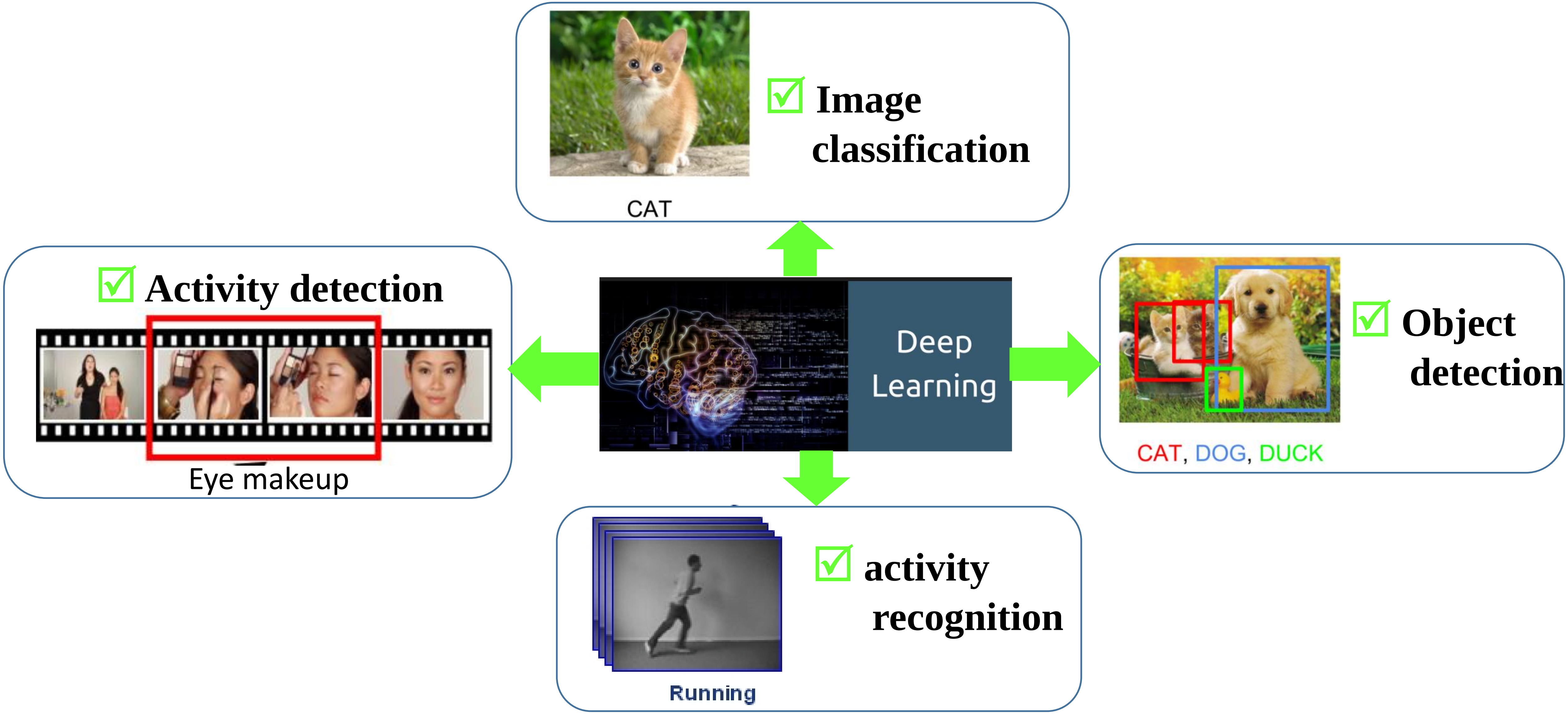
Weakly Supervised Learning (WSL)

