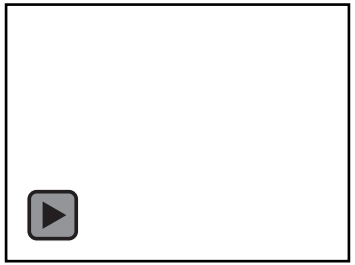
A Sparsely Labeled Dataset for Action Classification and Localization (SLAC)

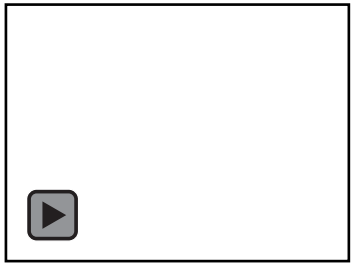
Presenter: Kamran Ali

Action Analysis: Background

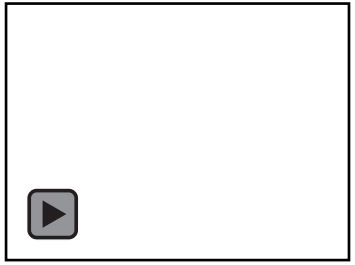
Action Recognition



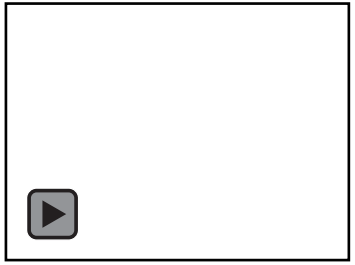
Biking



SalsaSpin

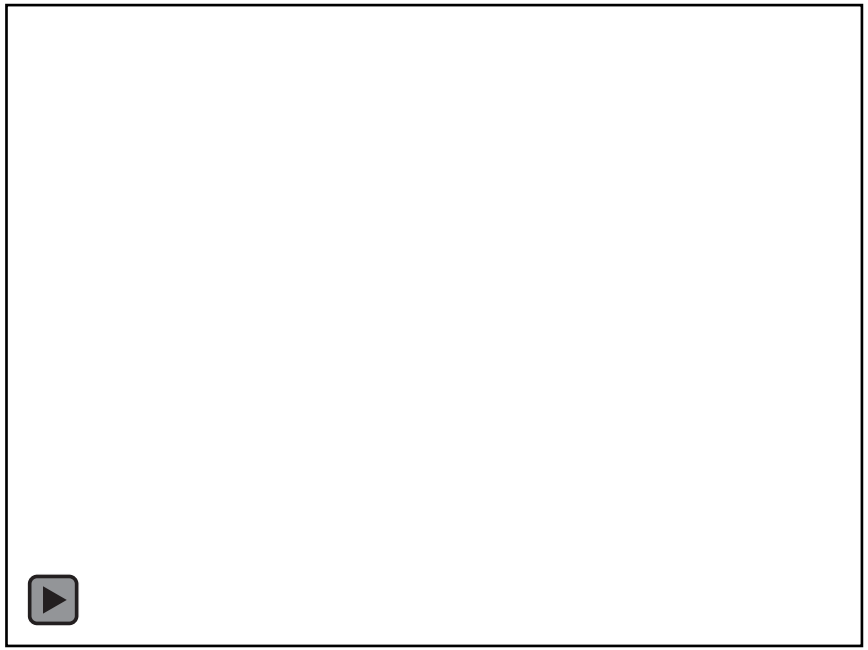


LongJump



TennisSwing

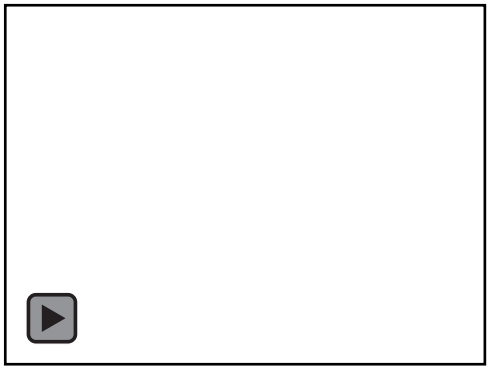
Action Detection



Diving

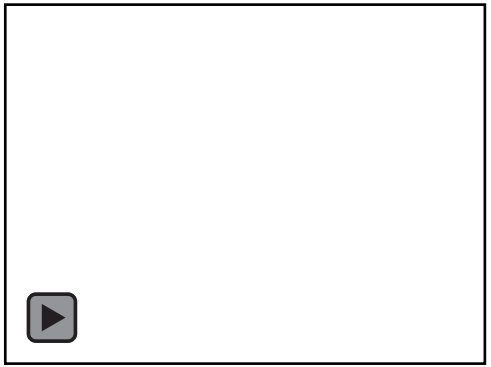
Action Analysis: Background

- Trimmed video



Diving

- Untrimmed video



Tennis

Action Analysis: Background

- **Temporal Action Localization**

  (Untrimmed Video)

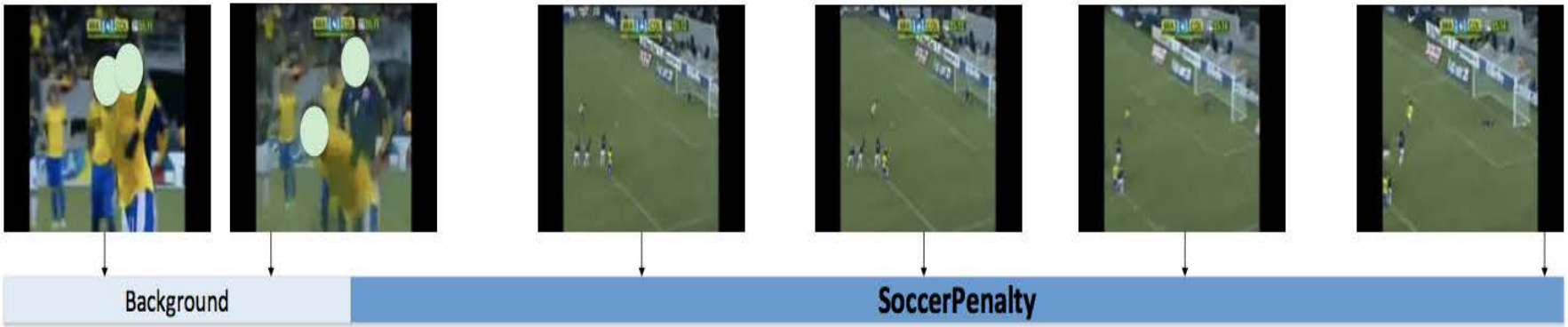
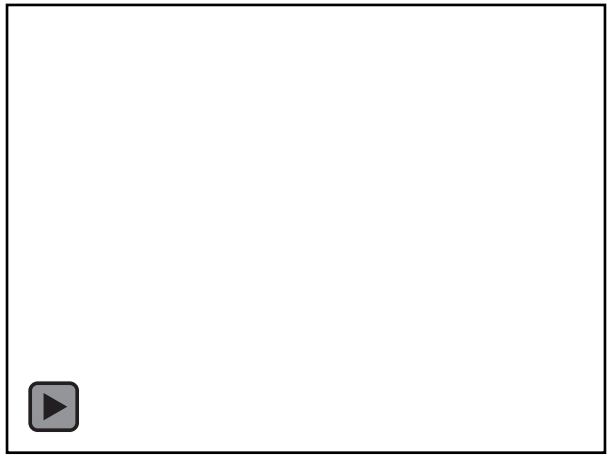


