Learning and Using the Arrow of Time (AoT)

Donglai Wei[1], Joseph Lim[2], Andrew Zisserman[3], William T. Freeman[4,5]

Problem Statement

Input: video frames

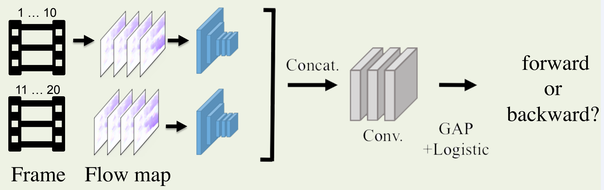
backward

Output: AoT

Learning the Arrow of Time (AoT)
- Data: are there artificial signals?
- Model: what cues does it learn to use?

Using the Arrow of Time
- How to apply it to other video tasks?

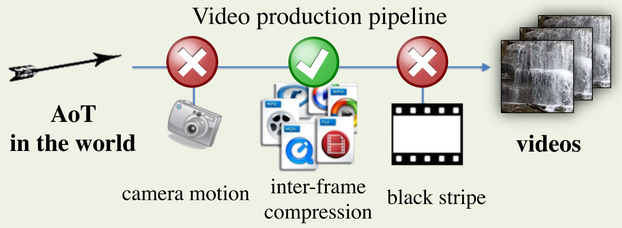
Model: Temporal-CAM



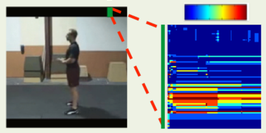
Model
- Input: optical flow in two chunks
- Final layer: global average pooling for Class Activation Map (CAM)

Training/Inference
- Augmentation: spatial (5 corners+flip), temporal (random crop)
- Outlier rejection: discard samples with subtle motion

Data: Remove Artificial Signals



A. Black stripe: non-zero intensity value



Train \ Test	original	zero-out
original	98.1%	87.9%

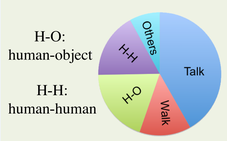
Zero out flow map, drops 10% acc

B. Camera motion: cameraman bias (e.g. zoom-in, tilt down)

Train \ Test	original	stabilization
original	88.3%	75.2%

Stabilize camera motion, drops 10% acc

C. Inter-frame compression: forward frame prediction



Train \ Test	Original	H.264-F	H.264-B
Original	59.1%	58.2%	58.6%
H.264-F	58.1%	58.9%	58.8%
H.264-B	58.3%	59.0%	58.8%

MJPEG-AoT dataset With or without inter-frame codec, similar result

Pre-processed Video Dataset (#clips)

[a] TA180(165k), [b] Flickr-AoT(147k), [c] Kinetics-AoT(58k)

Task 1: Self-supervision

Pre-training for Action Recognition
- train and test on UCF101 (A/B/C) or HMDB51 (A)
- control variable
  - = input: RGB, D-RGB, flow
  - = supervision: ImageNet, temporal order, none
  - = backbone architecture: AlexNet, VGG-16, ResNet-50

A. Comparison with the state-of-the-art (UCF101, HMDB51)

Method/Dataset	UCF101			HMDB51
	split1	split2	split3	
Wang et al. (2016)	85.7%	88.2%	87.4%	55.0%
AoT (ours)	**86.3%**	**88.6%**	**88.7%**	**55.4%**

Same flow input+VGG-16 architecture, different supervision (ImageNet)

B. Comparison with self-supervision methods (UCF101)

Input	Pre-train	Arch.	Accuracy
RGB	Rand.	AlexNet	38.6%
	Shuffle		50.9%
	AoT (ours)		**55.3%**
D-RGB	Odd-One		60.3%
	AoT (ours)		**68.9%**

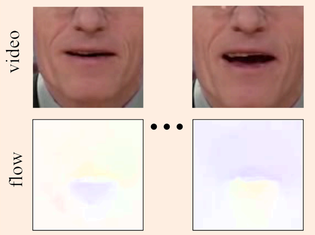
Same AlexNet architecture, different input+supervision

C. Comparison with different architectures (UCF101)

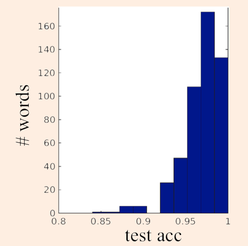
Input	Pre-train	Arch.	Accuracy
Flow	AoT	VGG-16	86.3%
		ResNet-50	**87.2%**
RGB		VGG-16	78.1%
		ResNet-50	**86.5%**
D-RGB		VGG-16	85.8%
		ResNet-50	**86.9%**