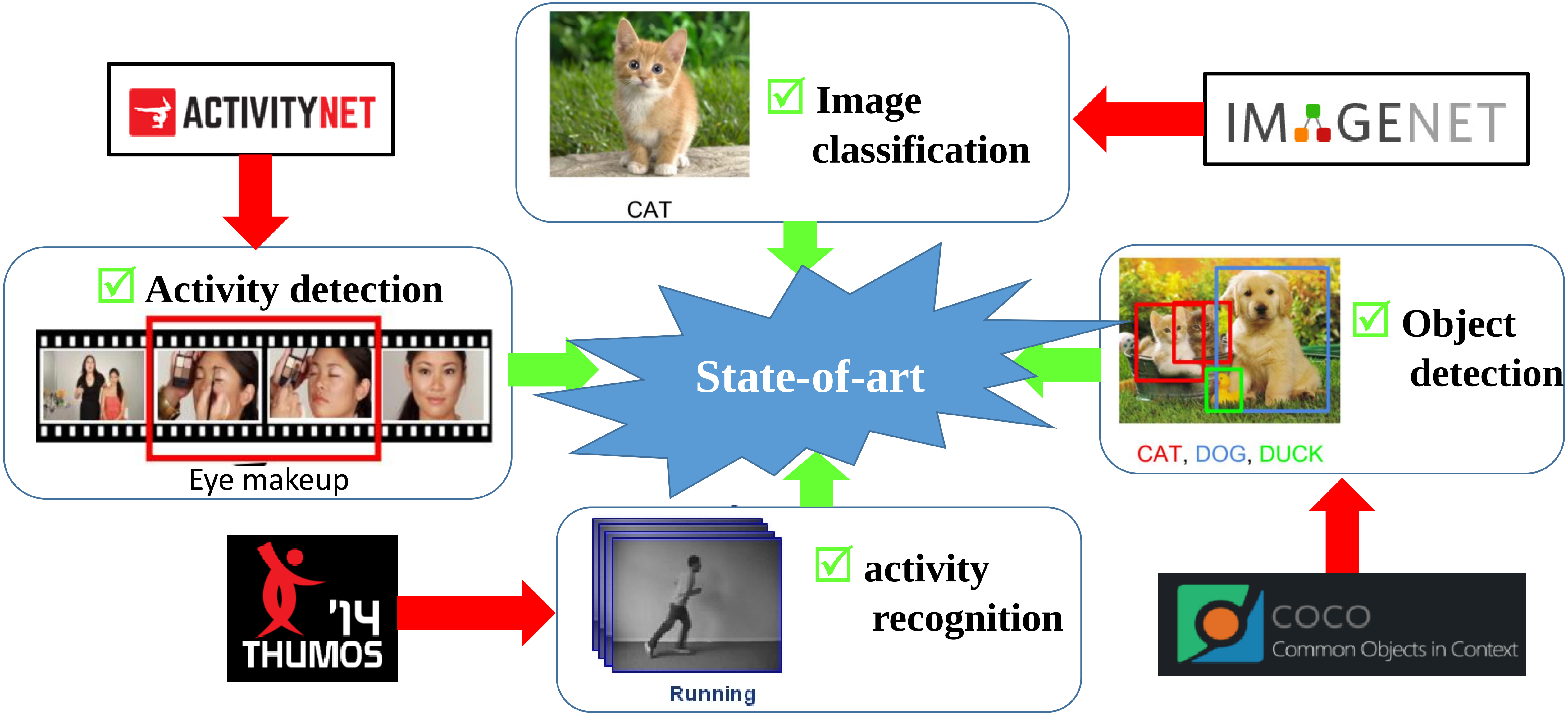
Yongqiang Zhang

July 6, 2017

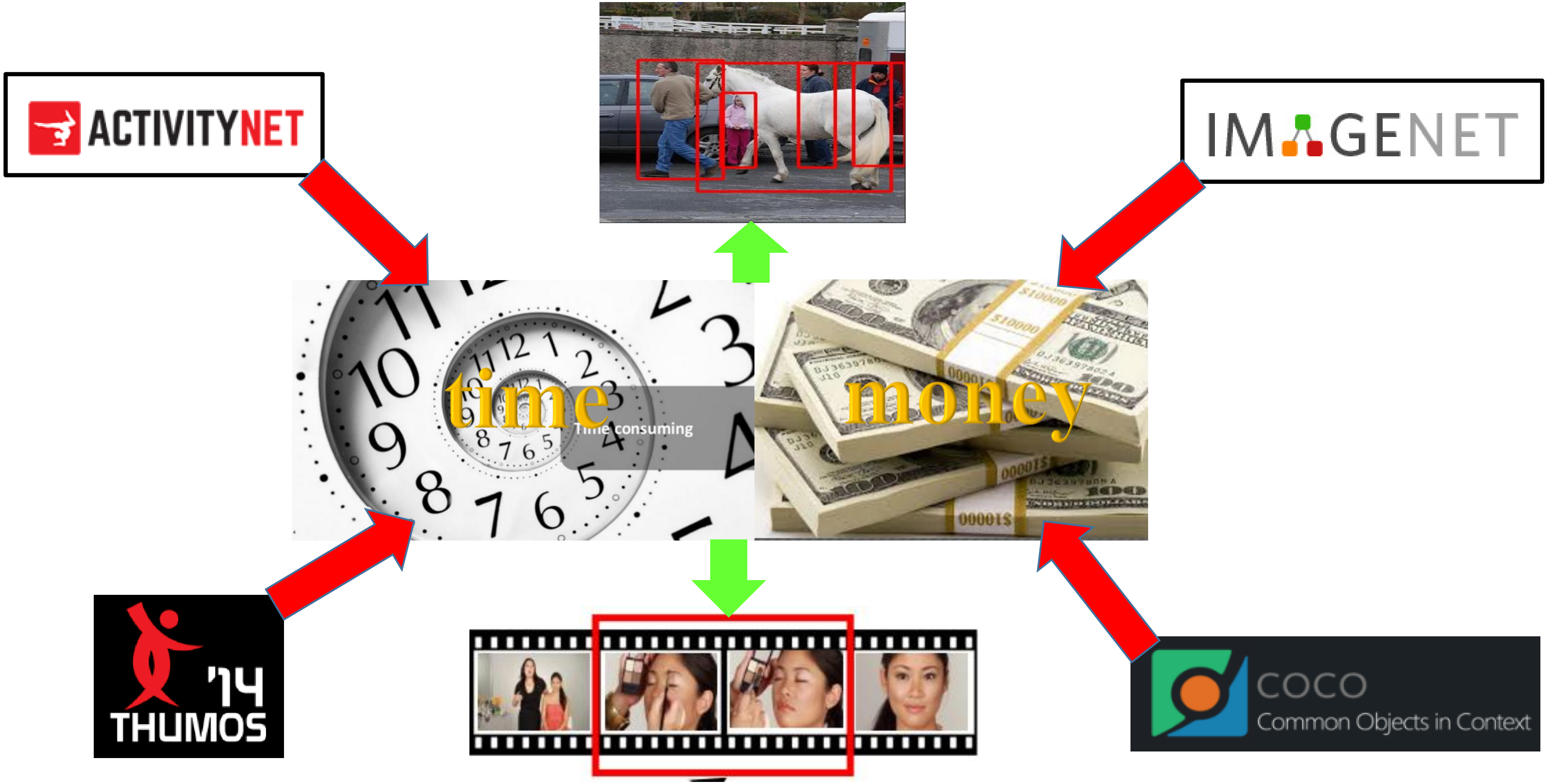
Motivated for Weakly Supervised Learning(WSL)



Motivated for Weakly Supervised Learning(WSL)



Motivated for Weakly Supervised Learning(WSL)



Motivated for Weakly Supervised Learning(WSL)

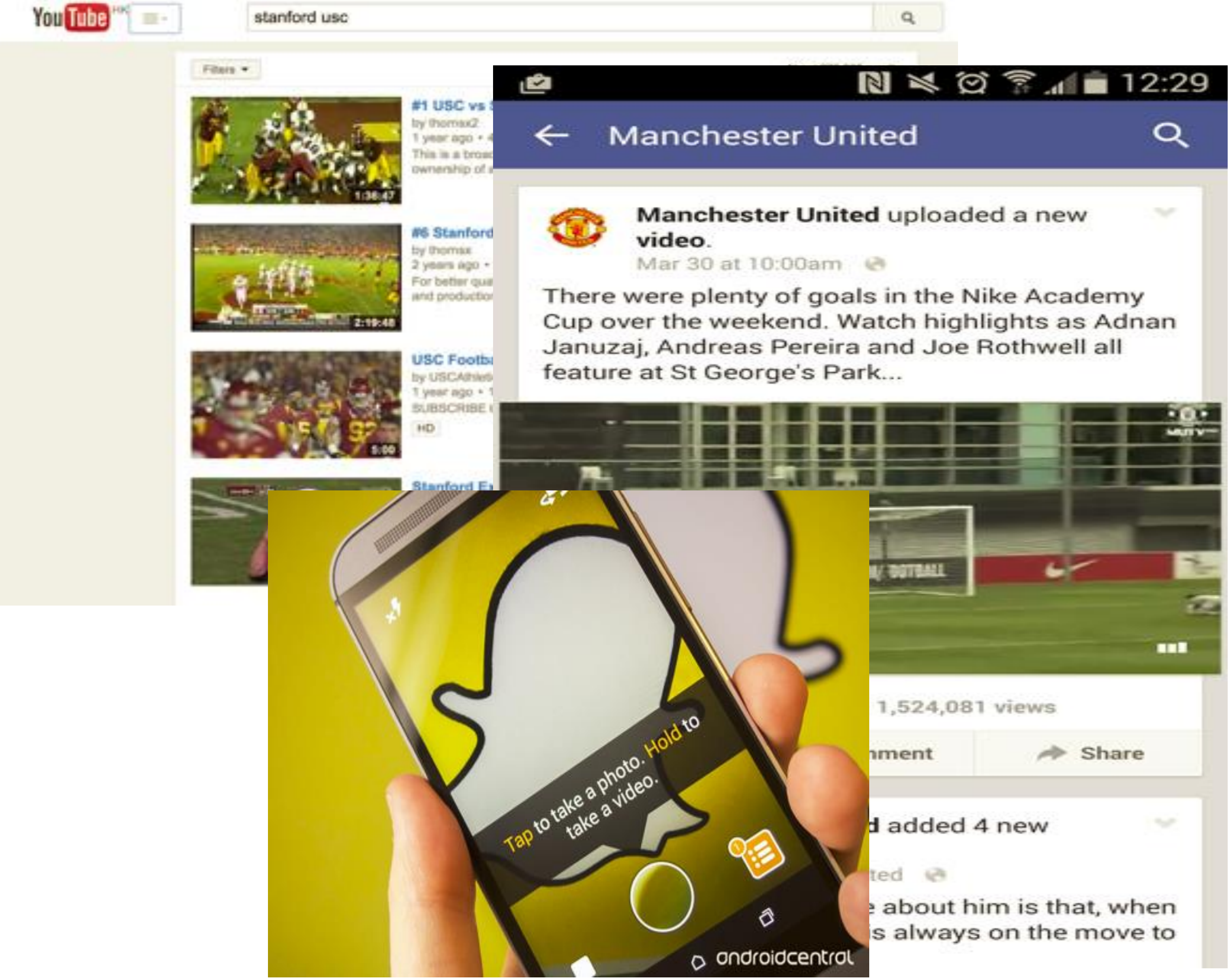
☑ Video capturing devices are more affordable and portable than ever.

☑ Almost every adults own a smartphone.

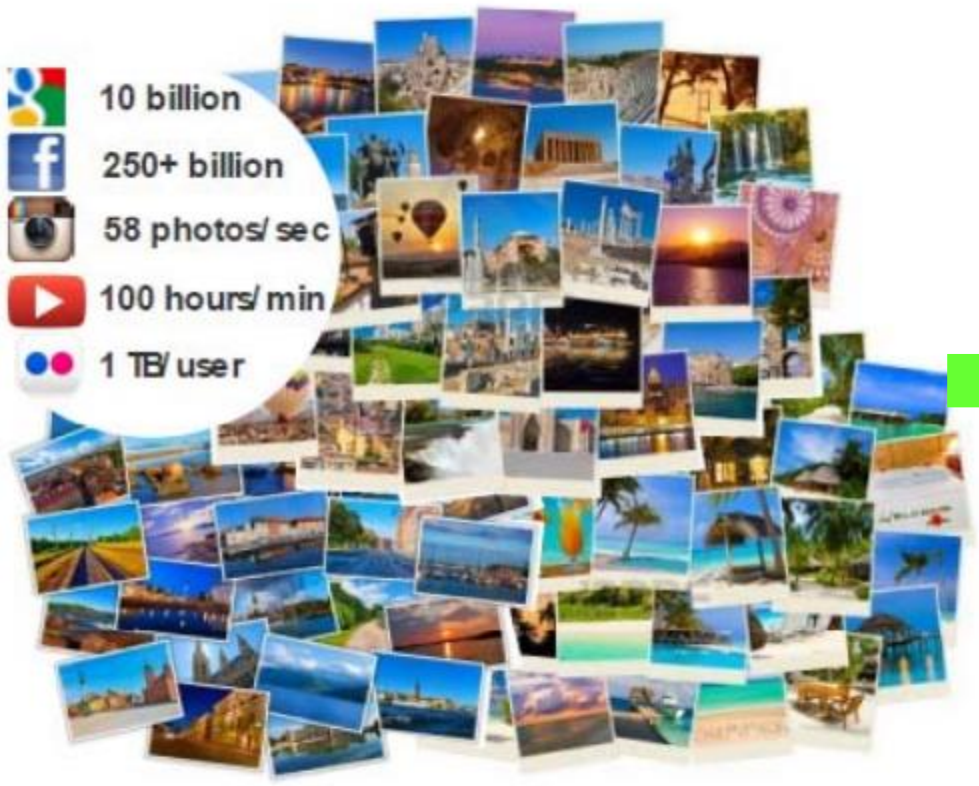
Motivated for Weakly Supervised Learning(WSL)

People also love to share their images and videos!