Image taken from the paper: "CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos" Zheng Shou

Action Analysis: Background

- Spatio-temporal localization



Tennis

Motivation

- **In Image Domain**
  - **Caltech101**
  - **ImageNet**
  - **COCO datasets**

Motivation

- **In Video Domain**
  - **UCF101**
  - **HMDB51**
  - **Kinetics**
  - **Sports 1M**
  - **ActivityNet**
  - **AVA**

Datasets for Action Classification

DataSets	Actions	Videos	Annotation	Source	Localization
UCF101	101	13K	13K	Youtube	No
HMDB51	51	7K	7K	Youtube/Movie	No
Kinetics	400	305K	305K	Youtube	No
Sports1M	487	1.1M	1.1M	Youtube	No

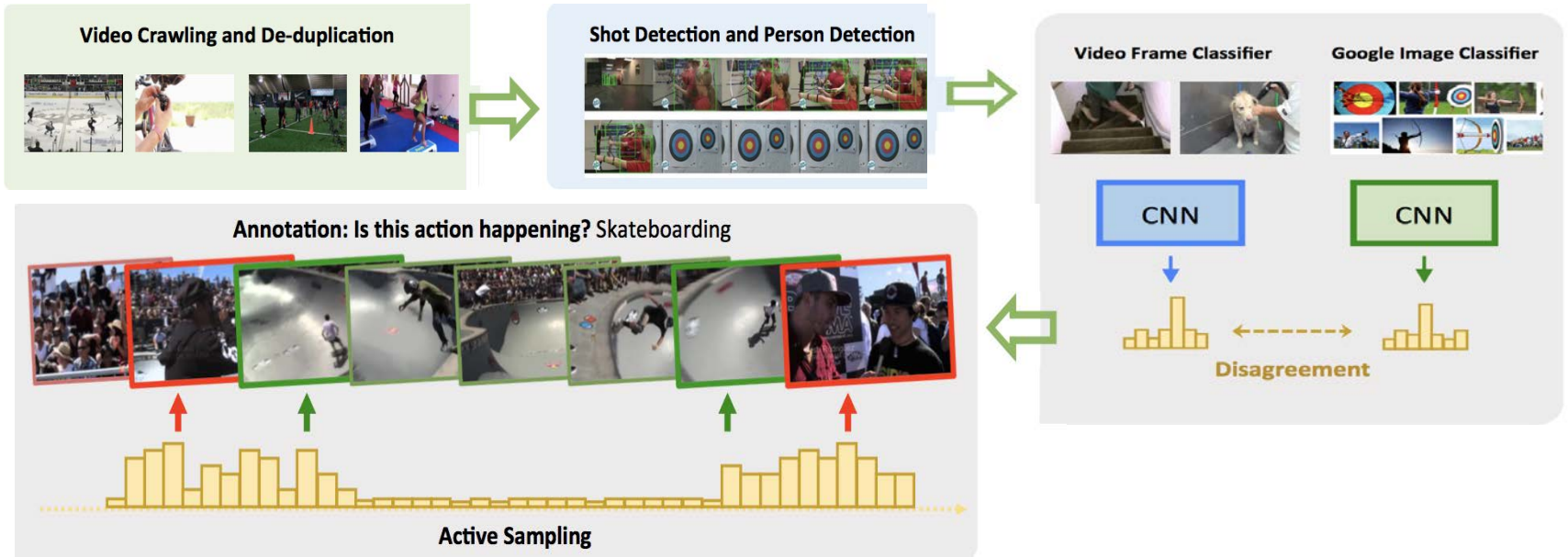
Datasets for Temporal Action Localization

DataSets	Actions	Videos	Annotations	Source	Localization
THUMOS 15	101	24K	21K	Youtube	Yes
ActivityNet	200	20K	23K	Youtube	Yes
Charades	157	10K	67K	267 Homes	Yes
AVA	80	214	197K	Movies	Yes
SLAC	200	520K	1.75M	Youtube	Yes

Common Trait of Datasets:

- **Exhaustive manual search of actions in videos**
- **Blurry action boundaries**
- **Time consuming and prone to error**

Data collection procedure to obtain SLAC



SLAC Dataset (http://slac.csail.mit.edu.)

Actions	Videos	Annotations	Source	Average length	Workforce/Time
200	520K	1.75M	Youtube	2.6 mins	70 annotators /10.5 days

SLAC Training, Validation and Testing Split

Training	Validation	Testing
502K	6K	12K

Challenges in Data Collection:

