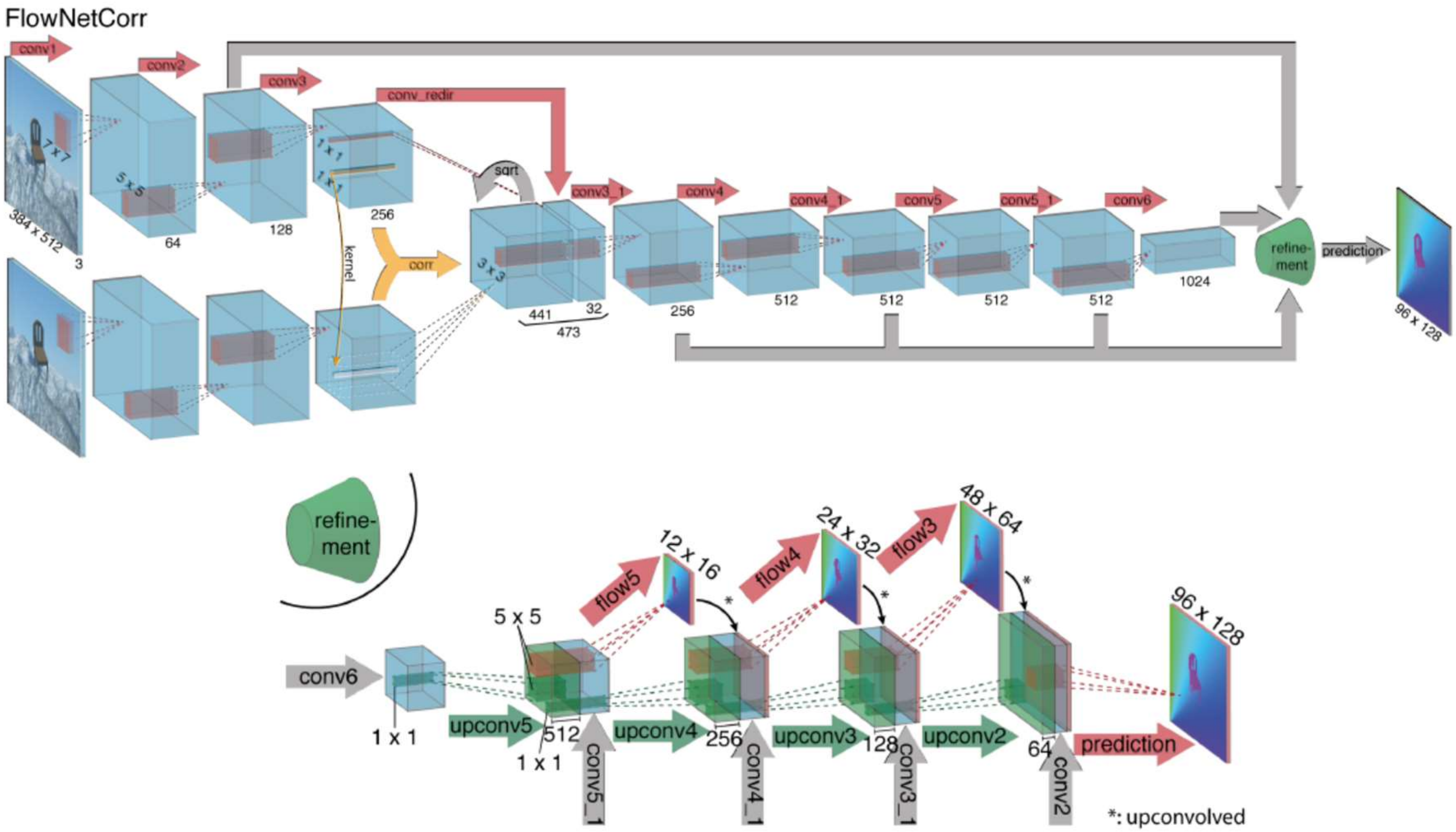


[A. Dosovitskiy et al. ICCV'15]

Architecture FlowNetSimple



Architecture FlowNetCorrelation



Synthetic dataset for training: Flying chairs

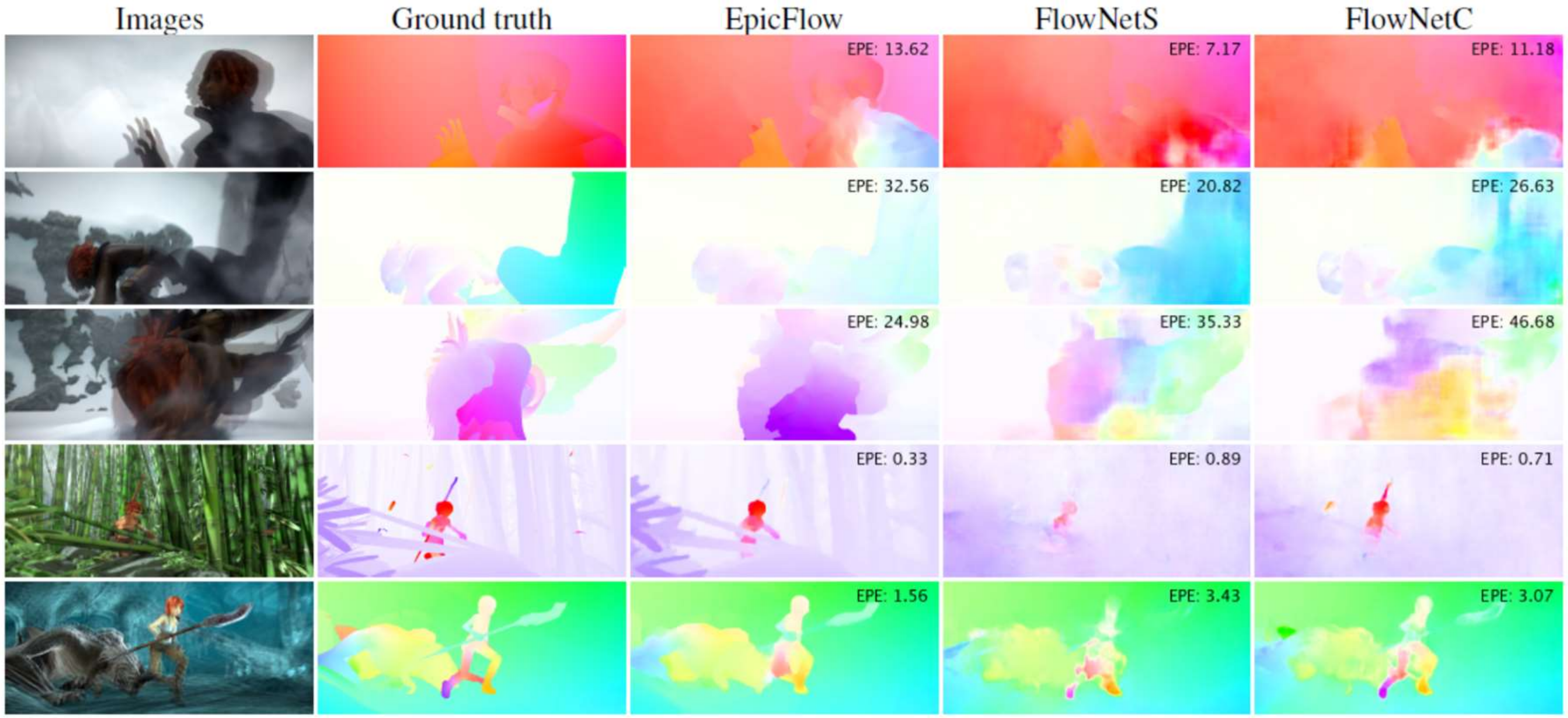
A dataset of approx. 23k image pairs

Experimental results

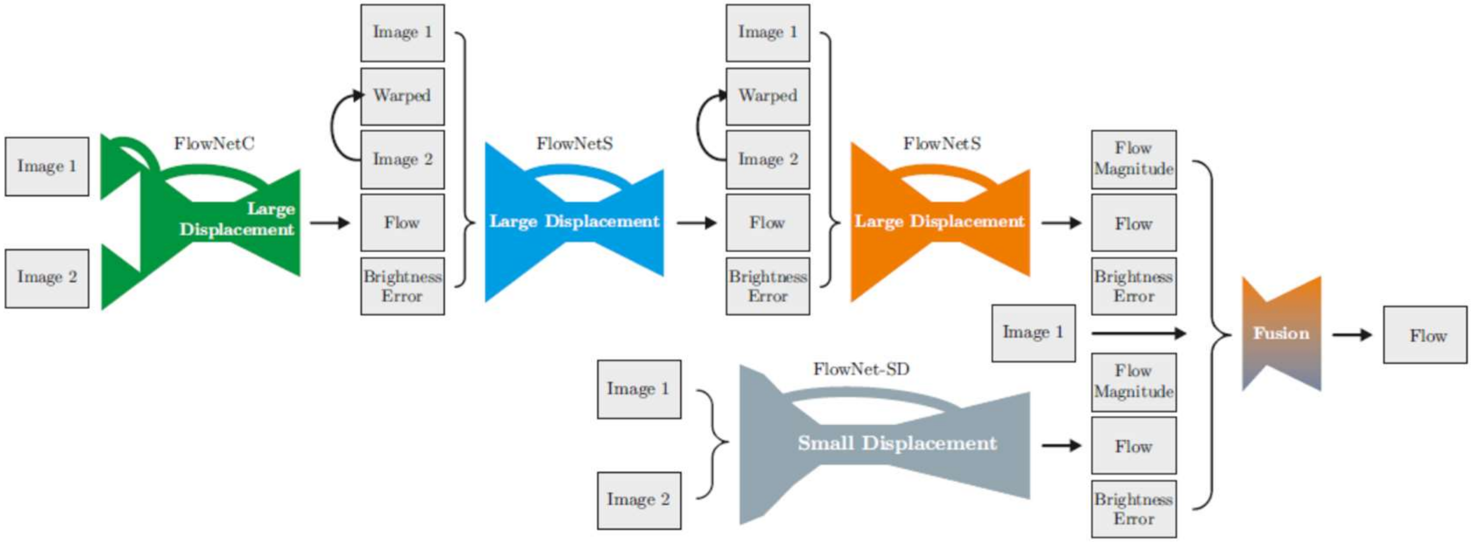
Method	Sintel Clean		Sintel Final	
	train	test	train	test
EpicFlow [30]	2.27	4.12	3.57	6.29
DeepFlow [35]	3.19	5.38	4.40	7.21
EPPM [3]	-	6.49	-	8.38
LDOF [6]	4.19	7.56	6.28	9.12
FlowNetS	4.50	7.42	5.45	8.43
FlowNetS+v	3.66	6.45	4.76	7.67
FlowNetS+ft	(3.66)	6.96	(4.44)	7.76
FlowNetS+ft+v	(2.97)	6.16	(4.07)	7.22
FlowNetC	4.31	7.28	5.87	8.81
FlowNetC+v	3.57	6.27	5.25	8.01
FlowNetC+ft	(3.78)	6.85	(5.28)	8.51
FlowNetC+ft+v	(3.20)	6.08	(4.83)	7.88

S: simple, C: correlation, v: variational refinement, ft:fine-tuning

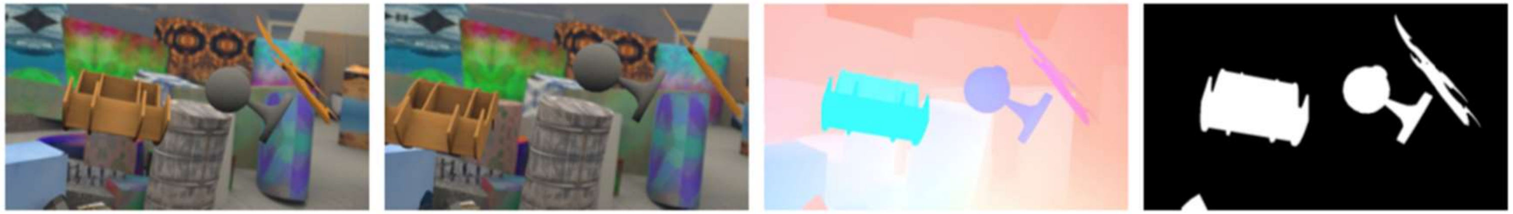
Experimental results



FlowNet2.0 [Ilg et al. CVPR'17]



FlyingThings3D [Mayer et al., CVPR'16]

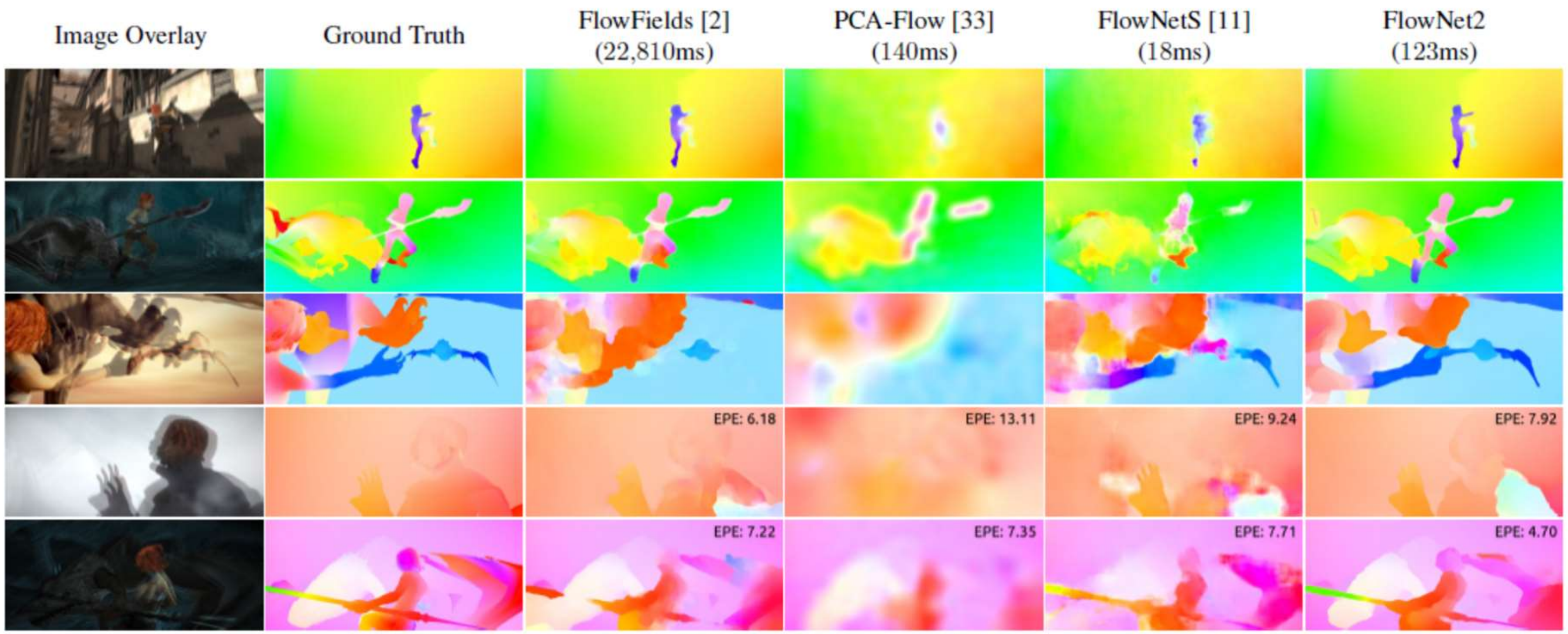


Stacking of networks

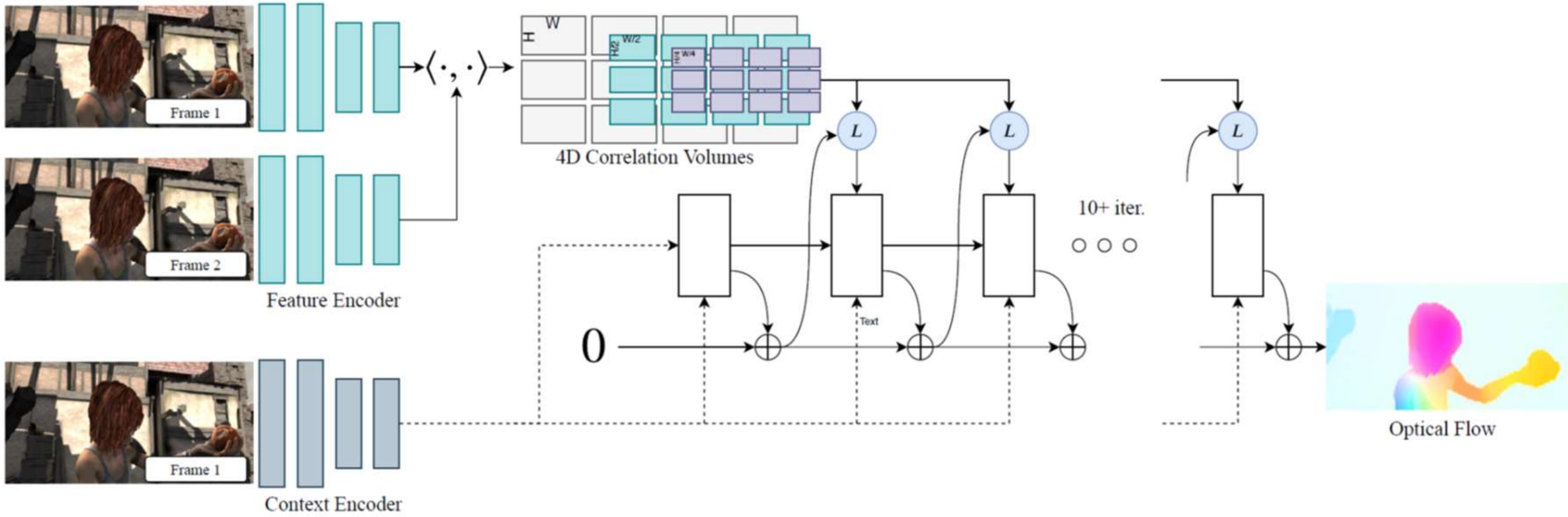
Stack architecture	Training enabled		Warping included	Warping gradient enabled	Loss after		EPE on Chairs test	EPE on Sintel train *clean*
	Net1	Net2			Net1	Net2		
Net1	✓	–	–	–	✓	–	3.01	3.79
Net1 + Net2	✗	✓	✗	–	–	✓	2.60	4.29
Net1 + Net2	✓	✓	✗	–	✗	✓	2.55	4.29
Net1 + Net2	✓	✓	✗	–	✓	✓	2.38	3.94
Net1 + W + Net2	✗	✓	✓	–	–	✓	1.94	**2.93**
Net1 + W + Net2	✓	✓	✓	✓	✗	✓	1.96	3.49
Net1 + W + Net2	✓	✓	✓	✓	✓	✓	**1.78**	3.33

Importance of warping

Optical flow results on Sintel



RAFT optical flow



- Feature extraction with CNNs
- Comparison between all features in the 2 images → 4D correlation volume
- Multi-scale representation of the 4D correlation volume
- Matching to the features of image 1
- Iterative updates which refine the current flow

[RAFT, Z. Teed and J. Deng, ECCV 2020]

RAFT optical flow – results

Fig. 3: Flow predictions on the Sintel test set.

Video object segmentation

- Segment the moving object in all the frames of a video

DAVIS (ground-truth)

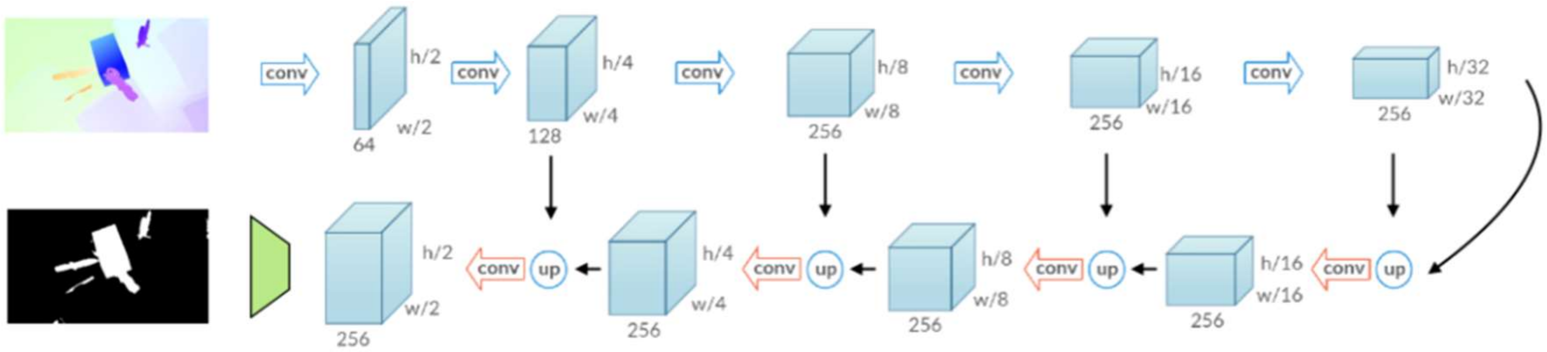
Challenges

- Strong camera or background motion

LDOF flow

DAVIS

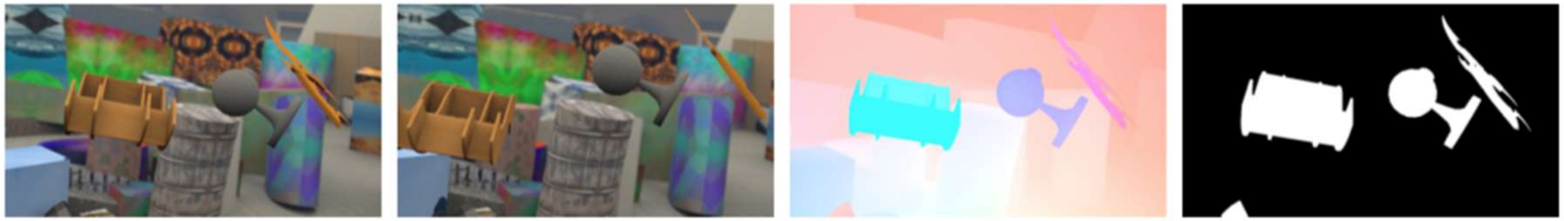
Network architecture – MP-Net



Convolutional/deconvolutional network, similar to U-Net

Training data

- FlyingThings3D dataset [Mayer et al., CVPR'16]
- 2700 synthetic, 10-frame stereo videos of random object flying in random trajectories (2250/450 training/test split)
- Ground-truth optical flow and camera data available
- Labels for moving object can be obtained from the data



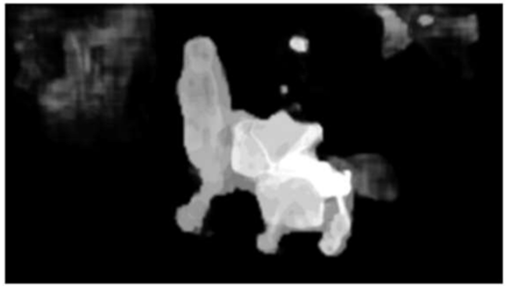
Results on FlyingThings3D test set

Motion estimation in real videos

- Flow estimation inaccuracies


DAVIS

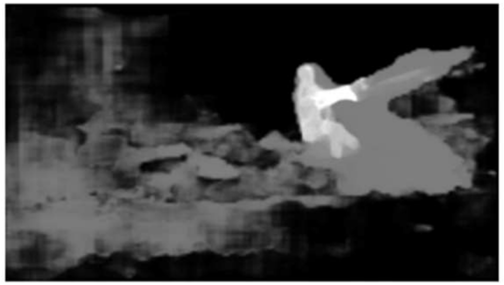

LDOF


MP-Net

- Background motion

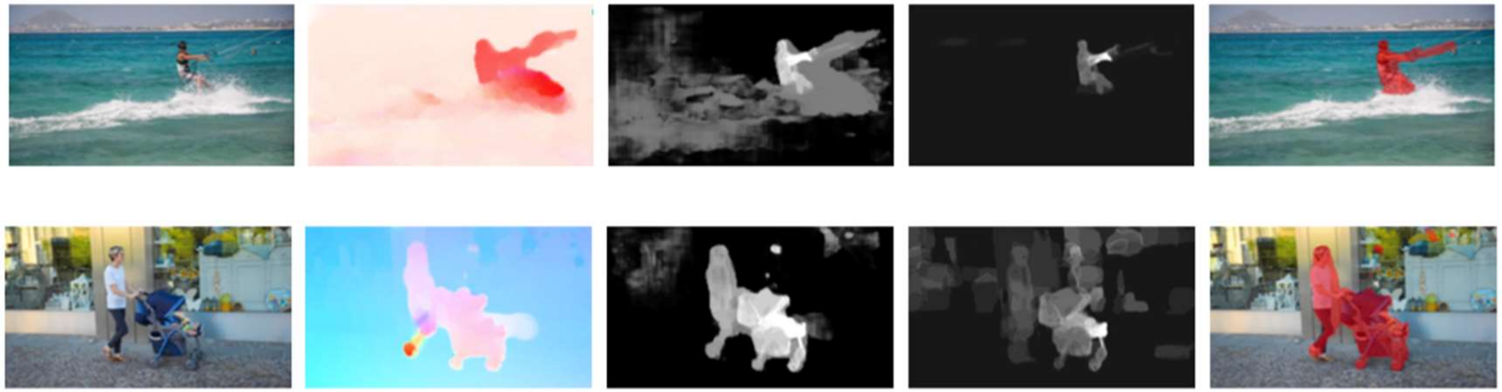

DAVIS


LDOF


MP-Net

Addition of an objectness measure

- Extract 100 object proposals per frame with SharpMask [Pinheiro et al., ECCV'16]
- Aggregate to obtain pixel-level objectness scores o_i
- Combine with the motion predictions m_i



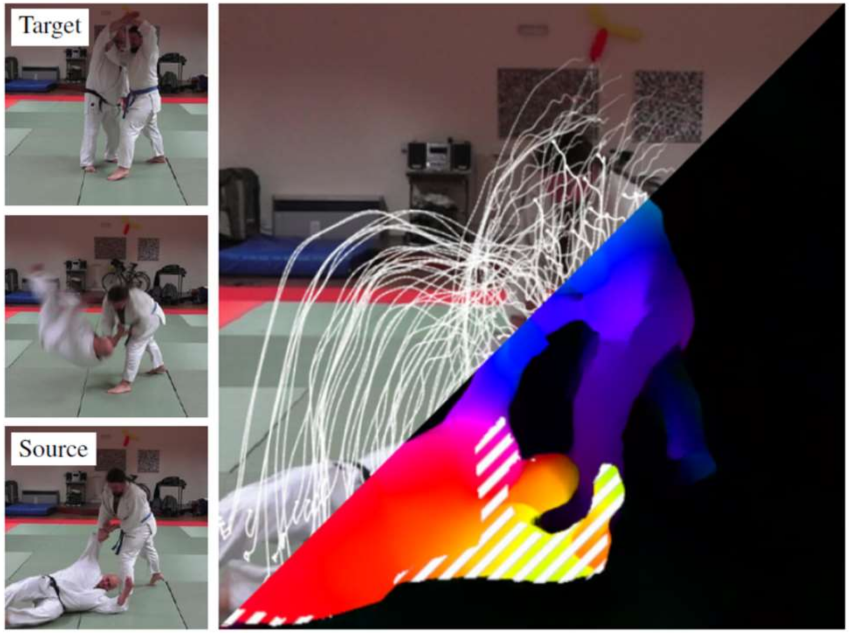
DAVIS LDOF MP-Net Objectness Result

FlowNet 2.0 Evaluation

Setting	LDOF flow	FLowNet 2.0 flow
MP-Net	52.4	62.6
MP-Net + Obj	63.3	69.0
MP-Net + Obj + CRF	69.7	72.5

Mean IoU on DAVIS trainval set

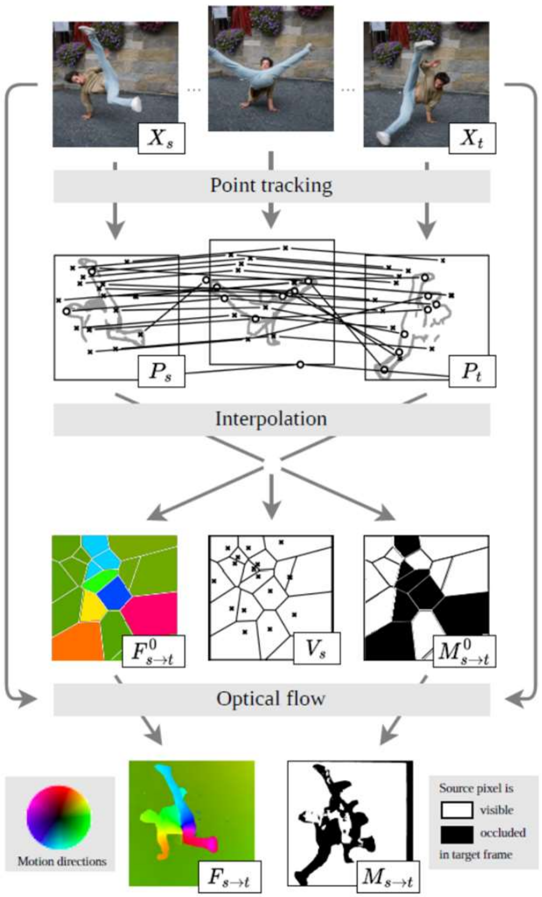
Dense point tracking



- Dense motion from source to target frames
- From a few point tracks (white)
  → dense flow (colors for directions, occlusion with stripes)

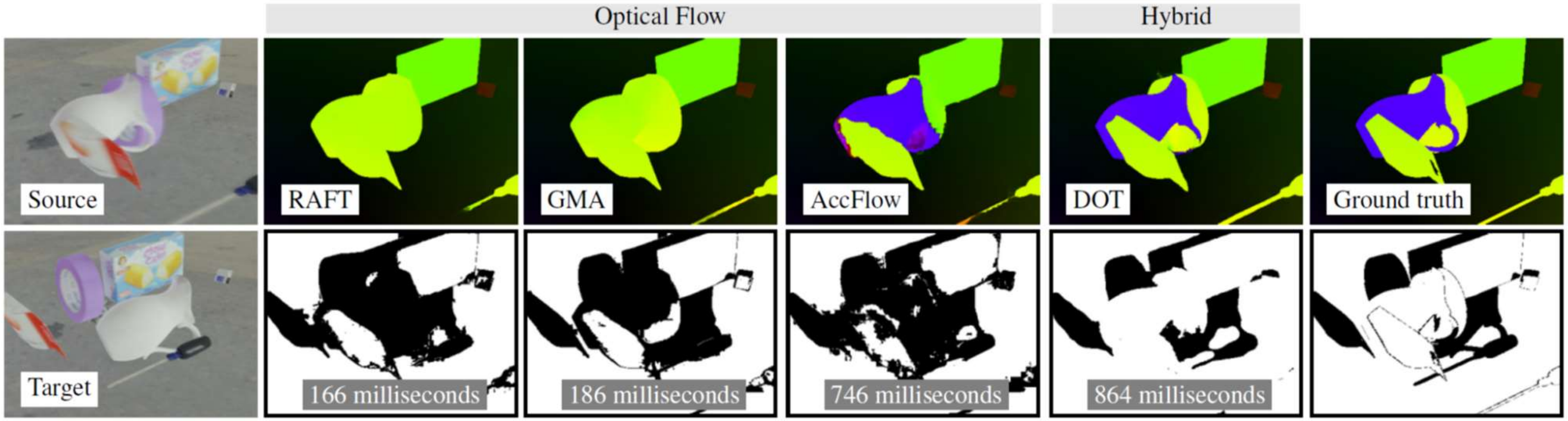
[Le Moing et al., Dense Optical Tracking: Connecting the Dots, arXiv'23]

