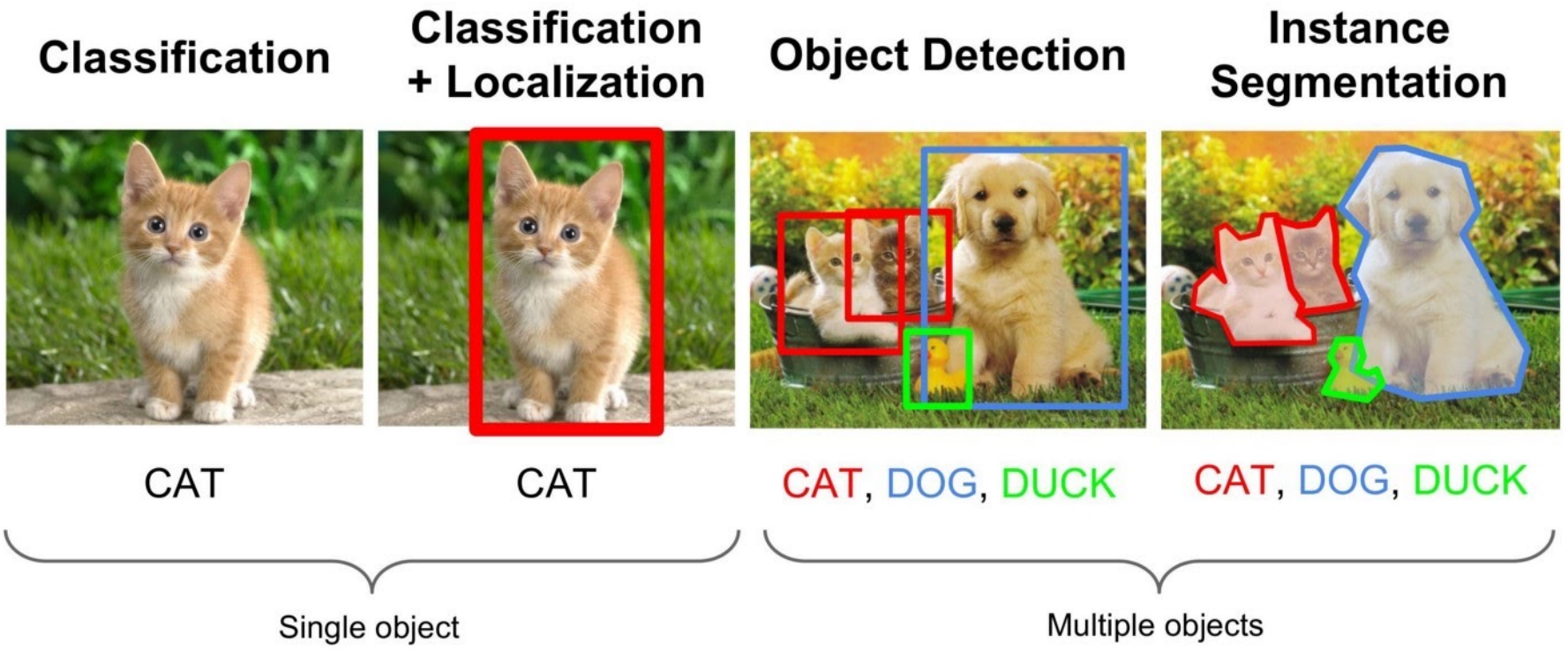
Weakly Supervised learning for Medical Imaging

Waqas Sultani
MedAI Research Group
Information Technology University

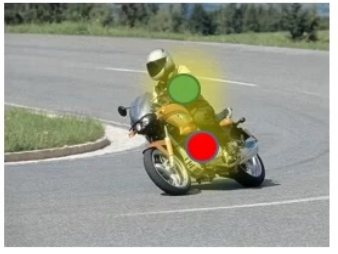
Presentations

3. Questions

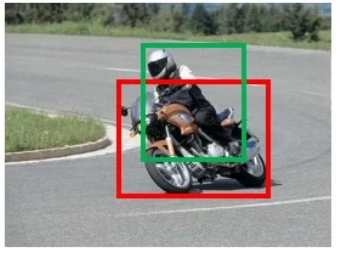

Classification
Classification + Localization
Object Detection
Instance Segmentation
CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
Single object
Multiple objects

Manual supervision for object recognition

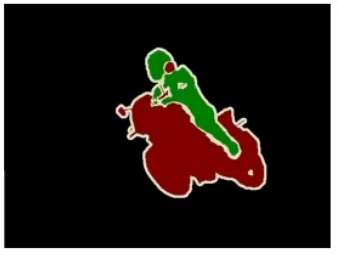
{motorbike,person}



{motorbike (point), person (point)}



{motorbike (b-box), person (b-box)}



{motorbike (pixel labels), person (pixel labels)}

1 sec per class

2.4 sec per instance

10 sec per instance

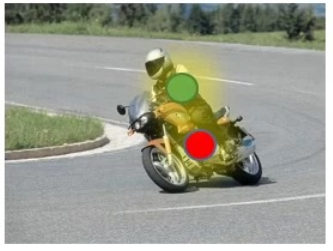
78 sec per instance

annotation time

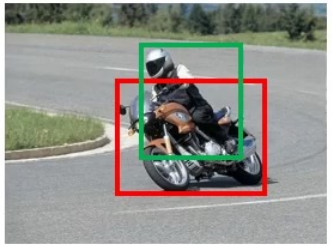
Berman et al., What's the Point: Semantic Segmentation with Point Supervision, ECCV 16