- **Imbalance samples**
- **Not all negative examples are equally useful**

Negative Sample Analysis

Positives

Action: Archery



Negative

No person



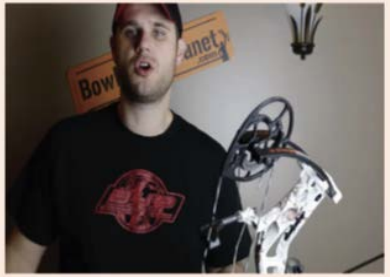
Negative

No context



Negative

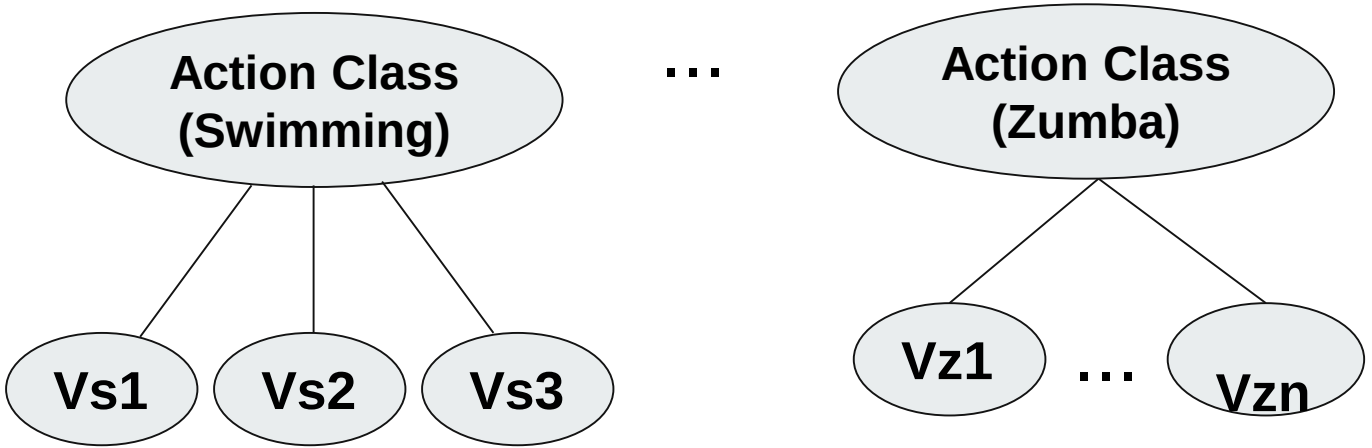
No action



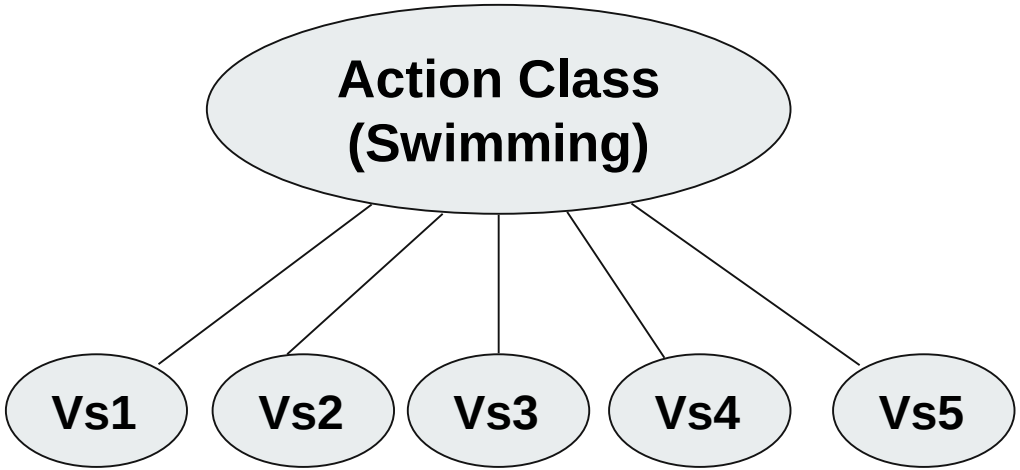
Video Retrieval and De-duplication

- **1.1M videos (200 classes)**
- **Internal De-duplication**
- **External De-duplication**