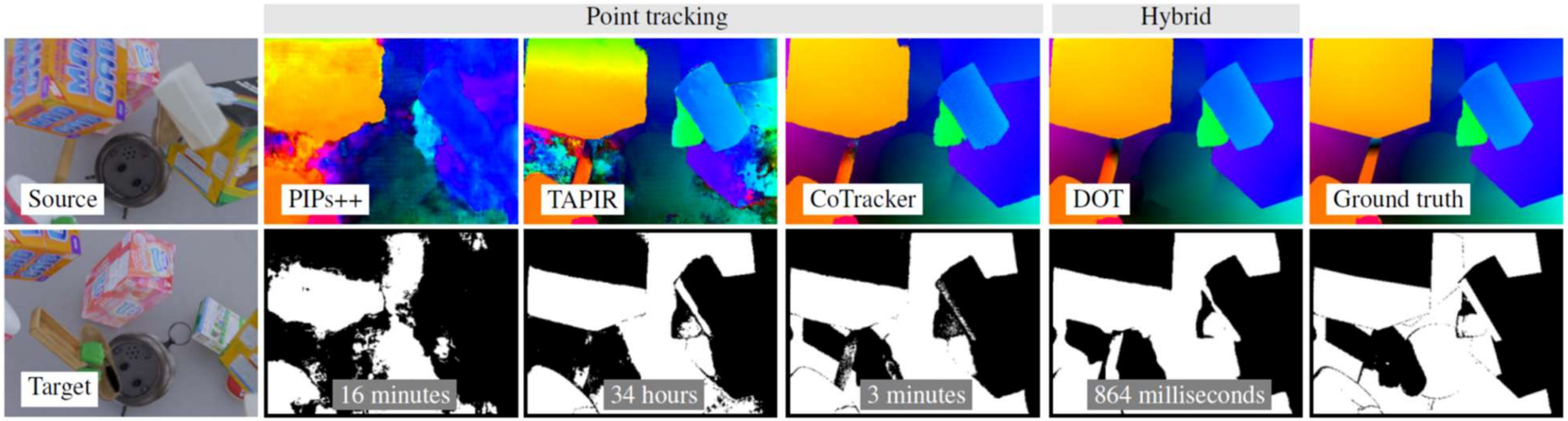
Dense point tracking



- Sparse point tracks (TAPIR, Co-Tracker)
- Near neighbor point interpolation
- Optical flow estimation to refine local neighborhood (RAFT)

Dense point tracking – results





Dense point tracking – results

	Method	N	CVO (*Clean*)		CVO (*Final*)			CVO (*Extended*)		
			EPE ↓ (all / vis / occ)	IoU ↑	EPE ↓ (all / vis / occ)	IoU ↑	Time* ↓	EPE ↓ (all / vis / occ)	IoU ↑	Time ↓
Optical flow	RAFT [57]	-	2.82 / 1.70 / 8.01	58.1	2.88 / 1.79 / 7.89	57.2	**0.166**	28.6 / 21.6 / 41.0	61.7	**0.166**
	GMA [28]	-	2.90 / 1.91 / 7.63	60.9	2.92 / 1.89 / 7.48	60.1	0.186	30.0 / 22.8 / 42.6	61.5	0.186
	RAFT (🔥) [57]	-	2.48 / 1.40 / 7.42	57.6	2.63 / 1.57 / 7.50	56.7	0.634	21.8 / 15.4 / 33.4	65.0	4.142
	GMA (🔥) [28]	-	2.42 / 1.38 / 7.14	60.5	2.57 / 1.52 / 7.22	59.7	0.708	21.8 / 15.7 / 32.8	65.6	4.796
	MFT [47]	-	2.91 / 1.39 / 9.93	19.4	3.16 / 1.56 / 10.3	19.5	1.350	21.4 / 9.20 / 41.8	37.6	18.69
	AccFlow [61]	-	1.69 / 1.08 / 4.70	48.1	1.73 / 1.15 / 4.63	47.5	0.746	36.7 / 28.1 / 52.9	36.5	5.598
Point tracking	PIPs++ [68]	262144	9.05 / 6.62 / 21.5	33.3	9.49 / 7.06 / 22.0	32.7	974.3	18.4 / 10.0 / 32.1	58.7	1922.
	TAPIR† [17]	262144	3.55 / 1.34 / 15.2	74.0	4.36 / 2.04 / 16.1	72.5	$\sim 10^5$	- / - / -	-	$\sim 10^6$
	CoTracker [30]	262144	1.51 / 0.88 / 4.57	75.5	1.52 / 0.93 / 4.38	75.3	191.5	5.20 / 3.84 / 7.70	70.4	1737.
Hybrid	*Dense optical tracking* (DOT)	1024	1.36 / 0.76 / 4.26	80.0	1.43 / 0.85 / 4.29	79.7	0.864	5.28 / 3.78 / 7.71	70.8	5.234
		2048	1.32 / 0.74 / 4.12	80.4	1.38 / 0.82 / 4.10	80.2	1.652	5.07 / 3.67 / 7.34	71.0	9.860
		4096	**1.29 / 0.72 / 4.03**	**80.4**	**1.34 / 0.80 / 3.99**	**80.4**	3.152	**4.98 / 3.59 / 7.17**	**71.1**	19.73

"†": evaluation is only performed on a random subset of 2% of the test videos due to extremely slow inference speed. "∗": the time is the same for *Clean* and *Final* sets.

Overview

- Optical flow
- *Video classification*
- Action localization

Action recognition - tasks

- Action classification: assigning an action label to a video clip

→

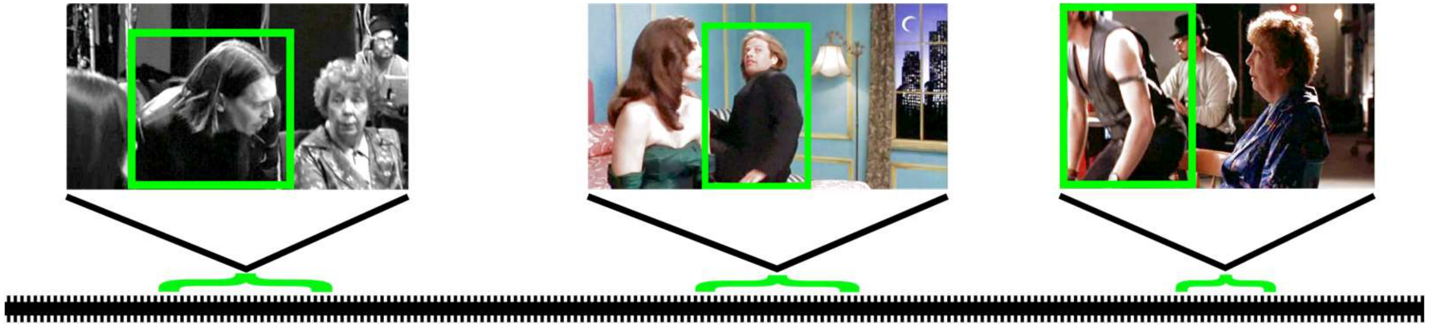
Making sandwich: present
Feeding animal: not present
...

Action recognition - tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
...

- Action localization: search locations of an action in a video



Action classification in videos

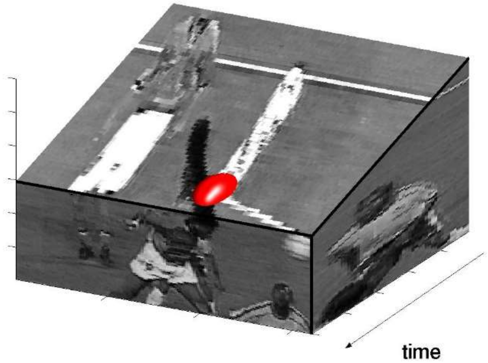
- Space-time interest points
- Dense trajectories
- Video-level CNN features
- Transformer-based approaches

Space-time interest points (STIP) [Laptev'05]

- Space-time corner detector
  [Laptev, IJCV 2005]

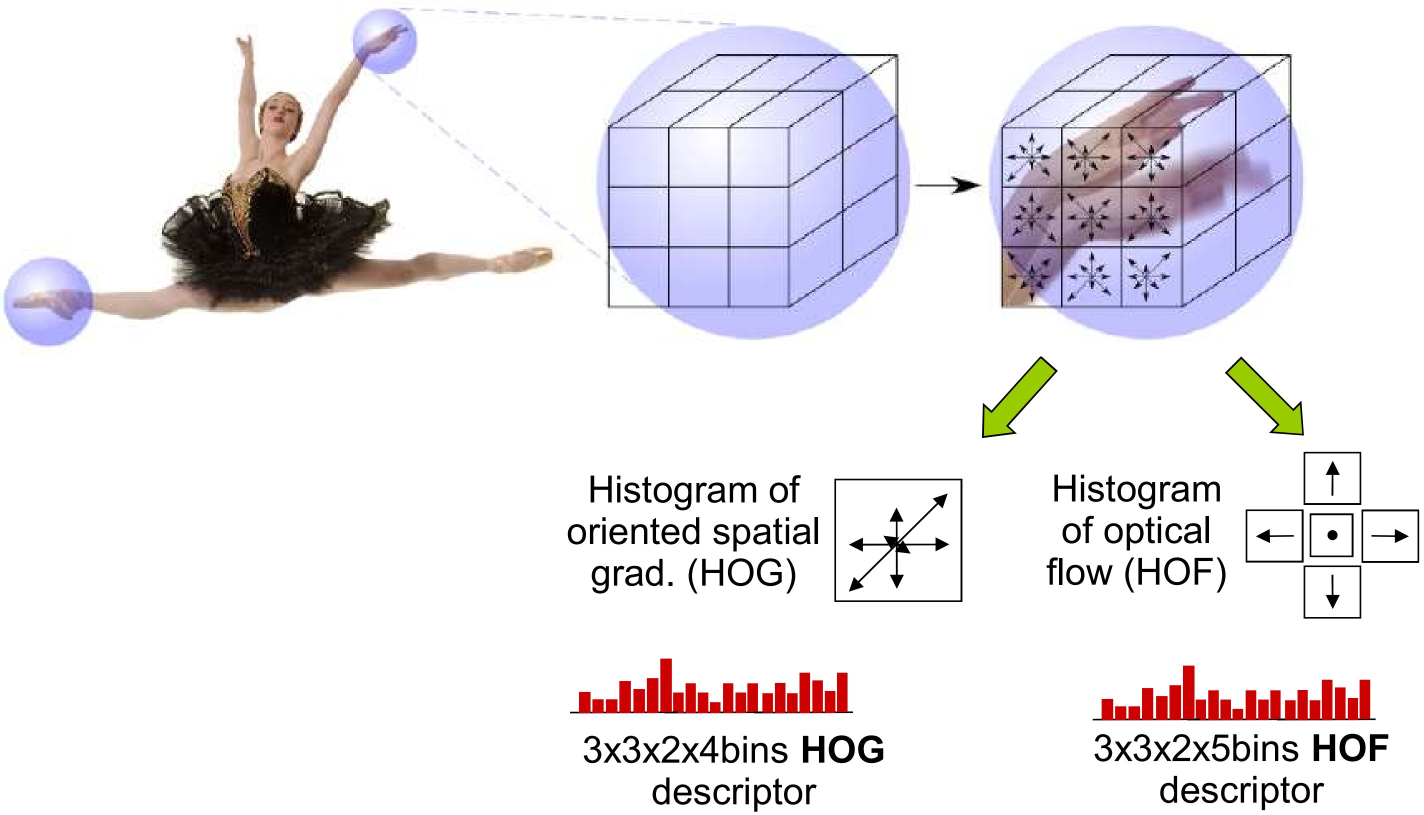
$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



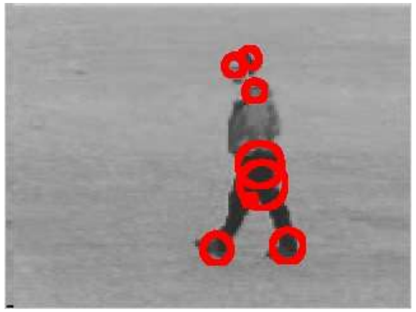
STIP descriptors

Space-time interest points

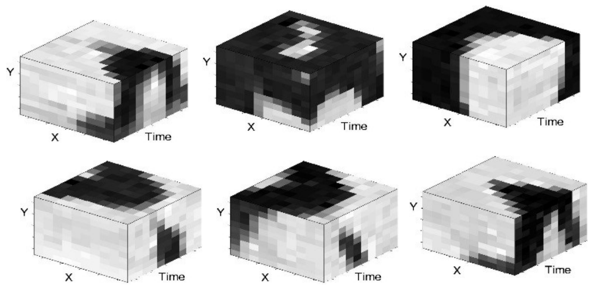


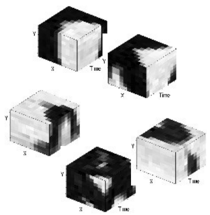
Action classification

- Bag of space-time features + support vector machine (SVM)
  [Schuldt'04, Niebles'06, Zhang'07]



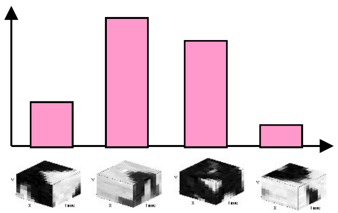
Collection of space-time patches





HOG & HOF patch descriptors

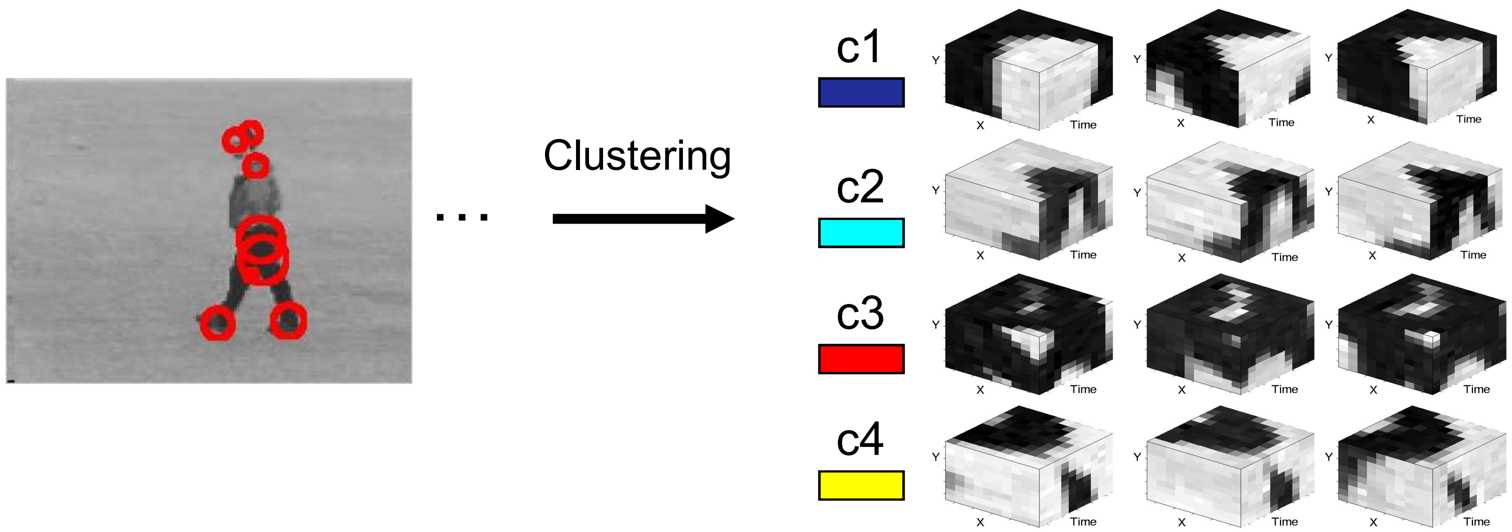
Histogram of visual words



SVM Classifier

Visual words: k-means clustering

- Group similar STIP descriptors together with k-means

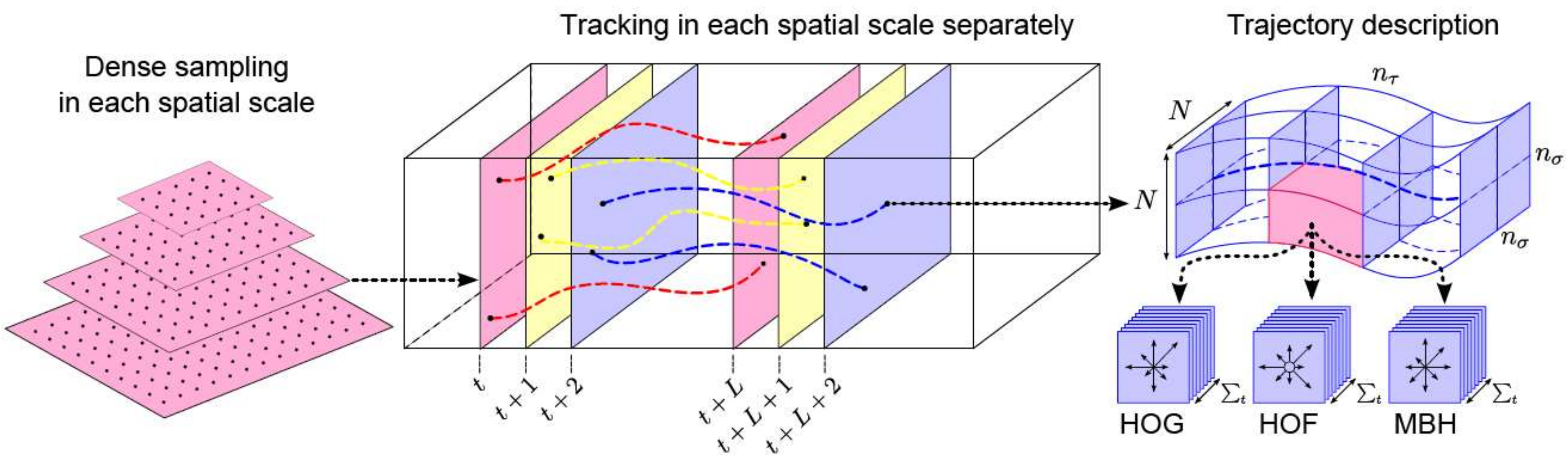


Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”,
“Indiana Jones and the Last Crusade”

Dense trajectories [Wang et al., IJCV'13]

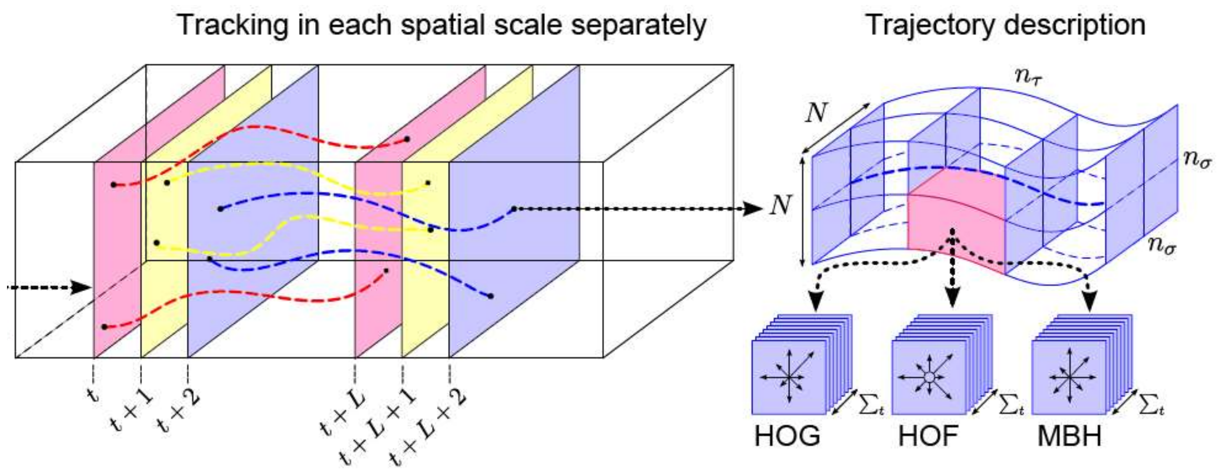
- Dense trajectories [Wang et al., IJCV'13] and Fisher vector encoding [Perronnin et al. ECCV'10]
  - Dense sampling at several scales
  - Feature tracking based on optical flow for several scales
  - Length 15 frames, to avoid drift



Example for dense trajectories

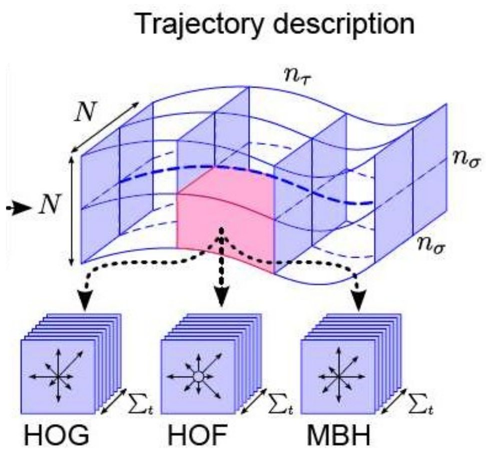
Descriptors for dense trajectory

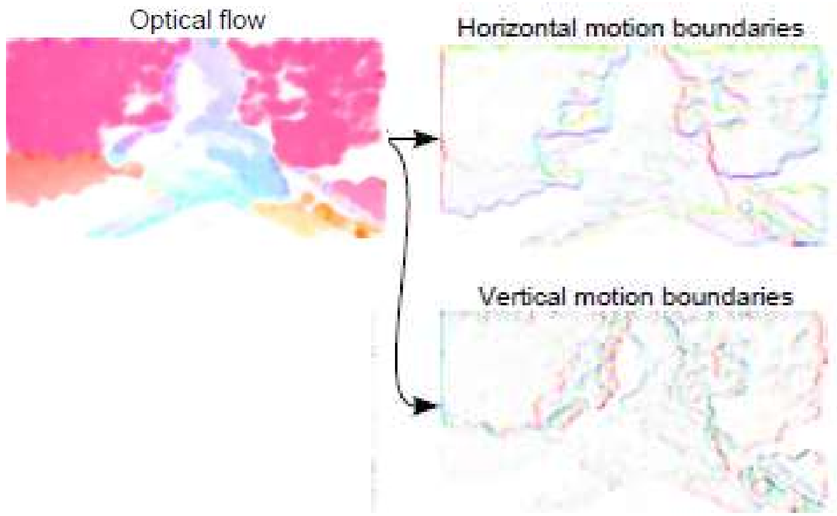
- Histogram of gradients (HOG: 2x2x3x8)
- Histogram of optical flow (HOF: 2x2x3x9)
- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)



Descriptors for dense trajectory

- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)
  - spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram
  - captures relative dynamics of different regions
  - suppresses constant motions





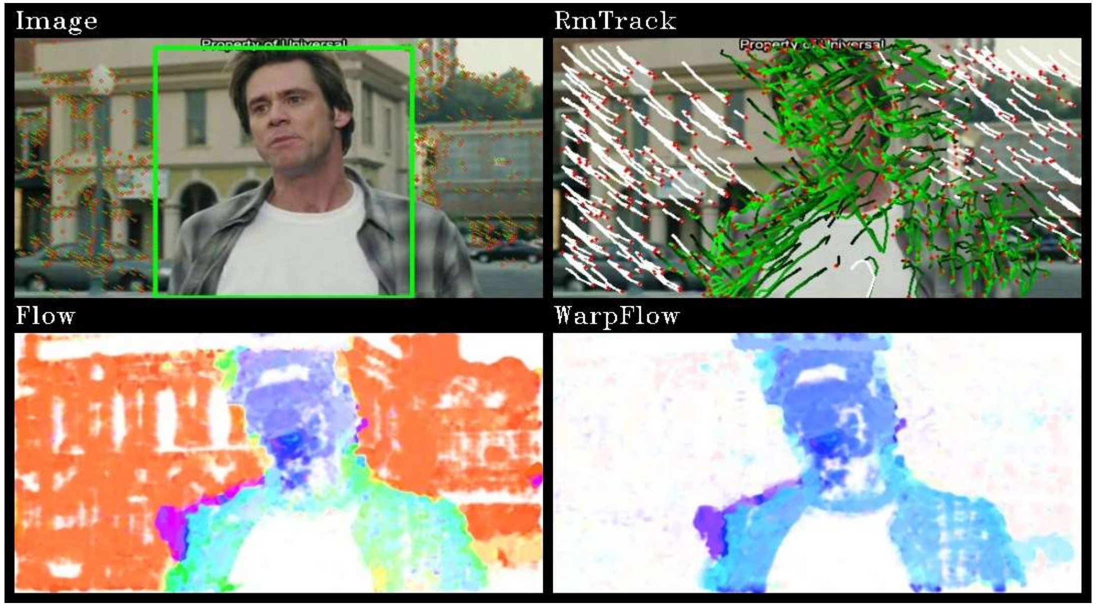
Dense trajectories

- Advantages:
  - Captures the intrinsic dynamic structures in videos
  - MBH is robust to certain camera motion

- Disadvantages:
  - Generates irrelevant trajectories in background due to camera motion
  - Motion descriptors are modified by camera motion, e.g., HOF, MBH

Improved dense trajectories

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion



[Wang and Schmid. Action recognition with improved trajectories. ICCV'13]

Camera motion estimation

- Find the correspondences between two consecutive frames:
  - Extract and match SURF features (robust to motion blur)
  - Use optical flow, remove uninformative points
- Combine SURF (green) and optical flow (red) results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches

Inlier matches of the homography

Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion, thus generates outlier matches

- Apply a human detector in each frame, and track the human bounding box forward and backward to join detections

- Remove feature matches inside the human bounding box during homography estimation

Inlier matches and warped flow, without or with HD

Remove background trajectories

- Remove trajectories by thresholding the maximal magnitude of stabilized motion vectors

- Our method works well under various camera motions, such as pan, zoom, tilt

Successful examples   Failure cases

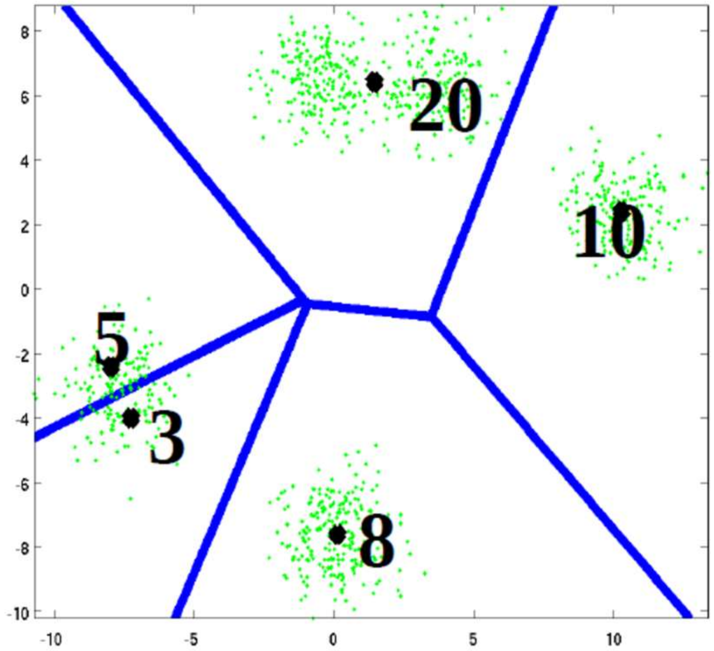
Removed trajectories (white) and foreground ones (green)

- Failure due to severe motion blur; the homography is not correctly estimated due to unreliable feature matches

Fisher Vector [Sanchez et al, 2013]

- Bag of features: stores the number of features assigned to each cluster center

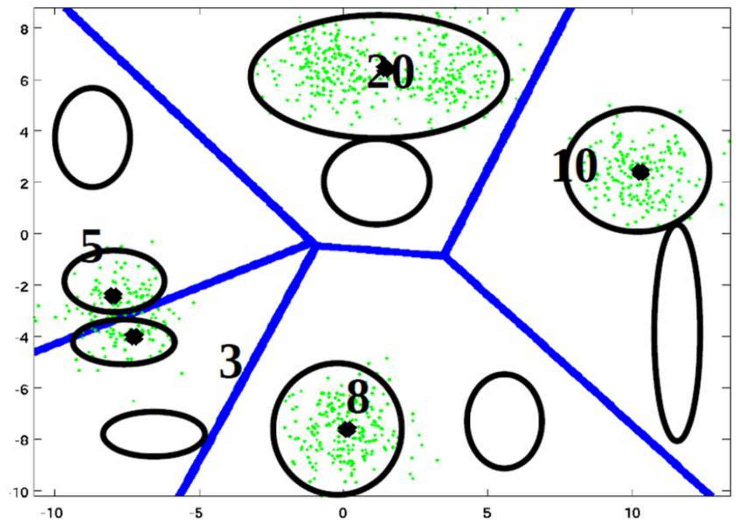
- Drawbacks:
  – Needs more words to refine the representation
  – This directly increases the computational cost
  – Also leads to many empty bins: redundancy



Fisher Vector [Sanchez et al, 2013]

- Fisher vector: also stores mean and variance of the features per cluster

- Even when the counts are the same,
the position can vary



- Advantages:
  - More information for the same visual word
  - Does not increase compute significantly
  - Leads for high dimensional features vectors

Evaluation datasets

Hollywood dataset [Marszalek et al.'09]


answer phone


get out of car


fight person

Hollywood2: 12 classes from 69 movies, report mAP

Evaluation datasets

HMDB 51 dataset [Kuehne et al.'11]

push-up

cartwheel

sword-exercice

HMDB51: 51 classes, report accuracy on three splits

Evaluation datasets

UCF 101 dataset [Soomro et al.’12]

haircut archery ice-dancing

UCF101: 101 classes, report accuracy on three splits

Evaluation of the intermediate steps

	HOG	HOF	MBH	HOF+MBH	Combined
DTF	38.4%	39.5%	49.1%	49.8%	52.2%
ITF	40.2%	48.9%	52.1%	54.7%	57.2%

Results on HMDB51 using Fisher vector

- Baseline: DTF = "dense trajectory feature"
- ITF = "improved trajectory feature”

- HOF improves significantly and MBH somewhat
- Almost no impact on HOG

- HOF and MBH are complementary, as they represent zero and first order motion information

Impact of feature encoding on improved trajectories

Datasets	Fisher vector		
	DTF	ITF wo human	ITF w human
Hollywood2	63.6%	66.1%	66.8%
HMDB51	55.9%	59.3%	60.1%
UCF101	83.5%	85.7%	86.0%

Compare DTF and ITF with and without human detection using HOG+HOF+MBH and Fisher encoding

- IDT significantly improvement over DT
- Human detection always helps. For Hollywood2 and HMDB51, the difference is more significant, as there are more humans present.