Same supervision, different input+architecture

Task 2: Time Symmetry Analysis



(a) example video and input flow

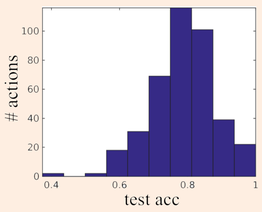


(b) histogram of acc for each word

Lip Motions
- Dataset: Lip Reading in the Wild (500 classes, 500k clips)
- AoT test accuracy: 97.6%

Top 5: Warning, Weekend, Today, Morning, Build
Bottom 5: System, National, Global, George, Enough

(a) top-3 action classes



(b) histogram of acc for each action

Human Actions
- Dataset: Kinetics (400 class, 58k clips)
- AoT test accuracy: 80.0%

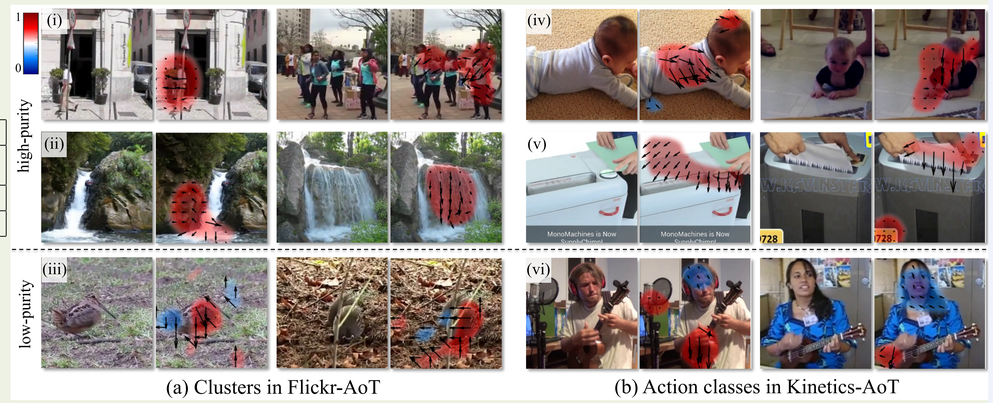
Top 3: Roller skating, Passing in football, Riding mule
Bottom 3: Brush painting, Doing nails, Shining shoes

Results: Classifying and Visualizing the Arrow of Time

Classification Result:
- pre-processed dataset: [a, b, c] as listed above

	[a]	[b]	[c]
Pickup et al.	82%	62%	59%
Ours	83%	81%	79%
Human	93%	81%	83%

Visualization Result
- Localization: CAM
- Clustering: K-means on second to last layer feature



(a) Clusters in Flickr-AoT (b) Action classes in Kinetics-AoT

Common Cues: (i, iv) Human body motion (ii) Gravity, (v) Human object interaction **Failure Cases:** (iii,vi) Repetitive motion

Task 3: Video Forensics

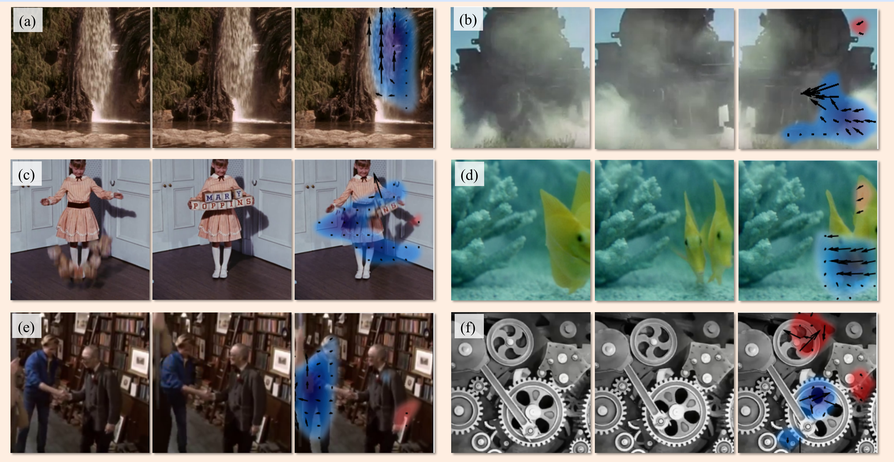
Reverse Film Dataset
- 67 clips from 25 movies

Test Result

Method	Acc.
Chance	50%
Pickup et al.	58%
Ours	76%
Human	80%

Failure Case:
- (f) Symmetric motion

Movie source:



(a) Anaconda, (b) The Railway Children, (c) Marry Poppins, (d) Bringing Out the Dead , (e) Top Secret! (f) Modern Times