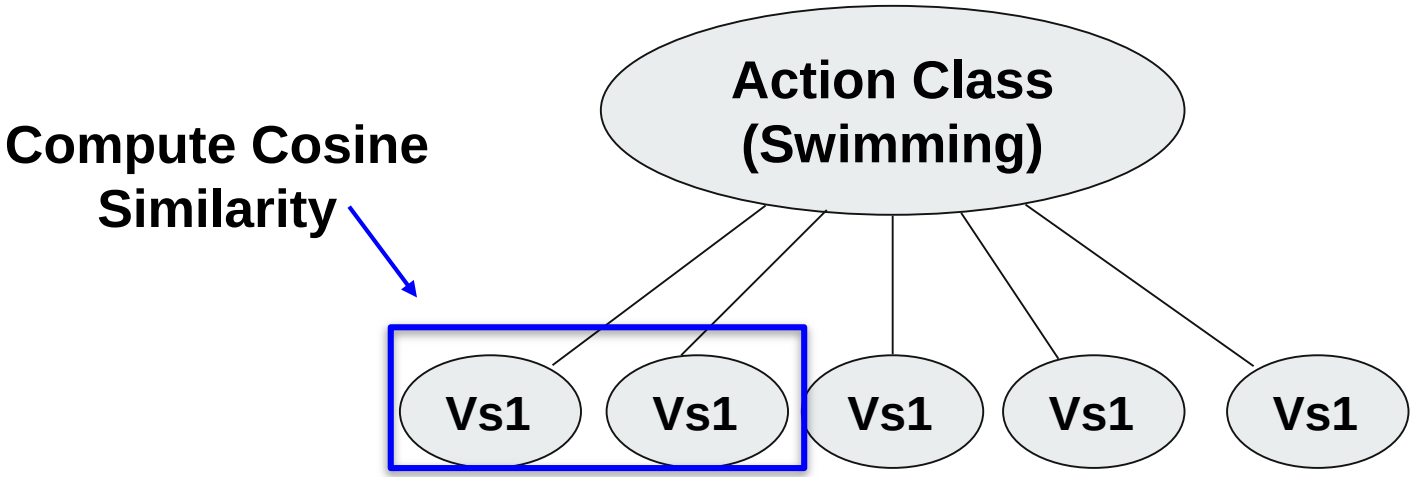
Internal De-duplication



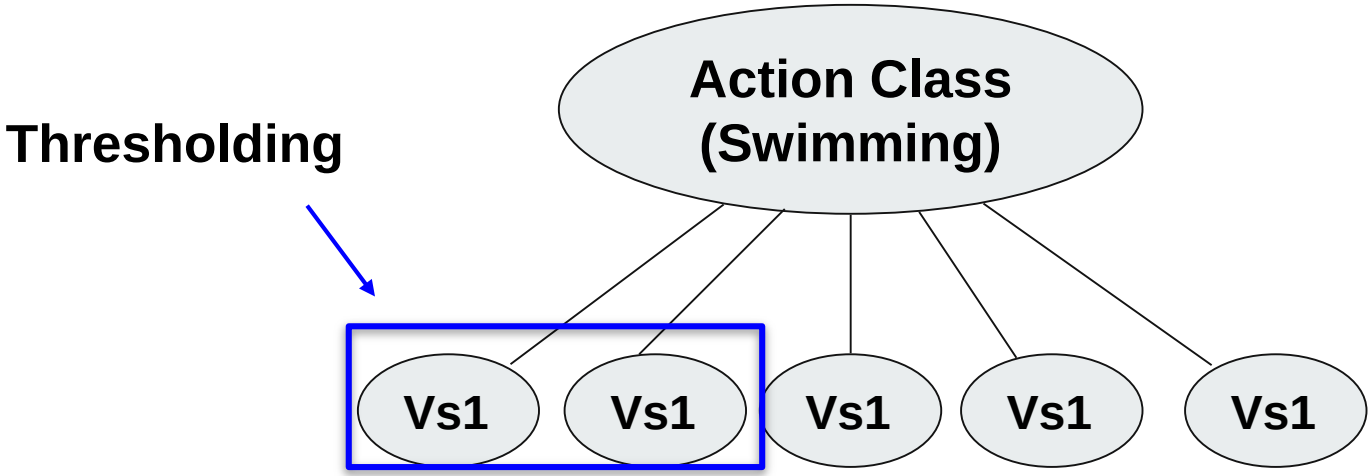
Internal De-duplication



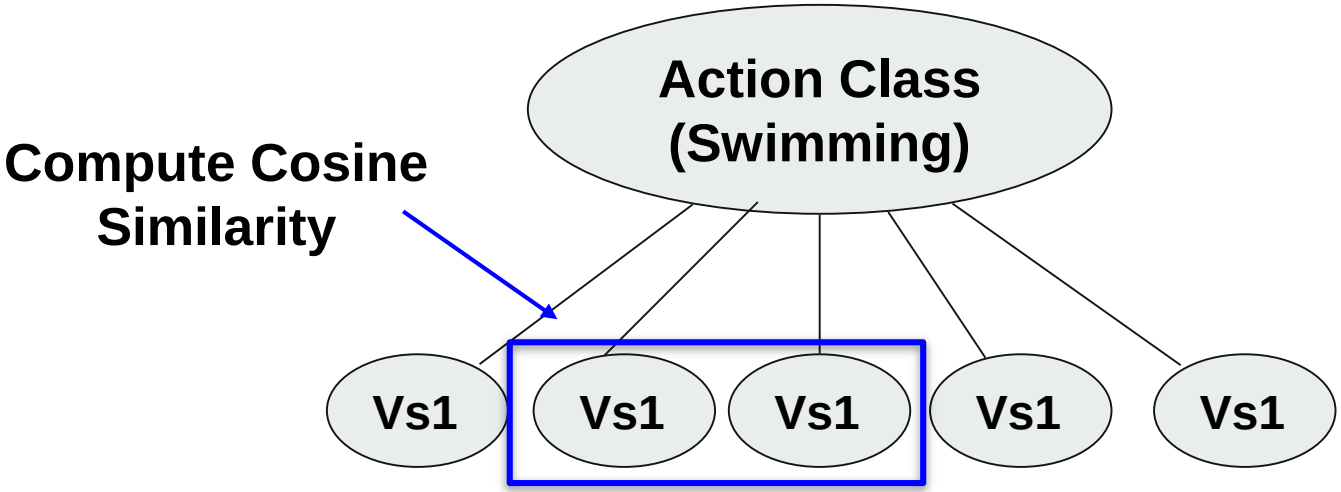
Internal De-duplication



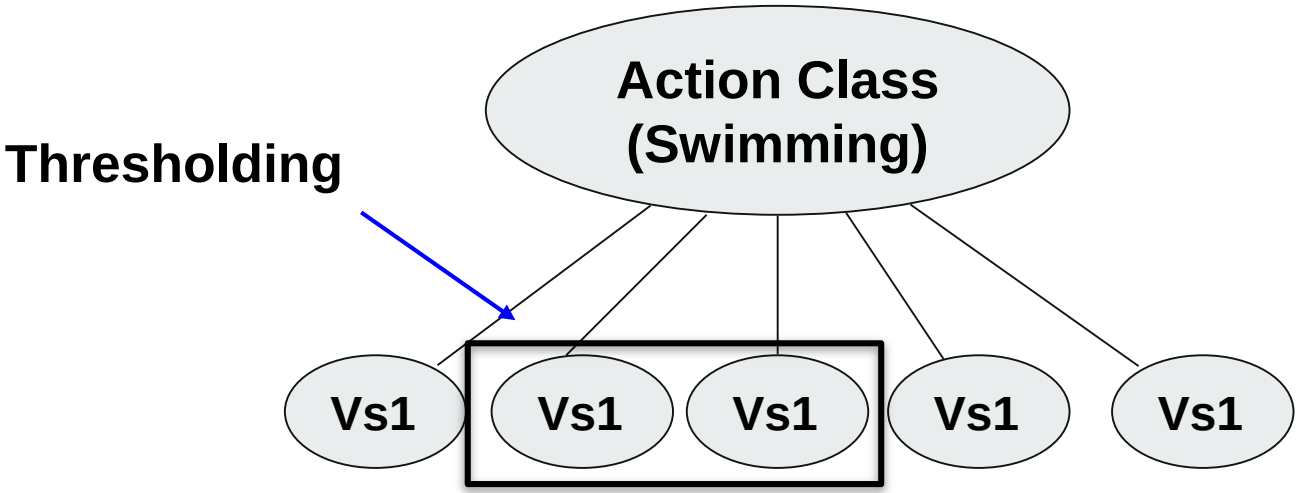
Internal De-duplication



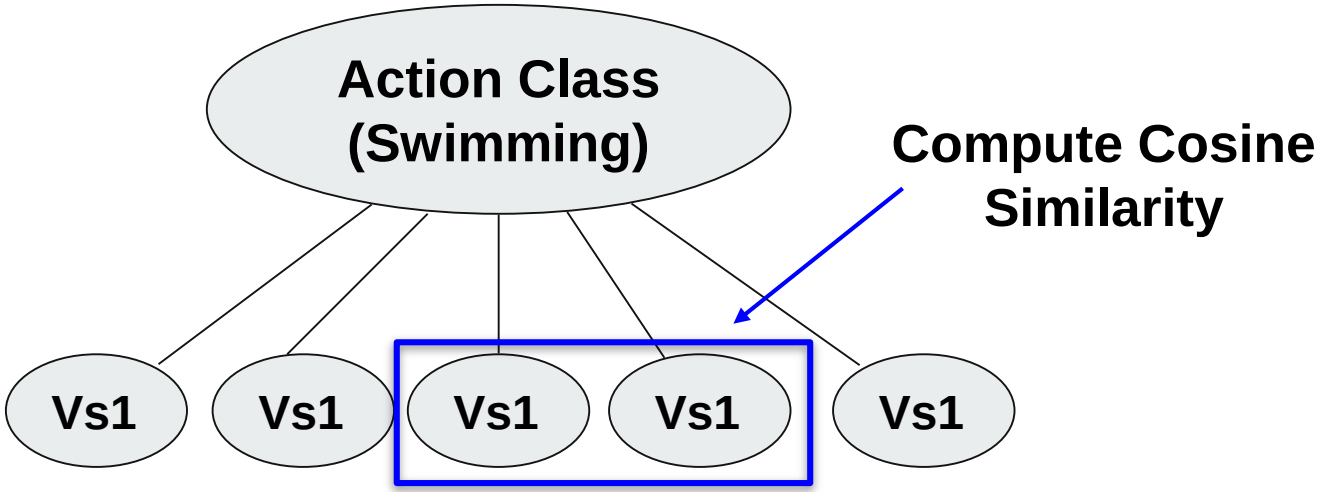
Internal De-duplication



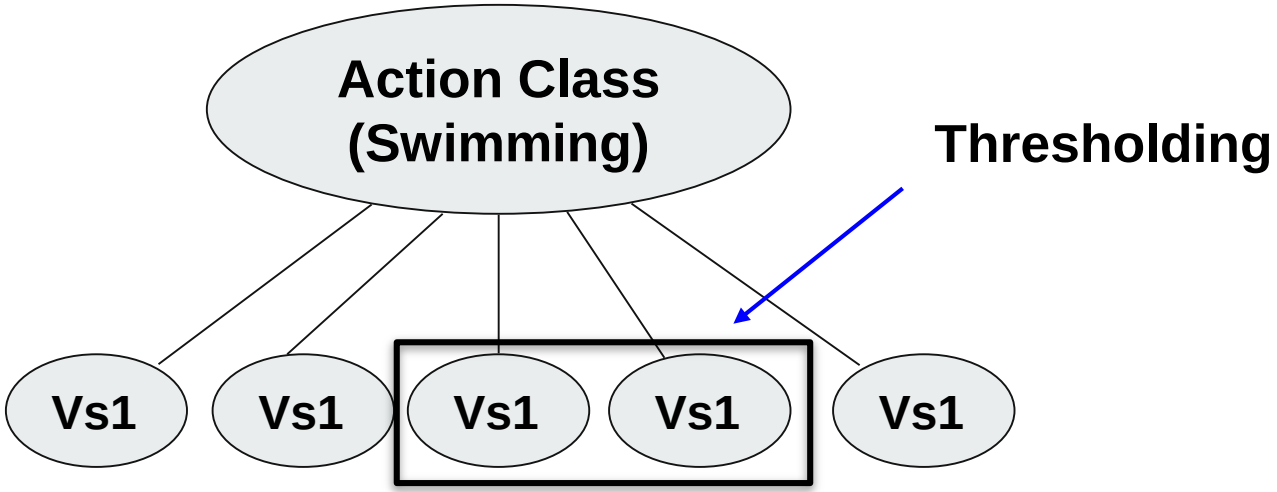
Internal De-duplication



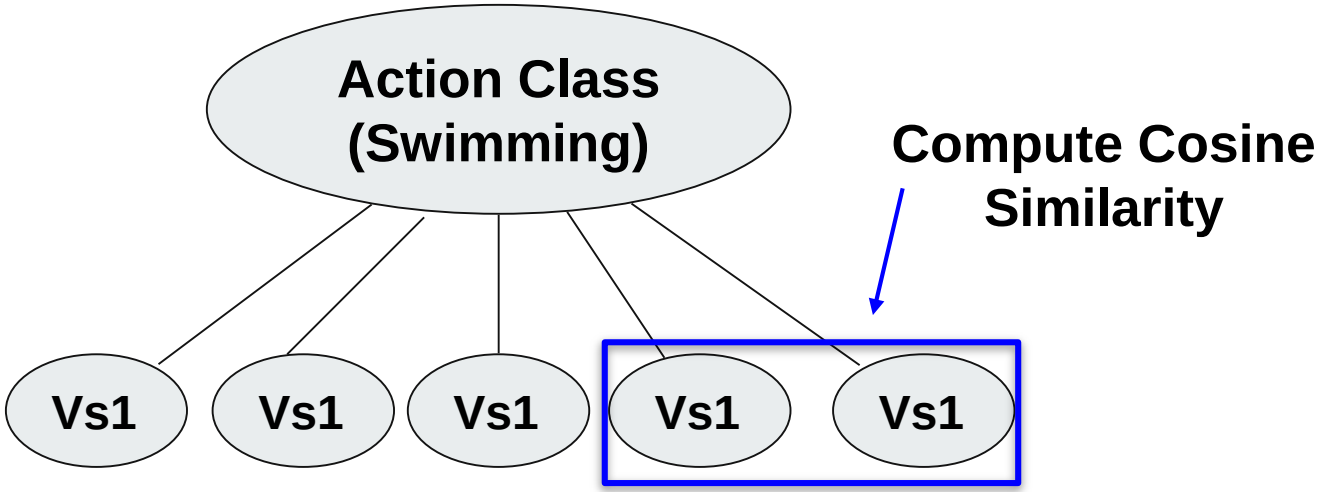
Internal De-duplication