400 hours of new YouTube video every minute.



Motivated for Weakly Supervised Learning(WSL)



Object detection

Action recognition

Can we use these webly image and video to train a deep model?

Motivated for Weakly Supervised Learning(WSL)

weakly supervised learning

Object detection

Action recognition

Can we use these webly images and videos to train a deep model?

Overview

- What is the weakly supervised learning?
- Weakly supervised for action recognition and event detection
- Weakly supervised for action detection
- conclusion

Overview

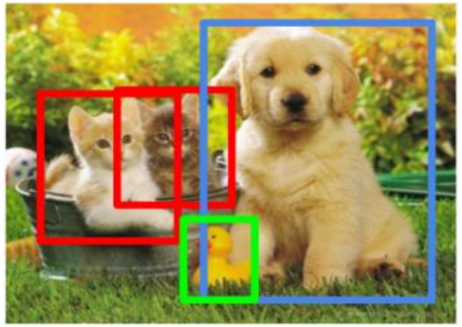
- What is the weakly supervised learning?
- Weakly supervised for action recognition and detection
- Weakly supervised for action detection
- conclusion

What is the weakly supervised learning?

weakly supervised learning means only use a limited amount of labeled data.

✓ **Image-level labels**: cat, dog, duck
✗ **No** position information(BBox)

Object detection



✓ cat, dog, duck
✓ BBox

✓ **Video-level labels**: eye makeup
✗ **No** temporal annotations(start time and end time)

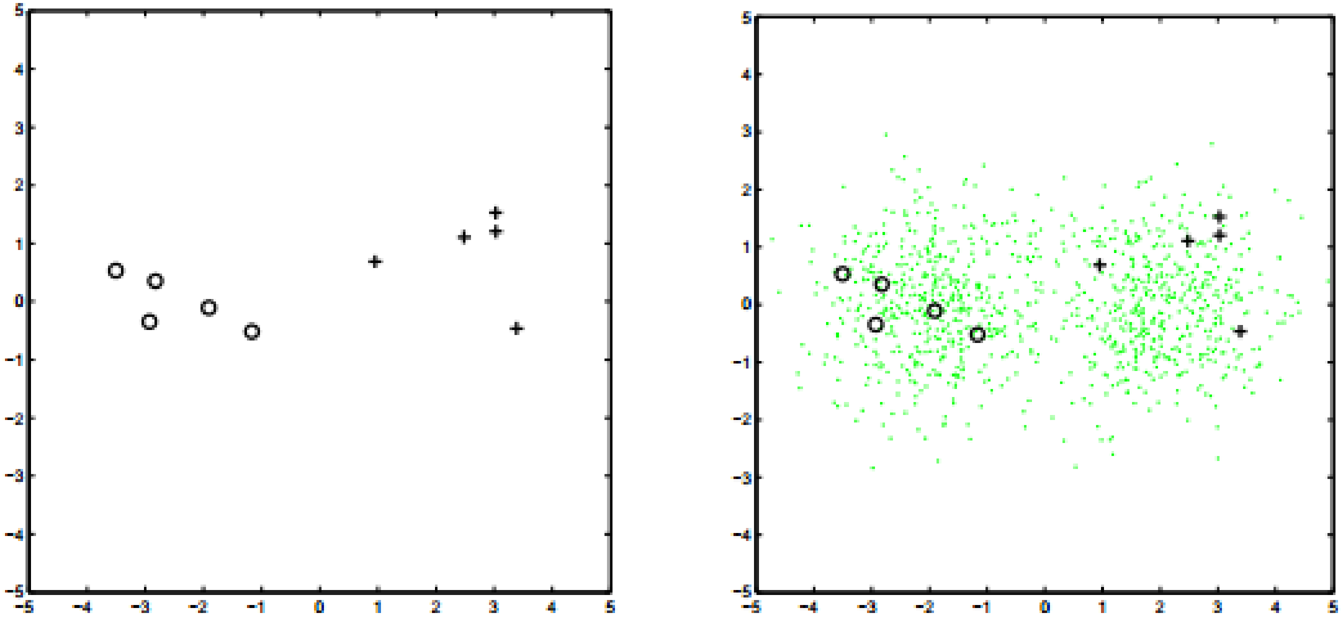
Activity detection

✓ eye makeup
✓ time and end time

What is the weakly supervised learning?

Compare with semi-supervised learning(SSL):

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.

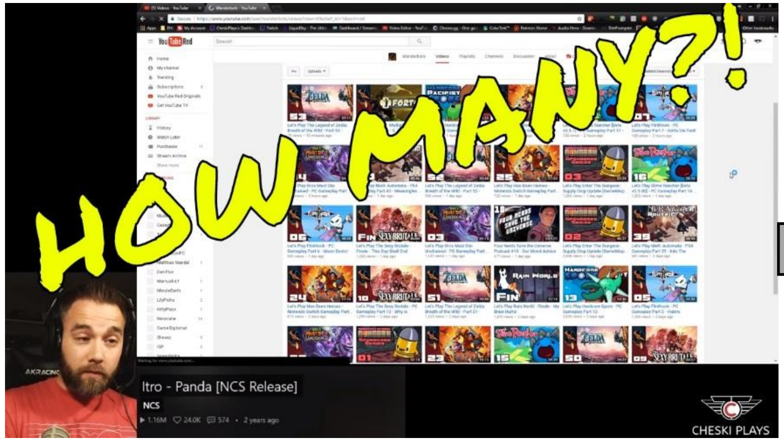


(a) labeled data　　(b) labeled and unlabeled data (small dots)

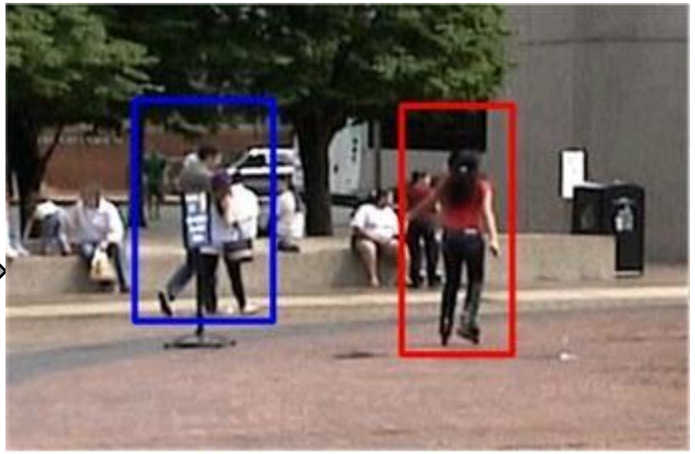
Overview

- What is the weakly supervised learning?
- Weakly supervised for action recognition and detection
- Weakly supervised for action detection
- conclusion

Weakly supervised for action recognition and event detection



A deep model



walk skate

Webly videos

Action recognition

Event detection

Weakly supervised for action recognition and event detection

Challenge for using the webly images and videos:

1. web videos are always untrimmed and contain large portion of irrelevant frames.

(a) Mopping floor

The irrelevant frames are indicated by green boxes in this figure

Weakly supervised for action recognition and event detection

2. Web images could be noisy due to

1) **semantic drift**, i.e. the mismatch between query and returned images. For example juggling balls in this figure(only returned ball, not juggling balls).