TrecVid MED 2011

- 15 categories

Attempt a board trick

Feed an animal

Landing a fish

…

Wedding ceremony

Working on a
wood project

Birthday party

TrecVid MED 2011

- 15 categories
- ~100 positive video clips per event category, 9600 negative video clips
- Testing on 32000 videos clips, i.e., 1000 hours
- Videos come from publicly available, user-generated content on various Internet sites

- Descriptors: MBH, SIFT, audio, text & speech recognition

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP
Motion	44.65
Static	33.97
Audio	18.15
OCR	10.85
ASR	8.21
Visual=Motion+Static	47.22
Visual+Audio	50.41
Visual+OCR	48.97
Visual+ASR	48.28
Visual+Audio+OCR+ASR	52.28

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party
Motion	44.65	30.7
Static	33.97	25.9
Audio	18.15	33.3
OCR	10.85	10.1
ASR	8.21	3.6
Visual=Motion+Static	47.22	34.8
Visual+Audio	50.41	47.7
Visual+OCR	48.97	35.8
Visual+ASR	48.28	35.0
Visual+Audio+OCR+ASR	52.28	48.4

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party	Repair appliance
Motion	44.65	30.7	42.6
Static	33.97	25.9	43.6
Audio	18.15	33.3	43.3
OCR	10.85	10.1	32.1
ASR	8.21	3.6	39.2
Visual=Motion+Static	47.22	34.8	47.5
Visual+Audio	50.41	47.7	54.5
Visual+OCR	48.97	35.8	50.8
Visual+ASR	48.28	35.0	54.5
Visual+Audio+OCR+ASR	52.28	48.4	57.2

Quantitative results on TrecVid MED'11

Performance of all channels (mAP)

Channel	mAP	Birthday party	Repair appliance	Make sandwich
Motion	44.65	30.7	42.6	22.5
Static	33.97	25.9	43.6	21.5
Audio	18.15	33.3	43.3	11.2
OCR	10.85	10.1	32.1	19.4
ASR	8.21	3.6	39.2	6.7
Visual=Motion+Static	47.22	34.8	47.5	27.8
Visual+Audio	50.41	47.7	54.5	27.3
Visual+OCR	48.97	35.8	50.8	35.7
Visual+ASR	48.28	35.0	54.5	28.8
Visual+Audio+OCR+ASR	52.28	48.4	57.2	35.4

Experimental results

- Example results

rank 1

rank 2

rank 3

Highest ranked results for the event «horse riding competition»

Experimental results

- Example results

rank 1

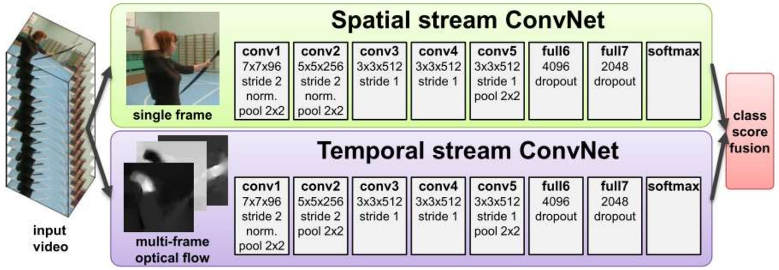
rank 2

rank 3

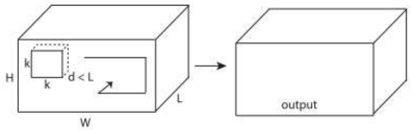
Highest ranked results for the event «tuning a musical instrument»

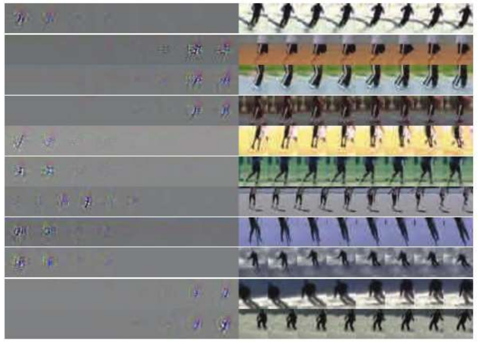
CNN based methods

Two-Stream Convolutional Networks
for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]

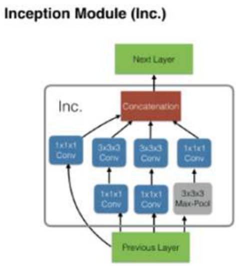


Learning Spatiotemporal Features with
3D Convolutional Networks
[Tran et al. ICCV15]



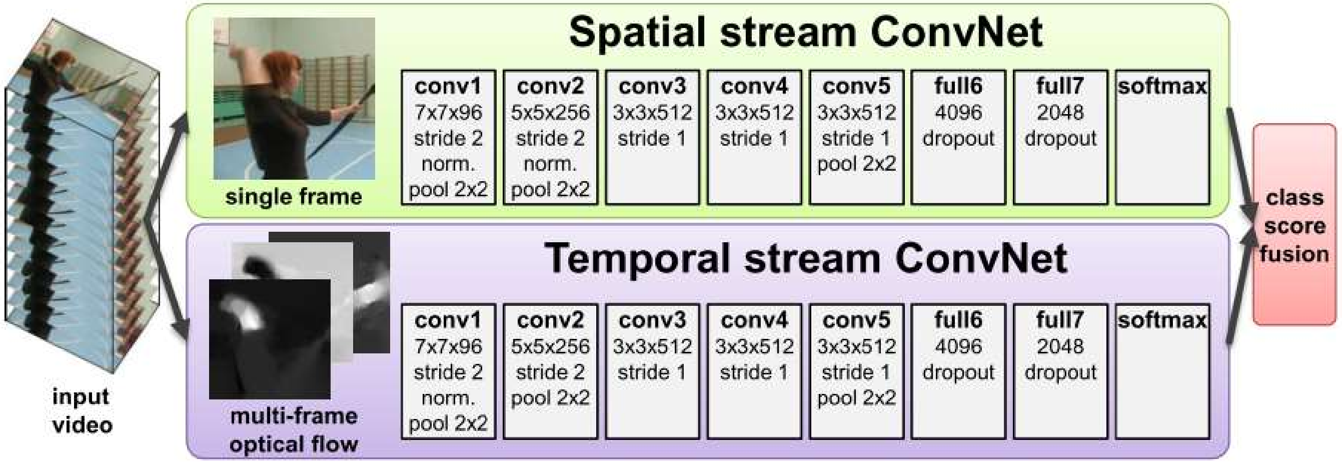


Quo vadis action recognition? A new
model and the Kinetics dataset
[Carreira et al. CVPR17]



Recent CNN methods

Two-Stream Convolutional Networks
for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]



CNN based methods

Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]

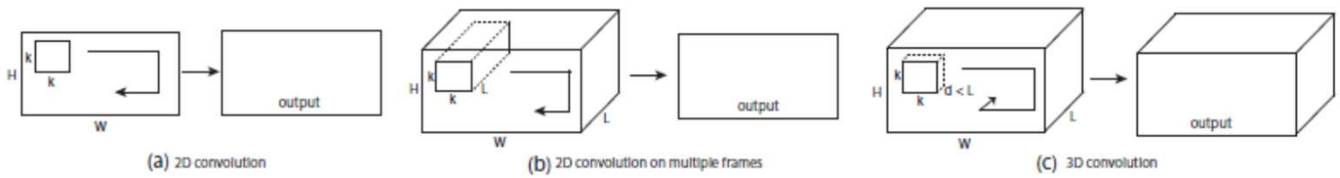
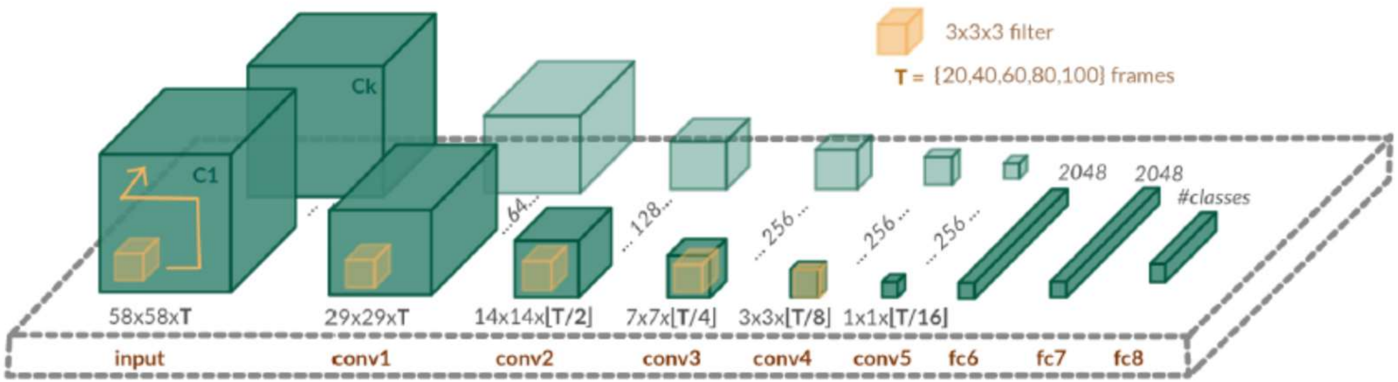
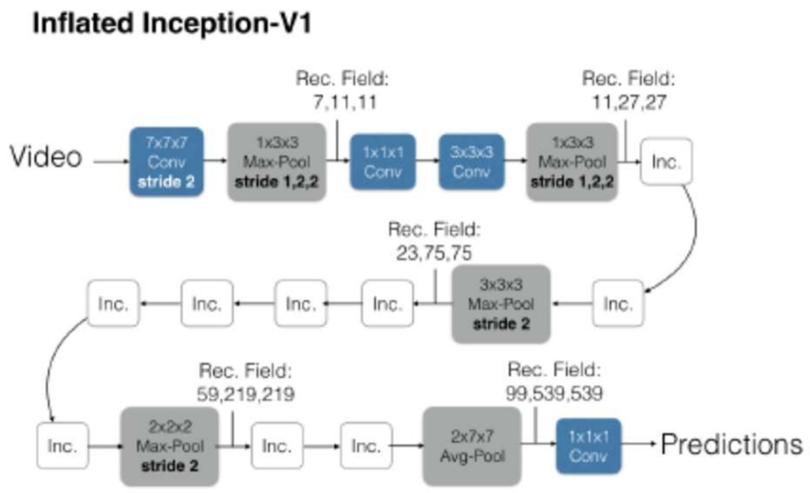


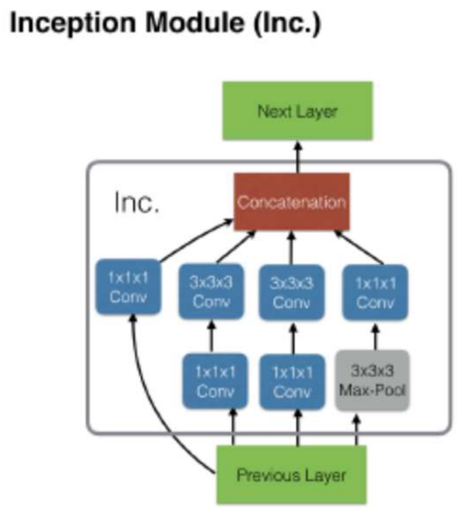
Figure 1. **2D and 3D convolution operations.** a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.



CNN based methods

Quo vadis, action recognition? A new model and the Kinetics dataset
[Carreira et al. CVPR17]





Pre-training on the large-scale Kinetics dataset 240k training videos
→ significant performance grain

Kinetics dataset

- Kinetics-700 dataset
  - 700 action classes
  - 650 00 clips
  - manual verification after automatic collection from YouTube

(c) shaking hands

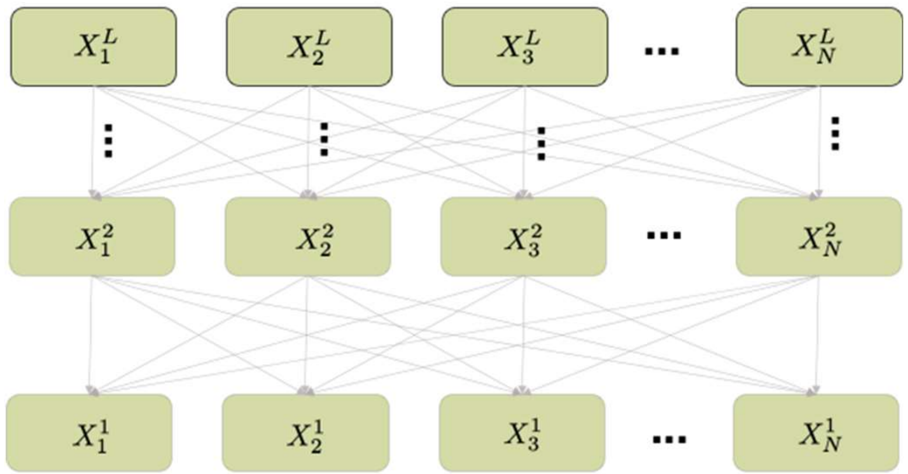
(n) dunking basketball

(j) playing trumpet

(l) brushing hair

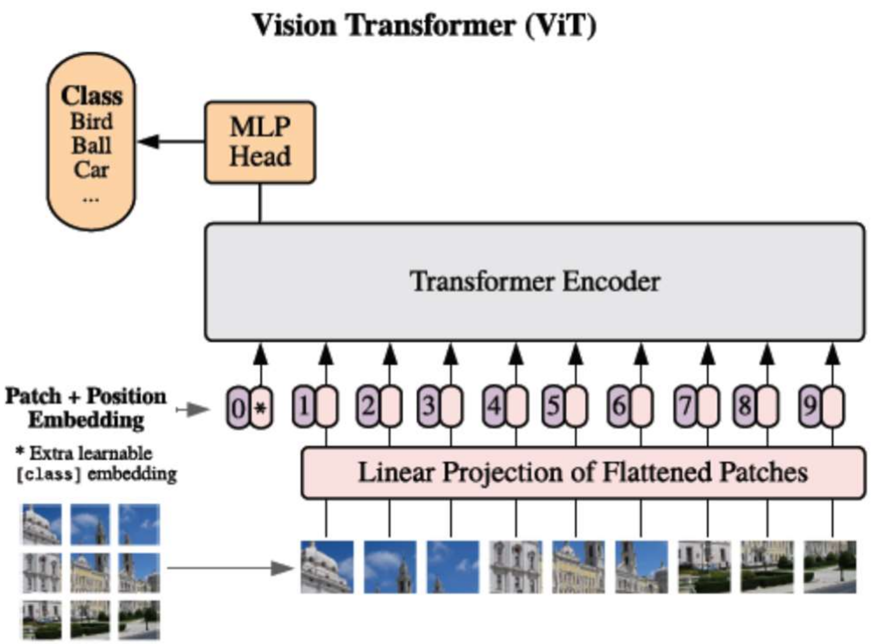
Transformer based models

- Transformer models are great for processing sequences
  - Text, images, videos can be expressed as sequences
  - Relies on self-attention between all tokens of a sequence [Vaswani et al., Neurips'17]



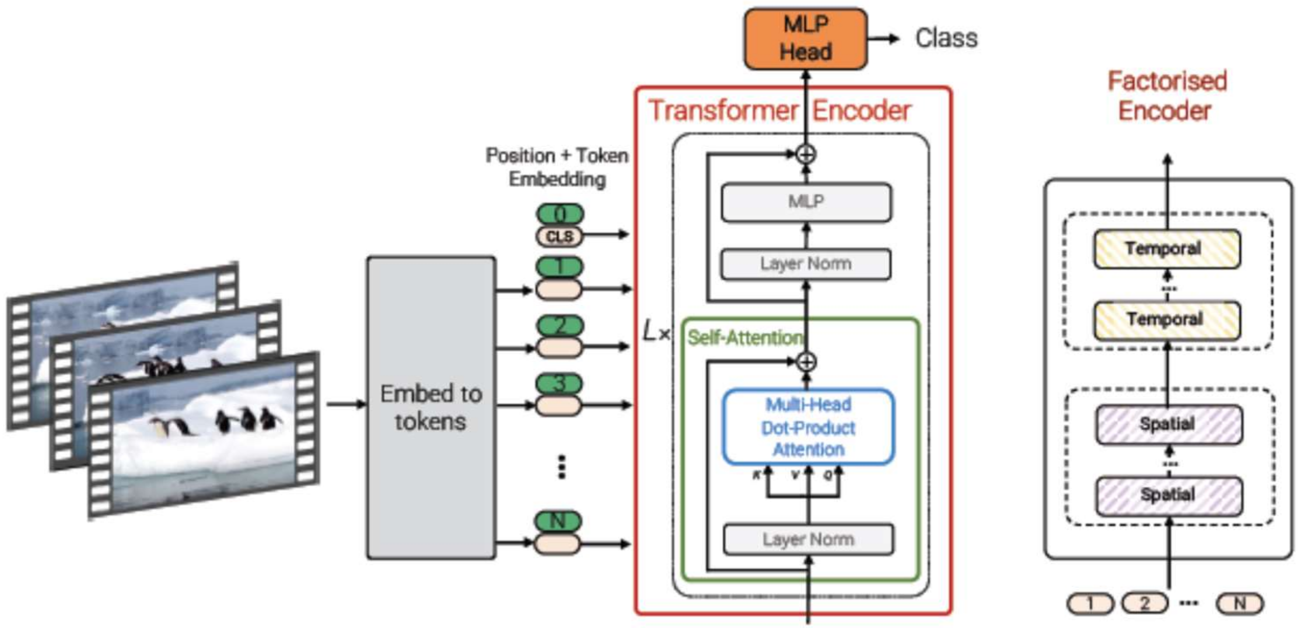
Vision Transformer (ViT)

- Fully transformer based architecture for image classification [A. Dosovitskiy et al., ICLR'21]
  - Image encoded as sequence of 16x16 patches
  - Tokenization by linear projection



ViViT: A Video Vision Transformer

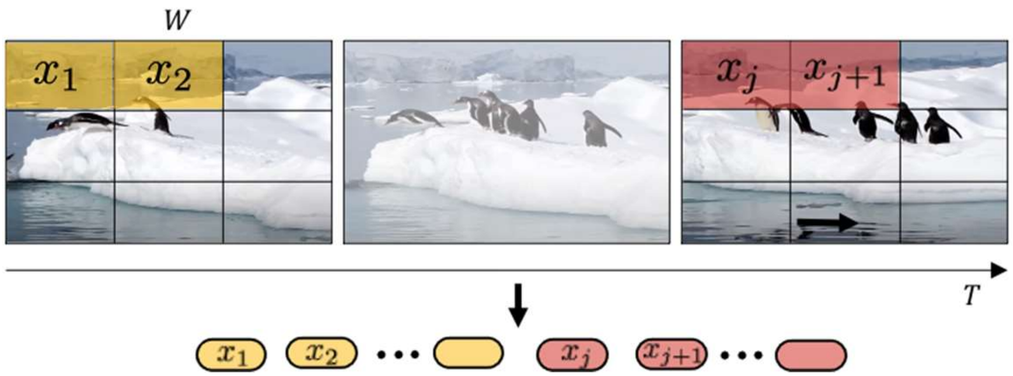
- Extend Vision Transformer ViT (for static images) to videos
- To handle large number of tokens, explore more efficient factorised attention variants



[ViViT, A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, C. Schmid, ICCV'21]

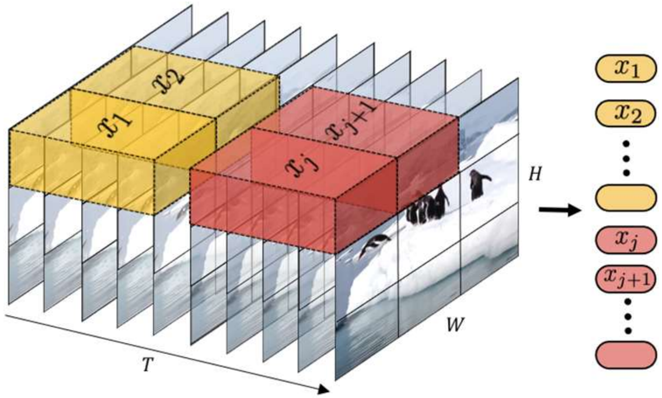
Input encoding – uniform frame sampling

- Sample frames, extract 2D patches and linearly project
- Effectively consider a video as a “big image”



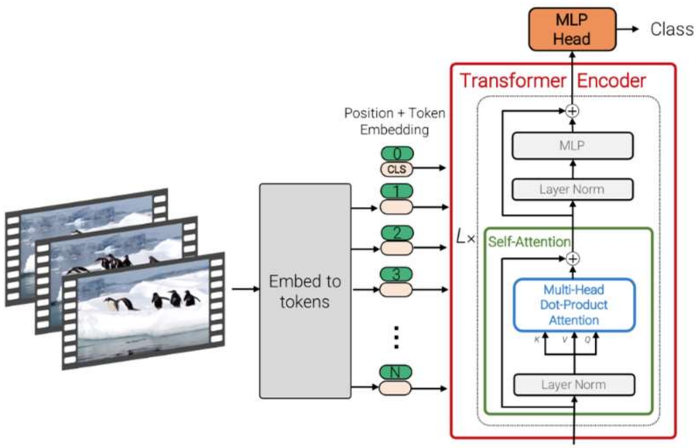
Input encoding – tubelet embedding

- Extract 3D spatio-temporal tubelets + linear project into tokens
- Captures temporal information in the tokenization stage
- Works better than uniform sampling



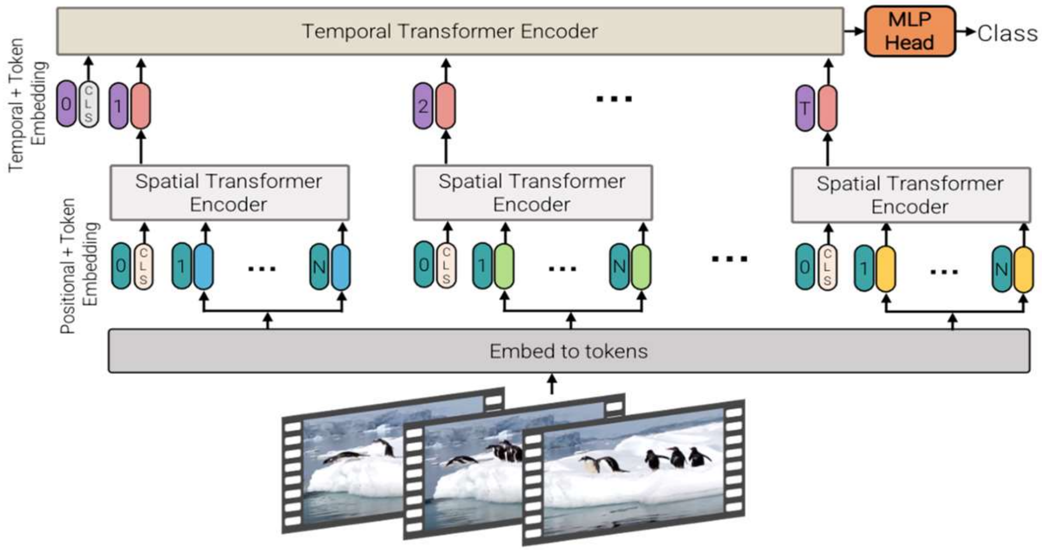
ViViT: A Video Vision Transformer

- An alternative to 3D convolutional neural networks
  - Extract 3D tubelets to encode spatio-temporal "tubes" into tokens
  - Encode tubes into embedding by linear project and add position
  - Train a transformer to predict classes
- Quadratic complexity in tokens



ViViT: Factorized Encoder

- Separate encoders for spatial and temporal information
  - Reduces complexity, compute, less overfitting
  - Spatial encoder is initialised from a pretrained-ViT model
  - "*Late fusion*" of spatial and temporal information



Comparison of model variants

	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 2: Ave. pool baseline	75.8	38.8	283.9	86.7	17.3

- Spatio-temporal model better for large datasets (K400)
- Factorized encoder faster than spatio-temporal model
- Factorized encoder better for small datasets (EK:EpicKitchen)
- Spatio-temporal model > average pooling

Impact of regularization

- Use pretrained ImageNet model for initialization
- Regularization with data augmentation and stochastic depth

	Top-1 accuracy
Random crop, flip, colour jitter	38.4
+ Kinetics 400 initialisation	39.6
+ Stochastic depth [28]	40.2
+ Random augment [10]	41.1
+ Label smoothing [58]	43.1
+ Mixup [79]	43.7

5.3% gain on Epic Kitchens

Comparison to state of the art

(a) Kinetics 400

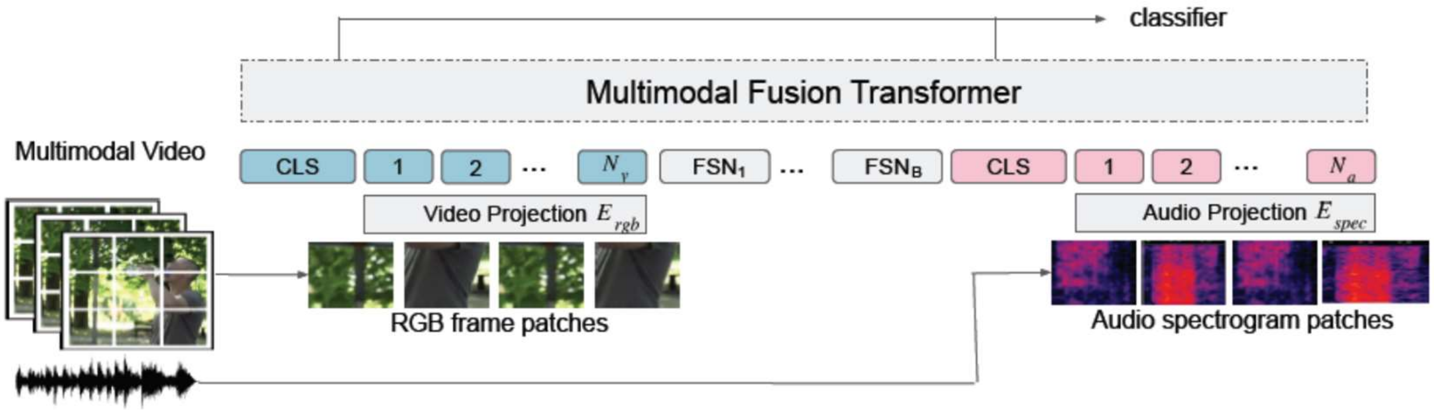
Method	Top 1	Top 5	Views
bIVNet [16]	73.5	91.2	–
STM [30]	73.7	91.6	–
TEA [39]	76.1	92.5	10 × 3
TSM-ResNeXt-101 [40]	76.3	–	–
I3D NL [72]	77.7	93.3	10 × 3
CorrNet-101 [67]	79.2	–	10 × 3
ip-CSN-152 [63]	79.2	93.8	10 × 3
LGD-3D R101 [48]	79.4	94.4	–
SlowFast R101-NL [18]	79.8	93.9	10 × 3
X3D-XXL [17]	80.4	94.6	10 × 3
TimeSformer-L [2]	80.7	94.7	1 × 3
ViViT-L/16x2	80.6	94.7	4 × 3
ViViT-L/16x2 320	**81.3**	**94.7**	4 × 3
Methods with large-scale pretraining			
ip-CSN-152 [63] (IG [41])	82.5	95.3	10 × 3
ViViT-L/16x2 (JFT)	82.8	95.5	4 × 3
ViViT-L/16x2 320 (JFT)	83.5	95.5	4 × 3
ViViT-H/16x2 (JFT)	**84.8**	**95.8**	4 × 3

(b) Kinetics 600

Method	Top 1	Top 5	Views
AttentionNAS [73]	79.8	94.4	–
LGD-3D R101 [48]	81.5	95.6	–
SlowFast R101-NL [18]	81.8	95.1	10 × 3
X3D-XL [17]	81.9	95.5	10 × 3
TimeSformer-HR [2]	82.4	**96.0**	–
ViViT-L/16x2	82.5	95.6	4 × 3
ViViT-L/16x2 320	**83.0**	95.7	4 × 3
ViViT-L/16x2 (JFT)	84.3	96.2	4 × 3
ViViT-H/16x2 (JFT)	**85.8**	**96.5**	4 × 3

A multimodal (audio-visual) transformer

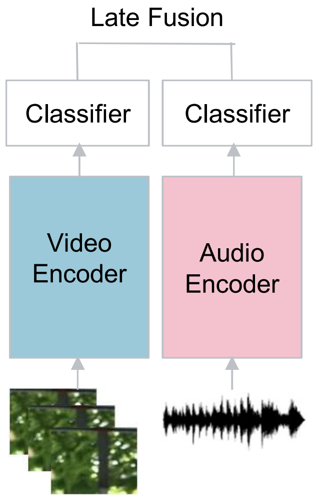
- Extend ViViT to multimodal information by adding audio
- Audio is represented by a spectrogram



[Attention bottlenecks for multimodal fusion, A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun, Neurips'21l]

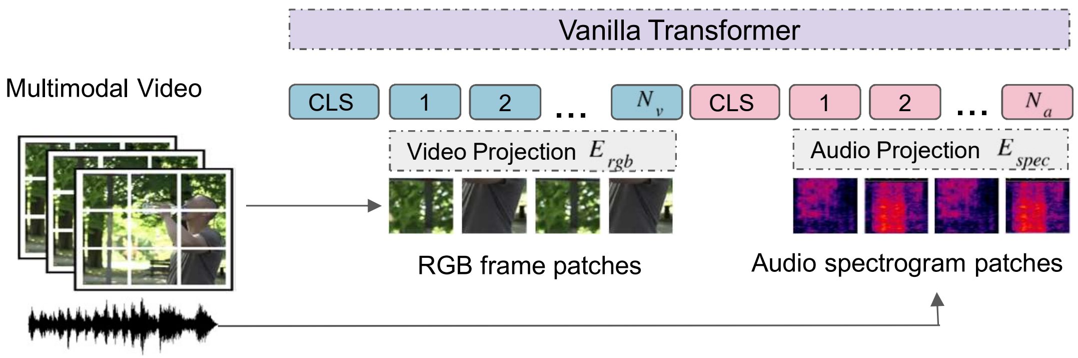
Late fusion

- Multimodal inputs
  - Heterogeneity of inputs (RGB frames, audio spectrograms)
  - Specialized architectures
  - Different datasets and evaluation benchmarks
- The “dominant” paradigm
  - Different encoders
  - Output scores a fused at the end



Vanilla Multimodal Transformer

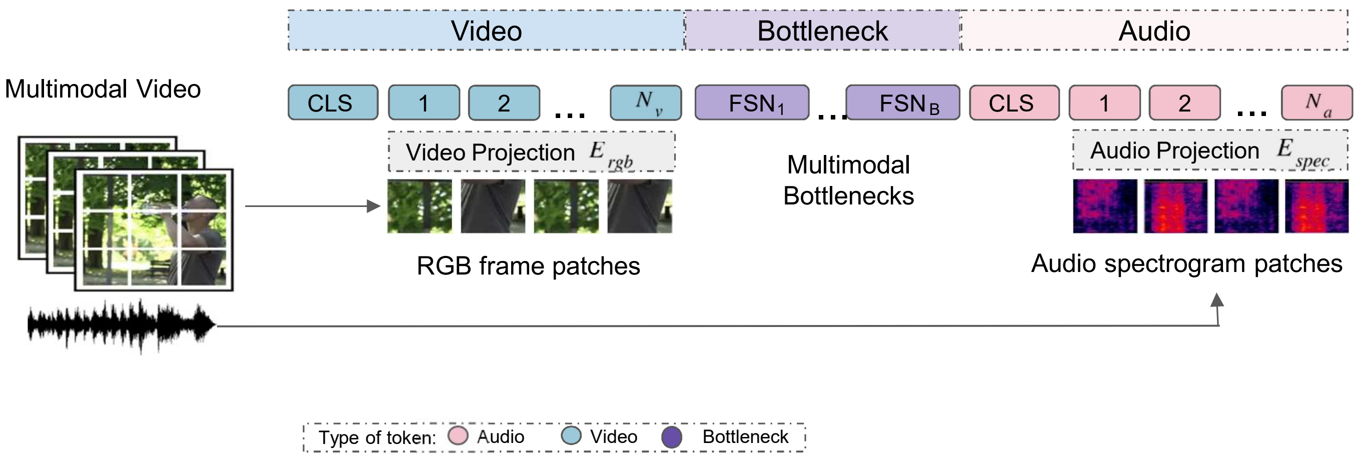
- Tokenize RGB frame and spectrogram patches
- Feed all tokens to a transformer
- Pairwise self-attention between all tokens (early fusion)



- Scales quadratically with sequence length
- Video has a lot of redundancy

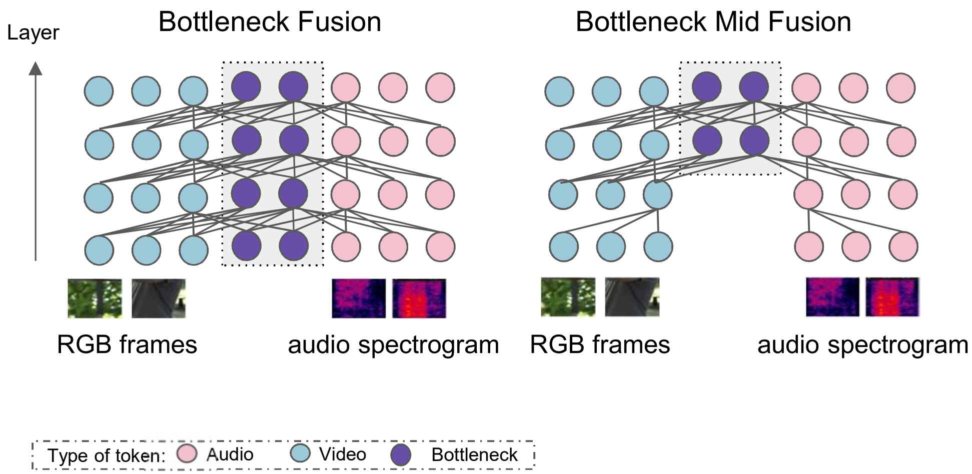
Multimodal Bottleneck Transformer

- Introduces a number of bottleneck tokens (B=4)
- Full pairwise self attention within a modality
- Attention between the vision/audio tokens and the bottleneck tokens



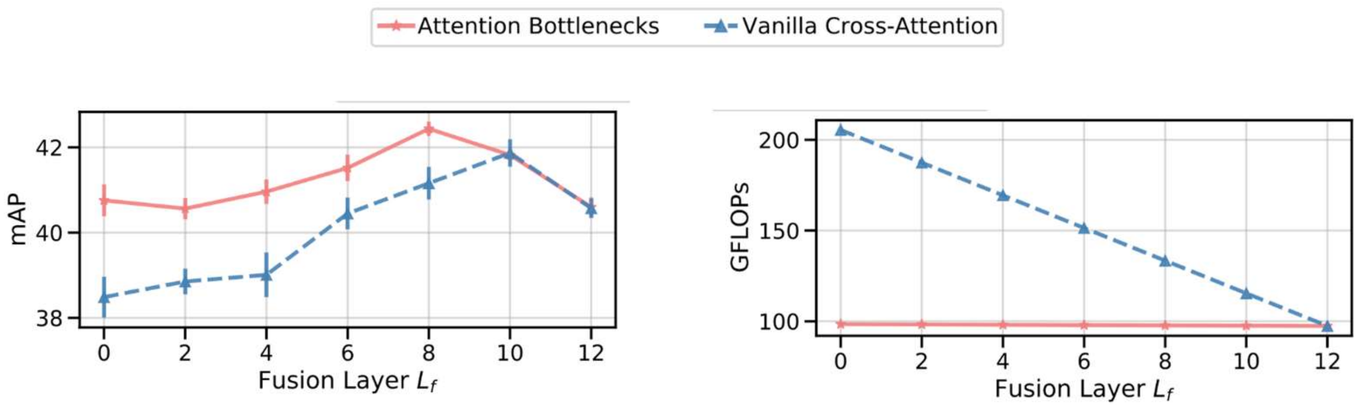
Do all layers need to be cross-modal?