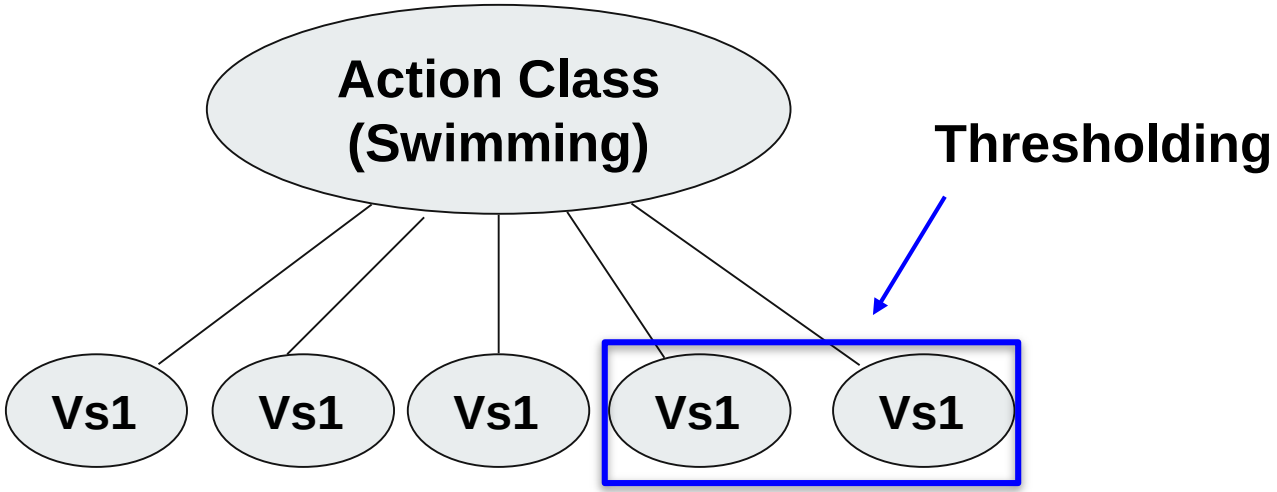
Internal De-duplication



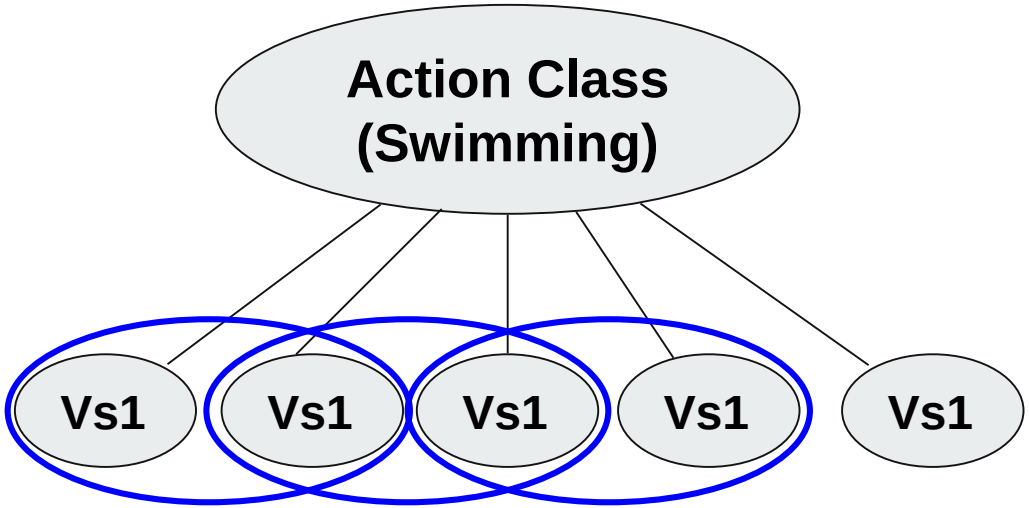
Internal De-duplication



Internal De-duplication

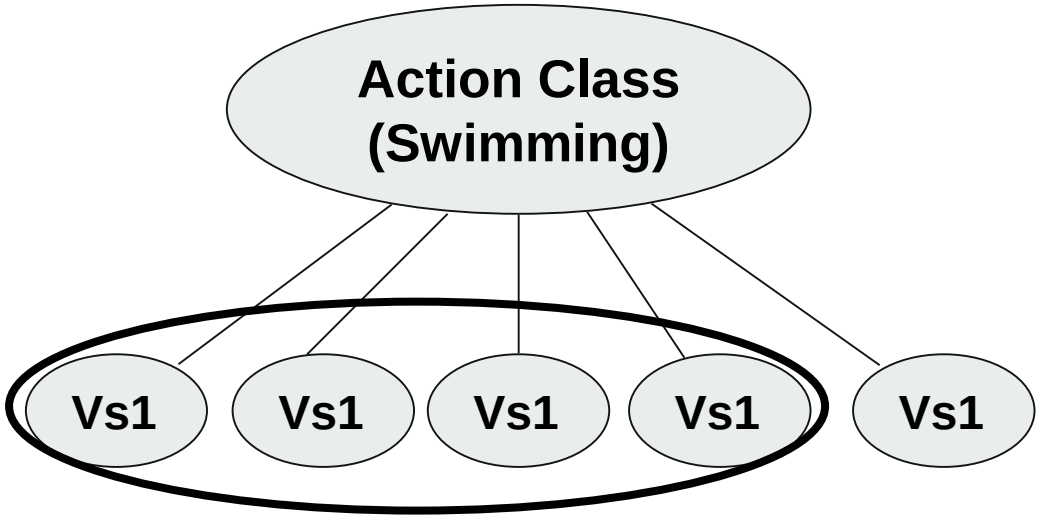


Internal De-duplication



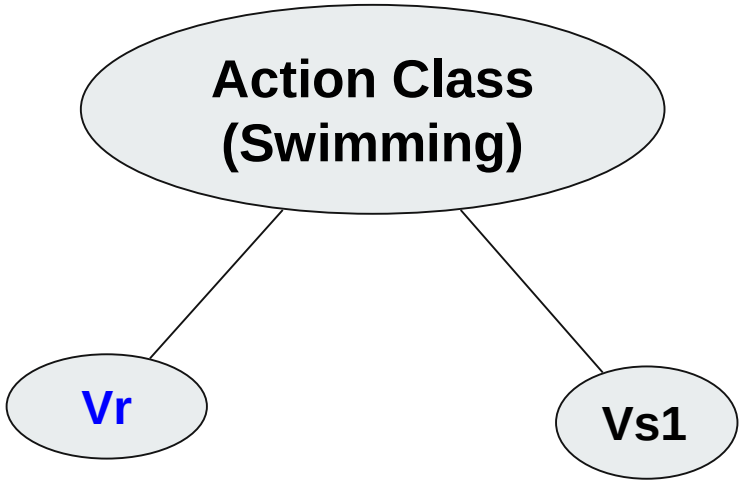
Connected Components

Internal De-duplication



Connected Component

Internal De-duplication



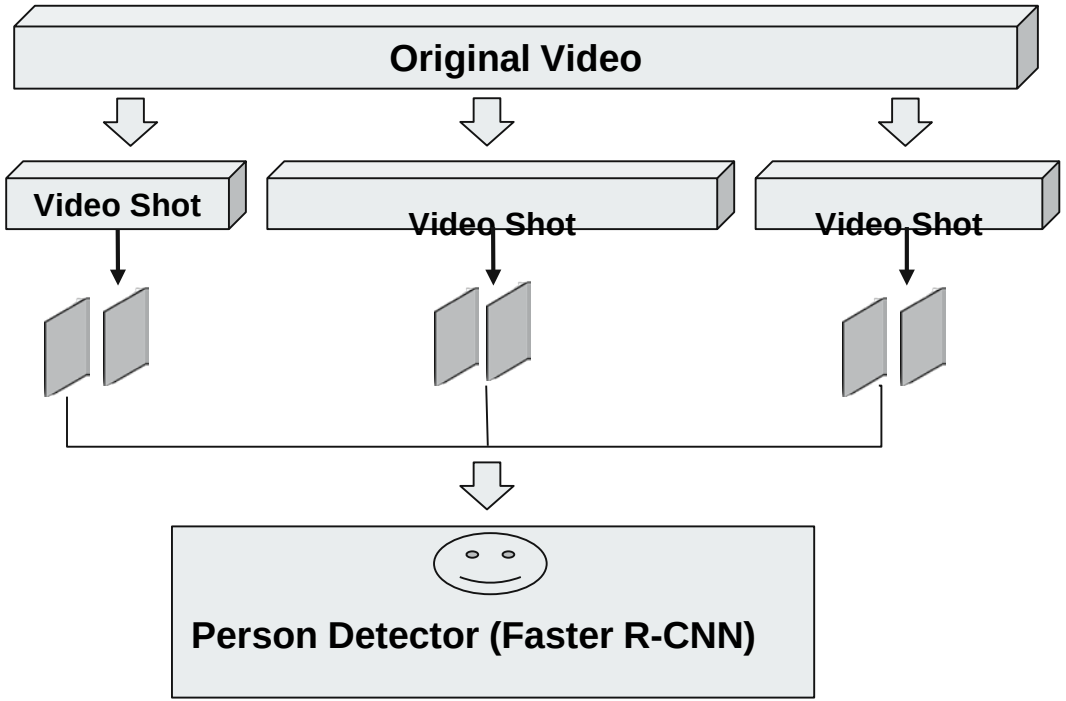
Result of De-duplication