Manual supervision for object recognition

{motorbike,person}

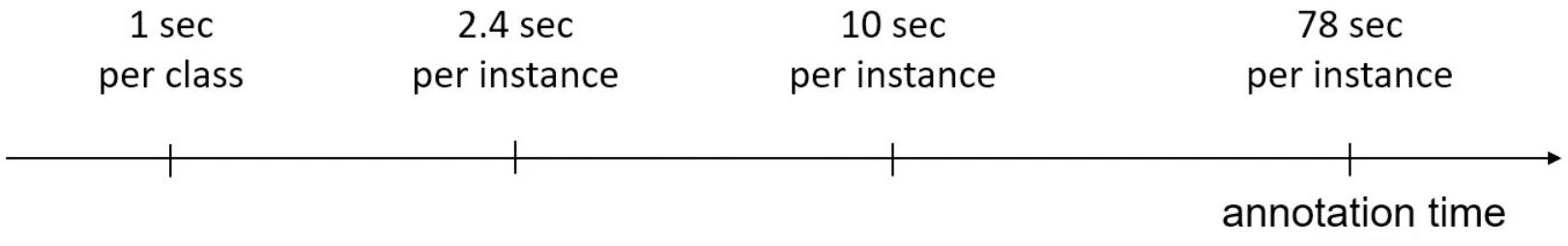


{motorbike (point), person (point)}



{motorbike (b-box), person (b-box)}

{motorbike (pixel labels), person (pixel labels)}

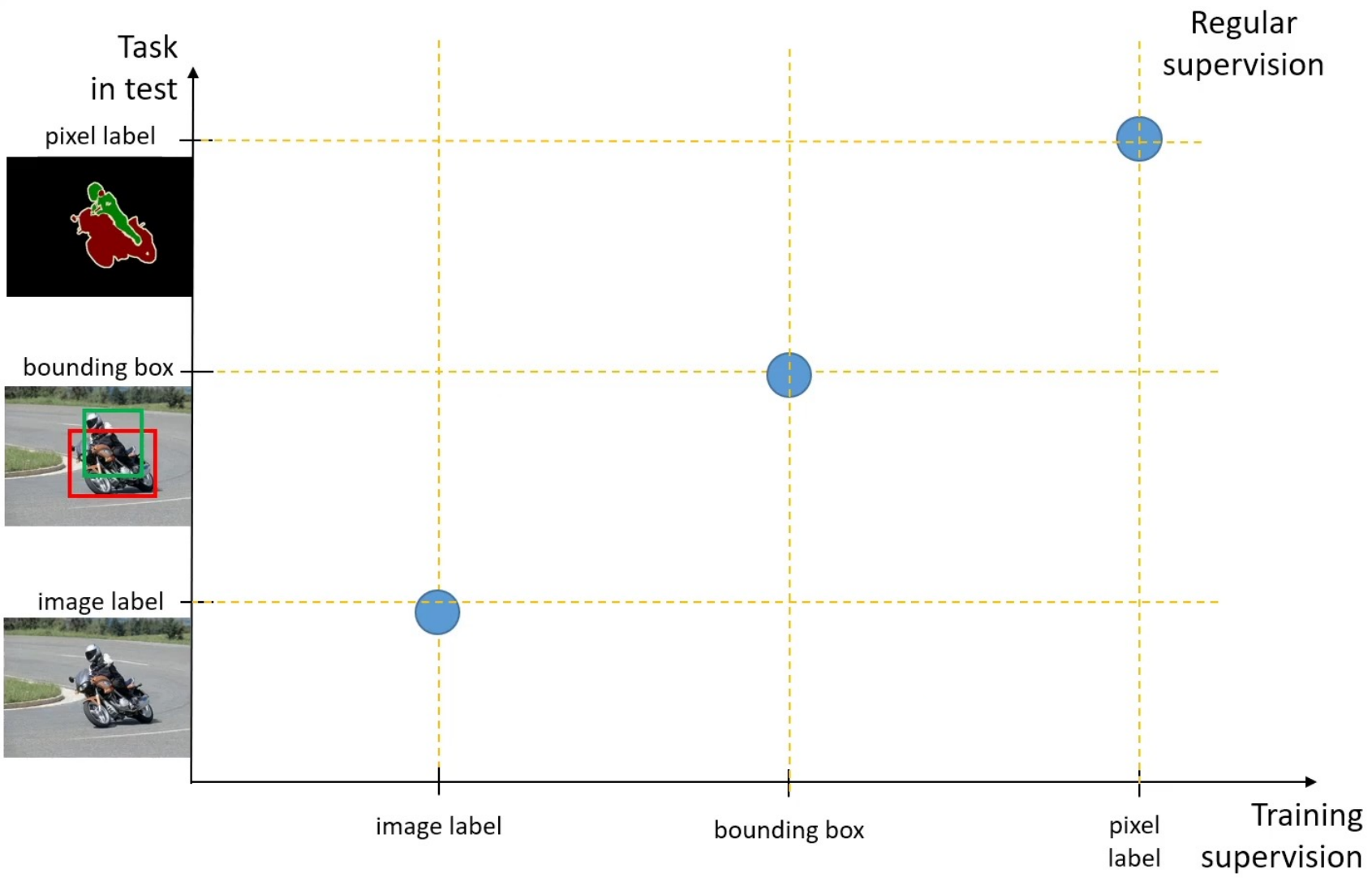


Berman et al., What's the Point: Semantic Segmentation with Point Supervision, ECCV 16

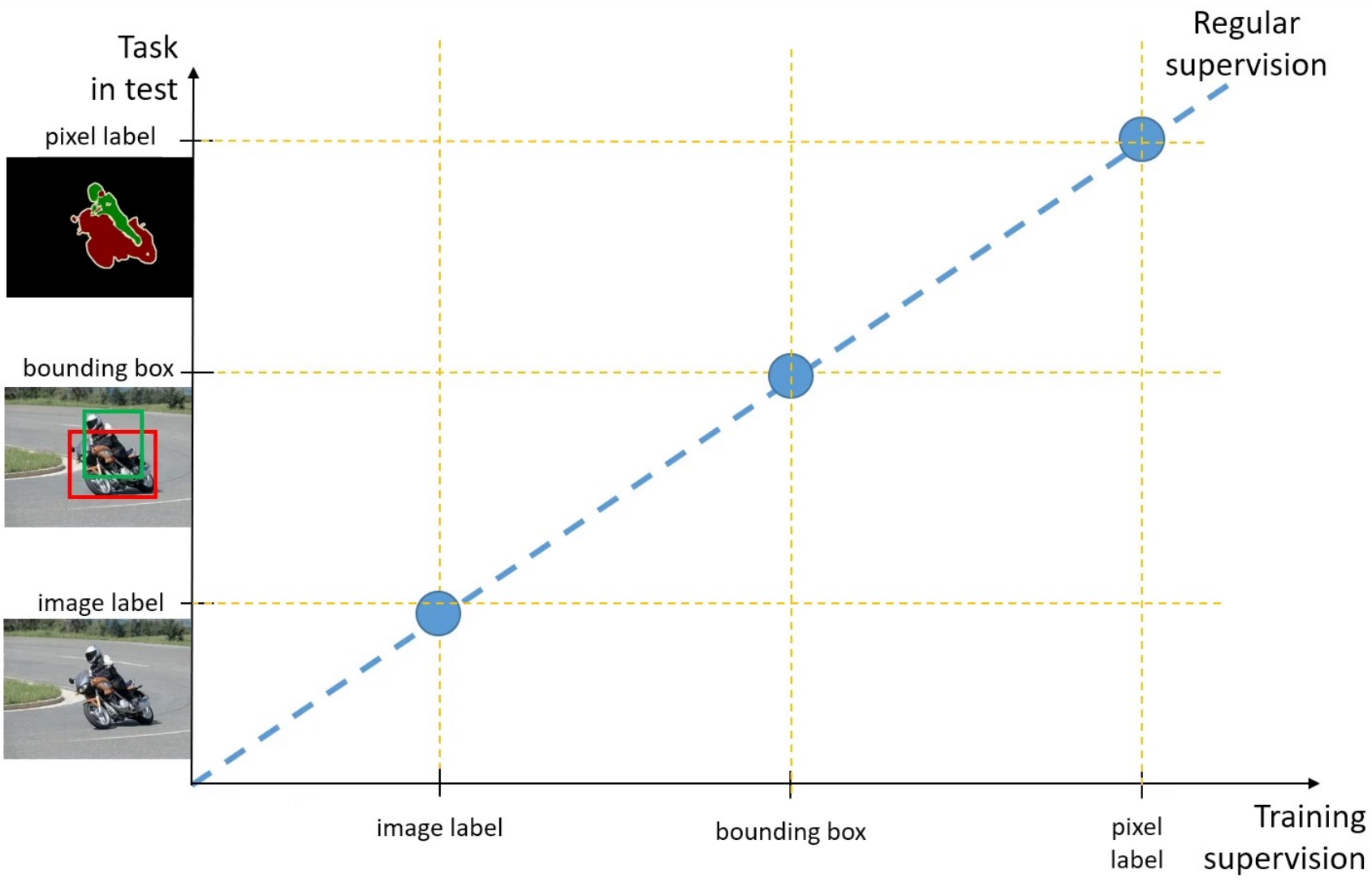
Weak supervision

Lower degree (or cheaper) annotation at train time than the required output at test time