- Restrict cross-modal information to later layers (mid-fusion)
- The layer we introduce cross-modal interactions is called the “fusion layer”
- Allows early layers to “specialize” to unimodal patterns



Improved performance and efficiency

- Mid Fusion outperforms early and late fusion on most datasets



Results for Audio-Set and 4 bottleneck tokens
- Improved performance, lower compute

Experimental results

- Two different video classification tasks

Action Recognition

Kinetics
Moments in Time

Epic Kitchens

Sound Event Classification

Audioset
VGGSound
Kinetics-Sou

Experimental results

Model	Training Set	A only	V only	AV Fusion
GBlend [58]	MiniAS	29.1	22.1	37.8
GBlend [58]	FullAS-2M	32.4	18.8	41.8
Attn Audio-Visual [19]	FullAS-2M	38.4	25.7	46.2
Perceiver [29]	FullAS-2M	38.4	25.8	44.2
MBT	MiniAS	31.3	27.7	43.9
MBT	AS-500K	**44.3**	**32.3**	**52.1**

Table 1: **Comparison to the state of the art on AudioSet [22].** We report mean average precision (mAP). For audio-visual fusion, our method outperforms others that use the entire AudioSet training set (almost 2M samples), while we train on only 500K.

Audioset

Late Fusion	49.2
MBT (ours)	**52.1**

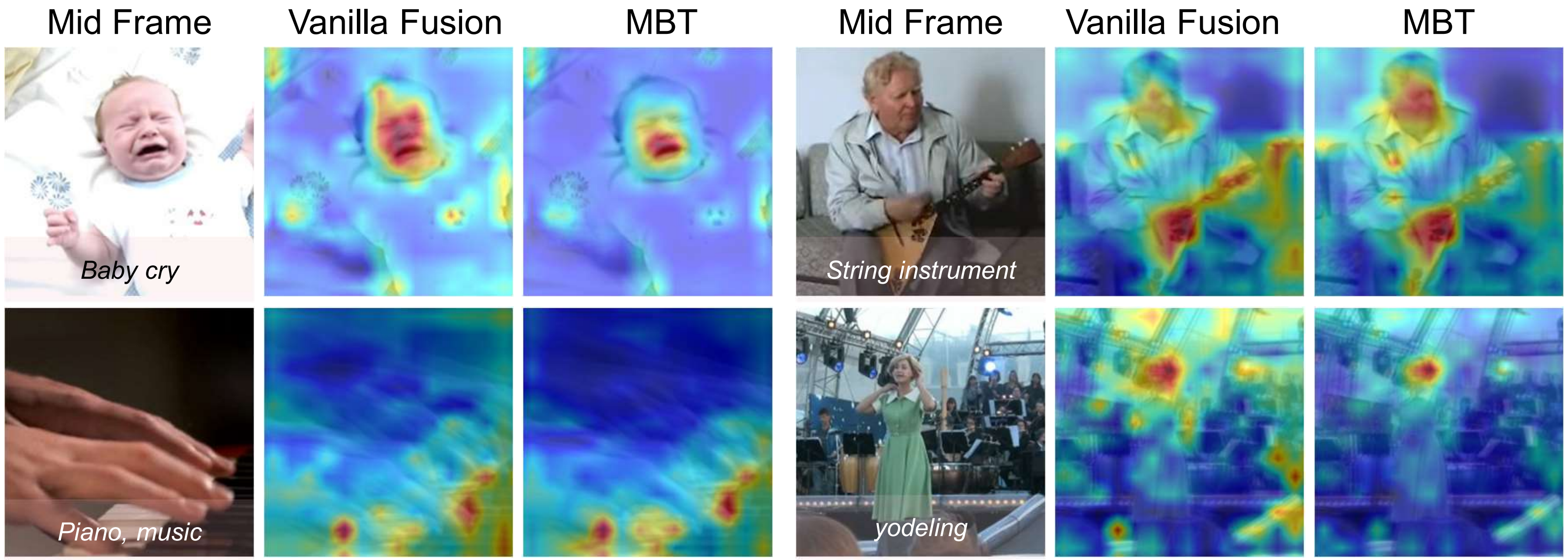
Model	Modalities	Verb	Noun	Action
Damen et al. [13]	A	42.1	21.5	14.8
AudioSlowFast [34]†	A	46.5	22.78	15.4
TSN [57]	V, F	60.2	46.0	33.2
TRN [63]	V, F	65.9	45.4	35.3
TBN [33]	A, V, F	66.0	47.2	36.7
TSM [42]	V, F	**67.9**	49.0	38.3
SlowFast [20]	V	65.6	50.0	38.5
MBT	A	44.3	22.4	13.0
MBT	V	62.0	56.4	40.7
MBT	A, V	64.8	**58.0**	**43.4**

Table 2: **Comparison to the state of the art on Epic Kitchens 100 [13].** Modalities (Mods) are **A:** Audio, **V:** Visual, **F:** Optical flow.

Epic-Kitchens

Late Fusion	37.9
MBT (ours)	**43.4**

Attention Heatmaps

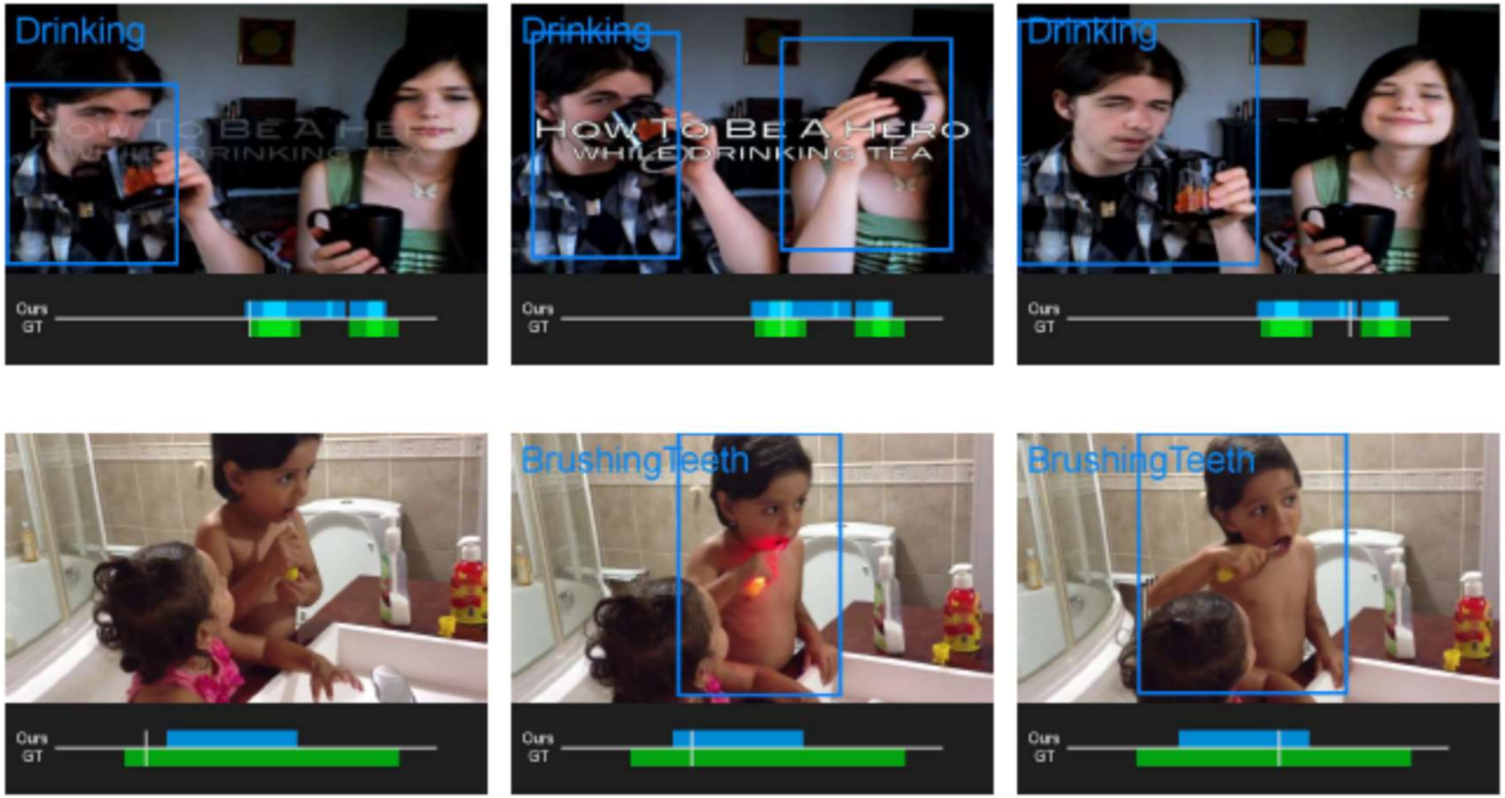


MBT: focus on smaller regions, sound sources (mouth, fingertips)

Overview

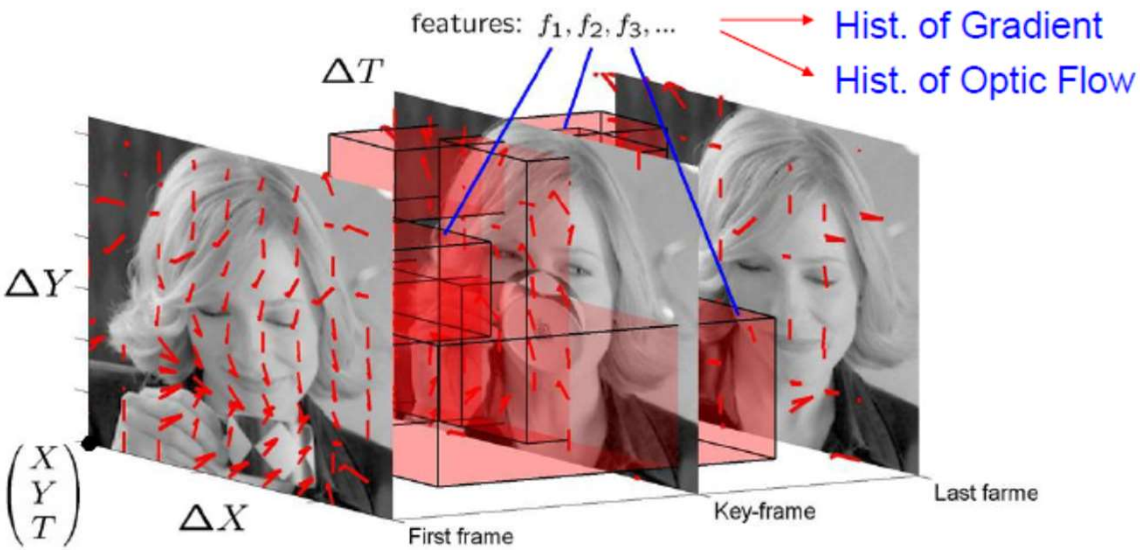
- Optical flow
- Video classification
- *Action localization*
- Multi-modal / LLM-based video understanding

Spatio-temporal action localization



Spatio-temporal action localization

- Space-time sliding window
  - Spatio-temporal features selection with a cascade, Laptev & Perez, ICCV'07

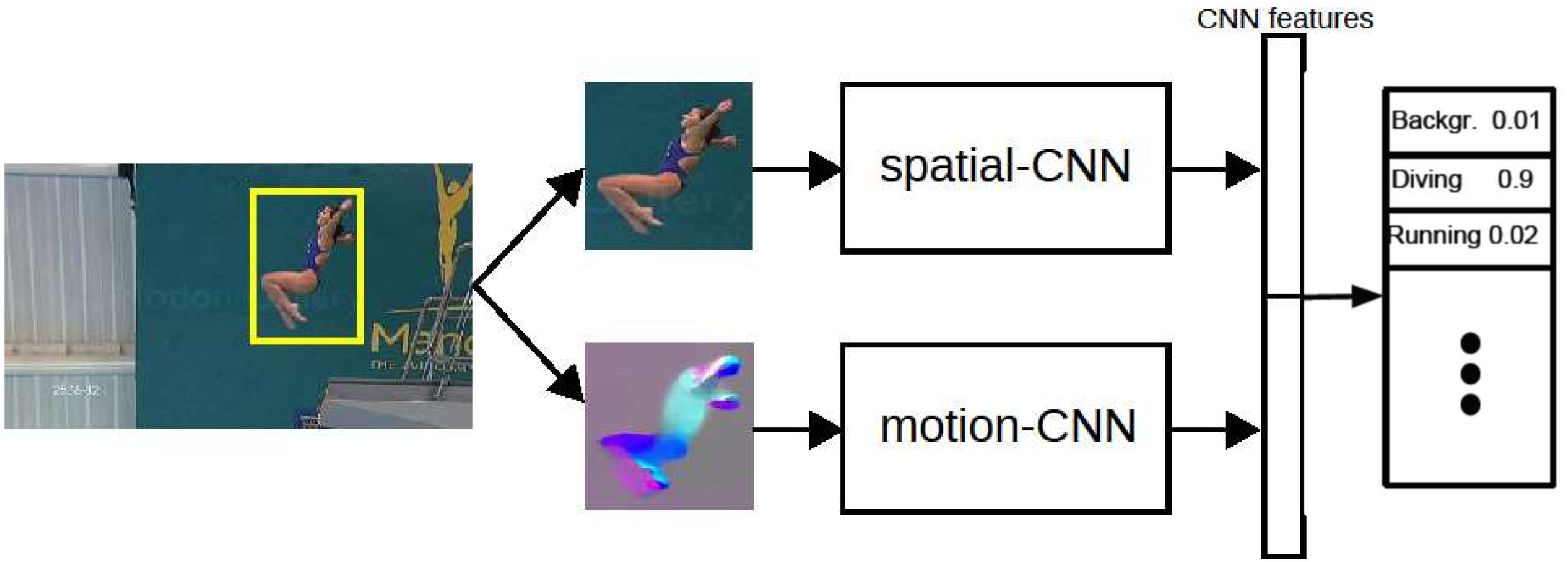


Spatio-temporal action localization

- Human tubes + tube classification
  - Human focused action localization in video, Kläser et al., SGA'10

Spatio-temporal action localization

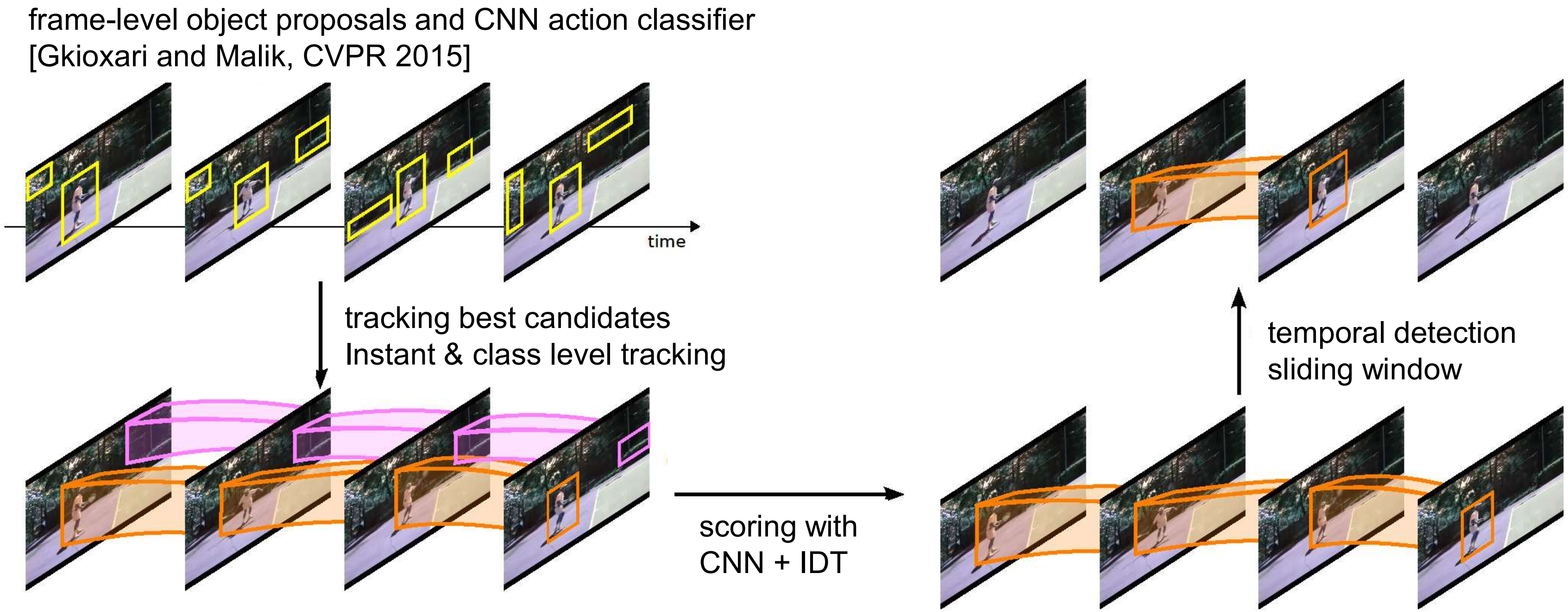
- Frame-level candidates
  - Compute object proposals (EdgeBoxes [Zitnick et al. 2014])
  - Extract CNN features (training similar to R-CNN [Girshicket al. 2014])
  - Score each object proposal



[Gkioxari and Malik'15, Simonyan and Zisserman'14]

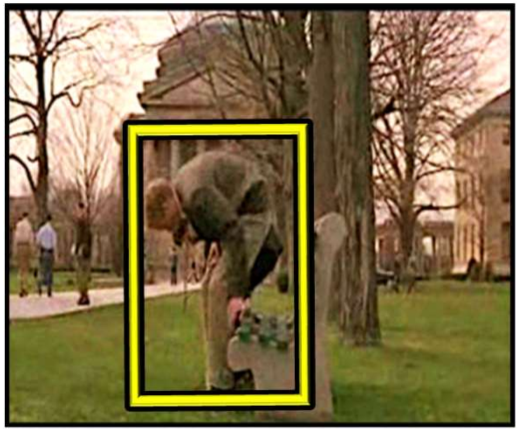
Spatio-temporal action localization

- Learning to track frame-based proposals [Weinzaephel et al., ICCV'15]



Action recognition - temporal context

Ambiguous action given only one frame



- Jump

? $\boxed{\text{• Sitting down}}$

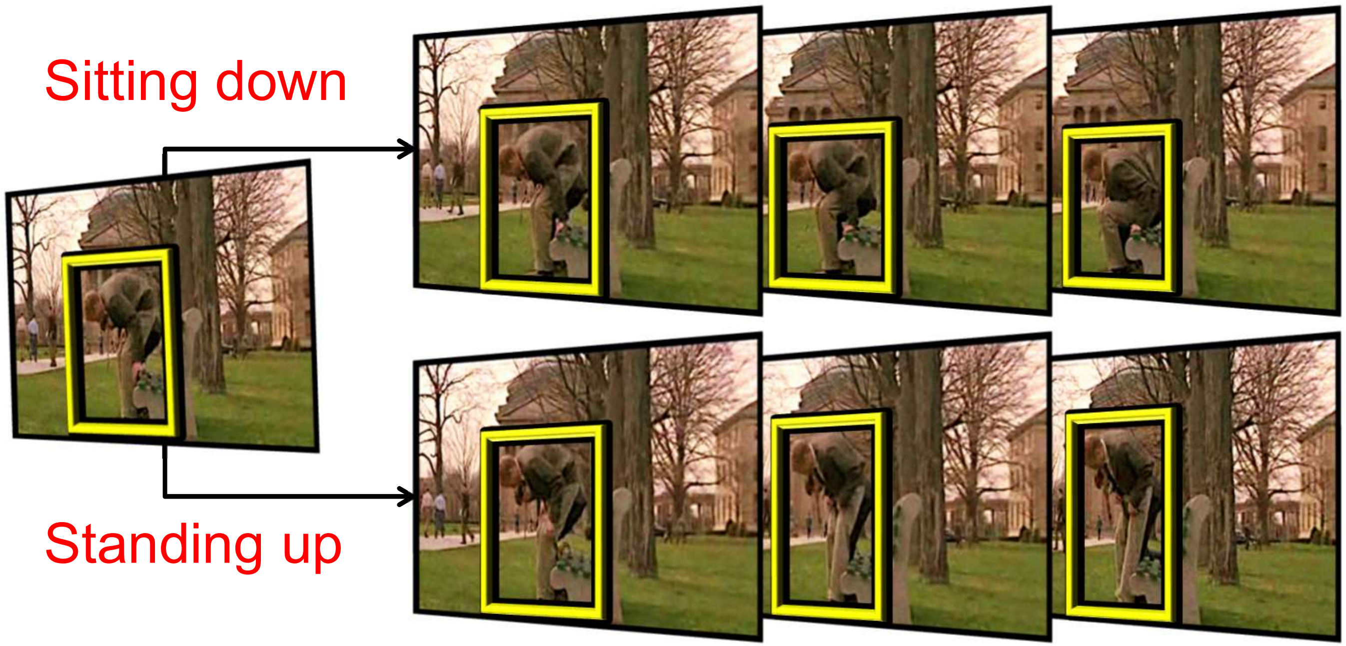
...

? $\boxed{\text{• Standing up}}$

- Walk

Action recognition - temporal context

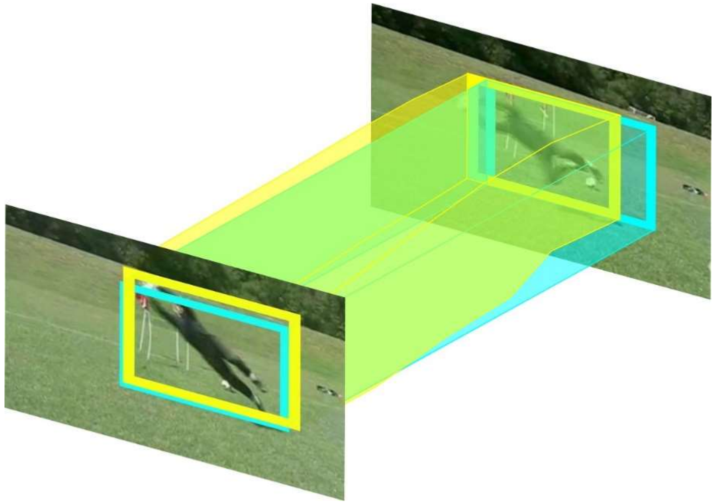
Ambiguity resolved given several frames

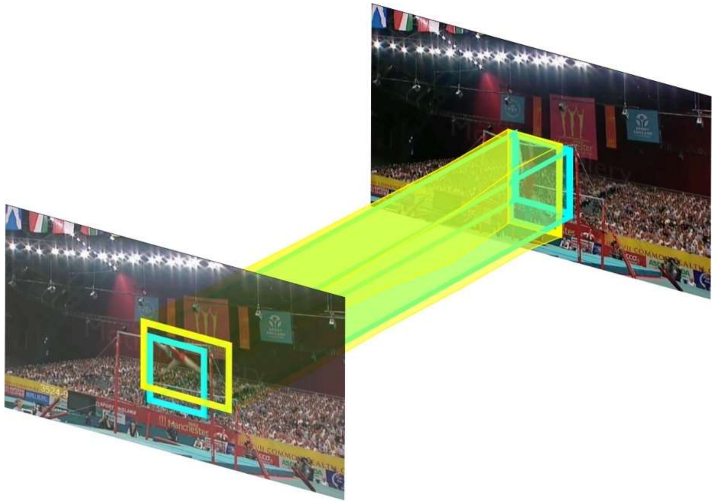


ACtion tubelet detector

Classify and regress spatio-temporal volumes

- *Anchor cuboids*: fixed spatial extent over time
- *Regressed tubelets*: score + deform the cuboid shape

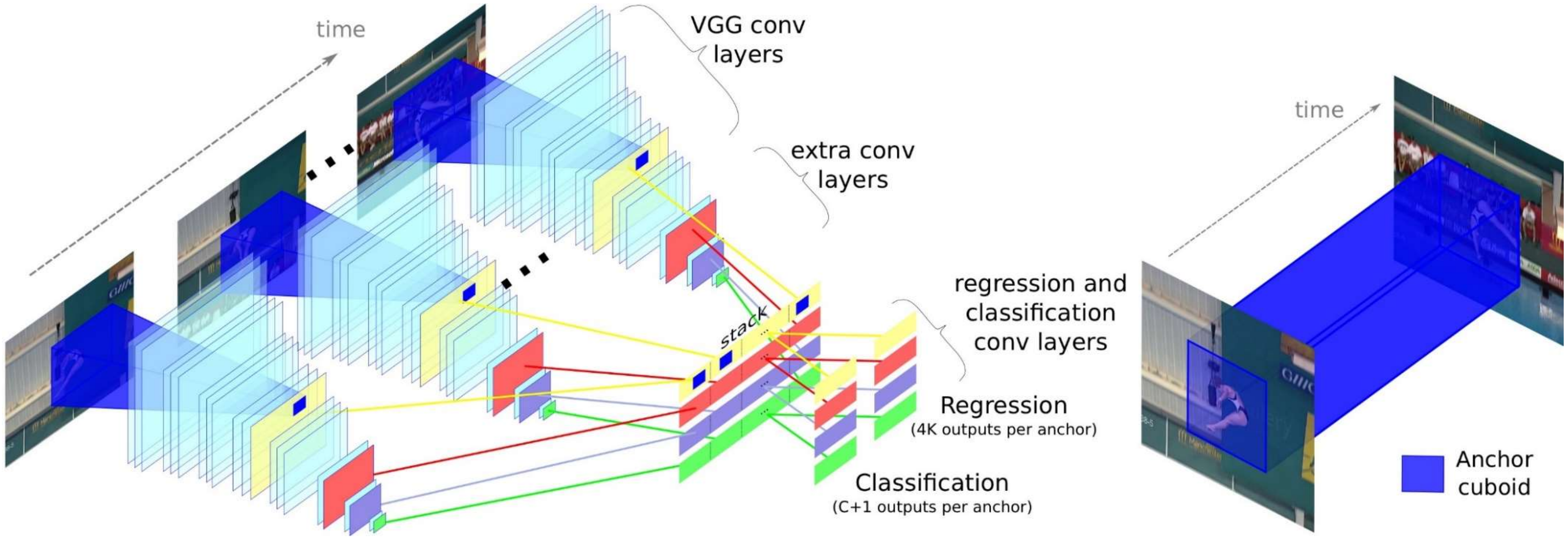




[Action tublet detector for spatio-temporal action localization, V. Kalogeiton et al., ICCV'17]

ACtion tubelet detector

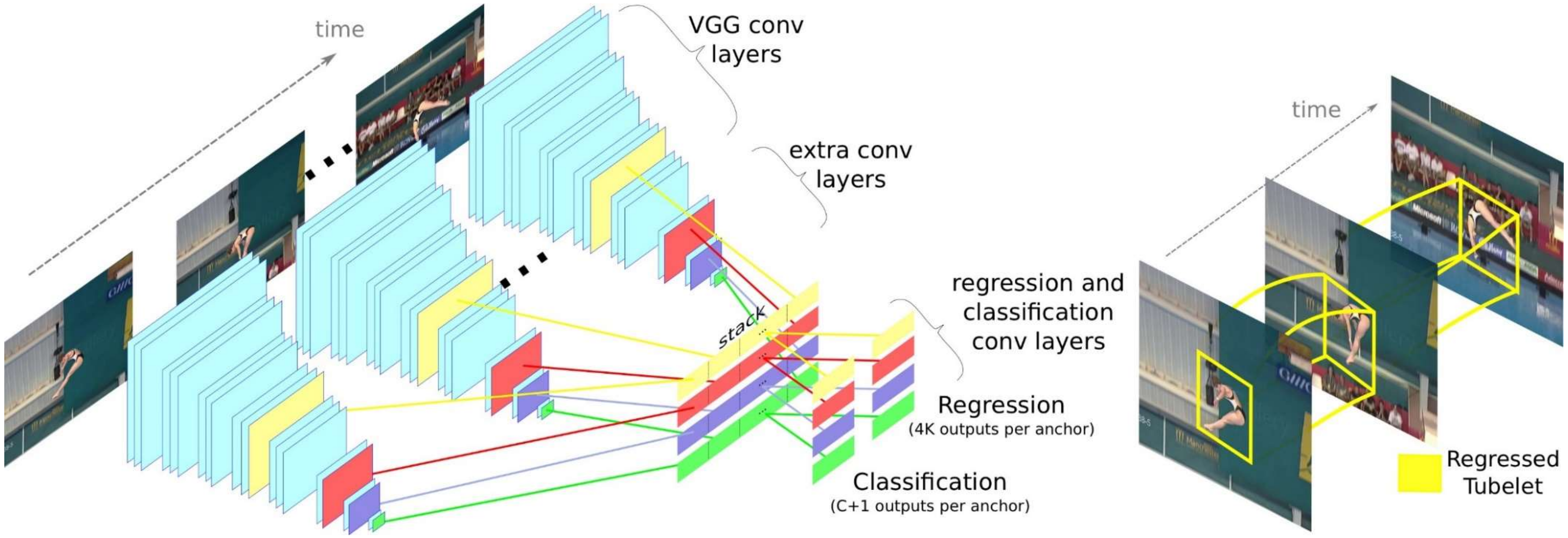
Use sequences of frames to detect *tubelets*: anchor cuboids