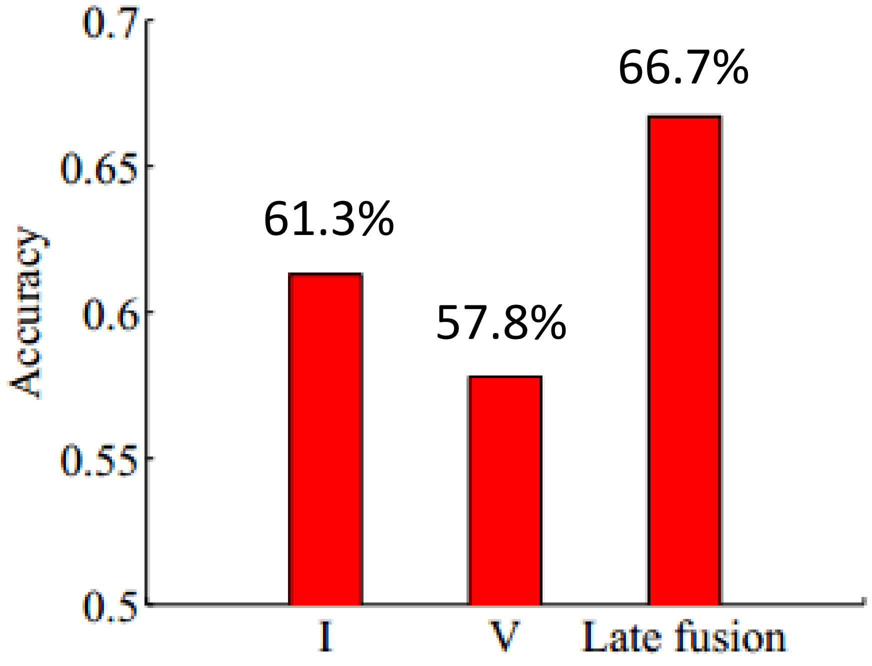
2) **domain gap**, i.e. the inconsistencies between videos and images, e.g. images of baby crawl usually post edited with clean white background.

(b) Juggling balls

(c) Baby crawl

Weakly supervised for action recognition and event detection

Preliminary experiment on action recognition



I: action recognition performance by using web images only

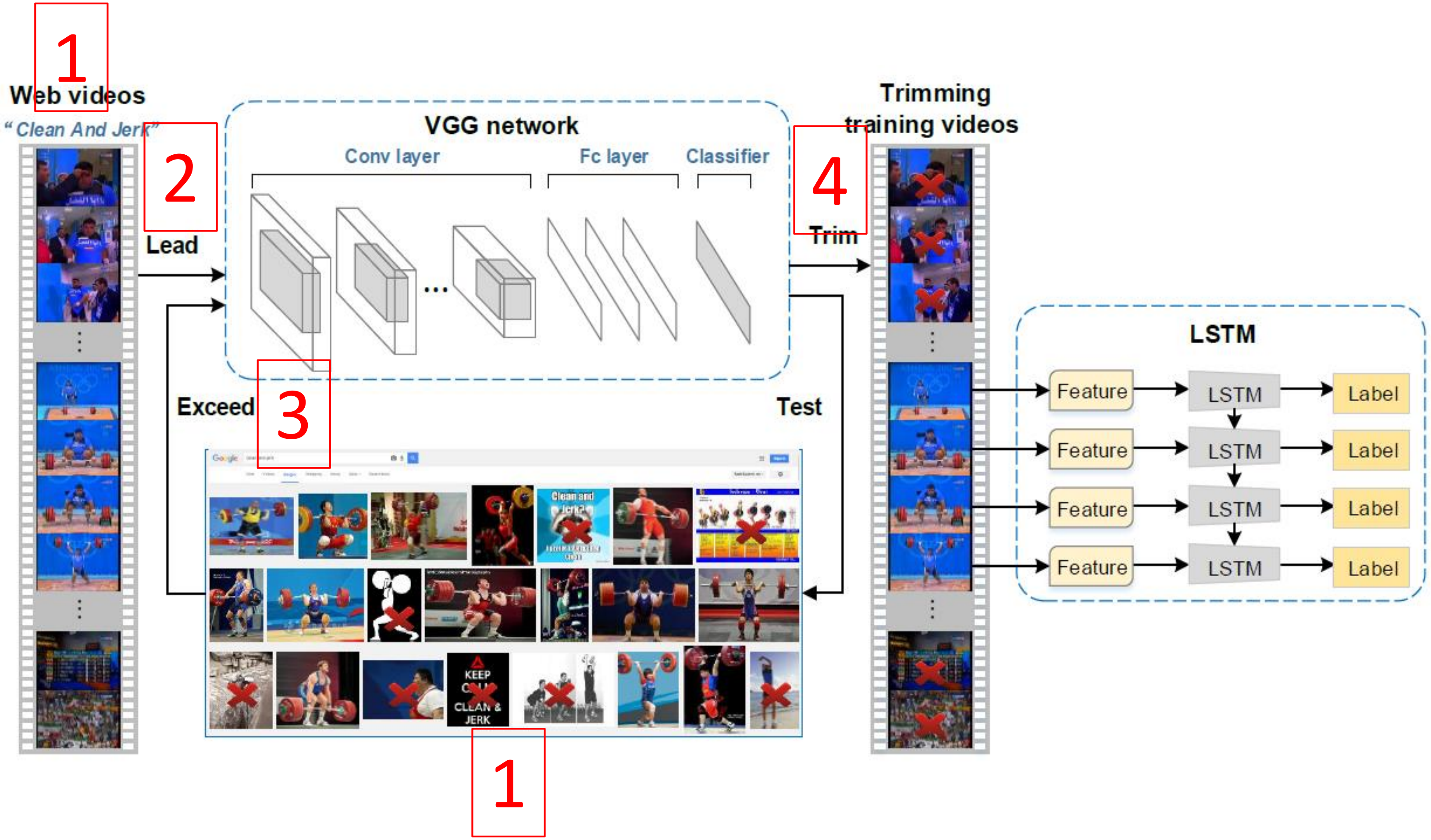
V: action recognition performance by using web videos only

Late fusion: a simple late fusion of the prediction scores of fine-tuned models on I and V.

Weakly supervised for action recognition and event detection

- Framework of Lead-Exceed Neural Network(LENN)

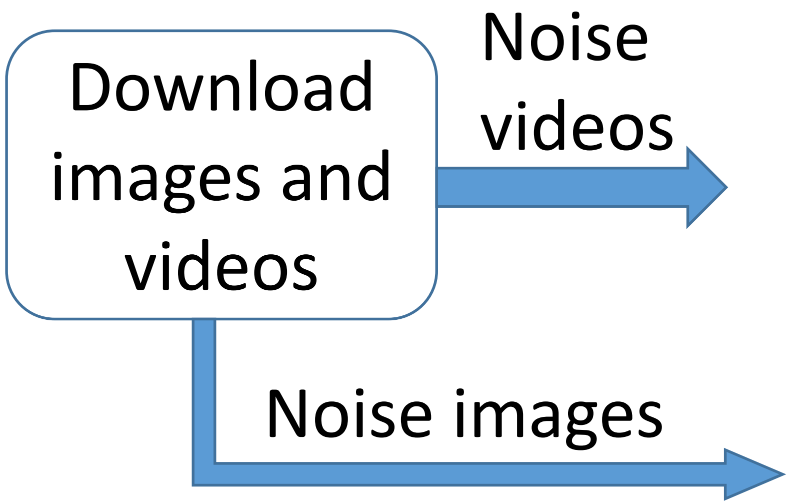
Step1: Web data gathering
Step2: Lead Network is trained by Web videos only.
Step3: The noise of Web images are filtered by Lead Network, fine-tune the lead network by adding image leading to **Exceed Network**.
Step4: Exceed Network is used to filter out noise frames of videos, the LSTM for temporal information.



Weakly supervised for action recognition and event detection

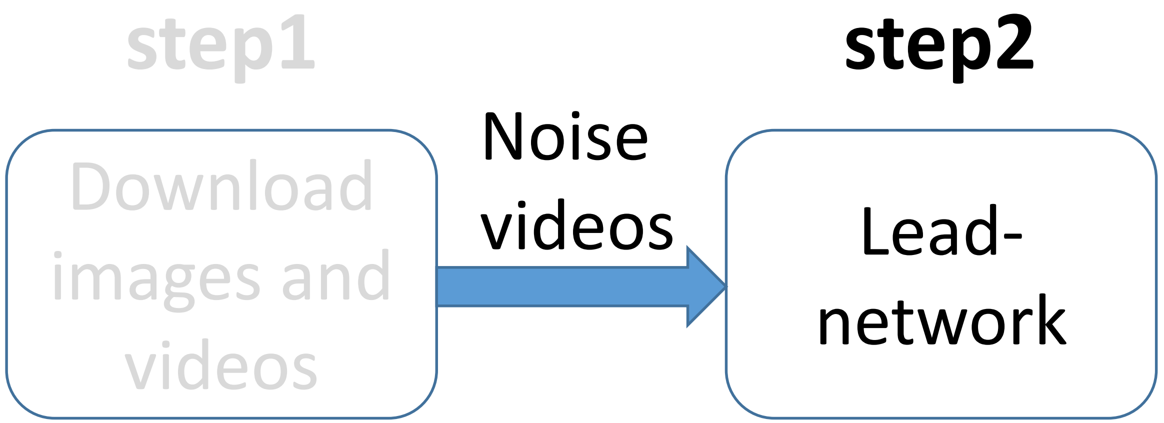
- Framework of Lead-Exceed Neural Network(LENN)