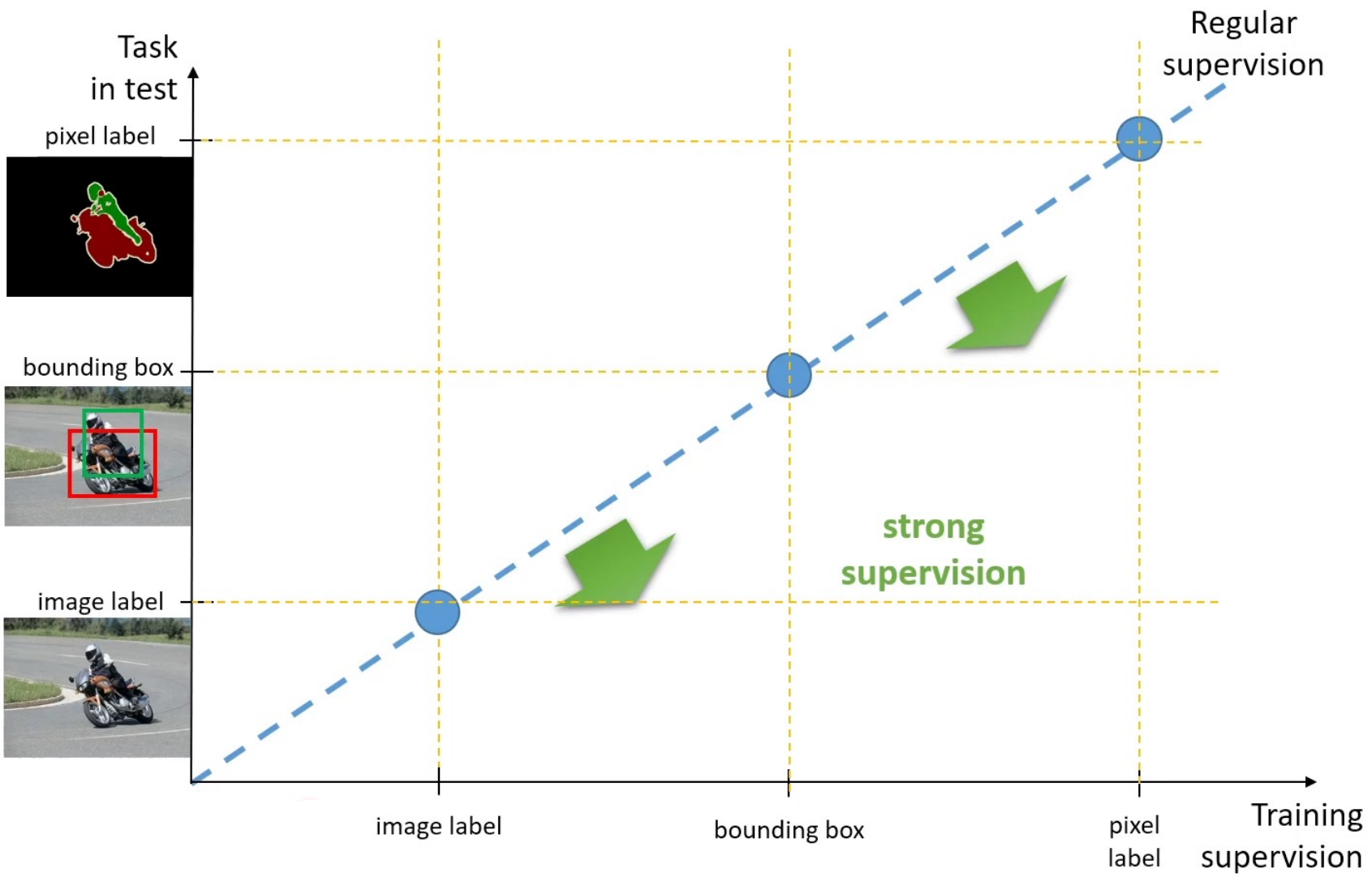
Manual supervision for object recognition



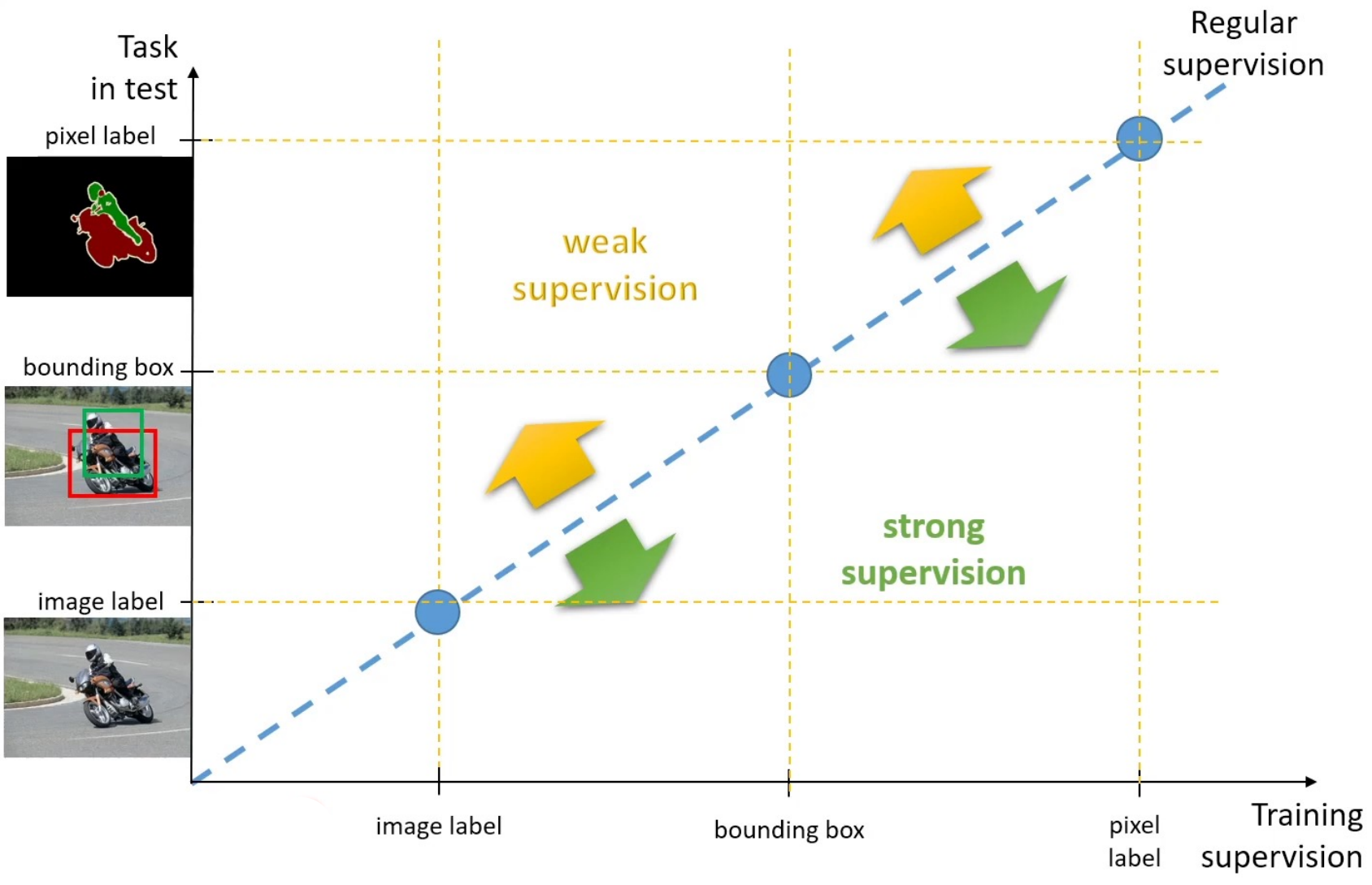
Manual supervision for object recognition