SSD detector [Liu et al., ECCV'16]

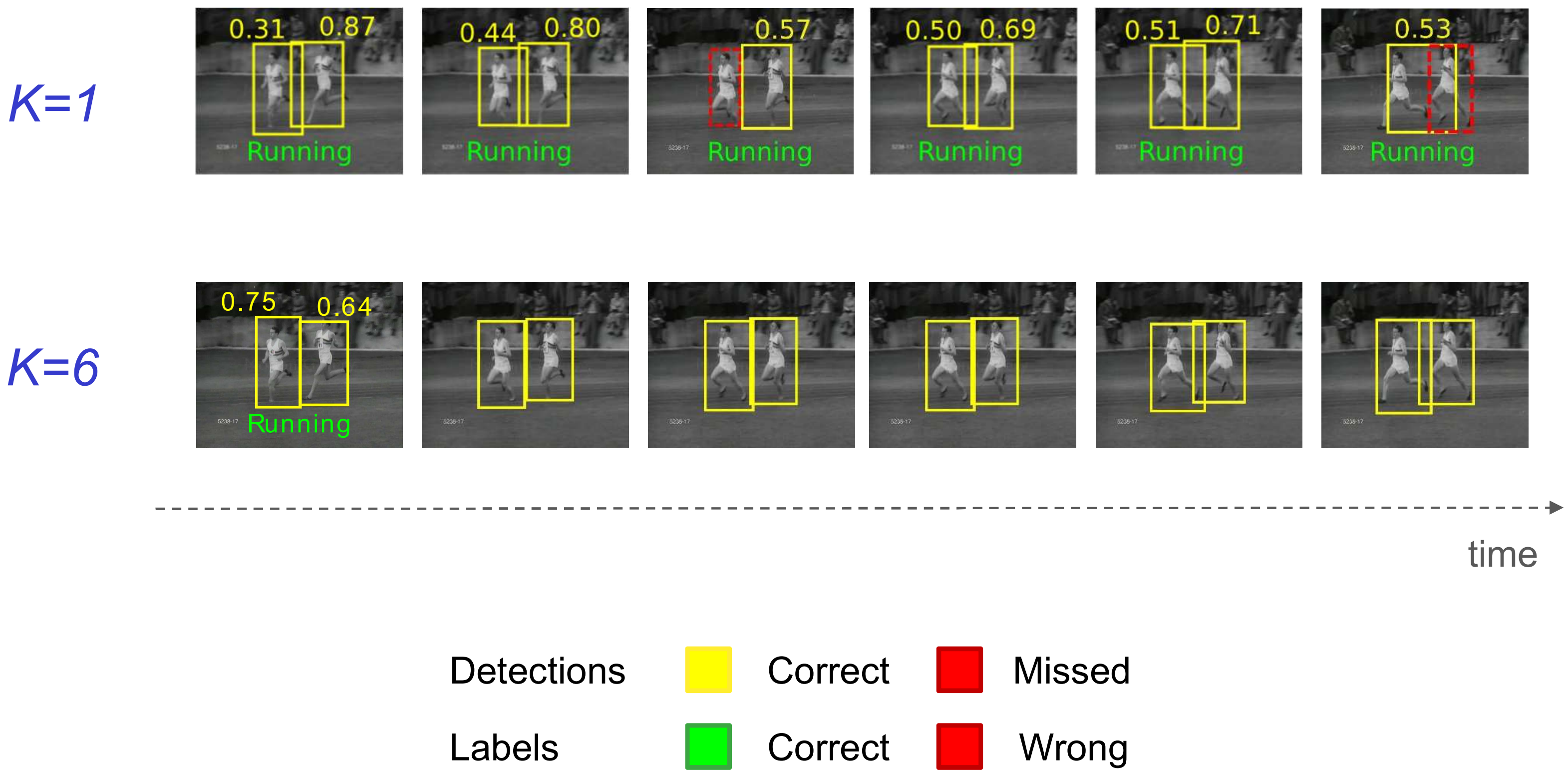
ACtion Tubelet detector

Use sequences of frames to detect *tubelets*



SSD detector [Liu et al., ECCV'16]

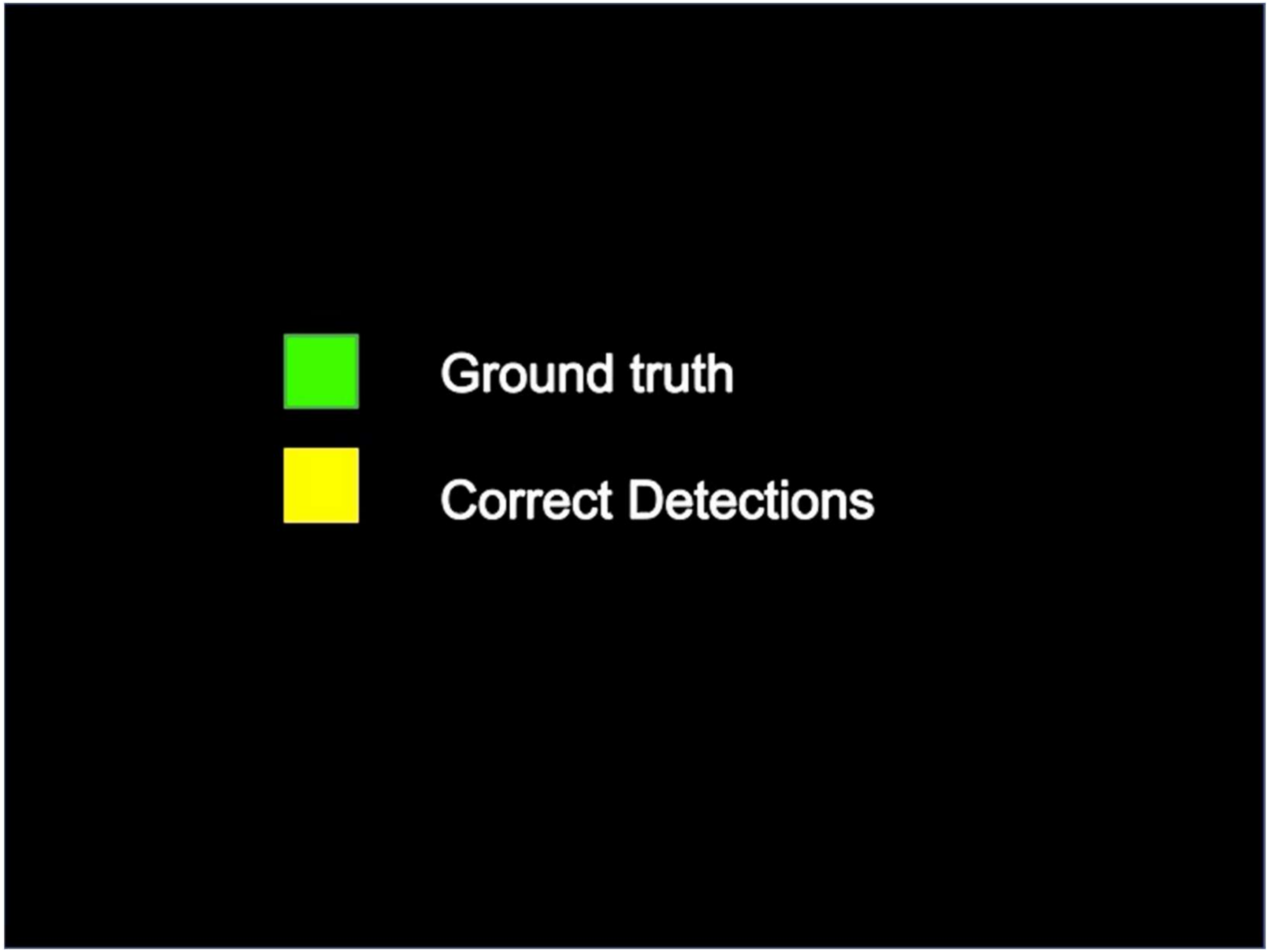
Example results



Example results

Detections Correct Missed

Labels Correct Wrong


Ground truth
Correct Detections

Datasets for action localization

– UCF-101 (24 sports actions, 3207 almost-trimmed low-res. videos)


basketball


long jump


rope climbing

– J-HMDB (21 daily actions, 928 trimmed videos, avg length: 1.5s, low resolution)


climbing stairs


jumping


pushing

– Limited by diversity, duration and resolution

Atomic Visual Actions (AVA) dataset

- Towards a definition of atomic actions + large scale collection
  → Atomic Visual Actions (AVA) dataset

Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Left: Stand, Carry/Hold, Read; Middle: Stand, Take (object) from; Right: Stand, Give (object) to

[AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions, CVPR'18]

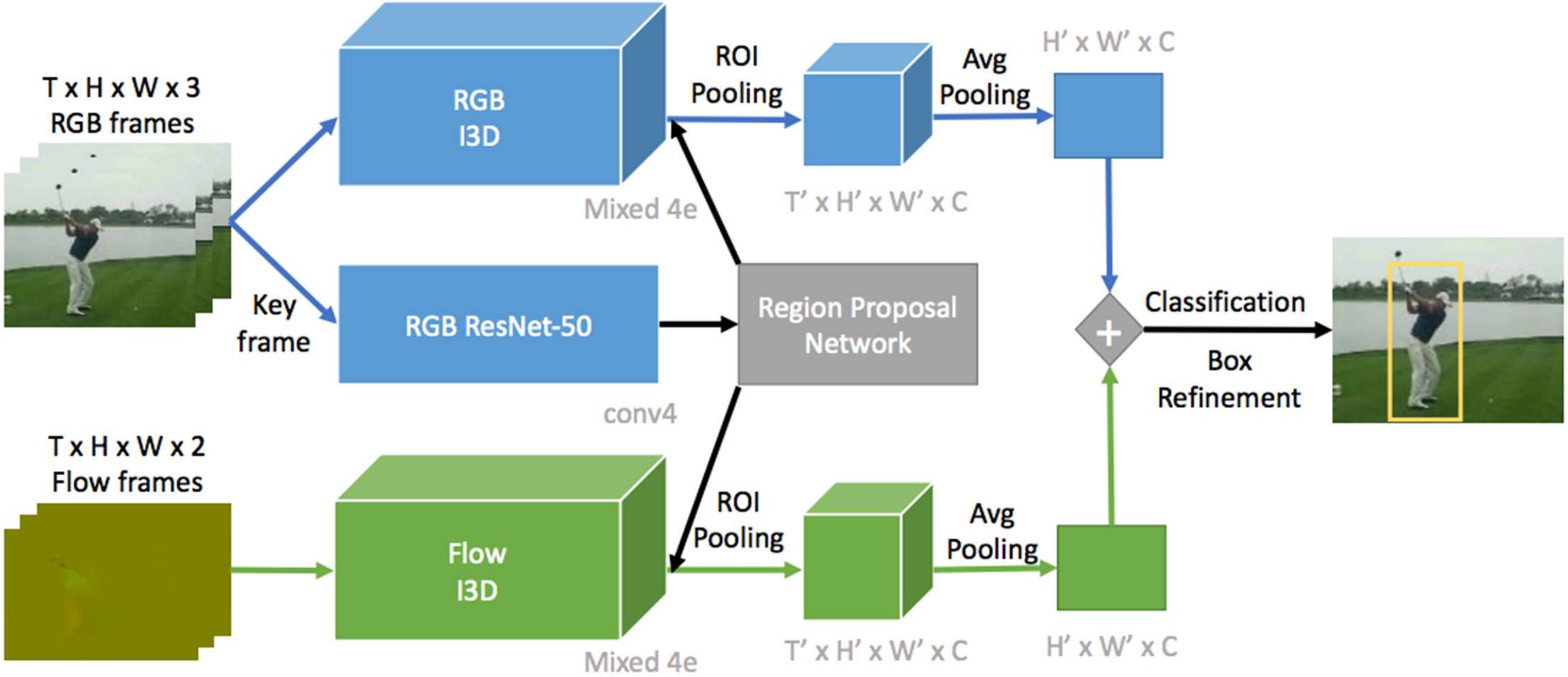
Ava dataset – atomic actions

- Three categories of atomic actions:
  1) Pose of the person, eg., stand, sit, walk, kneel, swim
  2) Interactions with objects, eg., drive, carry, pick up
  3) Human-human interactions, eg., talk to, hug, fight
- Multiple labels per person
- Exhaustive annotation of all humans

Ava dataset

- 192 videos with annotations for 15 minute intervals
- Annotation every 1 seconds
- 80 atomic actions in 107k movie segments with 740k labels with multiple labels per person
- Exhaustive annotation of all humans
  - Human are detected automatically and corrected manually

Action Detection Model – Faster R-CNN+3DCNN



[AVA, C. Gu, C. Sun et al. CVPR'18]

Impact of temporal extent on 3D convolutions

Temp. context	UCF101-24	AVA
5 RGB + 5 Flow	76.1%	13.4
10 RGB + 10 Flow	78.0%	13.9
20 RGB + 20 Flow	**78.3%**	14.9
40 RGB + 40 Flow	76.0%	**16.2**
50 RGB + 50 Flow	73.2%	15.8

Spatio-temporal action localization

Frame-mAP	JHMDB	UCF101-24
Actionness [41]	39.9%	-
Peng w/o MR [29]	56.9%	64.8%
Peng w/ MR [29]	58.5%	65.7%
ACT [40]	65.7%	69.5%
3D CNN + Faster-RCNN	**73.3%**	**76.3%**

Video-mAP	JHMDB	UCF101-24
Peng w/ MR [29]	73.1%	35.9%
Singh *et al.* [37]	72.0%	46.3%
ACT [40]	73.7%	51.4%
TCNN [16]	76.9%	-
3D CNN + Faster-RCNN	**78.6%**	**59.9%**

Failure modes on AVA

FA for “hand shake”:
Reaching out arm

FA for “smoke”:
Hand covering mouth

FA for “write”:
Looking downwards

Failure modes on AVA

FA for “hand shake”:
Reaching out arm

Other person does not reach out arm

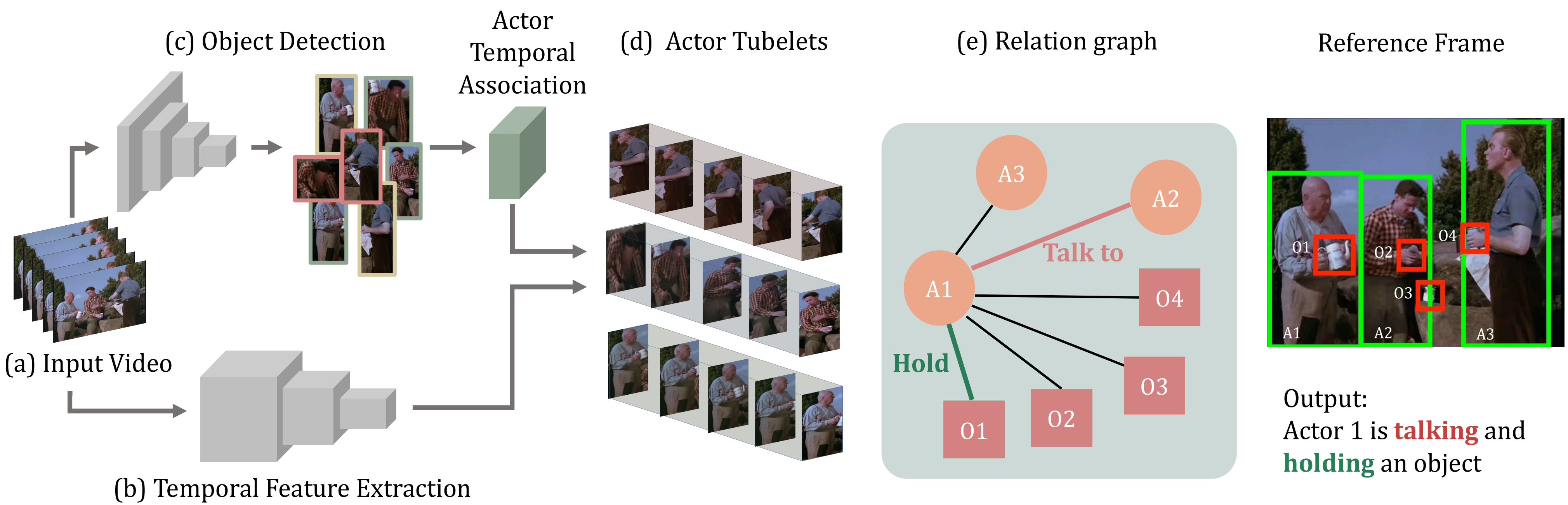
FA for “smoke”:
Hand covering mouth

No cigarette in hand

FA for “write”:
Looking downwards

Dining table with plates

A structured model for action detection



[A structured model for action detection. Y. Zhang et al., CVPR’19]

A structure model for action detection

Temporal dependency learning

- Construct tublets based on appearance similarity of actors
  - with Siamese network + triplet loss
- Learn how to combine features in the tublet with graph convolutions

Relation modeling

- Graph of model human-human and human-object interactions
- Soft-assignment to integrate the features

Quantitative results

Model	mAP
Single Frame model [1]	14.2
ACRN [2]	17.4
Our Baseline	16.7
Person similarity graph on ROIs [3]	20.1
Object similarity graph on ROIs [3]	20.3
Actor tubelet model	21.1
Actor tubelet + hard relation graph	21.5
Actor tubelet + soft relation graph	22.2

[1] C. Gu et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. CVPR , 2018.
[2] C. Sun et al. . Actor-centric relation network. ECCV, 2018.
[3] X. Wang and A. Gupta. Videos as space-time region graphs. ECCV, 2018.

Illustration of temporal dependency learning

Baseline

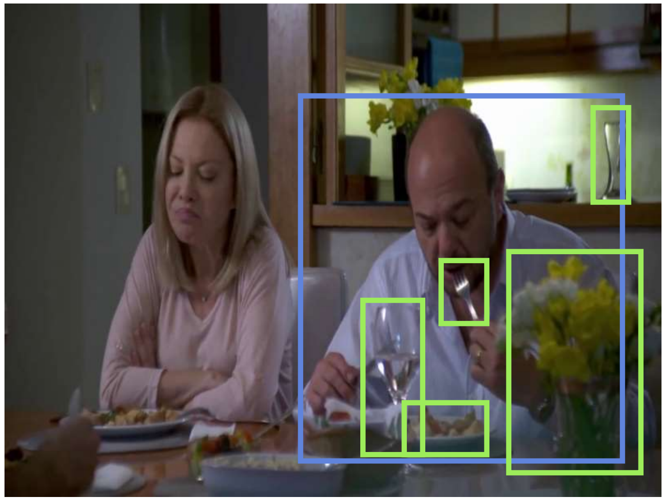
Incorrect label: sit

Our approach – temporal dependency learning

Correct label: fall down

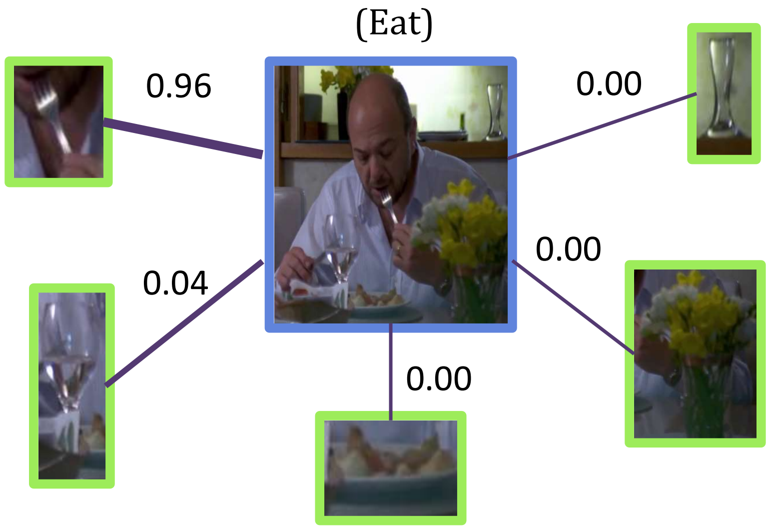
Illustration of relation modeling

Actor and Object Detection



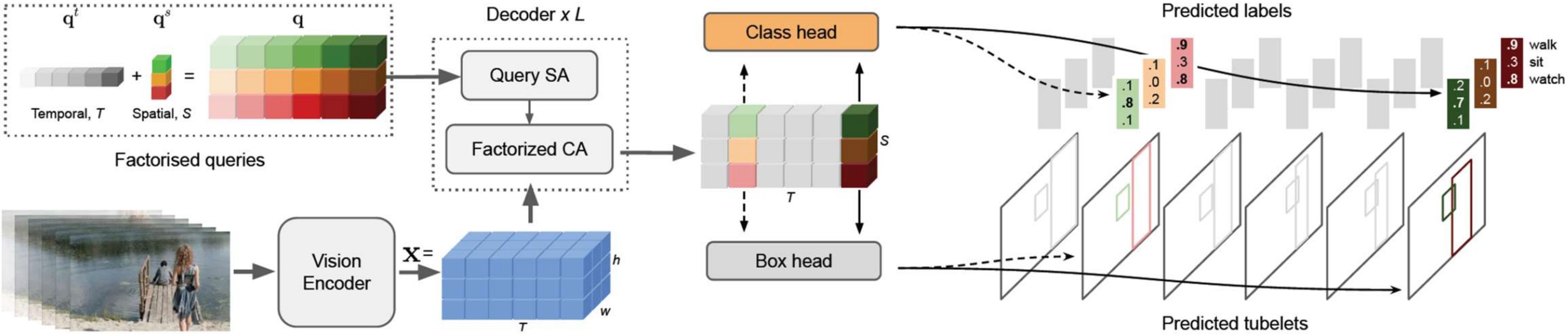
Baseline: hold

Soft Relation Graph



Relational model: eat

STAR - end-to-end training transformers

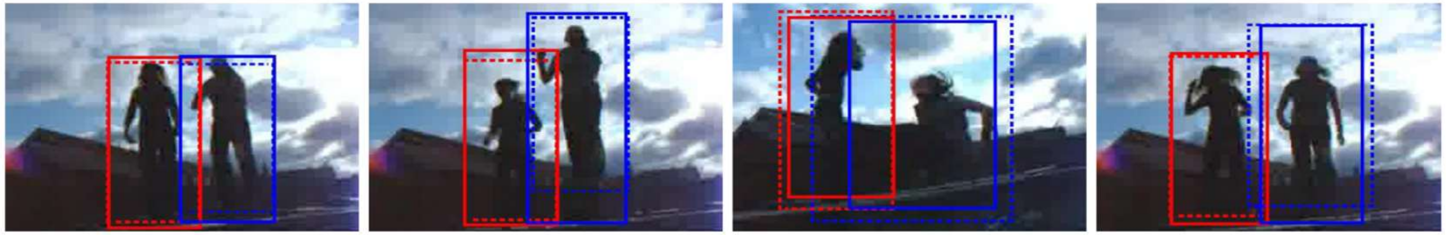


For each frame outputs tubelets, i.e., linked bounding boxes with action class probabilities

- Transformer-based vision encoder which outputs a video representation
- Learn queries, which are factorized into spatial and temporal components, similar to DETR for images
- Decoder (L layer with query self-attention and factorized cross-attention)
- Followed by a box and class prediction head

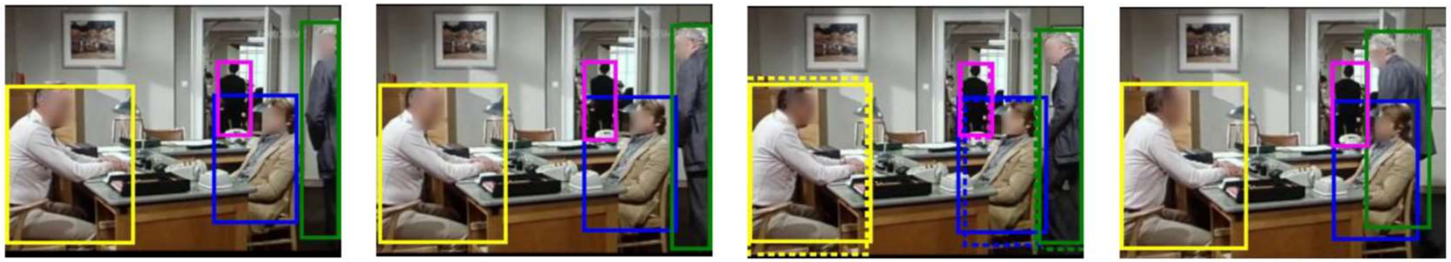
[STAR, Gritsenko et al., arXiv, 2023]

STAR – experimental results



trampoline jumping, trampoline jumping

Results on UCF 101



sit, talk to, watch, touch

watch, listen to, sit

stand, watch, listen to

walk, watch, listen to

Results on AVA

STAR – experimental results

	Pretraining	UCF101-24				JHMDB51-21			Backbone
		fAP	vAP20	vAP50	vAP50:95	fAP	vAP20	vAP50	
ACT [23]	IN1K	67.1	77.2	51.4	25.0	65.7	74.2	73.7	VGG
MOC [31]	IN1K → COCO	78.0	82.8	53.8	28.3	70.8	77.3	77.2	DLA34
Unified [2]	K600	79.3	–	–	–	–	–	–	SlowFast
WOO [8]	K600	–	–	–	–	80.5	–	–	SlowFast
TubeR [65]	IG65M→K400	83.2	83.3	58.4	28.9	–	87.4	82.3	CSN-152
TubeR with flow [65]	K400	81.3	85.3	60.2	29.7	–	81.8	80.7	I3D
STAR/B (ours)	IN21K→K400	87.3	87.7	66.2	30.9	86.6	89.1	88.5	ViViT/B
STAR/L (ours)	CLIP→K700	**90.3**	**88.0**	**71.8**	**35.2**	**92.1**	**93.1**	**92.6**	ViViT/L

Comparison to the state of the art

Overview

- Optical flow
- Video classification
- Action localization
- *Multi-modal / LLM-based video understanding*

Why multimodal data?

- Precise understanding of the video content
  → Requires access to all modalities simultaneously

Is this Indian?

Why multimodal video representation?

- Large-scale cross-modal supervision
  → No manual annotation required

Training on the **HowTo100M** [1] dataset

→ + 120M pairs clip-narration   → + 1,2M videos   → Uncurated

[HowTo100M. A. Miech, D. Zhukov, JB Alayrac, M. Tapaswi, I. Laptev and J. Sivic, ICCV 2019]

VideoBERT: learning multimodal video representation

- Learning from visual video and speech transcribed with ASR

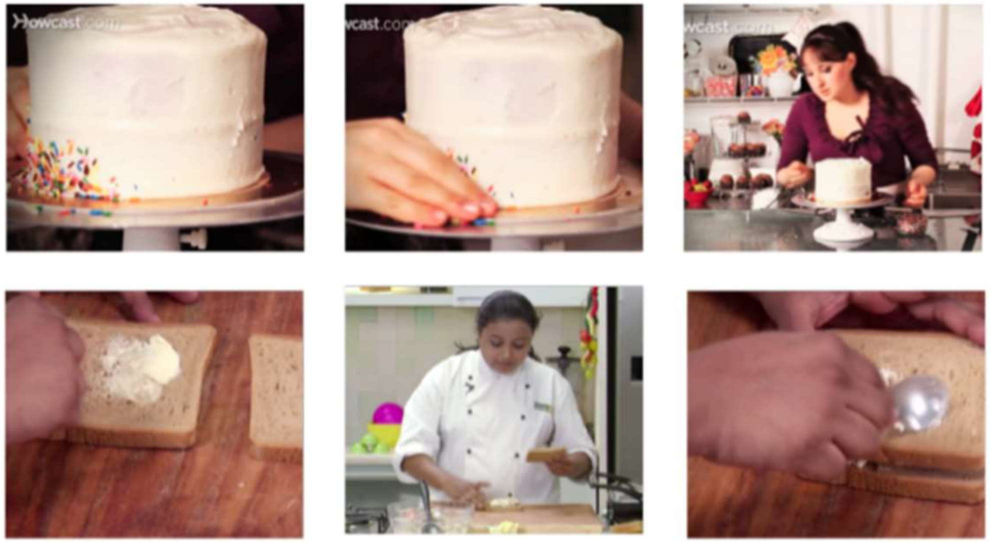
But in the meantime, you're just kind of moving around your cake board.