step1



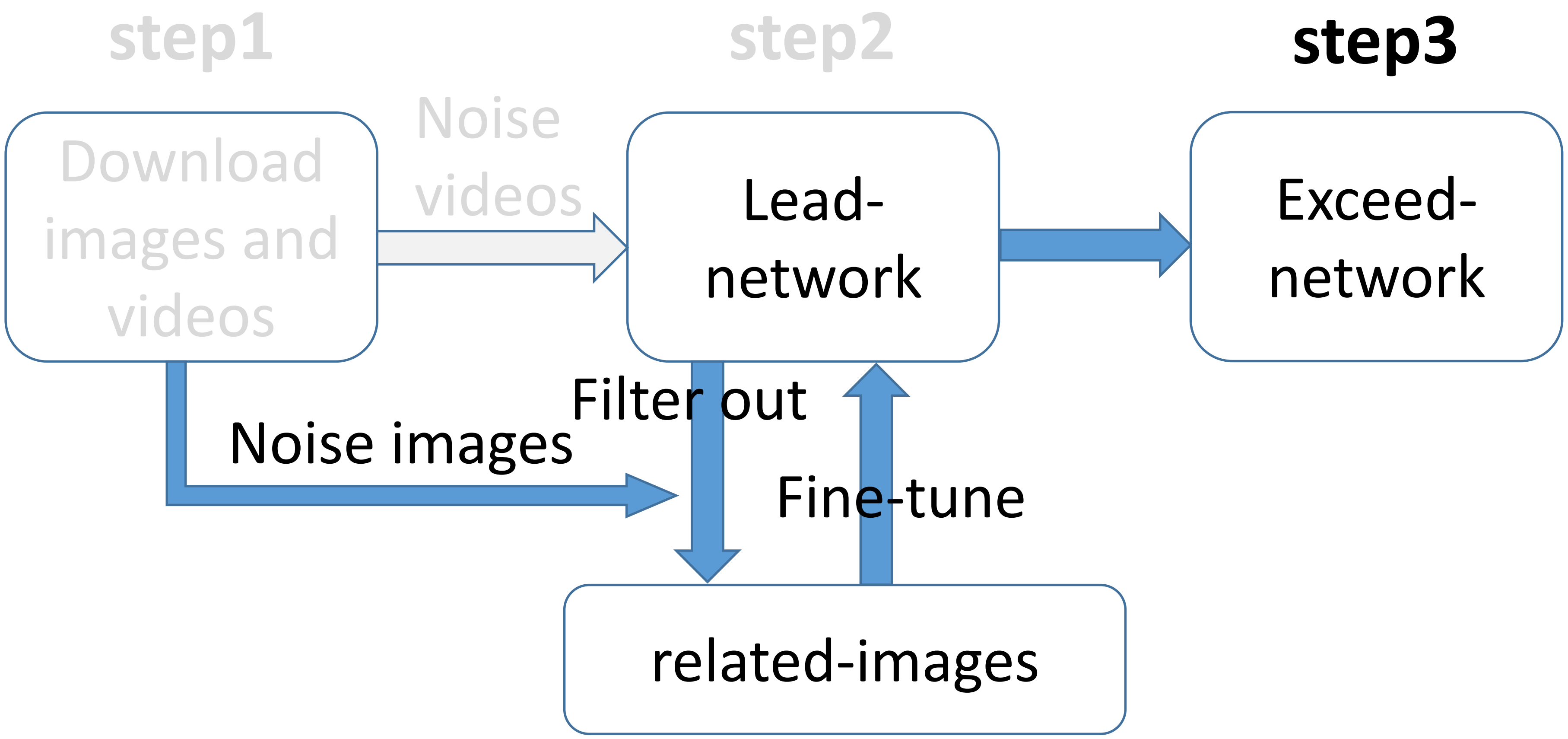
Weakly supervised for action recognition and event detection

- Framework of Lead-Exceed Neural Network(LENN)



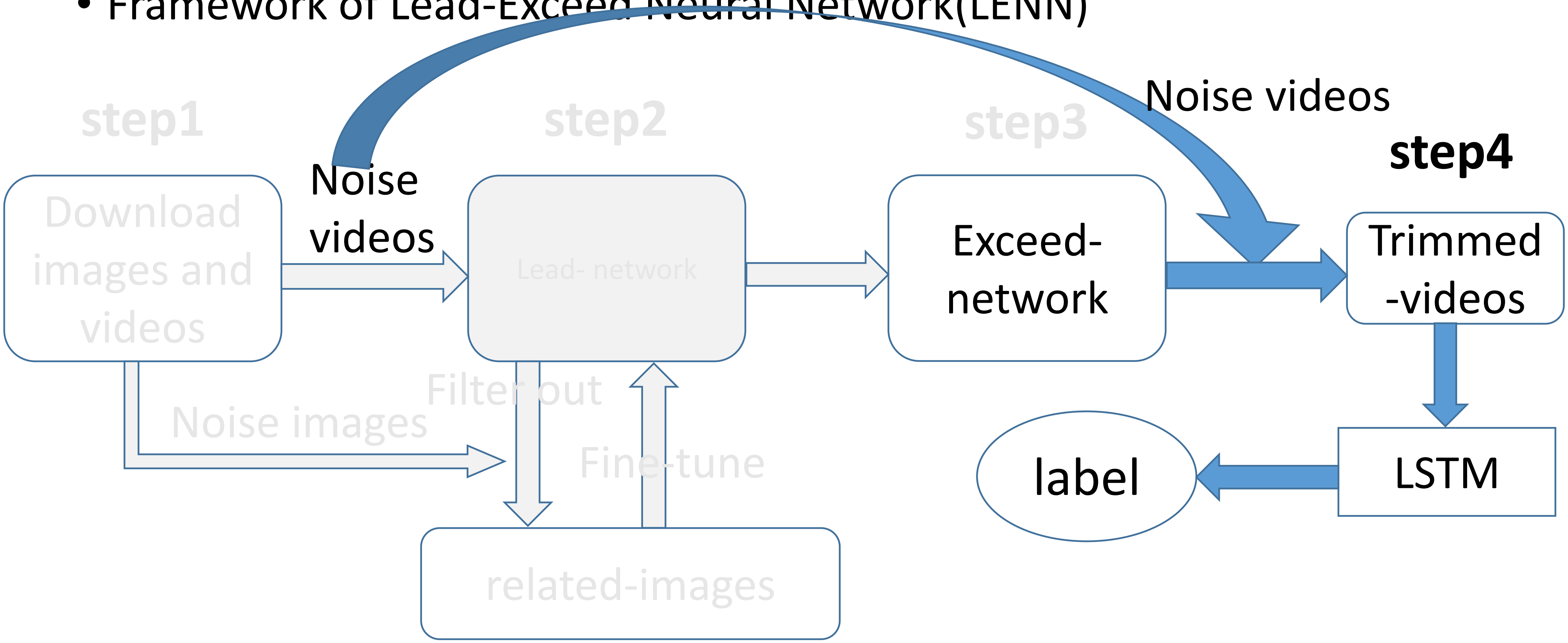
Weakly supervised for action recognition and event detection

- Framework of Lead-Exceed Neural Network(LENN)



Weakly supervised for action recognition and event detection

- Framework of Lead-Exceed Neural Network(LENN)



Weakly supervised for action recognition and event detection

- Data gathering

1. Web images

   About 600 images per category are downloaded from google image search.

2. Web videos

   About 20 videos per category are downloaded from YouTube.

   Video to be less than 15 minutes in length.

   90% of videos have a duration between 5 and 10 minutes.

   Around 60% of the videos are in resolution 1280 $\times$720.

   While the majority have a frame rate of 30 FPS.

Weakly supervised for action recognition and event detection

- Training lead Network

1. Each video is decomposed into a set of frames.
2. Selected the key frames by L1 distance between the previous color histogram and the current one. Around 200 key frames are extracted for a 5 minute video.

$$d_{L_1}(\vec{x},\vec{y}) = polynom\_abs(\vec{x}) = \sum_{i=1}^{I} |x_i - y_i|$$

$\vec{x}_i$: the previous color histogram

$\vec{y}$: the current color histogram

3. Initializing by VGG-16

Weakly supervised for action recognition and event detection

- Training exceed Network

1. To remove useless Web images and keep related ones, used the Lead Network to perform filtering.
2. The remain Web images are used to further fine-tune the Lead Network and obtain the Exceed Network.
3. The Exceed Network is further taken back to trim Web videos to keep related frames.