Manual supervision for object recognition



Manual supervision for object recognition



Evaluating Weakly Supervised Object Localization Methods Right (CVPR 2020)

- https://github.com/clovaai/wsolevaluation

Weakly Supervised Object Detection

•

Weakly Supervised Object Detection (WSOD) is the task of **training object detectors with only image tag** supervisions.

•This is different than the baseline supervised object detection since it contains **instance-level annotations**

•Fully supervised object detection methods have become state-of-the-art for object detection. However, due to the **inconvenience of gathering a large amount of data with accurate object-level annotations**, weakly supervised object detection (semi-supervised approach) is recently seeking a lot of attention in medical domains.

Multiple Instance Learning

Problem Motivation

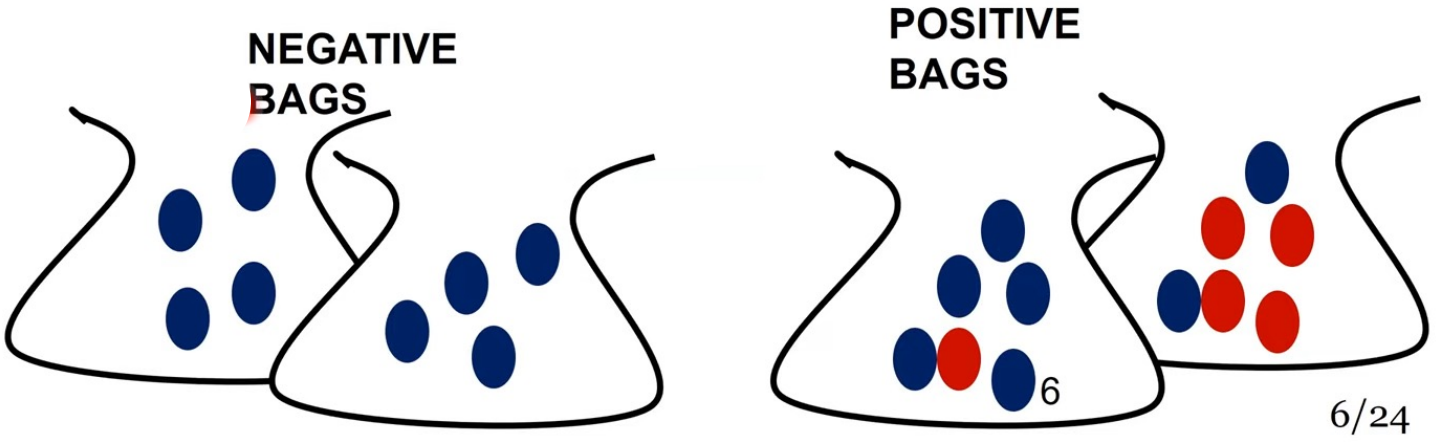
- Locksmith (Dietterich 1997)
  - All employees carry a key ring with many keys
    - Senior employees have access to the supply room.
    - Junior employees do not.
  - You are a locksmith and are required to copy a key for the supply room
    - The employees do not tell you which key opens the storage room, they simply give you their key ring
    - You do not have access to the supply room door
  - Which key do you copy?

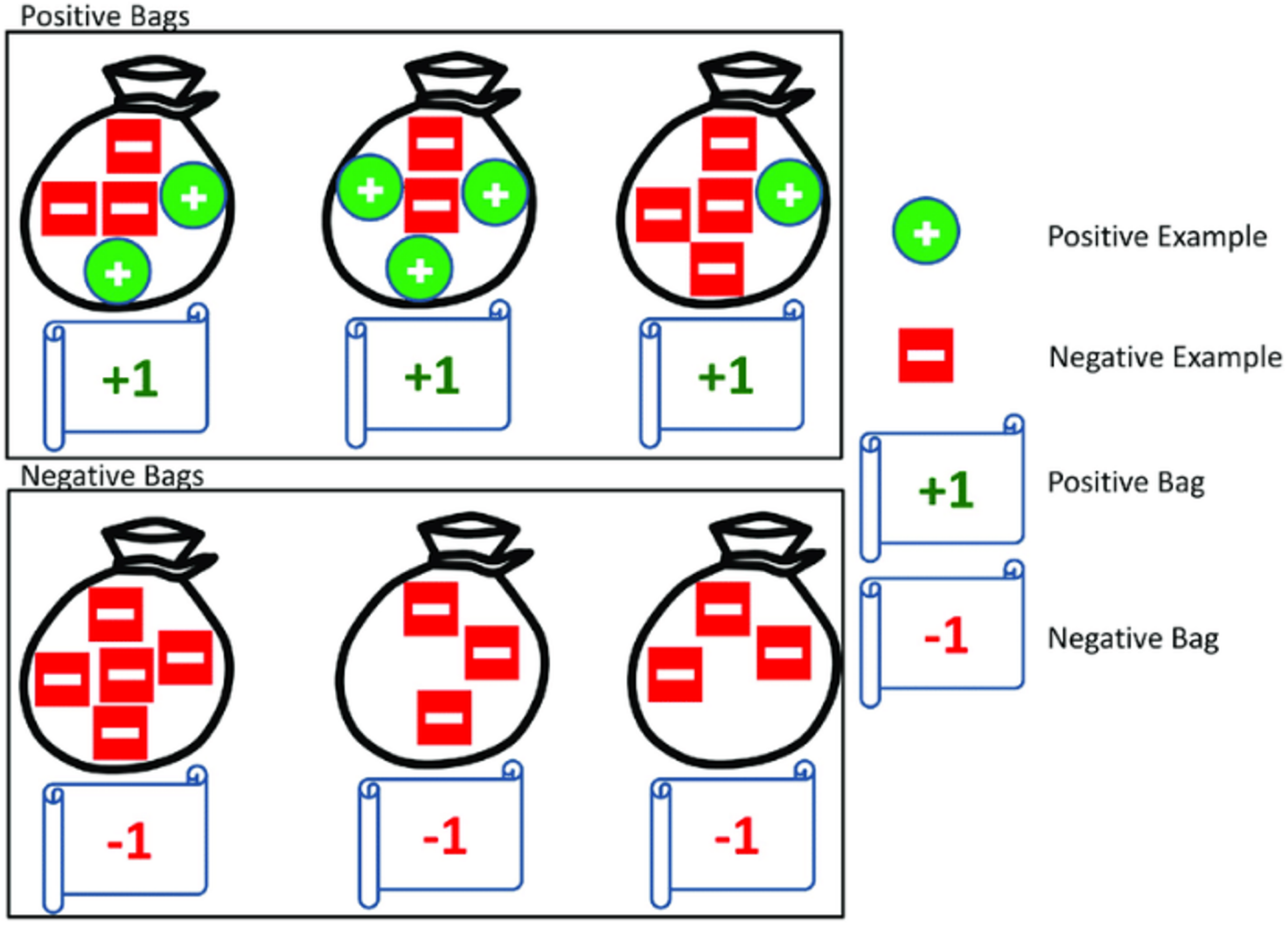
Multiple Instance Learning

- MIL is basically a variation to supervised learning which assigns a single label to a set (bag) of instances instead of labeling individual instances.
- A particular bag is labeled as **negative if all the instances** in the bag are negative
- If **at least one positive instance** is present then that bag is labeled positive.