- BERT model learns correspondence between video and speech

- Learning from large-scale data without manual annotations

[VideoBERT, C. Sun et a., ICCV'19]

Large-scale training data without manual annotations

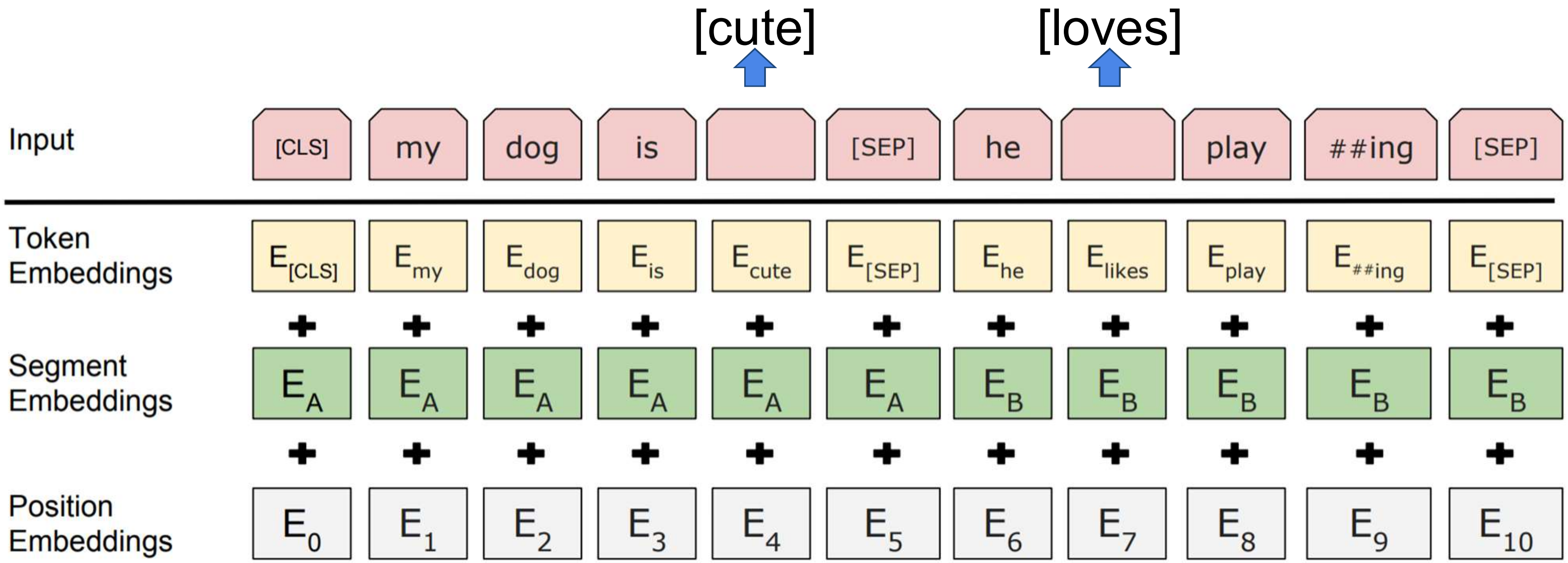


*“but in the meantime, you're just kind of **moving around** your **cake** board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party.”*

*“**apply** a little bit of **butter** on one side and place a portion of the stuffing and **spread** evenly **cover** with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches.”*

- ~320K *cooking/recipe* videos on YouTube
- ~1000 days in total, average length is ~4 mins
- ~120K videos with English ASR outputs

State-of-the-art for NLP: BERT



Two pre-training tasks:

- Masked language modeling
- Next sentence prediction

Network:

- Stacked Transformers
- Large amount of data

[1] Figure credit: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805

Google

Self-supervised pre-training for NLP

Input corpus:

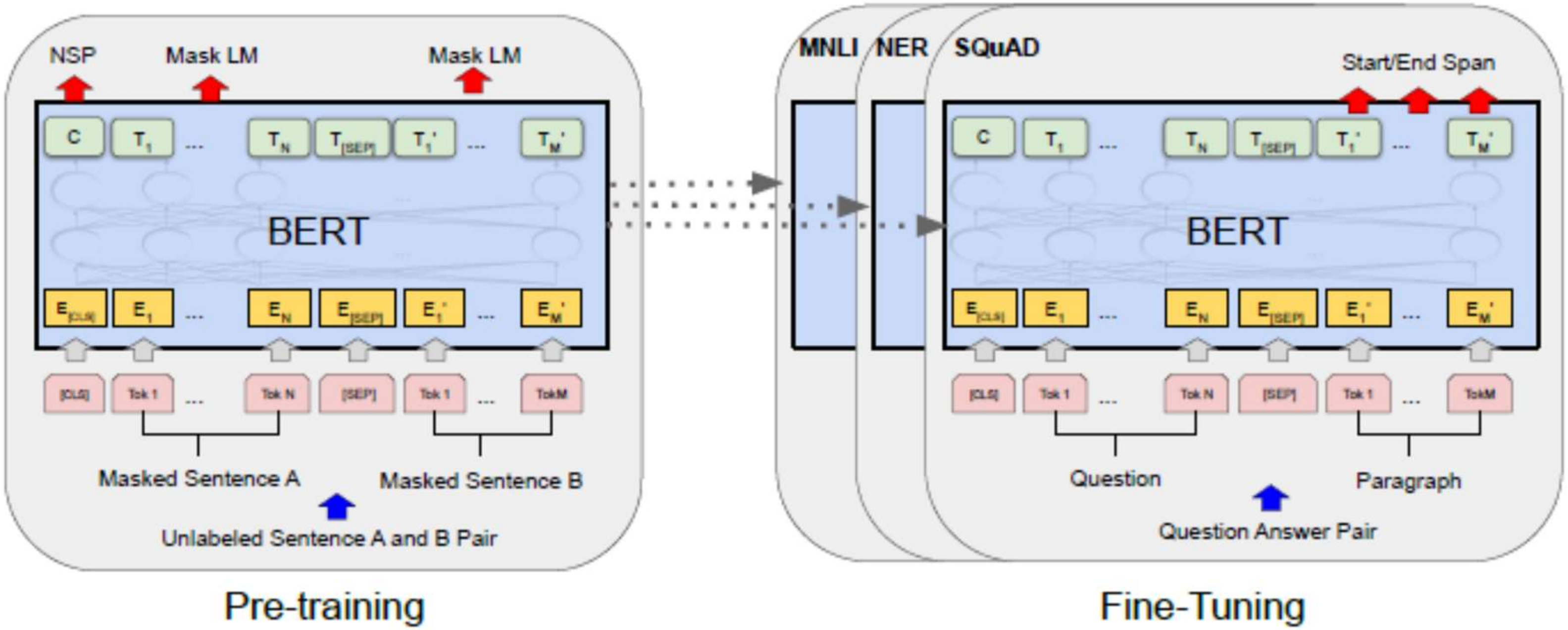
Apply a little bit of butter on one side and place a portion of the stuffing. Spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches.

Masked language modeling (MLM):

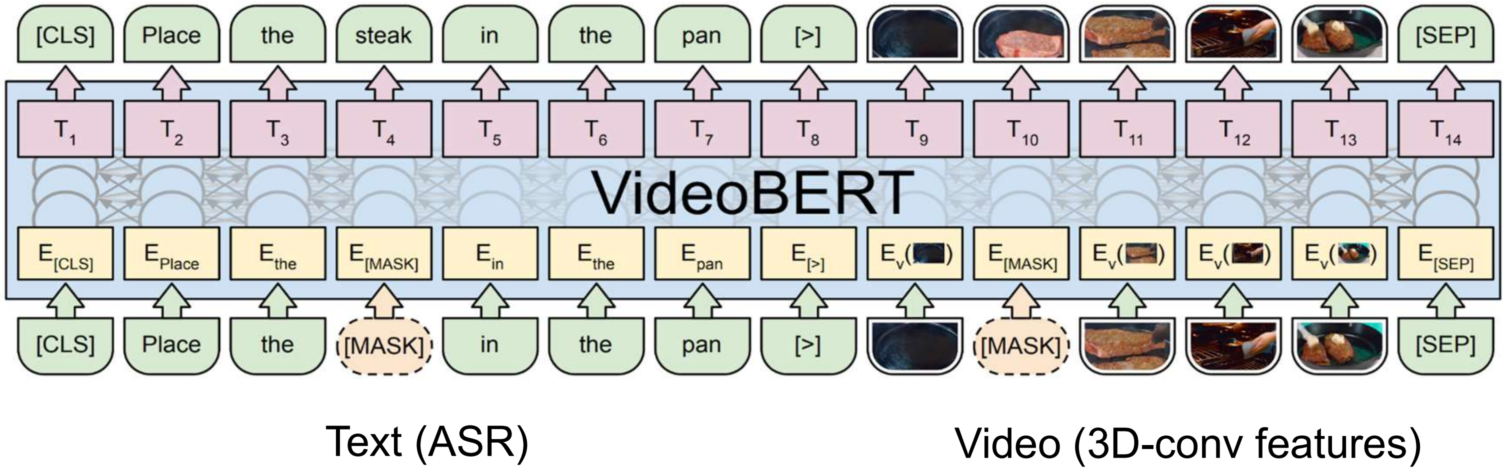
Apply a little bit of [mask] on [mask] side and place a portion of the stuffing. Spread [mask] cover with another slice of the [mask] and apply some more butter on top since we're gonna grill the [mask].

BERT model

- BERT: Bidirectional Encoder Representations from Transformers [Devlin et al., NAACL'19]



VideoBERT



- Multimodal transformer: excellent way of combining multiple modalities
- Masked ‘language’ modeling as in BERT, video-speech alignment
- Video representation with 3D-convolutions + clustering

Video representation

- 3D convolutions for 1.5 second video clips (S3D), 1024-dim features vector
- Video tokenization by clustering
- Hierarchical k-means: depth of 4, branch size of 12 (20736 clusters)
- High-level semantics preserved after tokenization

Original:

Centroids:

VideoBERT

Training on 300k cooking videos

"Keep rolling tight and squeeze the air out to its side"

Zero-shot prediction

Verb: make, **Noun:** pizza

Zero-shot prediction

Method	Verb (top-5 %)	Object (top-5 %)
S3D (supervised)	**46.9**	30.9
VideoBERT	43.3	**33.7**

Results on YouCook II dataset

Pre-training size	Verb (top-5 %)	Object (top-5 %)
10K	15.5	17.8
50K	15.7	27.3
100K	24.5	30.6
300K	**43.3**	**33.7**

- VideoBERT learns video-language correspondence
- Close to fully-supervised accuracy
- More data improves the performance (not saturated yet)

Fine-tuning on downstream tasks

- For captioning cooking video on YouCook2

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou et al. (CVPR'18)	-	1.42	11.20	-	-
S3D	6.12	3.24	10.00	26.05	0.35
VideoBERT	6.80	4.07	10.99	27.51	0.50

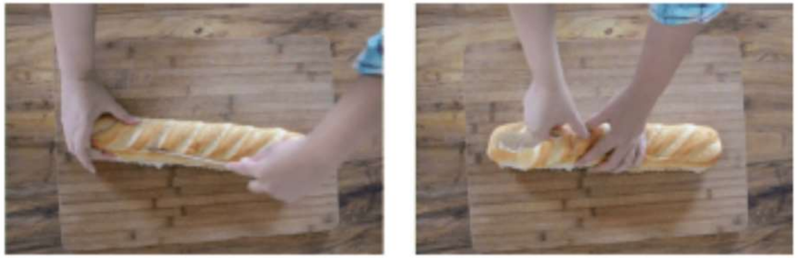
- Effective and outperforms S3D features
- Pre-training helps!

Video captioning - examples

GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices



GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan

GT: cut yu choy into diagonally medium pieces

VideoBERT: chop the cabbage

S3D: cut the roll into thin slices

GT: remove the calamari and set it on paper towel

VideoBERT: fry the squid in the pan

S3D: add the noodles to the pot

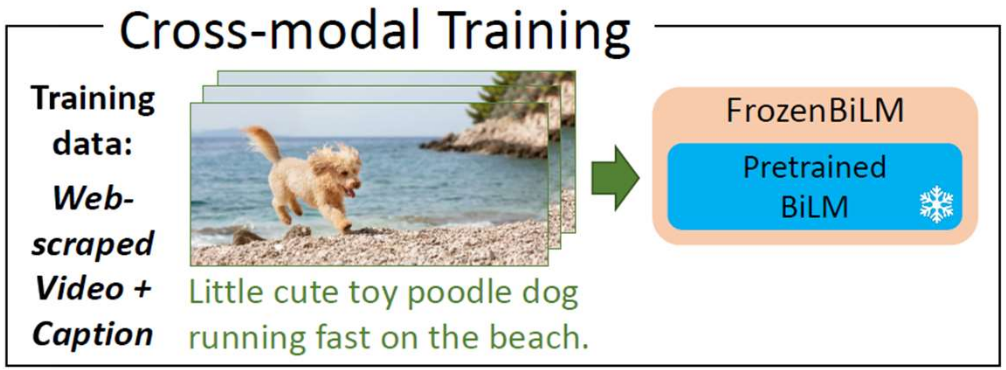
Multimodal transformers – different models / tasks

- Image / video question answering

Example model: FrozenBlim [A.Yang et al., Neurips'22]

Frozen Bidirectional Language Model (BiLM)

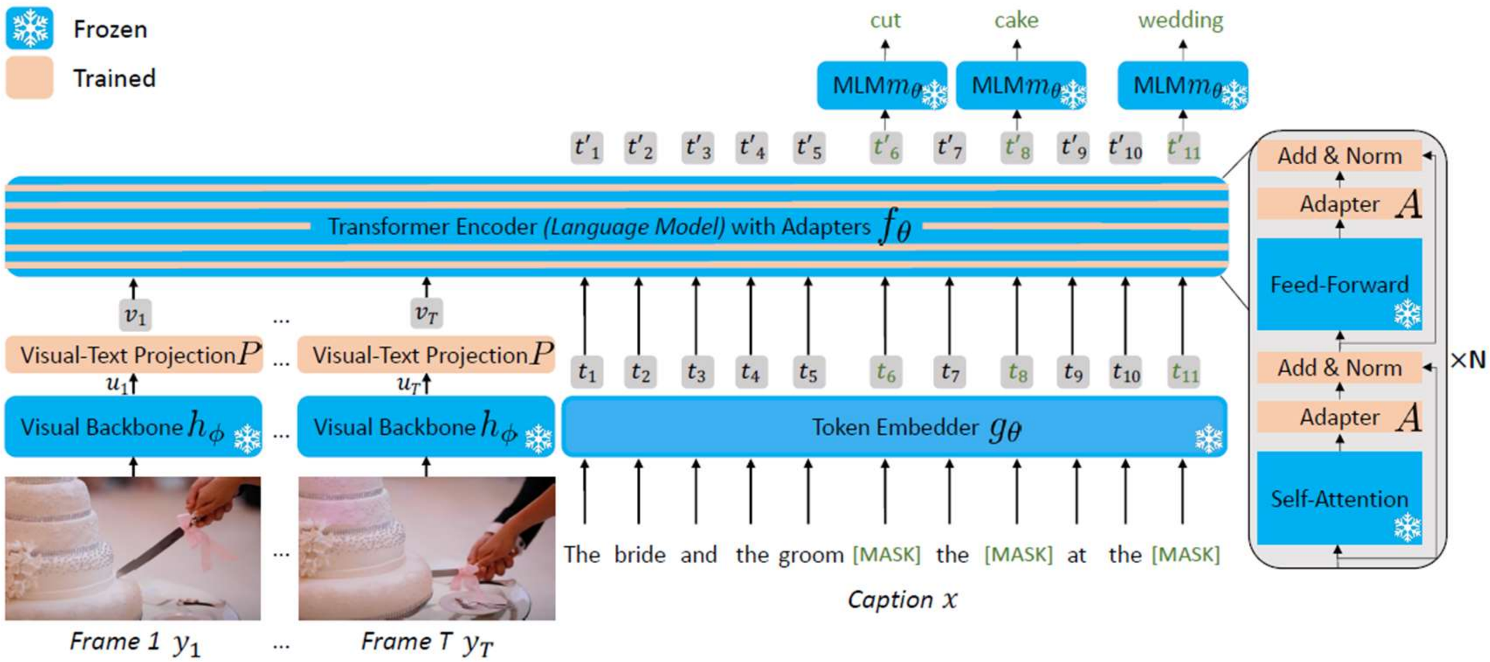
- Pre-trained large-scale language model + adapters



- Adapters are trained on web-scraped video/caption dataset
  - WebVid10M dataset with 10M video-text pairs

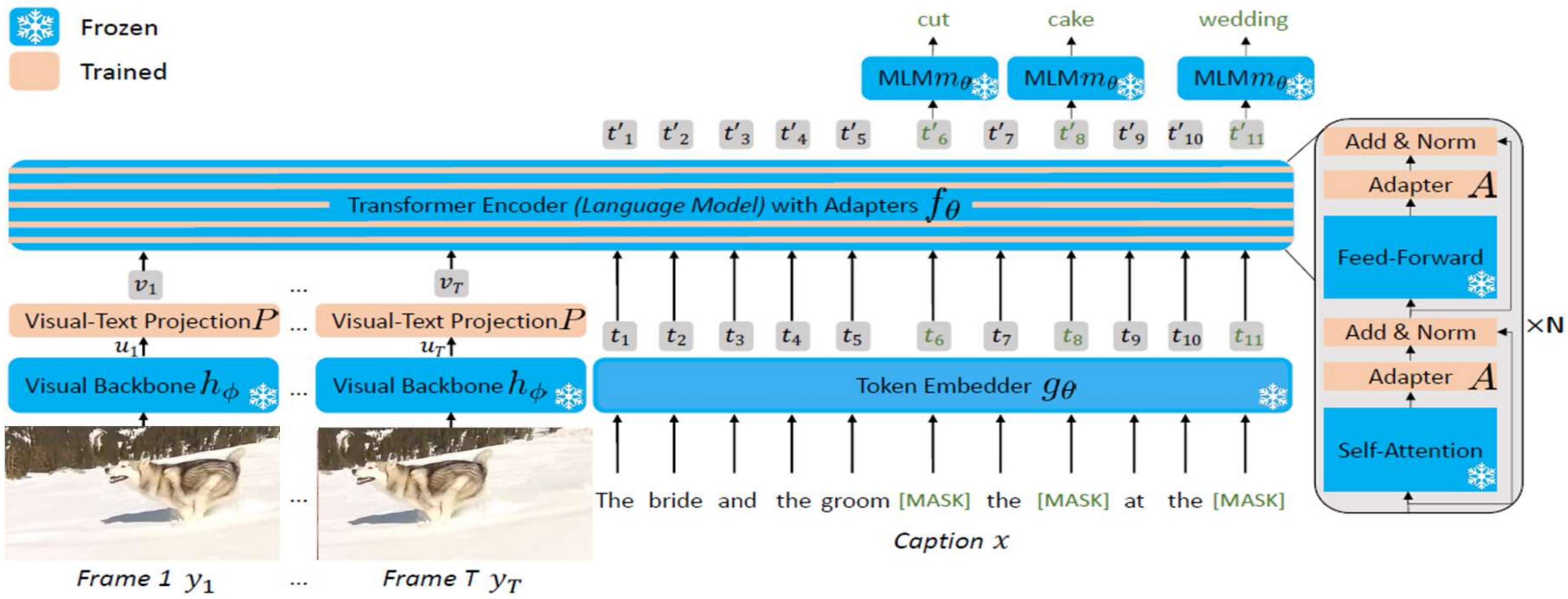
[Zero-shot video question answering via frozen bidirectional language models. A. Yang et al., Neurips'22l]

FrozenBiLM



- Linear mapping from the visual features to the text token embedding space
- Adapter: insert a multi-layer perceptron and add a residual connection
- Trained on web-scraped WebVid10M dataset with 10M video-text pairs

FrozenBiLM: Zero-Shot QA



Input prompt engineering

Open-ended VideoQA “[CLS] **Question:** <Question>? **Answer:** [MASK]. [SEP]”

Multiple-choice VideoQA “[CLS] **Question:** <Question>? **Is it** ’<Answer Candidate>’? [MASK]. [SEP]”

Video-conditioned fill-in-the-blank task “[CLS] <Sentence with a [MASK] token>. [SEP]”

Experimental results: ablation

- Zero-shot performance; no downstream training data is used; use of WebVid10M for training the adapter layers

- Ablation of different components of frozen BiLM

LM Pretraining	*Frozen* LM	Adapters	Fill-in-the-blank LSMDC	Open-ended iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	Multiple-choice How2QA	TVQA
✗	✗	✗	0.5	0.3	0.1	0.0	0.5	0.0	32.4	20.7
✓	✗	✗	37.1	21.0	**17.6**	31.9	20.7	30.7	45.7	45.6
✓	✓	✗	50.7	**27.3**	16.8	32.2	24.7	41.0	53.5	53.4
✓	✓	✓	**51.5**	26.8	16.7	**33.8**	**25.9**	**41.9**	**58.4**	**59.2**

- Pre-training is important
- Linear layer projection works well, adapter layers show additional gain

Experimental results: SOTA comparison

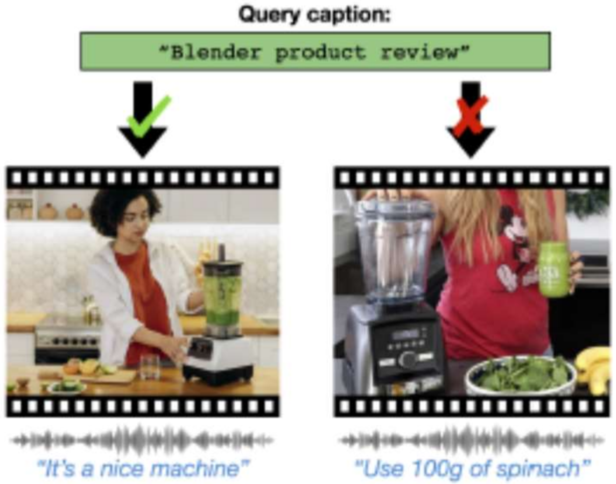
- Comparison to the SOTA on zero-shot VQA

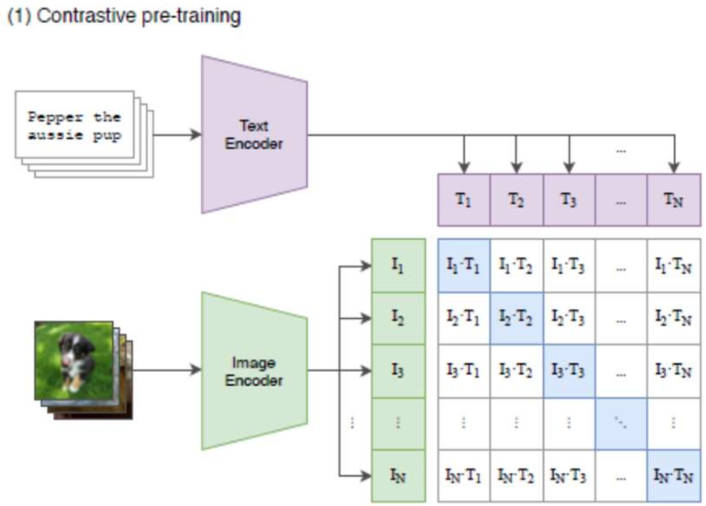
Method	Training Data	Fill-in-the-blank	Open-ended					Multiple-choice	
		LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
Random	—	0.1	0.1	0.1	0.1	0.1	0.1	25	20
CLIP ViT-L/14 [75]	400M image-texts	1.2	9.2	2.1	7.2	1.2	3.6	47.7	26.1
Just Ask [108]	HowToVQA69M + WebVidVQA3M	—	13.3	5.6	13.5	12.3	—	53.1	—
Reserve [116]	YT-Temporal-1B	31.0	—	5.8	—	—	—	—	—
FrozenBiLM (Ours)	WebVid10M	**51.5**	**26.8**	**16.7**	**33.8**	**25.9**	**41.9**	**58.4**	**59.7**

Question: where is the woman sitting on?
GT Answer: camal
JustAsk: horseyard
UnFrozenBiLM: desert
FrozenBiLM (text-only): chair
FrozenBiLM (ours): camel

Multimodal transformers – different models / tasks

- Text/ image/video retrieval (CLIP)





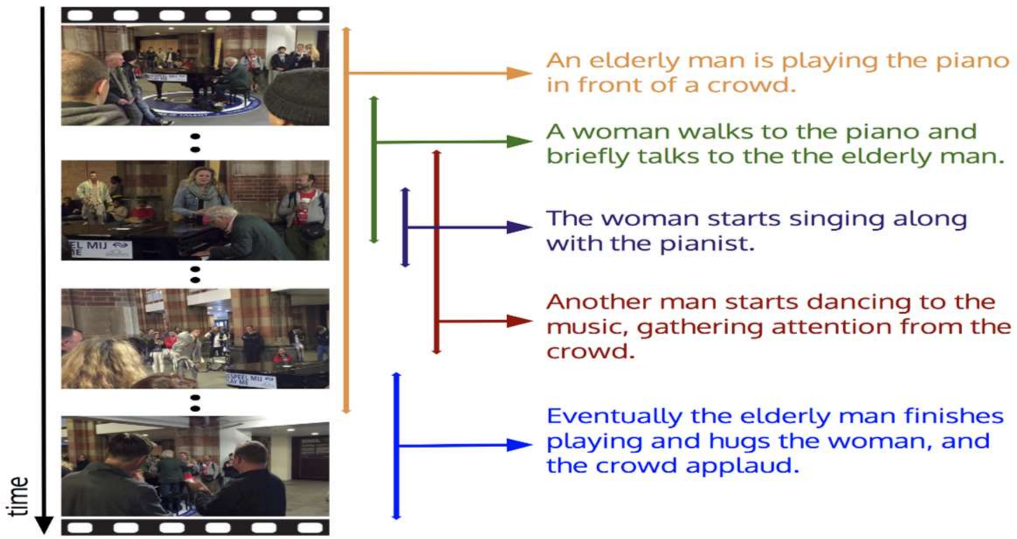
Contrastive training

- Image / video captioning (Vid2Seq)

Dense video captioning - task

Video captioning models for long videos with multiple events

- Captions are grounded in the video
- Combines localization and text generation



Example of dense, overlapping captions from the ActivityNet dataset

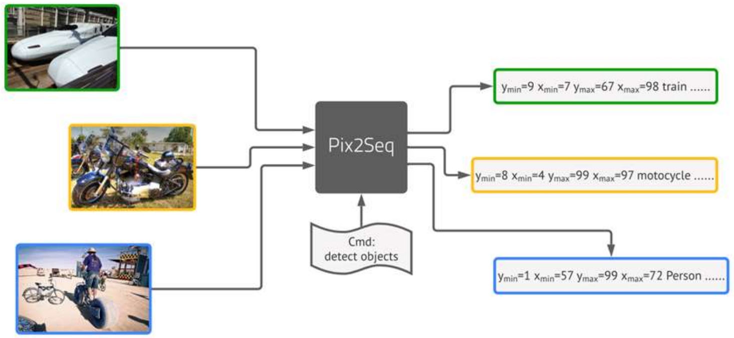
Dense video captioning – SOTA

Current approaches for dense video captioning

- Train separate networks for localization and captioning
- Require task-specific components like event counters
- Train on manually annotated datasets (small)
- Cannot reason over *long* videos

Localization as language modeling

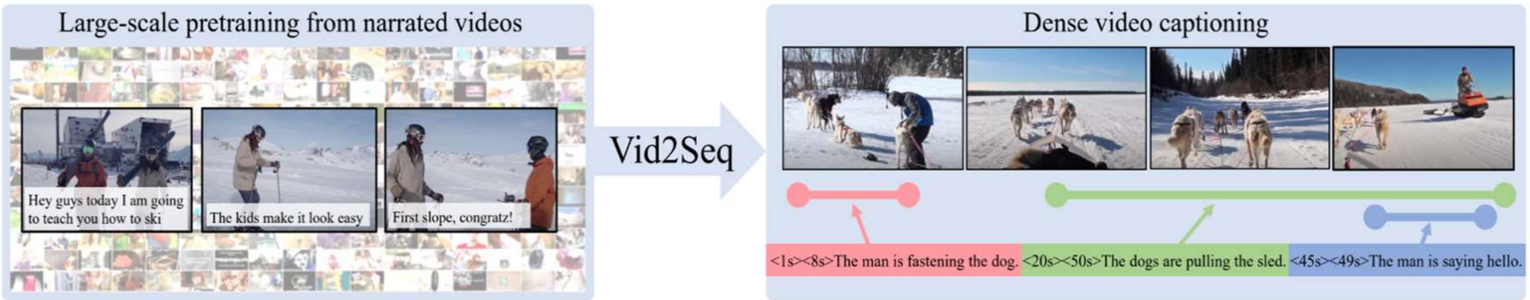
- Pix2seq casts object detection as sequence generation
- Spatial coordinates are quantized and tokenized



Vid2Seq approach

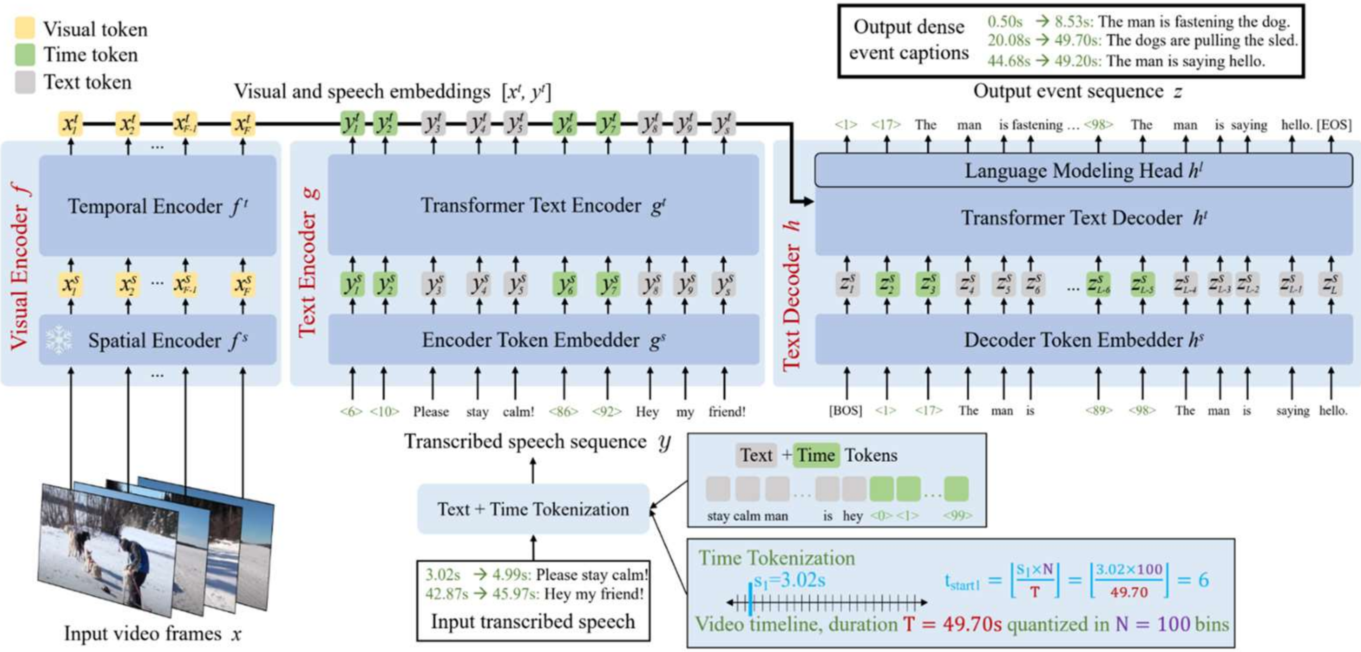
- Single target sequence consists of **Text + Time tokens combining localization + captioning**

- Large-scale pretraining from narrated untrimmed videos



[Vid2Seq, A. Yang et al., CVPR 2023]

Vid2Seq – model



- Frozen Visual backbone (CLIP)
- Temporal Encoder for video
- Speech is cast as a single sequence of text and time tokens
- T5 Encoder & Decoder

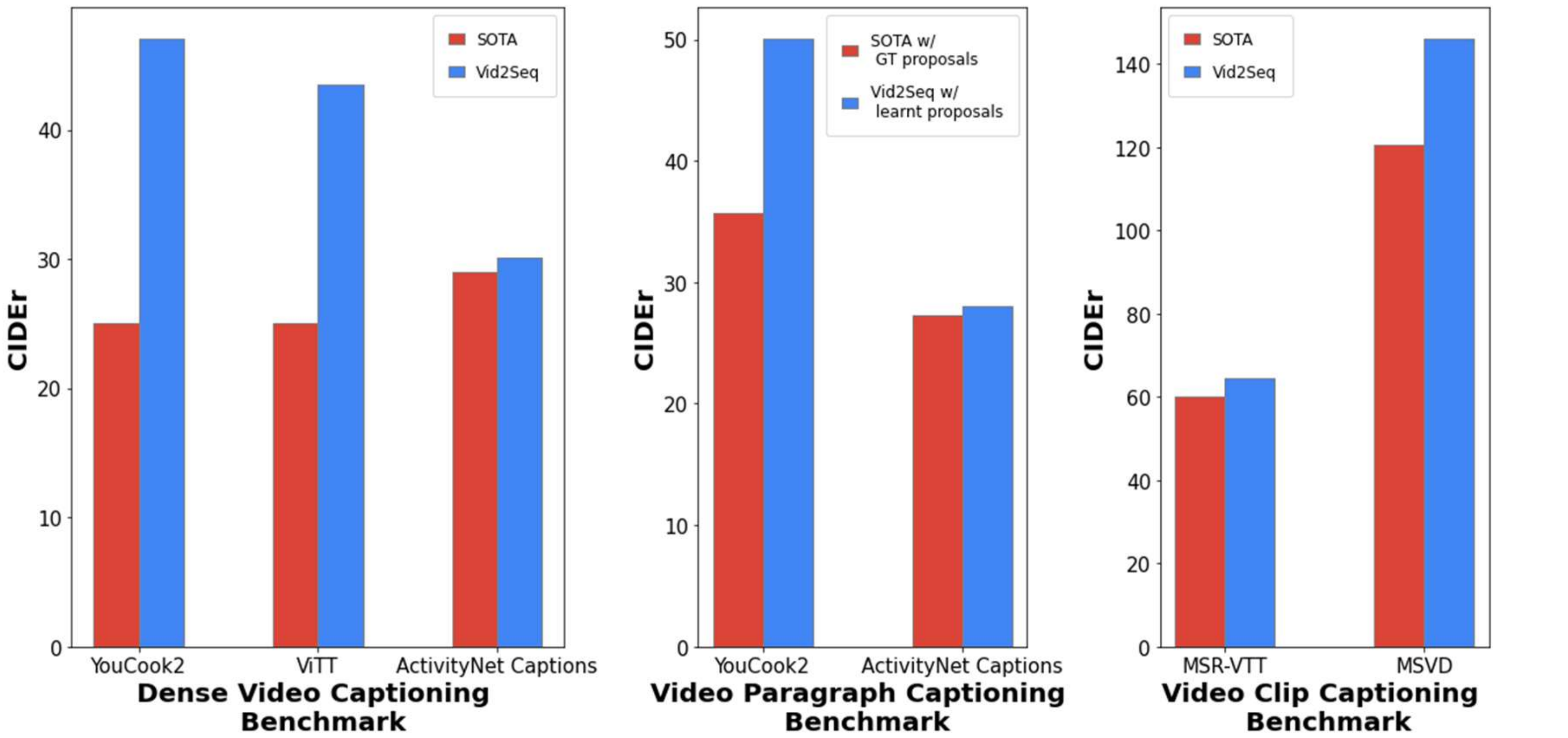
Vid2Seq – large-scale pretraining

- Pretraining dataset is 15 million YouTube narrated videos from YT-Temporal-1B
- ASR sentence boundaries used

as event boundaries

- Generative loss: given visual input predict speech
- Denoising loss: given visual input and masked speech, predict the masked tokens