SLAC	Kinetics	ActivityNet	UCF-101	HMDB-51
171,378	4029	1925	660	11

Excluding Videos Without Humans

- **Person Detector**
  - **Shot Detection**
  - **Person Classifier**

Person Detection



Negative Sample Analysis

Negative:

No person



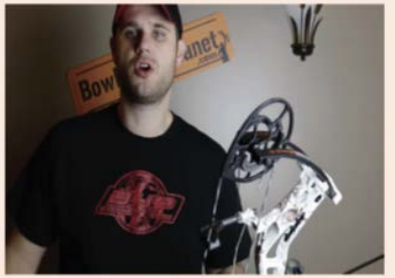
Negative:

No context



Negative:

No action



Negative Sample Analysis

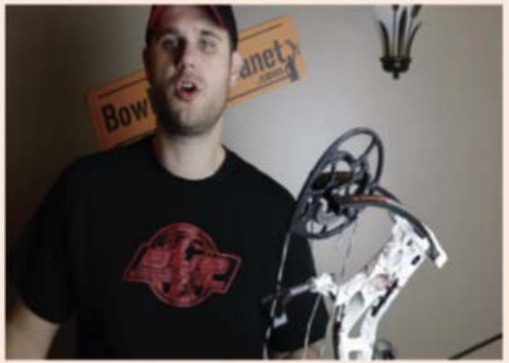
Negative:

No context

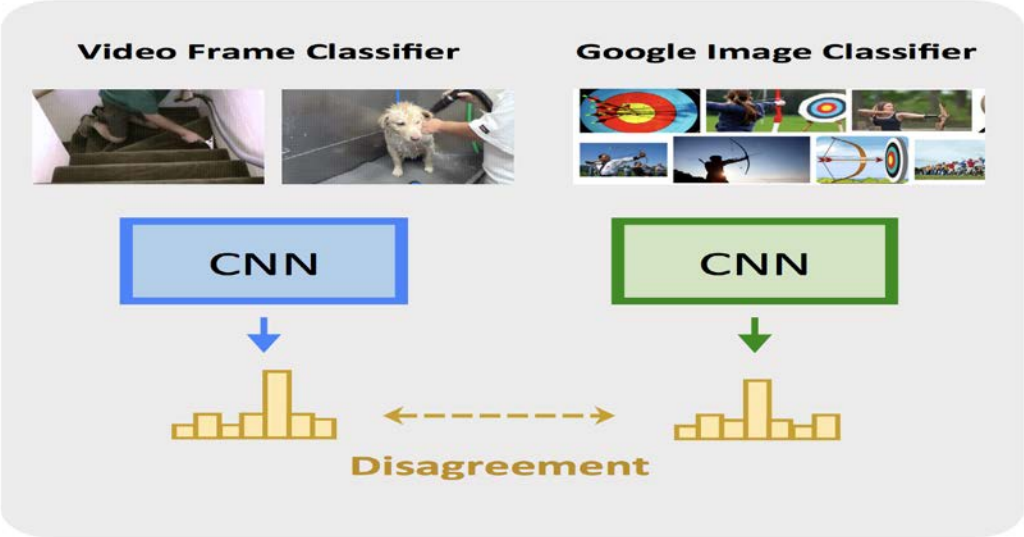


Negative:

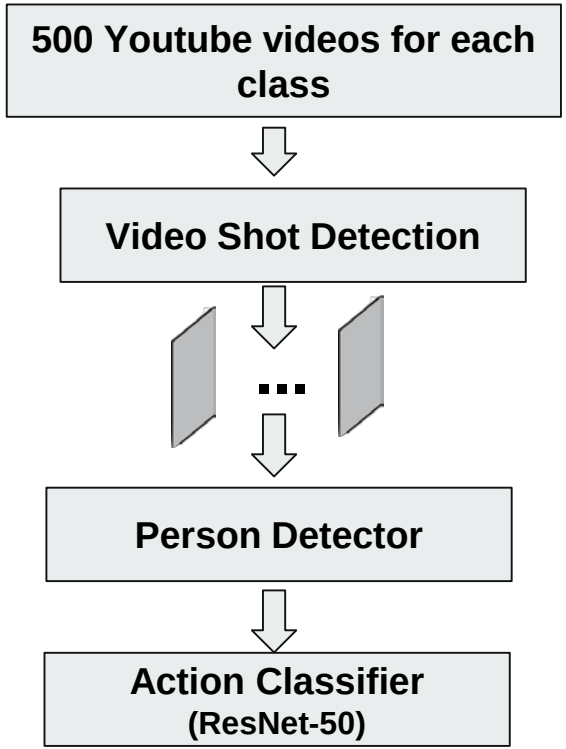
No action



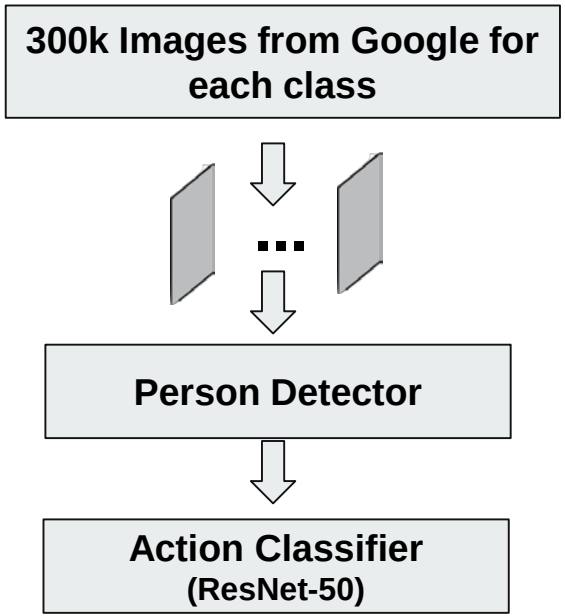
Proposal Models and Active Frame Sampling



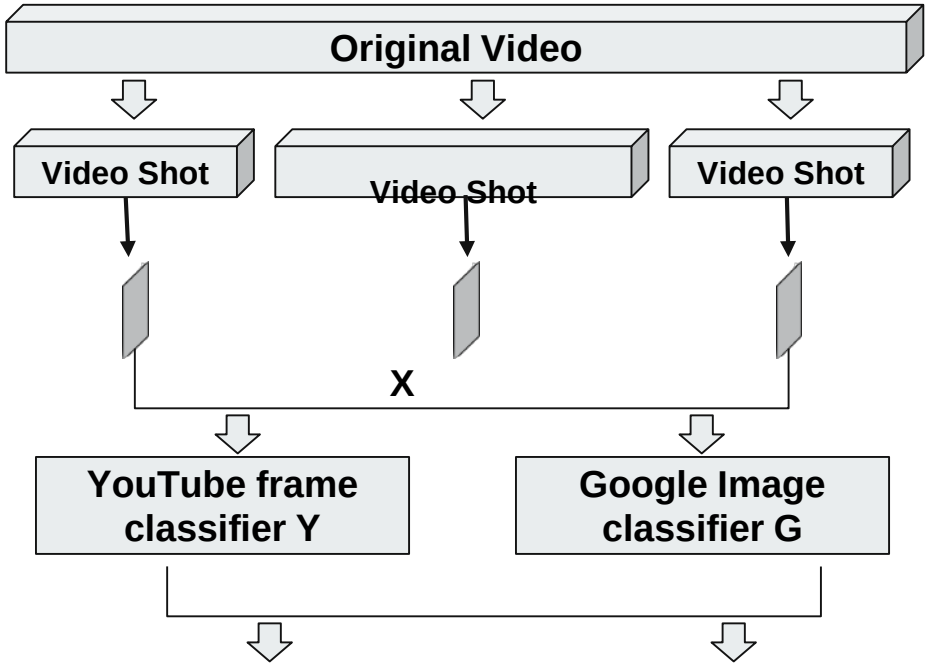
YouTube Frame-based Model



Google Image-based Model



Query-by-Committee (QBC)



$$R(x) = \left(P_c^Y(x) + P_c^G(x) \right) * D_{KL}(P^Y(x), P^G(x)) * \left(\frac{1}{|X|-1} \sum_{\hat{x} \in X, \hat{x} \neq X} S(f(x), f(\hat{x})) \right)$$

Annotation guideline:

- **Action: Clean and Jerk**
- **Definition: a two-movement weightlifting exercise in which a weight is raised above the head following an initial lift to shoulder level**