Learning from Bags

- In MIL, a label is attached to a set of samples.
- A bag is a set of samples
- A sample within a bag is called an instance.
- A bag is labeled as positive if and only if at least one of its instances is positive.




Positive Bags
+1
+1
+1
Negative Bags
-1
-1
-1
Positive Example
Negative Example
+1
Positive Bag
-1
Negative Bag

Multiple Instance Learning (MIL)

- Given:
  - Set of I bags
    - Labeled + or -

$$\mathbf{B} = \{B_1^+, ..B_i^+, B_{i+1}^-, ..., B_I^-\}$$

  - The i[th] bag is a set of J_i samples in some feature space

$$B_i = \{x_{i1}, ..., x_{iJ_i}\}$$

  - Interpretation of labels

$$B_i^+ \Rightarrow \exists j : label(x_{ij}) = 1$$
$$B_i^- \Rightarrow \forall j, label(x_{ij}) = 0$$

- Goal: learn concept
  - What characteristic is common to the positive bags that is not observed in the negative bags

Standard Learning vs. Multiple Instance Learning

- Standard supervised learning
  - Optimize some model (or learn a target concept) given training samples and corresponding labels

$$X = \{x_1,...,x_n\}, Y = \{y_1,...,y_n\}$$

- MIL
  - Learn a target concept given multiple sets of samples and corresponding labels (for the sets).
  - 2 interpretations
    - Learning with uncertain labels / noisy teacher

$$X_1 = \{x_1,...,x_n\}, Y = 1, \{y_1 = ?,...,y_n = ?\}$$

    - Learning target concepts induced by sets of samples
      - Direct analysis of sets

Key Idea:

Iterative Training

Recent approaches in literature

Many are inspired by Class Activation Maps

- Training detection and segmentation branches together

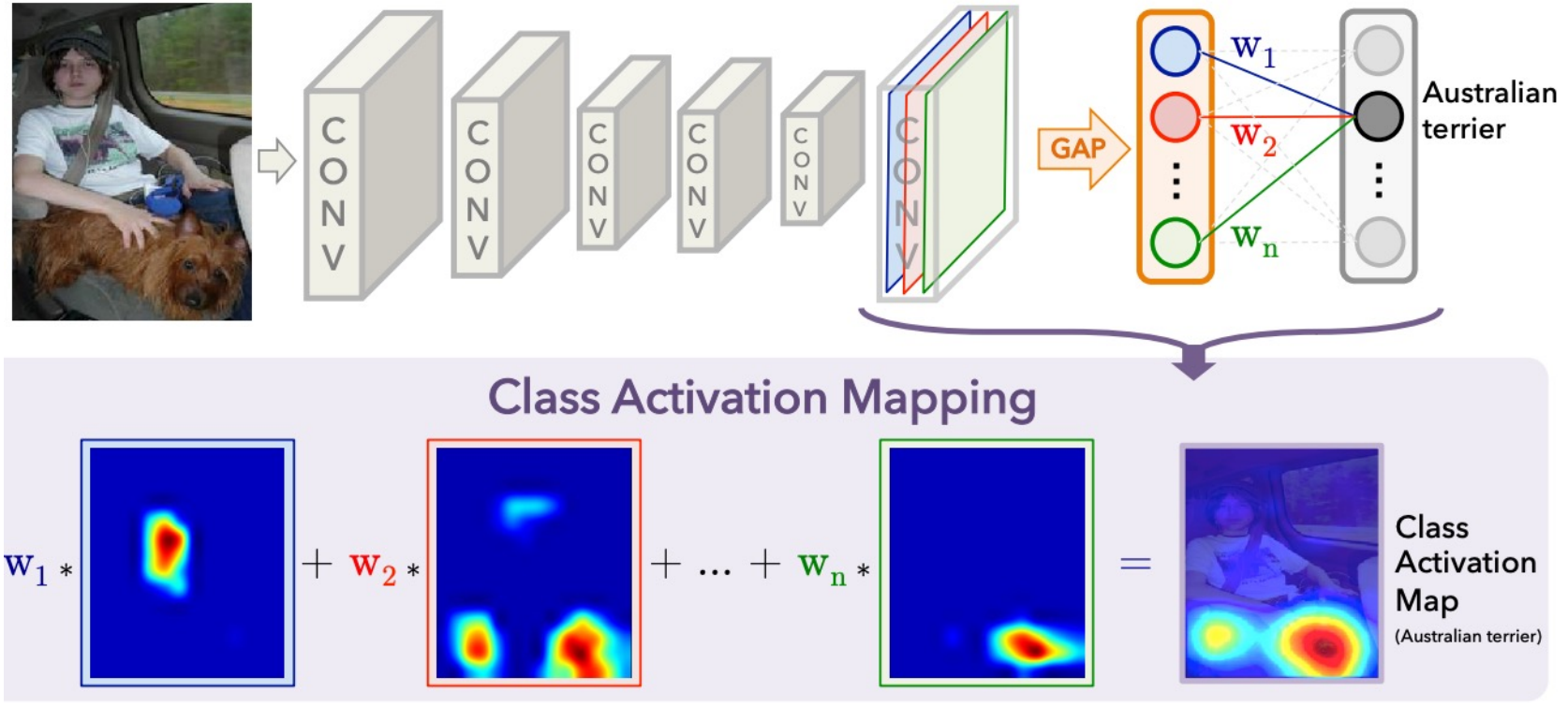
- Building hierarchies to produce better bounding boxes. Going from coarse classes to fine classes level by level.