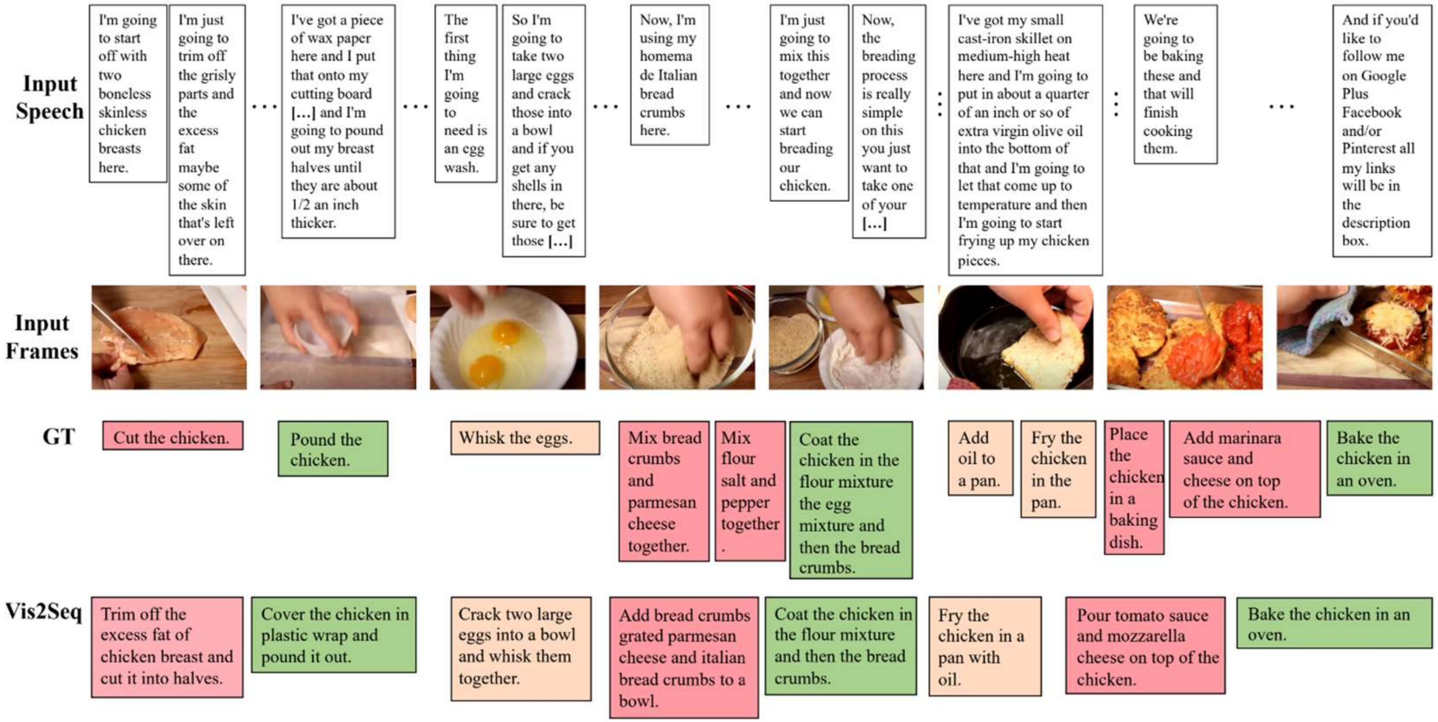
Vid2Seq – SOTA results



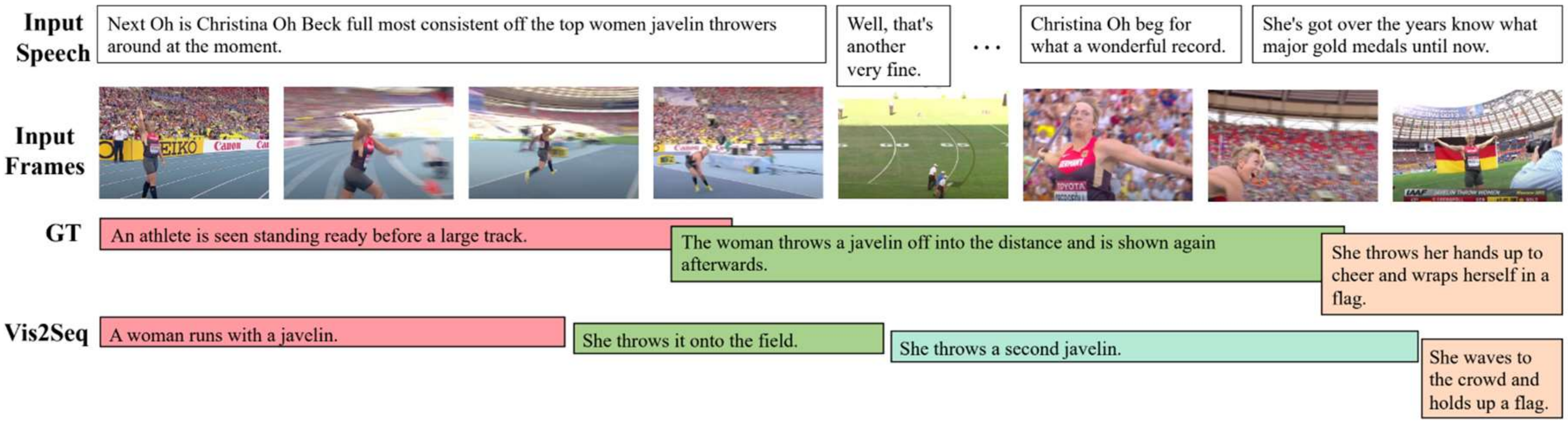
Ablation studies

- Pretraining is important, datasize and quality matter
- Time tokens help when pretraining on untrimmed videos
- Visual and speech information is complementary
- Importance of losses: denoising loss is important if we use speech during pretraining

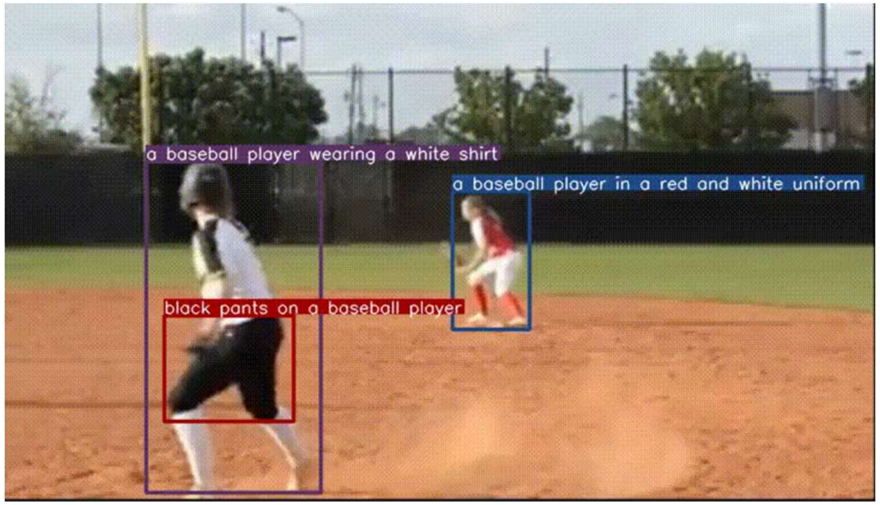
Qualitative results



Qualitative results



Dense Video Object Captioning



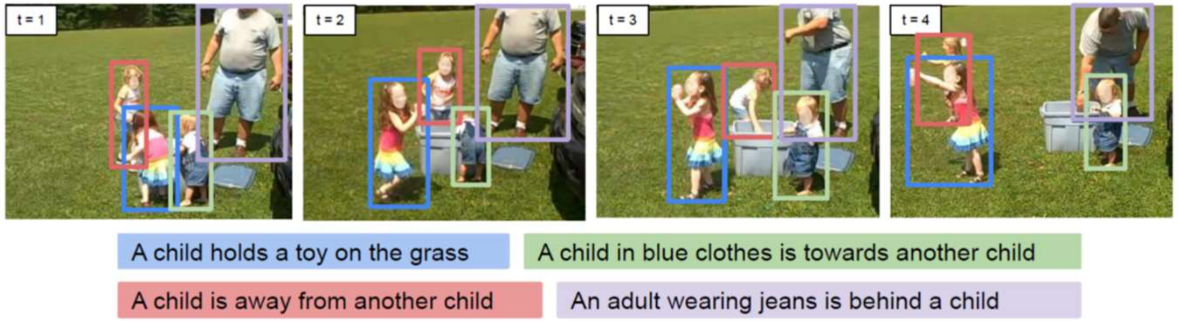
Detect, track and describe all objects in a video

→Object-centric video description / captioning

→Video object grounding

Dense video object captioning - task definition

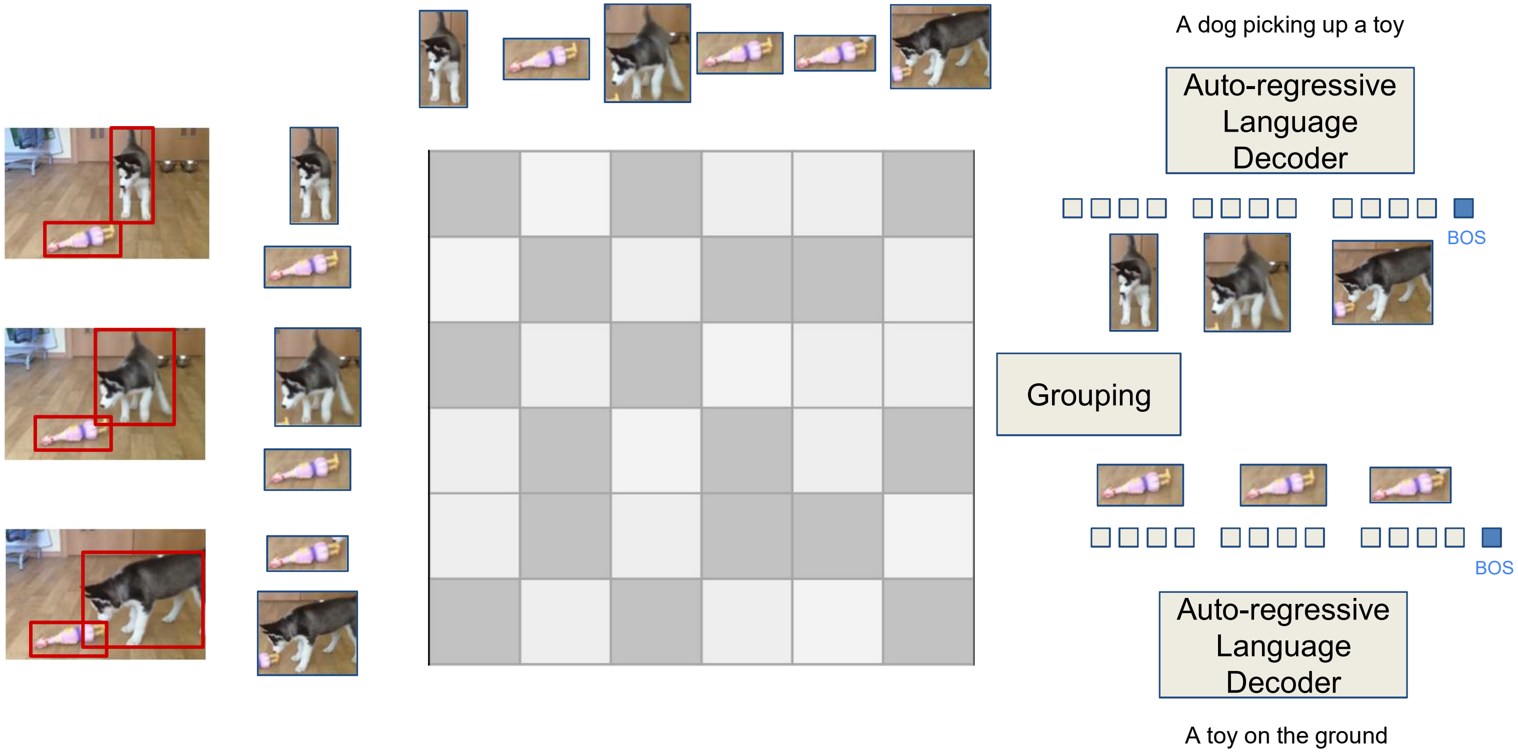
- Detect, track and caption objects



- Extension of the state-of-the-art multi-object tracking metric HOTA to include a captioning accuracy

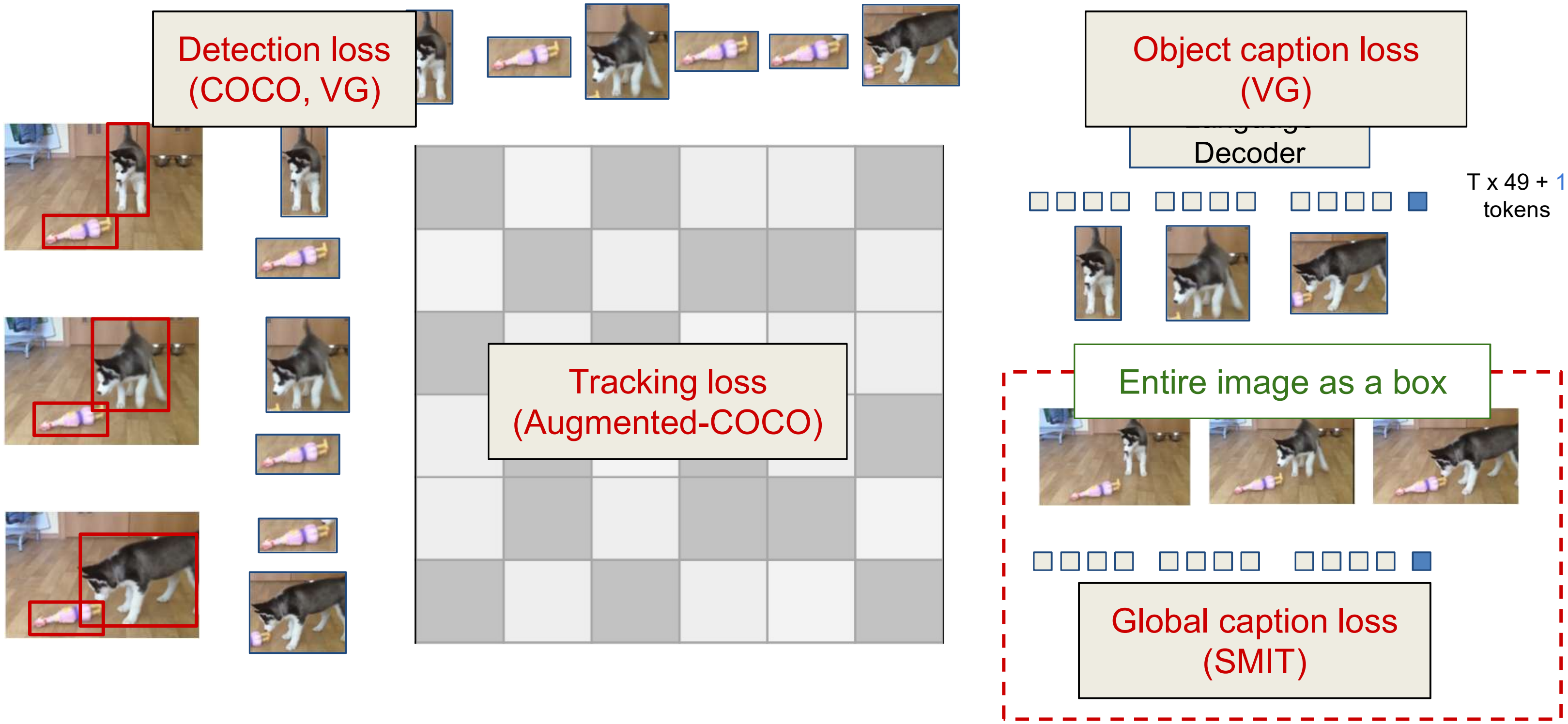
[Dense Video Object Captioning from Disjoint Supervision, X. Zhou et al., arXiv'23]

End-to-end video object tracking & captioning

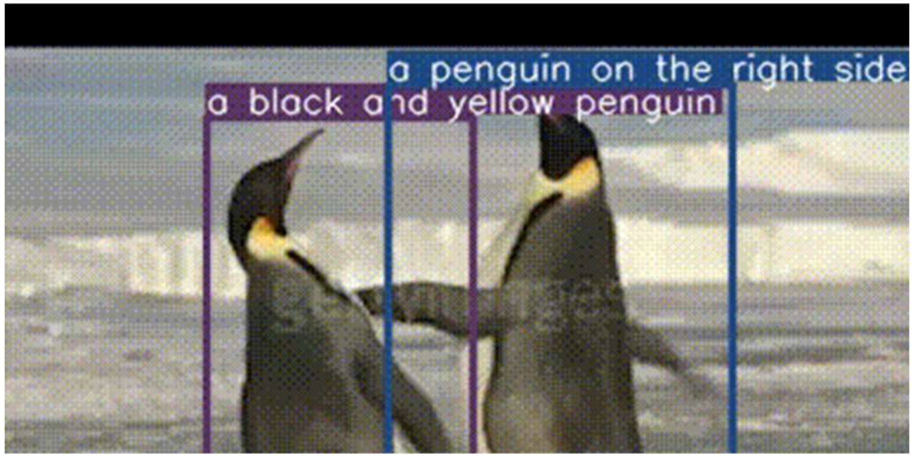


[Wu et al, GRiT: A Generative Region-to-text Transformer for Object Understanding, arXiv 2022]

End-to-end video object tracking & captioning

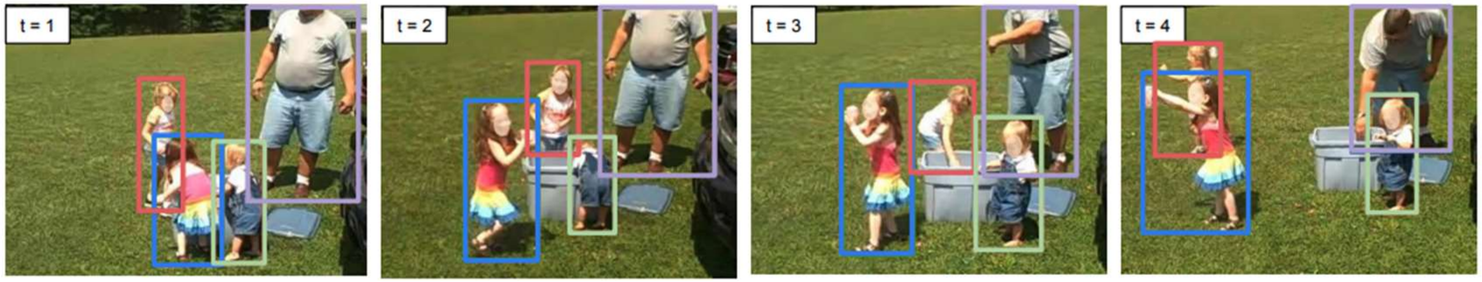


Qualitative results



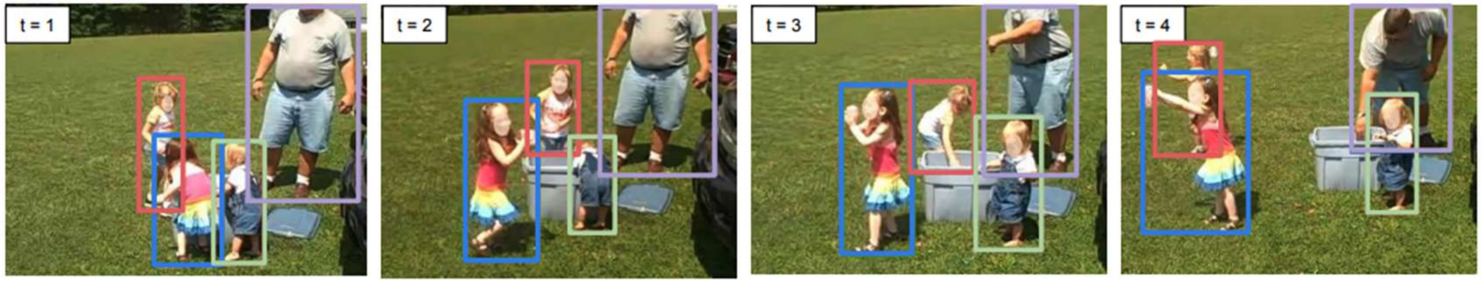
Application to video grounding

Query: q = “A child holds a toy on the grass”



Application to video grounding

Query: q = “A child holds a toy on the grass”

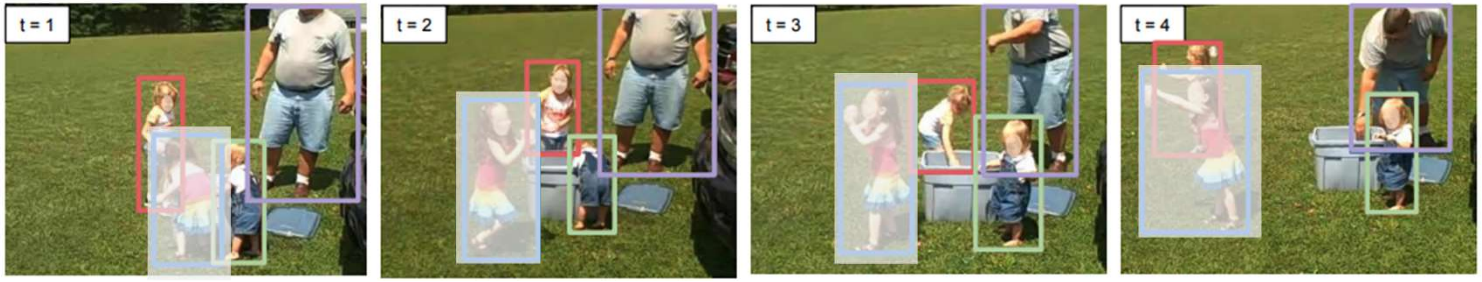


likelihood(, q) = 0.9 likelihood(, q) = 0.5

likelihood(, q) = 0.4 likelihood(, q) = 0.1

Application to video grounding

Query: q = “A child holds a toy on the grass”



likelihood(, q) = 0.9

likelihood(, q) = 0.5

likelihood(, q) = 0.4

likelihood(, q) = 0.1

Video grounding results

	Finetuned	Zero-shot
STVGBert [52]	47.3	-
TubeDETR [66]	59.0	-
STCAT [29]	61.7	-
Ours	**61.9**	54.1

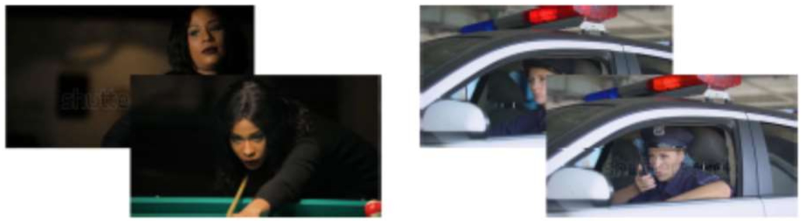
VidSTG spatial-grounding

Average intersection over union with GT (IoU)

Multimodal data for generating automatic training data

- Large-scale weakly supervised data
  - HowTo100M dataset with 100M video-ASR pairs [HowTo100M. A. Miech et al., ICCV'19]

  - WebVid10M dataset with 10M video-text pairs [Frozen In Time, M. Bain et al., ICCV'21]



"Billiards, concentrated young woman playing in club"

"Female cop talking on walkie-talkie, responding emergency call, crime prevention"

Multimodal data for generating automatic training data

- Cross-modal supervision
  - Speech2Action for mining clips
  - Levering text model for annotating clips with question/answers
- Data Mining
  - Transfer of image captions to video

Speech2Action: Cross-modal supervision

Train Speech to Action Classifier with Movie Screenplays

```
                    PETER
          Yes, it is him.
    Agent #1 hands him the phone.
                    PETER
          Hello, yes, operator, I accept the
          charges.
    Agent #1 gestures to Agent #3 to take a look around the
    apartment. Agent #3 slips away.
                    AGENT #1
          Would you mind very much if I
          listened?
                    PETER
          Please, go right ahead.
```

Speech2Action classifier

Speech is input to the action classifier

Scene description

Actions labels are obtained from scene descriptions

Weak label with Speech2Action Classifier

Hello, it's me

Speech2Action classifier

Weak label: [answer] phone

[Speech2Action, A. Nagrani et al., CVPR'20]

run

don’t move, hey!

mike, run, run!

he was running after

he is running away.

Chase him!

They ran into the

phone

[beeps] hello.

rebekah is not answering her calls.

skinner's not answering his phone.

dad, are you there ?

hey, it's me.

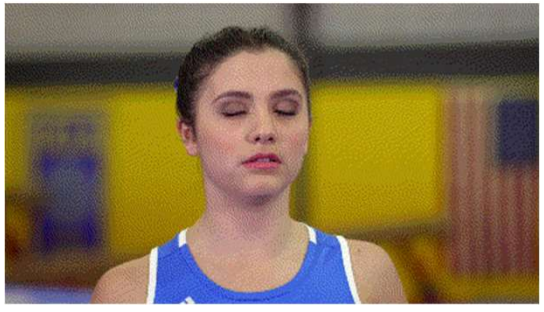
(phone line ringing) hello.

hit

i'm gonna smash that camera to bits!

you gotta hit him in the solar plexus!

backhand, snap down, round off reach into the back handspring, and then tuck.



hit him right between the eyes.

drive

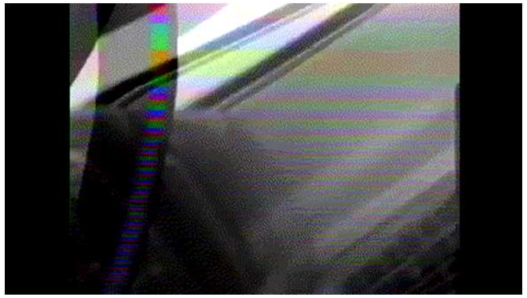
camaro headed east on ocean park.

they stopped under the brooklyn queens expressway.

he made a u turn on an empty street.

he got back in his car and chased after her.

my wife gets in the car i start driving down my block to the corner.



shoot

you got 10 seconds to come out, or we start shooting.

you need more arc in that shot.

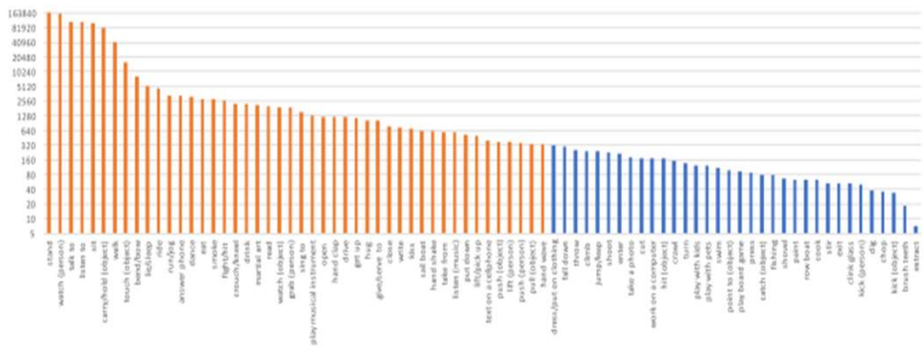
with the sharps carbine, that is within range.

kincaid ordered not to shoot.

Result - many examples of rare actions

- Long tail of natural distribution of actions
- Mines 2 orders of magnitude more training examples for rare/mid classes in AVA

Log scale!



AVA dataset

Log scale!

Results - directly evaluate on AVA

Data	Per-Class AP													
	drive	phone	kiss	dance	eat	drink	run	point	open	hit	shoot	push	hug	enter
AVA (fully supervised)	0.63	0.54	0.22	0.46	0.67	0.27	0.66	0.02	0.49	0.62	0.08	0.09	0.29	0.14
KS-baseline †	0.67	0.20	0.12	0.53	0.67	0.18	0.37	0.00	0.33	0.47	0.05	0.03	0.10	0.02
`S2A-mined` (zero-shot)	0.83	0.79	0.13	0.55	0.68	0.30	0.63	0.04	0.52	0.54	0.18	0.04	0.07	0.04
`S2A-mined` + AVA	0.84	0.83	0.18	0.56	0.75	0.40	0.74	0.05	0.56	0.64	0.23	0.07	0.17	0.04
AVA (few-shot)-20	0.82	0.83	0.22	0.55	0.69	0.33	0.64	0.04	0.51	0.59	0.20	0.06	0.19	0.13
AVA (few-shot)-50	0.82	0.85	0.26	0.56	0.70	0.37	0.69	0.04	0.52	0.65	0.21	0.06	0.19	0.15
AVA (few-shot)-100	0.84	0.86	0.30	**0.58**	0.71	0.39	**0.75**	**0.05**	0.58	0.73	0.25	**0.13**	0.27	0.15
AVA (all)	**0.86**	**0.89**	**0.34**	**0.58**	**0.78**	**0.42**	**0.75**	0.03	**0.65**	**0.72**	**0.26**	**0.13**	**0.36**	**0.16**

- For 8 out of 14 classes, exceed fully supervised performance without a single training example
- With fine-tuning, exceed supervised performance for all classes

More abstract actions

two quarters, three dimes, one nickel, two pennies

twenty four thousand four hundred

thirty six thousand four hundred, five hundred

COUNT

after you

follow me quick!

come right behind me!

FOLLOW

Cross-model supervision: JustAsk

- Learning zero-shot video question answering with cross-modal supervision

Question: What type of animal do we see?