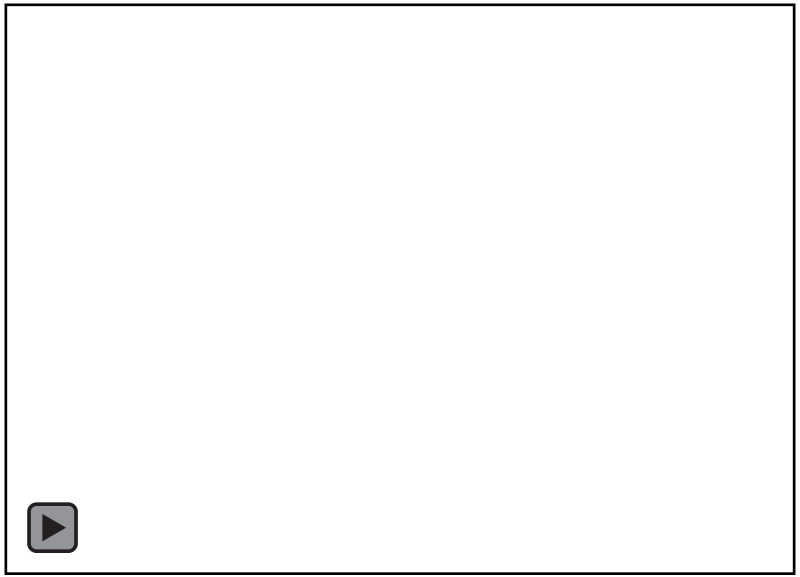
Positive



Negative in preparation stage



Annotation Tool



Annotation Quality Control

- **Same action clips, same annotator**

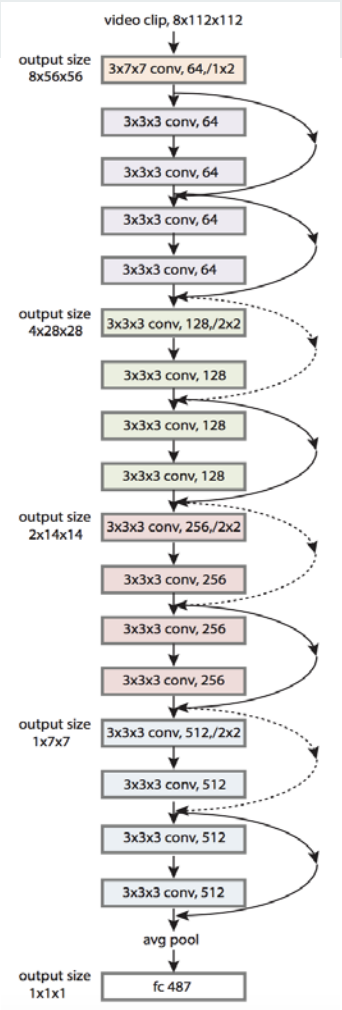
Golden Set:

- **Labeled 10K clips to measure the accuracy**
- **The final average accuracy is 94%**

Action Recognition Benchmark

Evaluations on SLAC:

- **Network Architecture:**
  - **Res3D-34**



Action Recognition Benchmark

	Validation Set		Test Set	
Input	Class-Agnostic, Clip@1	Class-Aware, Clip@1	Class-Agnostic, Clip@1	Class-Aware, Clip@1
RGB	54.3	76.5	53.7	76.2
Flow	47.3	64.9	46.3	63.6
RGB + Flow	**62.4**	**78.3**	**61.5**	**77.8**

Action Recognition Benchmark

Evaluations on Transfer Learning Tasks:

Target Dataset		UCF-101		HMDB-51		Kinetics	
Input	Training	Clip@1	Video@1	Clip@1	Video@1	Clip@1	Video@1
RGB	w/o pre-training	72.7	75.6	34.0	35.9	57.0	68.2
	w/ pre-training	**93.4**	**95.7**	**67.3**	**71.3**	**58.5**	**70.2**
Flow	w/o pre-training	68.7	80.8	34.0	45.1	42.2	59.2
	w/ pre-training	**87.9**	**93.6**	**60.0**	**69.5**	**45.0**	**62.6**

Action Recognition Benchmark

Comparisons with other pre-training datasets:

Target Dataset	UCF-101		HMDB-51		Kinetics	
Dataset pre-train	Clip@1	Video@1	Clip@1	Video@1	Clip@1	Video@1
w/o pre-training	72.7	75.6	34.0	35.9	57.0	68.2
Kinetics	89.9	92.5	64.9	67.4	57.0	68.2
Sports1M	90.9	92.6	64.1	65.9	57.4	68.4
SLAC	**93.4**	**95.7**	**67.3**	**71.3**	**58.5**	**70.2**

Action Recognition Benchmark

Comparisons with other methods:

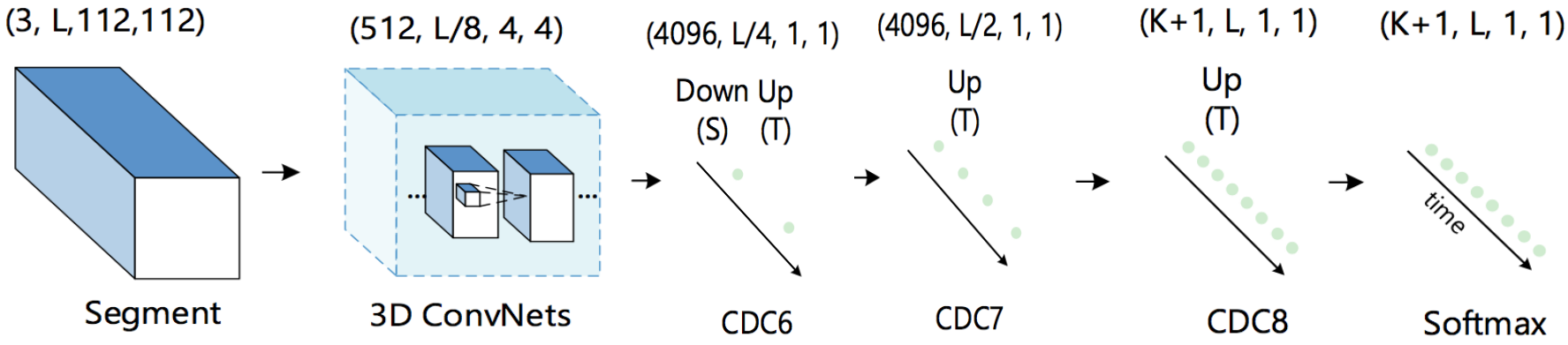
Method	UCF-101	HMDB-51	Kinetics
Long-term temporal convolutions [33]	92.7	67.2	N/A
ST-Multiplier Net + IDT [7]	94.9	72.2	N/A
Temporal Segment Networks [35]	94.2	69.4	N/A
Two-Stream I3D, Kinetics pre-training [14]	**98.0**	**80.7**	**75.7**
Two-Stream Res3D-34, SLAC pre-training, video@1	96.6	75.6	73.5
Two-Stream Res3D-34, SLAC pre-training, dense prediction, w/ multi-cropping@1	**97.6**	**76.7**	**74.8**

Temporal Action Localization Benchmark

Evaluations:

- **THUMOS14**
- **ActivityNet**
- **Network Architecture:**
  - **CDC Network**

Temporal Action Localization: CDC Network



Temporal Action Localization Benchmark

Comparison:

mAP @ IoU	0.3	0.4	0.5	0.6	0.7	Average
w/o pre-training	33.0	27.1	18.9	10.5	4.0	18.7
pre-train on Sports1M [26]	40.1	29.4	23.3	13.1	7.9	22.8
pre-train on SLAC	**45.0**	**35.8**	**29.2**	**17.2**	**9.5**	**27.3**
R-C3D [37]	44.8	35.6	28.9	N/A	N/A	N/A
SSN [40]	**51.9**	**41.0**	**29.8**	N/A	N/A	N/A

Temporal Action Localization Benchmark

Comparison:

mAP @ IoU	0.5	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	Average
before refinement [36]	43.3	39.9	36.7	16.8	8.7	3.9	1.9	0.8	0.2	0.0	15.3
after refinement, w/o pre-training	43.0	**39.5**	**36.0**	**32.0**	**28.6**	24.6	13.5	5.4	1.4	0.2	22.4
after refinement, pre-train on Sports1M [26]	**45.3**	N/A	N/A	N/A	N/A	**26.0**	N/A	N/A	N/A	0.2	23.8
after refinement, pre-train on SLAC	42.9	38.9	34.7	31.0	28.0	24.6	**20.9**	**16.9**	**9.7**	**1.4**	**24.9**



Thank You