- Dropout: Remove the most discriminative parts at training time to force the model to learn whole object boundary

- Self-Attention maps: Average pooling of the input feature maps

- Pre-training at very large scale despite on noisy social media videos (label noise and temporal noise)

Class Activation Maps



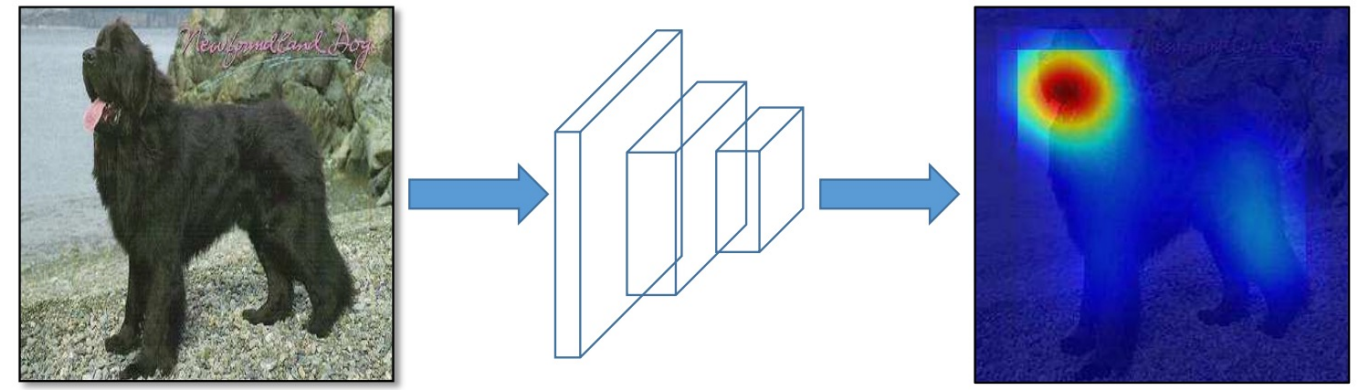
Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

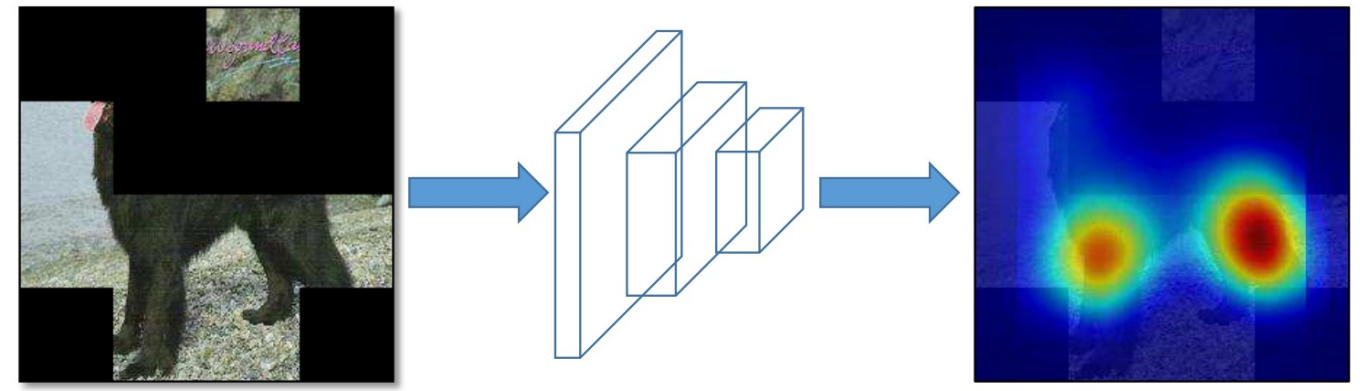
Major Problem with WSOD

Local Minima Problem:

The detector may focus on very discriminative parts of the objects, for eg, the head of a cat.



Full image



Randomly hidden patches

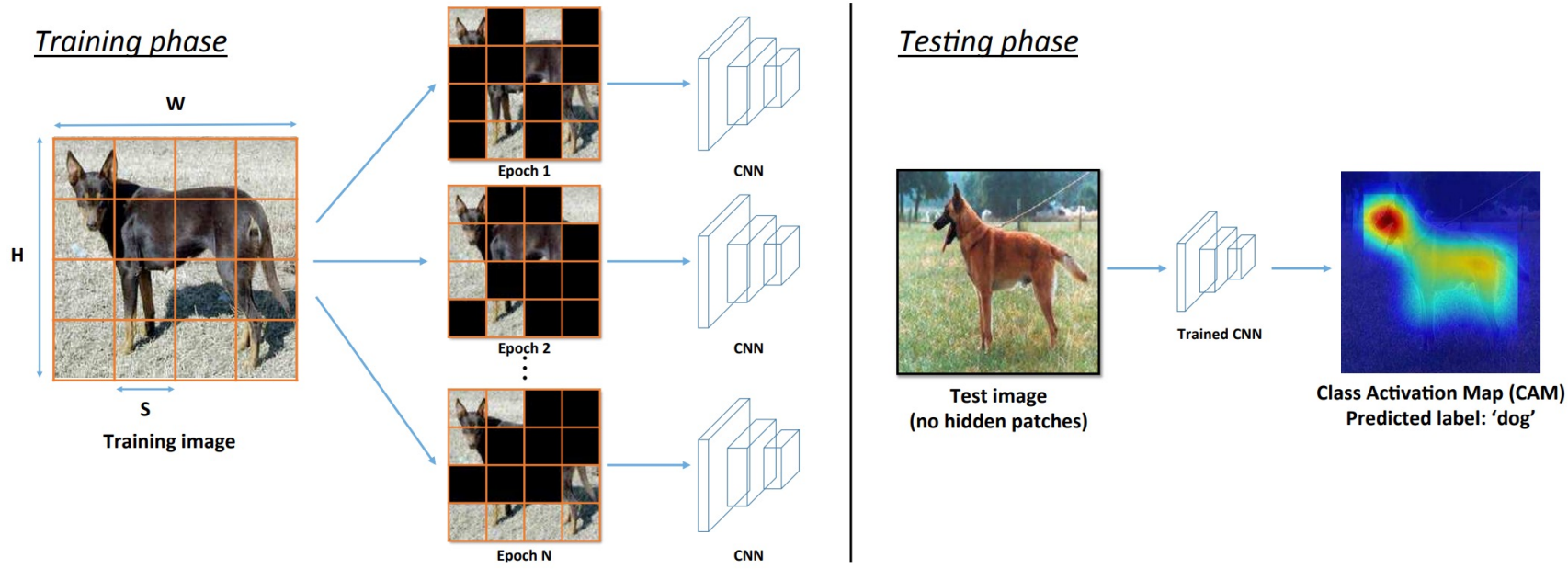
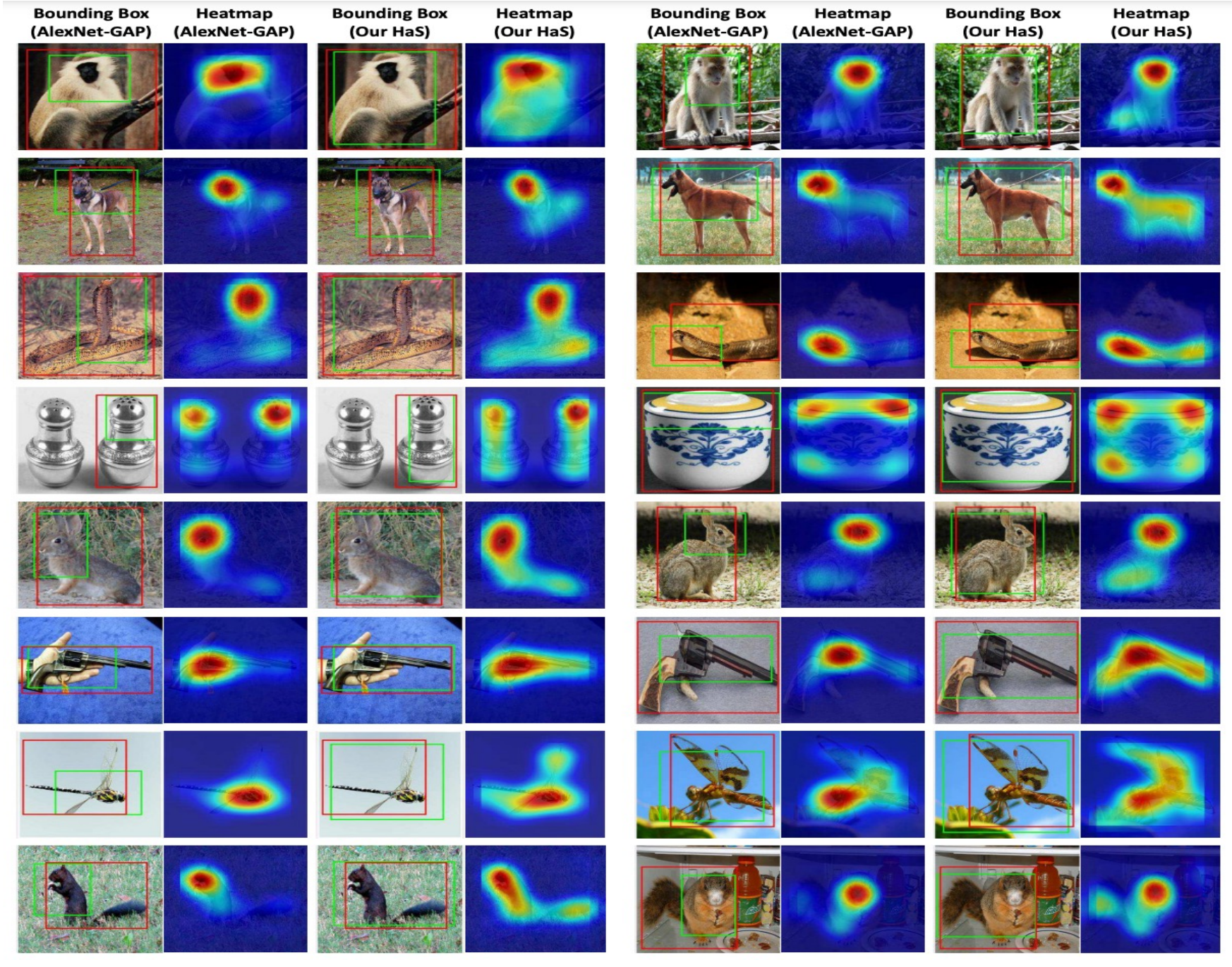