Weakly supervised for action recognition and event detection

- Training LSTM

1. Input : $\{x_1, x_1, \cdots, x_T\}$, , key frames selected by exceed network.
2. top layer is a soft-max classifier

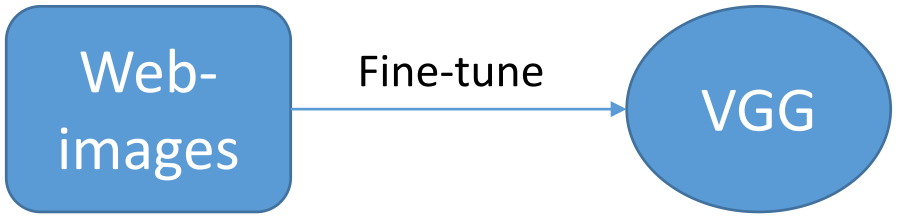
   rolling time k as 25

   the number of hidden state as 256
3. output: $\{y_1, y_2, \cdots, y_T\}$, y $\in \{1, 2, ..., C\}$, labels

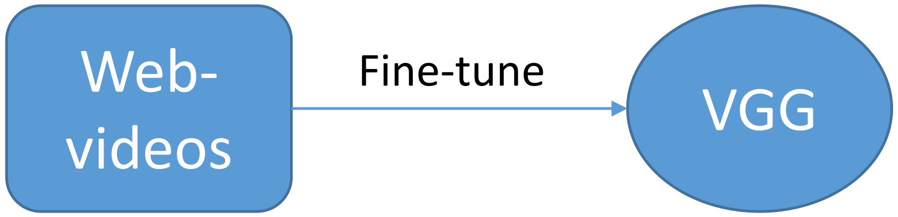
Experiment

- Experiment Result on Action Recognition(UCF101)

Image: fine-tune only by images



video: fine-tune only by videos

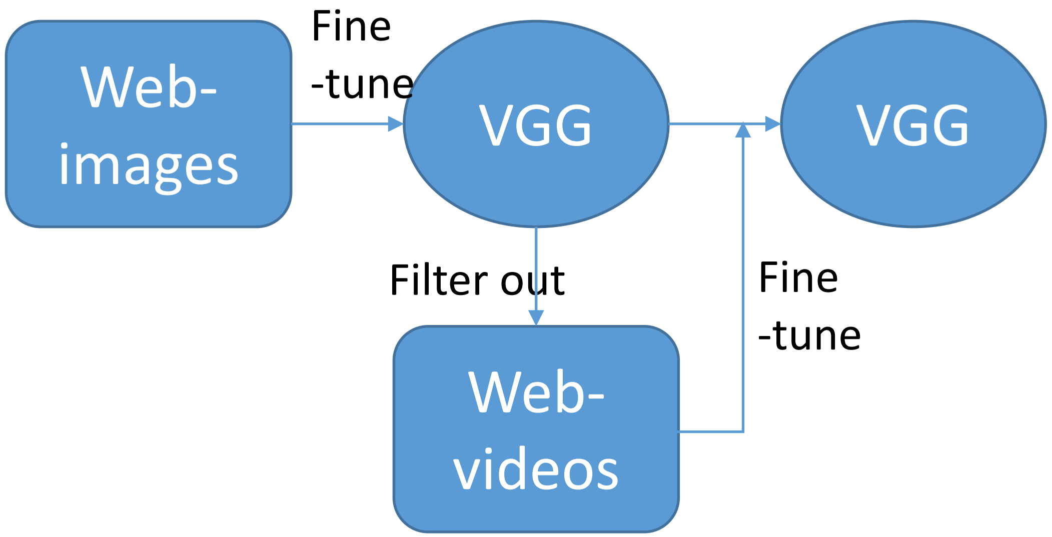


Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Experiment

- Experiment Result on Action Recognition(UCF101)

Image + Video: Using Web images to fine-tune the VGGNet first, then using the fine-tuned model to select key frames from videos

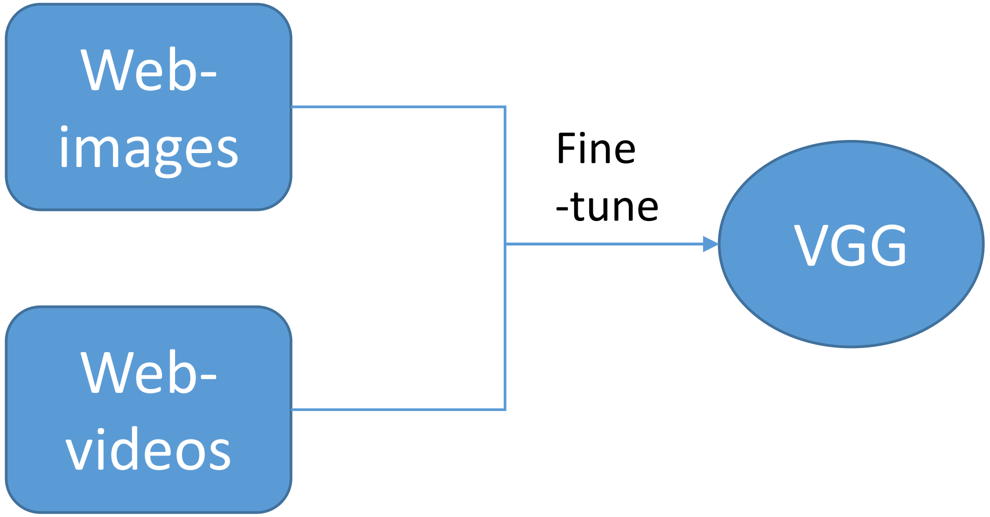


Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Experiment

- Experiment Result on Action Recognition(UCF101)

Noise Mixing: Directly mixing the Web image and video key frames together.

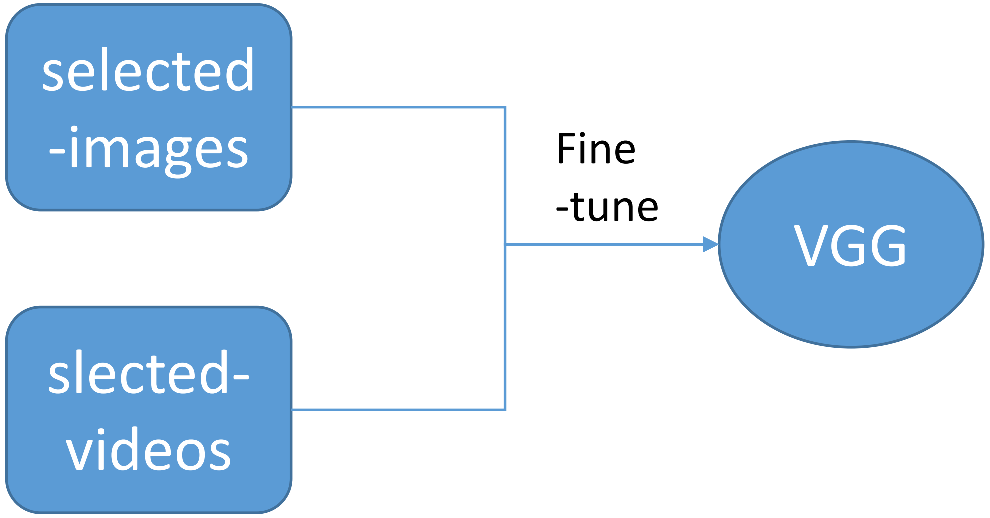


Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Experiment

- Experiment Result on Action Recognition(UCF101)

Mixing: Mixing the selected Web image and video key frames .

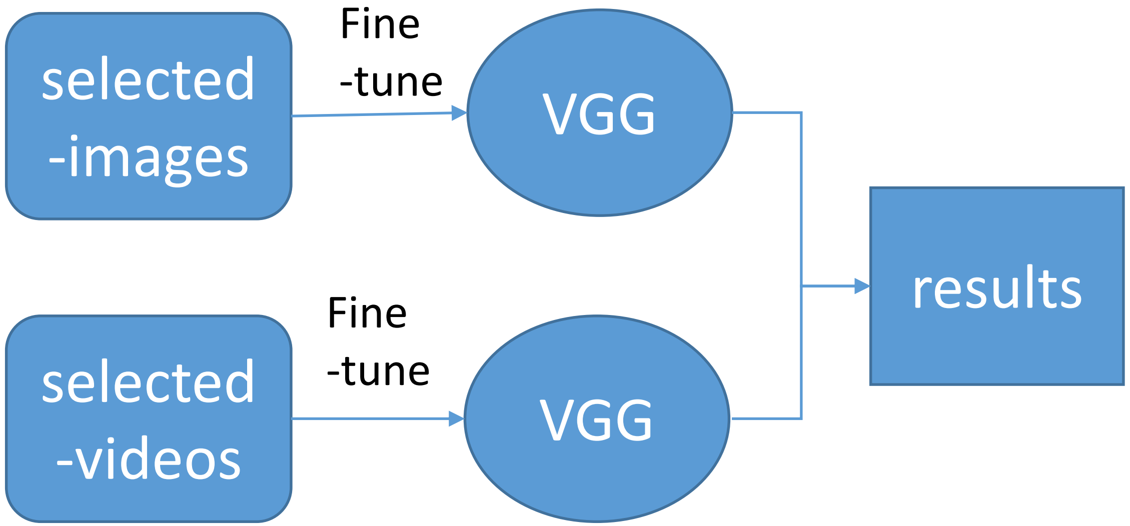


Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Experiment

- Experiment Result on Action Recognition(UCF101)

Late Fusion: Using the selected Web images and videos separately to fine-tune two VGGNETs and then average their scores as final prediction.



Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Experiment

- Experiment Result on Action Recognition(UCF101)

Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Experiment

- Experiment Result on event detection(TRECVID MED 2013 and 2014)

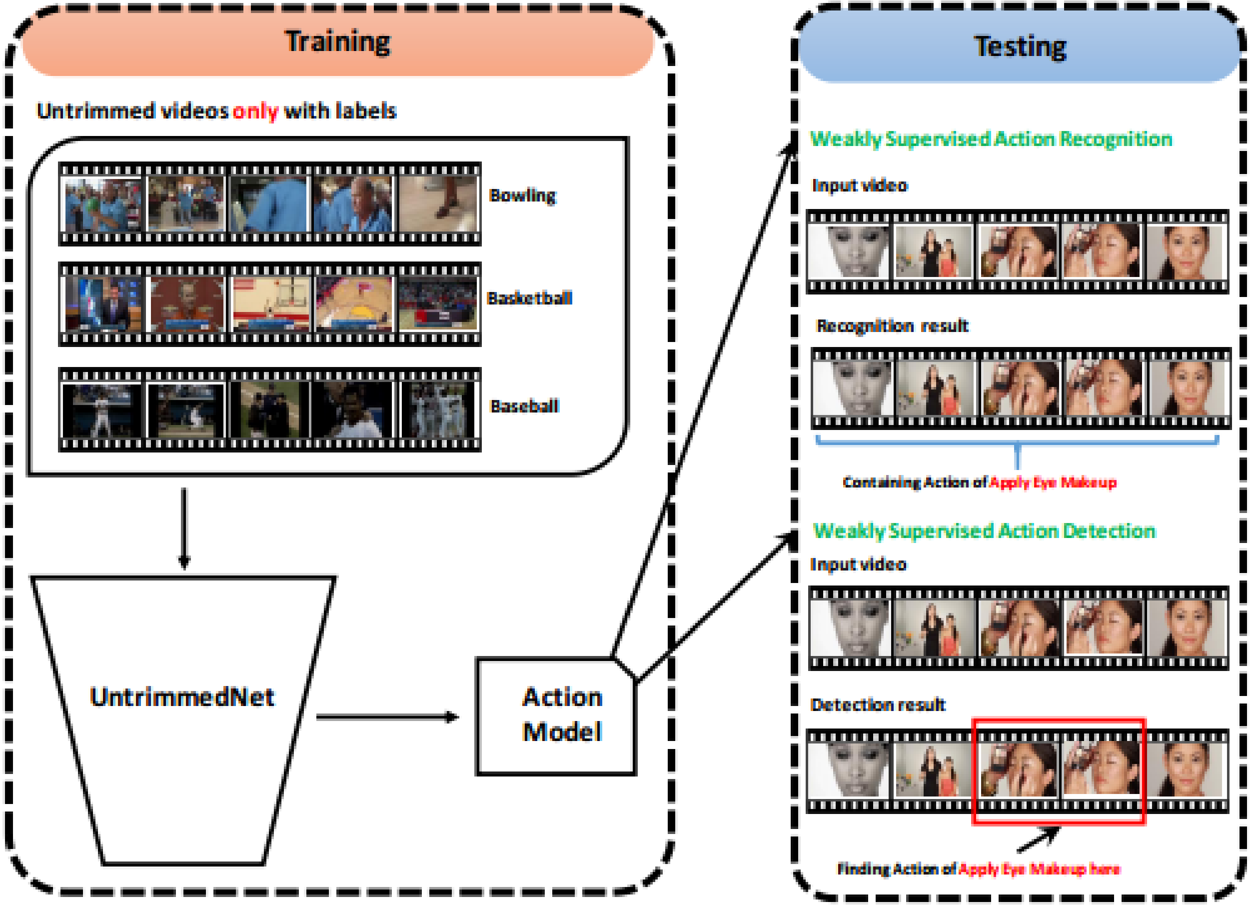
Method	mAP (%)
Concept Discovery [3]	2.3
Bi-concept [16]	6.0
Composite Concept [16]	6.4
EventNet [45]	8.9
Selecting [32]	11.8
Lead-Exceed (Ours)	**16.3**
Lead-Exceed + LSTM (Ours)	**16.7**

Overview

- What is the weakly supervised learning?
- Weakly supervised for action recognition and detection
- Weakly supervised for action detection
- conclusion

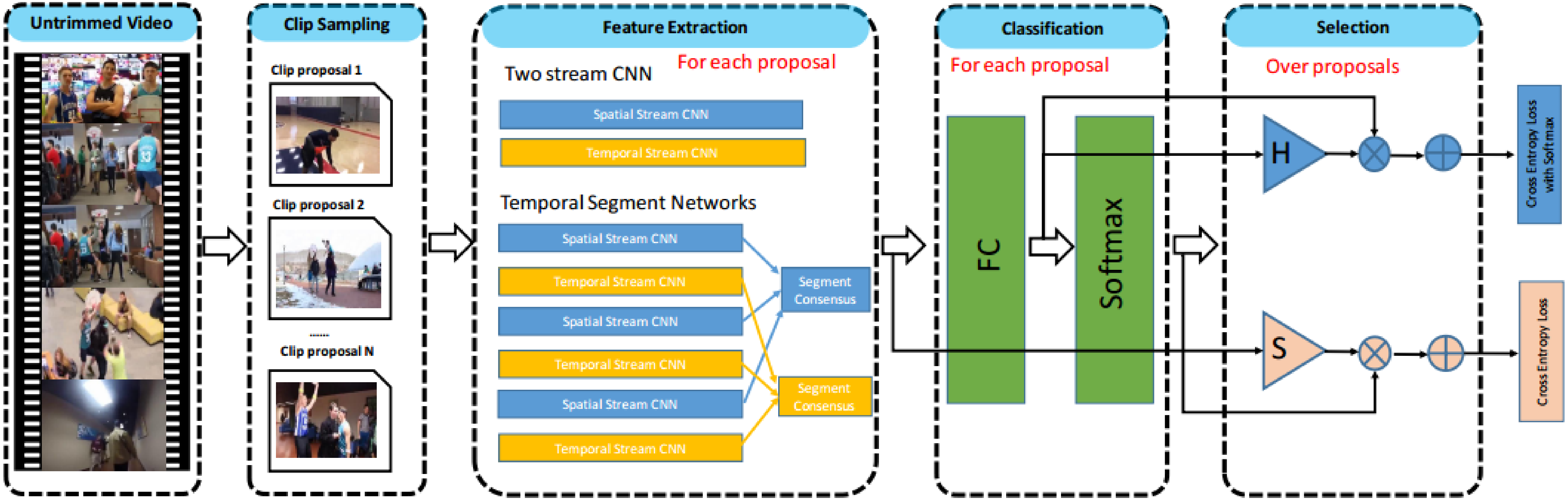
UntrimmedNets for Weakly supervised action detection

- Weakly supervised detection:



UntrimmedNets for Weakly supervised action detection

- The structure of learning from untrimmed videos



UntrimmedNets for Weakly supervised action detection

- Clip sampling

given an untrimmed video V with the duration of T frames, our method generates a set of clips $C = \{c_i\}_{i=1}^{N}$, where N is the number of clips

And $c_i = (b_i, e_i)$ is the beginning and ending location of the ith clips ci.