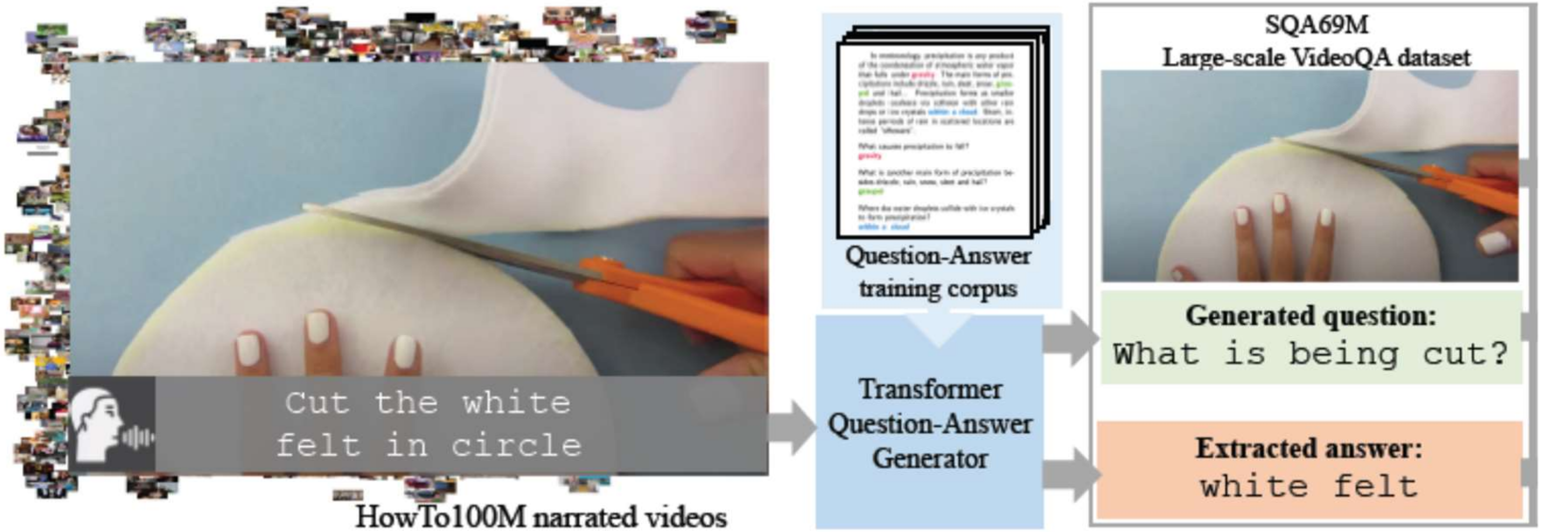
Our answer: Fish.

- Generate a large-scale video question answering dataset automatically (HowToVQA69M)

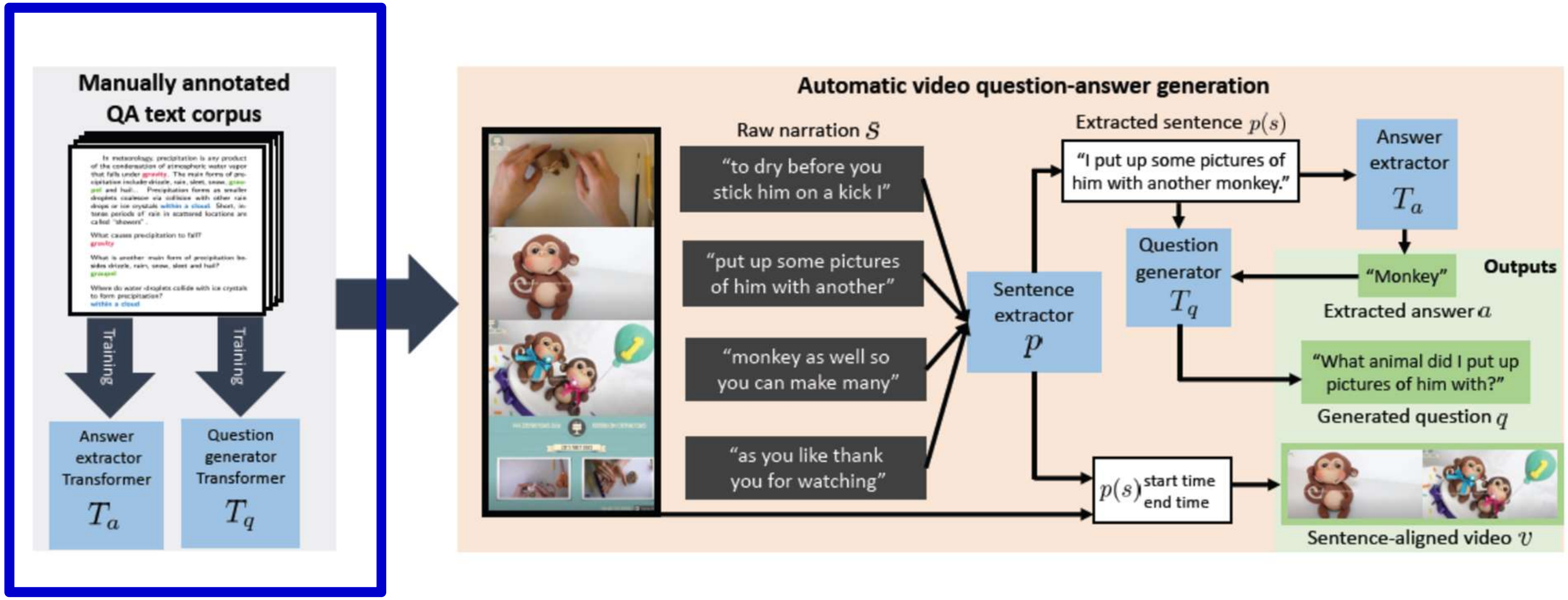
[JustAsk, A. Yang et al., ICCV'21]

Cross-modal supervision: JustAsk

- HowTo100M dataset with ASR captions
- Textual question-answer training corpus + transformer model
- Transformer extracts answer + question from ASR caption

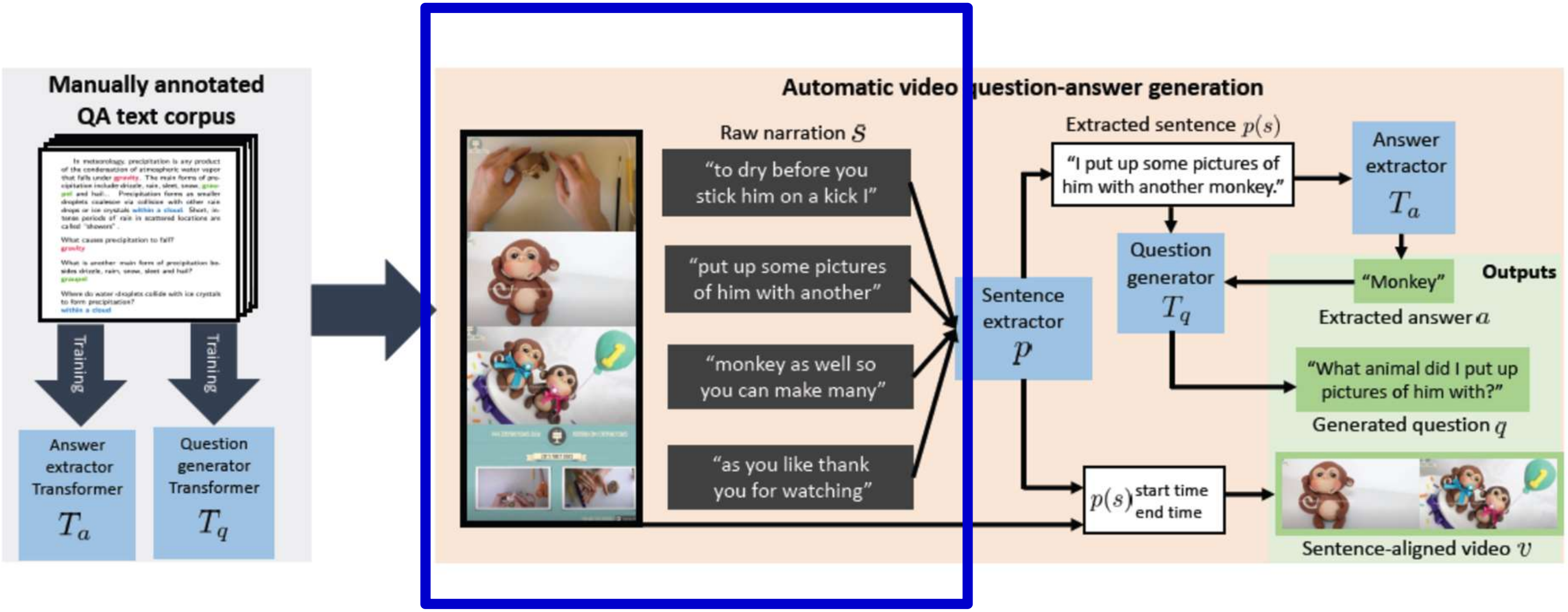


Cross-model training



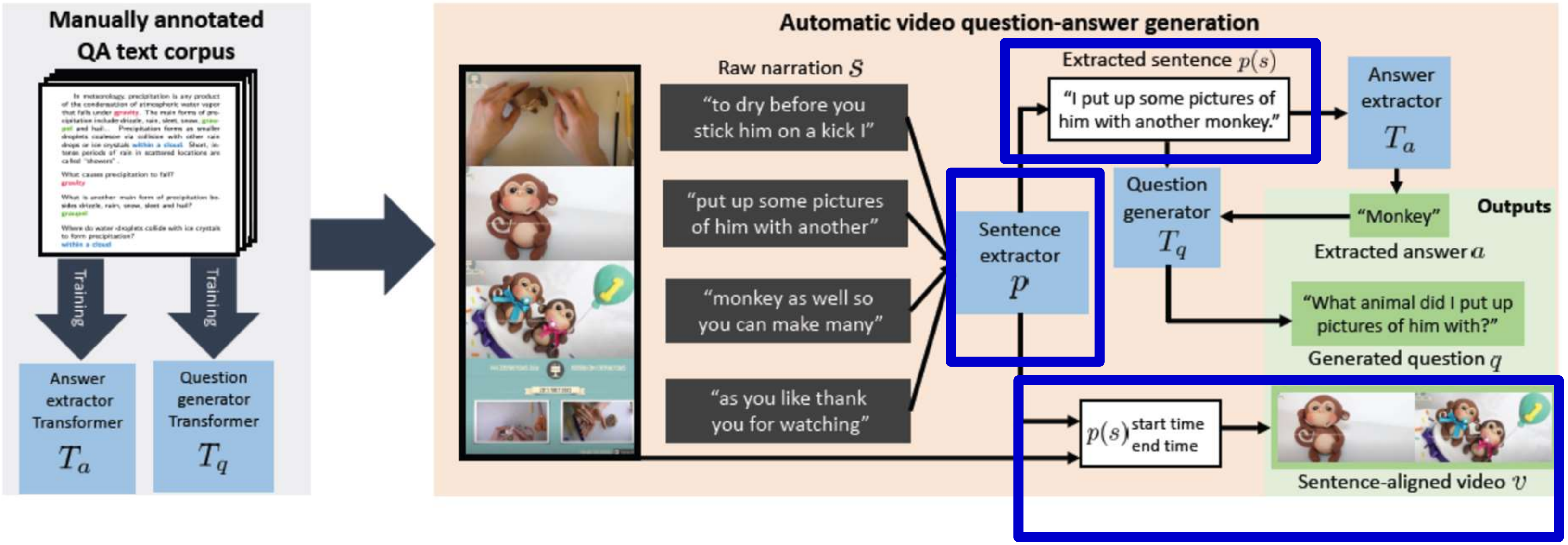
- Manually annotated QA text corpus: SQuADv1
  - 100k question-answer pairs for paragraphs from Wikipedia articles
- Transformers Ta and Tq are trained for answer extraction and answer-aware question extraction on SquADv1

Cross-model training



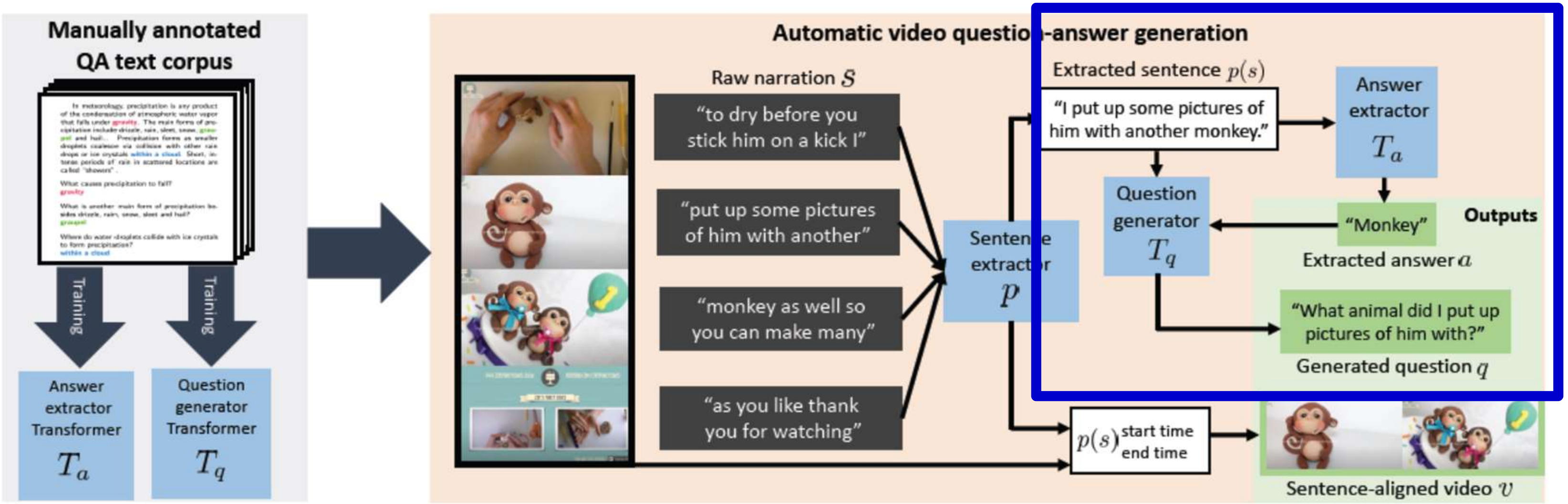
- HowTo100M clips + speech transcribed with ASR

Cross-model training



- HowTo100M clips + speech transcribed with ASR
- Sentence / punctuation extraction with recurrent network
  - Sentence aligned video

Cross-model training



- HowTo100M clips + speech transcribed with ASR
- Sentence / punctuation extraction with recurrent network
  - Sentence aligned video
- Answer + Question extraction with Ta and Tq

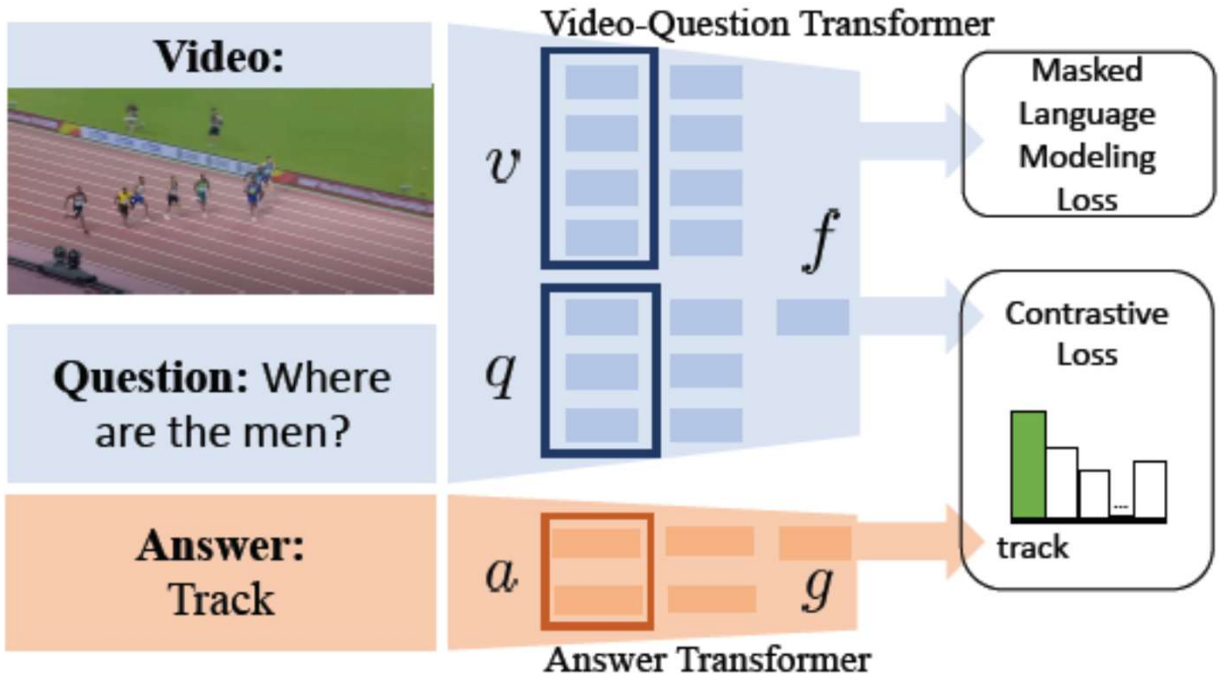
Example of generated question-answer

ASR: Add some of your favorite sprinkles give it a mix.

Generated question: What can you add to the mix?

Generated answer: Sprinkles.

VideoQA architecture



$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(v_i,q_i)^\top g(a_i)}}{e^{f(v_i,q_i)^\top g(a_i)} + \sum_{(v',q',a')\sim\mathcal{N}_i} e^{f(v',q')^\top g(a')}} \right)$$

- Multi-modal transformer
- Contrastive loss with positive and negative answers
  - Can deal with large-scale data, here 16M different answers

Zero-shot VQA

- No use of any annotated examples for training
- Results on state-of-the-art datasets, use of test data only

Pretraining	iVQA Top 1	iVQA Top10	MSVD-QA Top 1	MSVD-QA Top 10
Random	0.09	0.9	0.05	0.5
HowToVQA69M	12.2	43.3	7.5	22.4

Zero-shot results

Question: What is the largest object at the right of the man?

Our answer: Wheelbarrow.

[Text only: Statue.]

Impact of training data

- Results on state-of-the-art dataset with training data

Pretraining	iVQA Top 1	iVQA Top10	MSVD-QA Top 1	MSVD-QA Top 10
Zero-shot HowToVQA69M	12.2	43.3	7.5	22.4
Training w/o pretraining	23.0		41.2	
Training with pretraining HowTOVQA69M	35.4		46.3	

Impact of pretraining data size

Pretraining data size	Zero-shot		Finetune	
	iVQA	MSVD-QA	iVQA	MSVD-QA
0%	—	—	23.0	41.2
1%	4.5	3.6	24.2	42.8
10%	9.1	6.2	29.2	44.4
20%	9.5	6.8	31.3	44.8
50%	11.3	7.3	32.8	45.5
100%	**12.2**	**7.5**	**35.4**	**46.3**

- Amount of pretraining data impacts performance
- Not yet saturated

Video/audio – text dataset

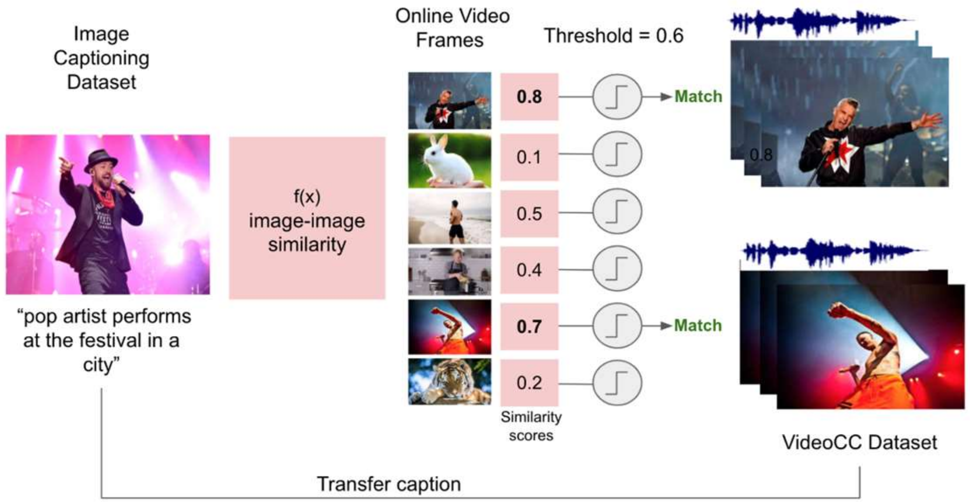
Existing datasets

	Video - Text	*Audio - Text*
Manually Labelled Expensive, time-consuming, => small	ActivityNet-captions, MSR-VTT, MSVD, YouCook2, etc	AudioCaps, CLOTHO
Semi-automatic/automatic Weak, noisy => require millions of samples to get good performance => text is not really a 'caption'	HowTo100M, WebVideoText, Instagram Hashtags,	*None*

Image captioning datasets, however, such as Conceptual Captions are large (millions), and relatively clean

Transfer image captions to video and audio

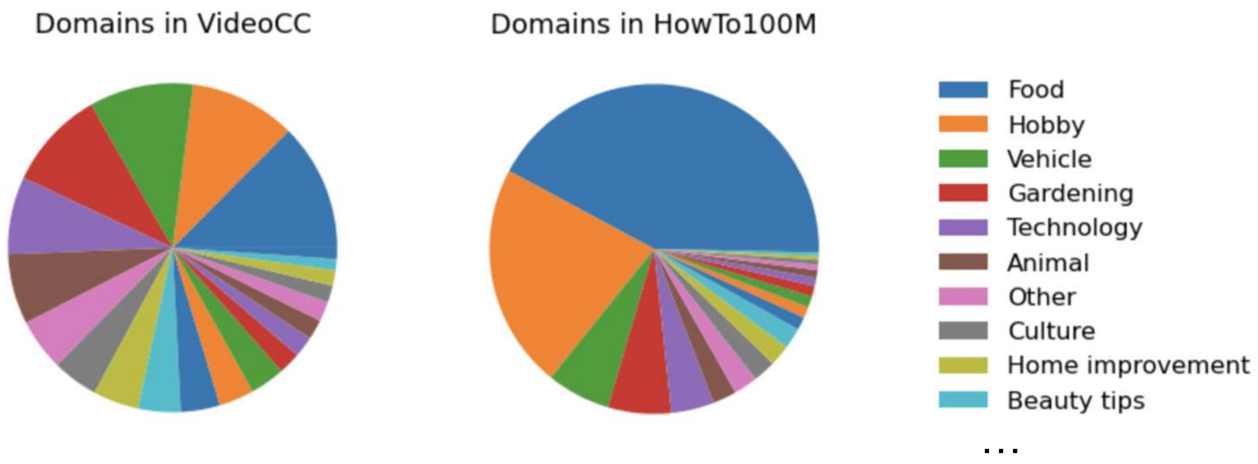
- Start with a seed image-captioning dataset, large-scale relatively clean image caption dataset available, i.e., Conceptual Captions
- Find frames in videos with high similarity scores to the seed image
- Extract short video clips around the matching frames and transfer the caption



[Learning Audio-Video Modalities from Image Captions, A. Nagrani et al., ECCV'22]

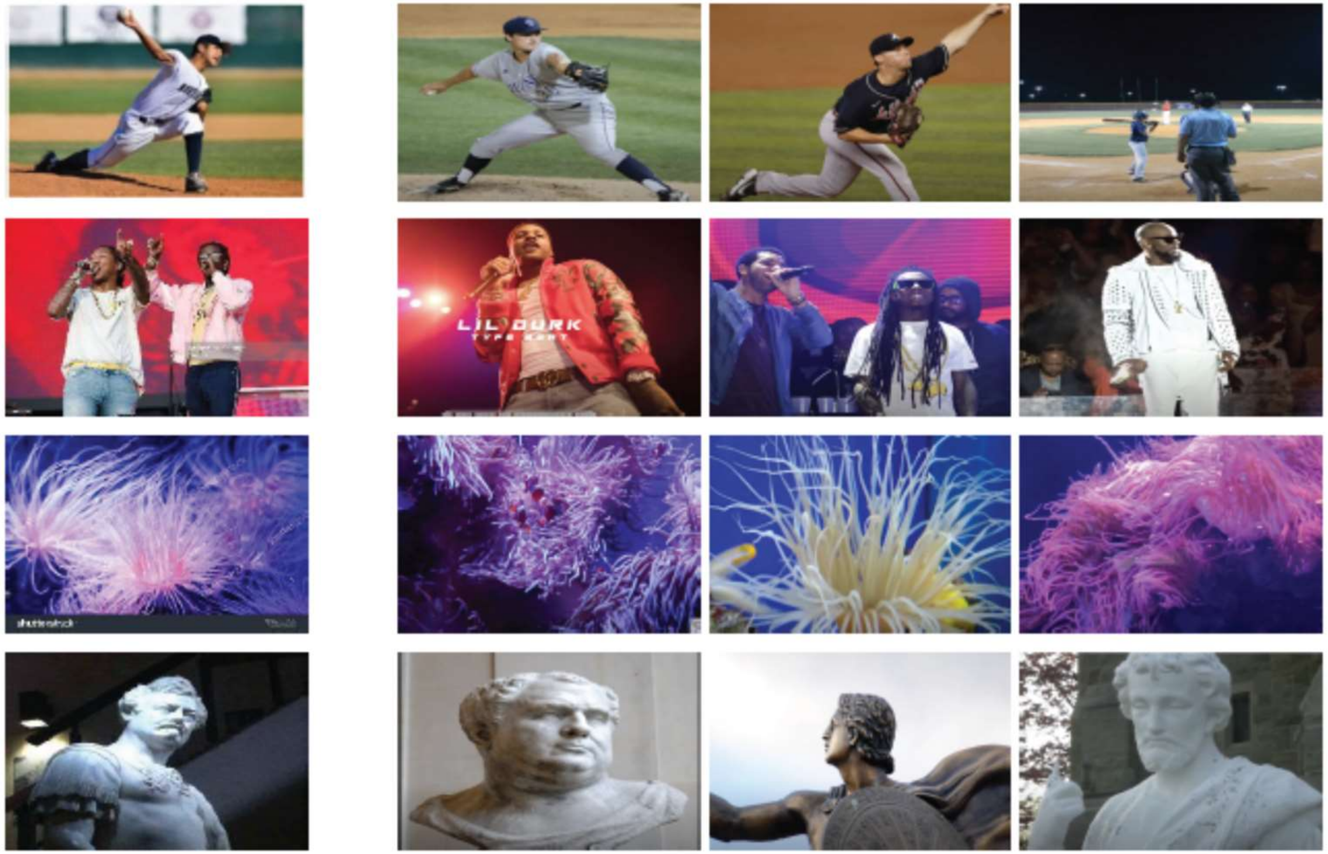
VideoCC3M

- Use the Conceptual Caption 3M dataset as seed
- Size: 10.3M; possibly multiple captions per video clip and multiple captions per video
- Multimodal: Both video and audio (unlike WebVid-2M)
- Diversity: more balanced that HowTo100M



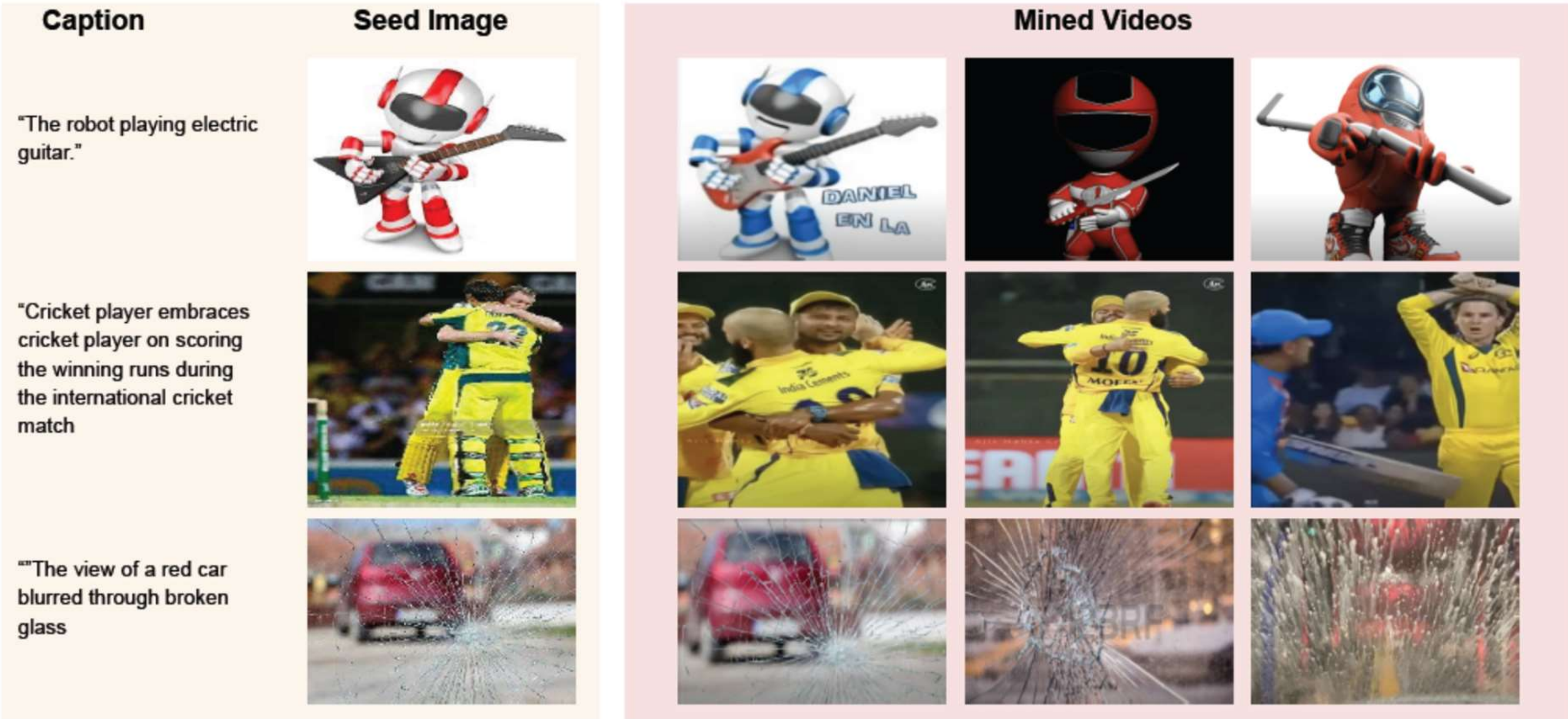
VideoCC3M – examples

Caption	Seed Image	Mined Videos
"Person throws a pitch during a game against university"		
"Rap artist perform onstage during day at festival"		
"Sea anemone in a dark blue water of aquarium"		
"And this is a statue"		



VideoCC3M – level of noise in the data

- ❖ Manual Study of 100 samples: 91/100 are relevant
  - ➢ 9 not relevant, 31 somewhat relevant, 60 highly relevant



Zero-shot results - Video retrieval

PT Data	Modality	#Caps	R@1	R@5	R@10
				Zero-shot	
-	V	-	-	-	-
HowTo100M [54]	V	130M	8.6	16.9	25.8
VideoCC3M	V	970K	18.9	37.5	47.1
VideoCC3M	A+V	**970K**	**20.4**	**39.5**	**50.3**

Zero-shot results on MSR-VTT text-video retrieval

Method	V-T PT	#Caps	R@1	R@5	R@10
MIL-NCE [54]	HT100M	136M	7.5	21.2	29.6
SupportSet [60]	HT100M	136M	8.7	23.0	31.1
EAO [68]	HT100M	136M	9.9	24.0	32.6
VideoCLIP [79]	HT100M	136M	10.4	22.2	30.0
FIT [9]	WebVid2M*	2.5M	15.4	33.6	44.1
Ours	VideoCC3M	**970K**	**20.4**	**39.5**	**50.3**

Zero-shot results - Video Captioning

- First results for zero-shot video captioning
- Outperforms HowTo100M by a large margin

Method	PT	Modality	B-4	C	M
Zero-shot					
Ours	HowTo100M	V	7.5	0.5	8.23
Ours	VideoCC3M	V	**13.23**	**8.24**	**11.34**

Table 4. **Results on the MSR-VTT dataset for video captioning.** Zero-shot results are obtained without any annotated video-text data. Modalities: **V:** RGB frames. **T:** ASR in videos.

GT:	a man is discussing the parts in an engine compartment in a vehicle	clouds are moving in the sky	this is about sports players making big plays during the game
HowTo100M:	So I'm going to go ahead and remove this	It's a great place to live and it's a great place to work.	I don't know if you can see that but there's a little bit of a gap in the middle of the field.
VideoCC3M:	the engine bay of an automobile model	clouds moving in the blue sky	american football player scores a touchdown against sports team