Figure 2. **Approach overview.** *Left:* For each training image, we divide it into a grid of $S \times S$ patches. Each patch is then randomly hidden with probability p_{hide} and given as input to a CNN to learn image classification. The hidden patches change randomly across different epochs. *Right:* During testing, the full image without any hidden patches is given as input to the trained network.


Bounding Box (AlexNet-GAP)
Heatmap (AlexNet-GAP)
Bounding Box (Our HaS)
Heatmap (Our HaS)
Bounding Box (AlexNet-GAP)
Heatmap (AlexNet-GAP)
Bounding Box (Our HaS)
Heatmap (Our HaS)

In Defence of Negative Mining for Annotating Weakly Labelled Data

Parthipan Siva, Chris Russell*, and Tao Xiang

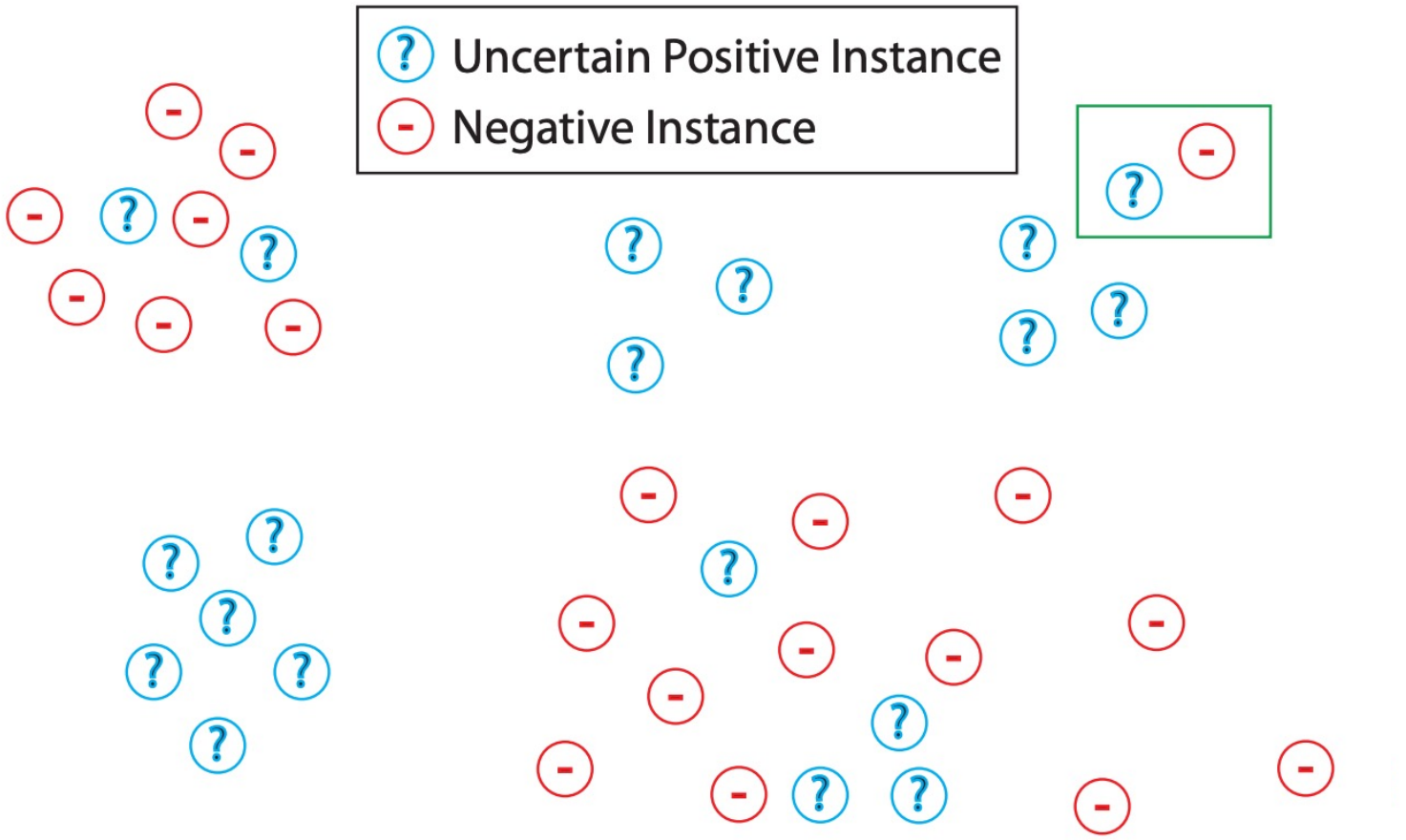
Queen Mary, University of London,
Mile End Road, London E1 4NS, United Kingdom
{psiva,chrisr,txiang}@eecs.qmul.ac.uk

Intuitively, distinctive concept segments are identified as those among testing data whose nearest neighbours among N is as far as possible.