- method

    Uniform Sampling

    shot-based sampling

    Any other method

UntrimmedNets for Weakly supervised action detection

- Feature learning model

1. Two-Stream CNN
2. Temporal Segment Network(TSN)
3. Any other methods

UntrimmedNets for Weakly supervised action detection

- Classification module

$$\mathbf{x}^c(c) = \mathbf{W}^c \phi(c)$$

$\mathbf{W}^c$ are the model parameters
$\mathbf{x}^c(c)$, a *C*-dimensional score vector
$\phi(c)$ are extracted features

Output from a soft-max layer as follow:

$$\bar{x}_i^c(c) = \frac{\exp(x_i^c(c))}{\sum_{k=1}^{C} \exp(x_k^c(c))},$$

UntrimmedNets for Weakly supervised action detection

- Selection module

1. **hard selection** based on the principle of multiple instance learning
    - Choose top *k* instances with the highest classification scores
    - then average among these selected instances
2. **soft selection** based on the attention-based modeling

    combining the classification scores of all clips and learn an importance weight to rank different clip proposals.

$$x^s(c) = \mathbf{w}^{sT}\phi(c) \qquad \mathbf{w}^s \in \mathcal{R}^D \text{ is the model parameter.}$$

output from a soft-max layer as follow:

$$\bar{x}^s(c_i) = \frac{\exp(x^s(c_i))}{\sum_{n=1}^{N}\exp(x^s(c_n))},$$

UntrimmedNets for Weakly supervised action detection

- Video prediction

$$x_i^p(V) = \sum_{n=1}^{N} x_i^s(c_n) x_i^c(c_n),$$

$$\bar{x}_i^p(V) = \frac{\exp(x_i^r(V))}{\sum_{k=1}^{C} \exp(x_k^r(V))},$$

$x^s(c_n)$: the selection indicator score for clip proposal c_n

$x^c(c_n)$: the classification score for clip proposals c_n

$\bar{x}_i^p(V)$: the softmax operation to normalize the aggregated video-level score

UntrimmedNets for Weakly supervised action detection

- Training

employing the standard back propagation method with cross-entropy loss:

$$\ell(\mathbf{w}) = \sum_{i=1}^{M} \sum_{k=1}^{C} y_{ik} \log \bar{x}_k^p(V_i),$$

y_{ik} is set to 1 if video *Vi* contains action instances of *kth* category, and set to 0 otherwise.

M is the number of training videos.

UntrimmedNets for Weakly supervised action detection

- Experiments on weakly supervised action recognition(WSR)

Method	THUMOS14	ActivityNet (a)	ActivityNet (b)
TSN (3 seg) [50]	67.7%	85.0%	88.5%
TSN (21 seg)	68.5%	86.3%	90.5%
UntrimmedNet (hard)	73.6%	**87.7%**	**91.3%**
UntrimmedNet (soft)	**74.2%**	86.9%	90.9%

UntrimmedNets for Weakly supervised action detection

- Experiments on weakly supervised action recognition(WSR) comparing with state of art method

THUMOS14		ActivityNet	
iDT+FV [45]	63.1%	iDT+FV [45]	66.5%*
Two Stream [40]	66.1%	Two Stream [40]	71.9%*
EMV+RGB [56]	61.5%	C3D [42]	74.1%*
Objects+Motion [19]	71.6%	Depth2Action [57]	78.1%*
TSN (3 seg) [50]	78.5%	TSN (3 seg) [50]	88.8%*
UntrimmedNet (hard)	81.2%	UntrimmedNet (hard)	**91.3%**
UntrimmedNet (soft)	**82.2%**	UntrimmedNet (soft)	90.9%

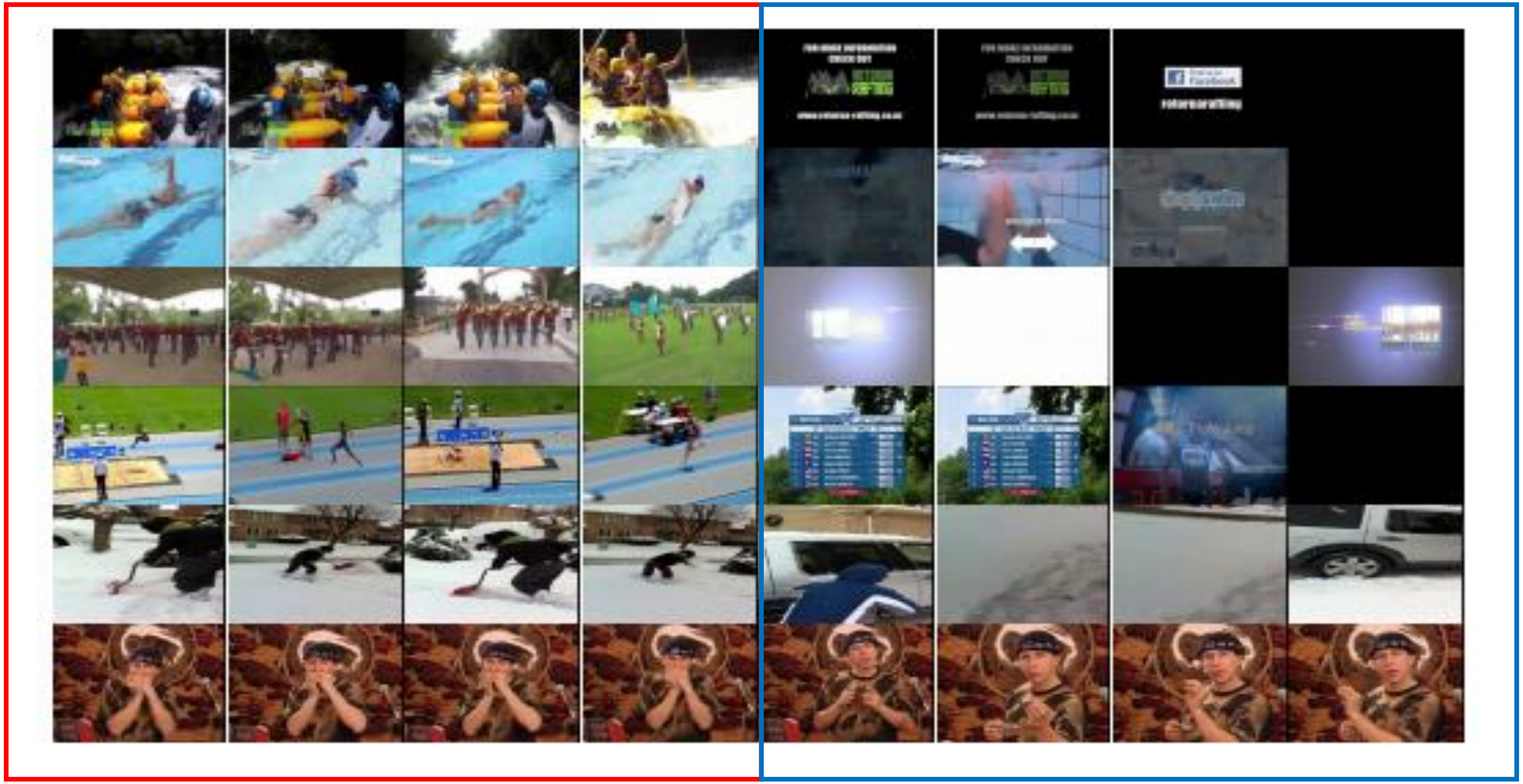
UntrimmedNets for Weakly supervised action detection

- Experiments on weakly supervised action detection(WSD) THUMOS14

Fully supervised method

IoU (α)	$\alpha = 0.5$	$\alpha = 0.4$	$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.1$
Oneata *et al.* [33]*	14.4	20.8	27.0	33.6	36.6
Richard *et al.* [35]*	15.2	23.2	30.0	35.7	39.7
Shou *et al.* [39]*	19.0	28.7	36.3	43.5	47.7
Yeung *et al.* [54]*	17.1	26.4	36.0	44.0	48.9
Yuan *et al.* [55]*	18.8	26.1	33.6	42.6	51.4
UntrimmedNet (soft)	13.7	21.1	28.2	37.7	44.4

UntrimmedNets for Weakly supervised action detection



Overview

- What is the weakly supervised learning?
- Weakly supervised for action recognition and detection
- Weakly supervised for action detection
- conclusion

conclusion

- Weakly supervised learning is a method to solve the problem of time-consuming and expensive for image and vide annotation.
- Weakly supervised learning can use the simple labels (image-level, video-level) for action recognition and action detection.
- Weakly supervised for action recognition get a better performance than some fully supervised methods.
- Weakly supervised for action detection get comparable performance to that of those fully supervised method(with temporal annotation).

Reference

- You Lead, We Exceed: Labor-Free Video Concept Learning by Jointly Exploiting Web Videos and Images**, CVPR2016**
- UntrimmedNets for Weakly Supervised Action Recognition and Detection**, CVPR2017**
- Webly-supervised Video Recognition by Mutually Voting for Relevant Web Images and Web Video Frames**, ECCV2016**
- Weakly Supervised Deep Detection Networks**, CVPR2016**
- Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection**, CVPR2016**

Thank you !