In Defence of Negative Mining for Annotating Weakly Labelled Data

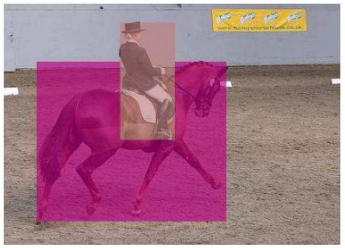
Parthipan Siva, Chris Russell*, and Tao Xiang

Queen Mary, University of London,
Mile End Road, London E1 4NS, United Kingdom
{psiva,chrisr,txiang}@eecs.qmul.ac.uk

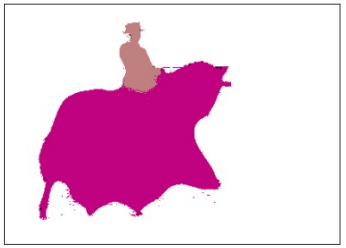


Discriminative Segment Annotation in Weakly Labeled Video

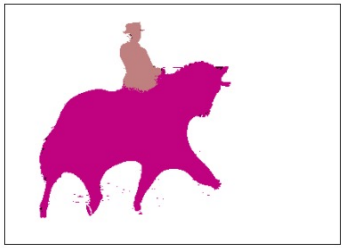
Kevin Tang[1,2] kdtang@cs.stanford.edu
Rahul Sukthankar[1] rahuls@cs.cmu.edu
Jay Yagnik[1] jyagnik@google.com
Li Fei-Fei[2] feifeili@cs.stanford.edu
[1]Google Research [2]Computer Science Department, Stanford University
https://sites.google.com/site/segmentannotation/



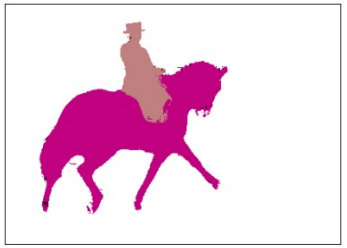
Example
input rectangles



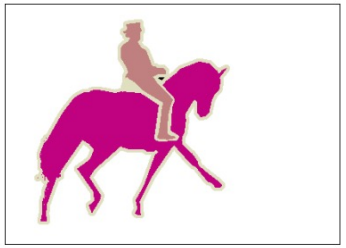
Output after
1 training round



After
5 rounds



After
10 rounds



Ground
truth

Simple Does It: Weakly Supervised Instance and Semantic Segmentation

Anna Khoreva[1] Rodrigo Benenson[1] Jan Hosang[1] Matthias Hein[2] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Saarland University, Saarbrücken, Germany

C1 Background: Since the bounding boxes are expected to be exhaustive, any pixel not covered by a box is labelled as background.

C2 Object extent: The box annotations bound the extent of each instance. Assuming a prior on the objects shapes (e.g. oval-shaped objects are more likely than thin bar or full rectangular objects), the box also gives information on the expected object area. We employ this size information during training.

C3 Objectness: Other than extent and area, there are additional object priors at hand. Two priors typically used are spatial continuity and having a contrasting boundary with the background. In general we can harness priors about object shape by using segment proposal techniques [35], which are designed to enumerate and rank plausible object shapes in an area of the image.

Simple Does It: Weakly Supervised Instance and Semantic Segmentation

Anna Khoreva[1] Rodrigo Benenson[1] Jan Hosang[1] Matthias Hein[2] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Saarland University, Saarbrücken, Germany

Real-World Anomaly Detection in Surveillance videos

Waqas Sultani, Chen Chen, Mubarak Shah

Computer Vision and Pattern Recognition (CVPR), 2018

Real-world Anomaly Detection in Surveillance Videos

Waqas Sultani[1]
[1]Department of Computer Science
Information Technology University, Pakistan

Chen Chen[2], Mubarak Shah[2]
[2]Center for Research in Computer Vision
University of Central Florida, Orlando, FL,USA

waqas5163@gmail.com, chenchen870713@gmail.com, shah@crcv.ucf.edu

Abstract

Surveillance videos are able to capture a variety of realistic anomalies. In this paper, we propose to learn anoma-

etc. to increase public safety. However, the monitoring capability of law enforcement agencies has not kept pace. The result is that there is a glaring deficiency in the utilization of

US police testing AI that learns to spot crimes in CCTV footage

TECHNOLOGY 14 August 2018

By Richard Kemeny

Just like humans, machines can be trained to spot illicit activity
MediaProduction/Getty

SURVEILLANCE cameras are already ubiquitous, but you still need trained guards to spot crime. Keeping that level of attention up can be hard, even for the most focused individual.

Now police in Orlando, Florida, have been testing a system that automatically scans CCTV looking for potentially illicit activity.

Previous AIs have been trained to spot specific activities, such as violence. Yet as crime comes in many forms, these systems are inherently limited. Waqas Sultani at the Information Technology University in Pakistan and his colleagues tried to incorporate a ...

Contributions

- Generic framework for real world anomaly detection
- New Multiple Instance Ranking loss with Smoothness and Sparsity constraints.
- State-of-the-art anomaly detection results
- Large-scale benchmark dataset
  - 1,900 real-world surveillance videos of 128 hours
  - 27 times more videos than the largest dataset.
  - Average video length is over four times
  - 13 real world anomalies
- Anomalous activity recognition results of baseline methods

Comparisons with other datasets

	# of videos	Average # of frames	Dataset length	Example anomalies
UCSD Ped1 [27]	70	201	5 min	Bikers, small carts, walking across walkways
UCSD Ped2 [27]	28	163	5 min	Bikers, small carts, walking across walkways
Subway Entrance [3]	1	121,749	1.5 hours	Wrong direction, No payment
Subwa Exit [3]	1	64,901	1.5 hours	Wrong direction, No payment
Avenue [28]	37	839	30 min	Run, throw, new object
UMN [2]	5	1290	5 min	Run
BOSS [1]	12	4052	27 min	Harass, disease, panic
Abnormal Crowd [31]	31	1408	24 min	Panic, fight, congestion, obstacle, neutral
Ours	**1900**	**7247**	**128 hours**	**Abuse, arrest, arson, assault, accident, burglary, fighting, robbery**

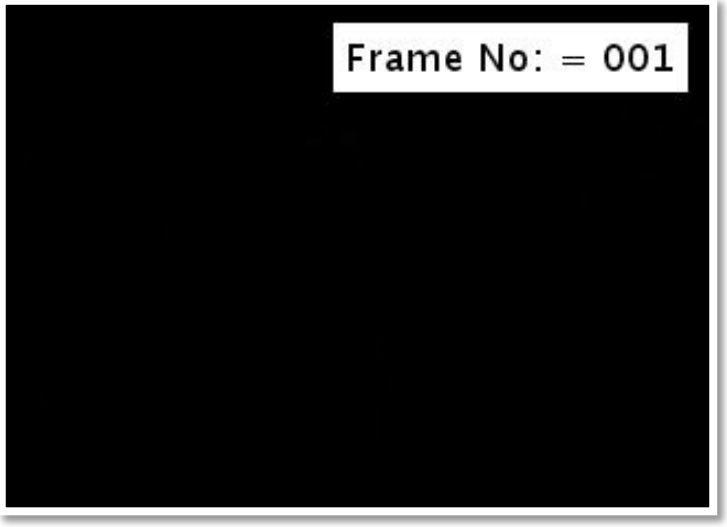
Our Anomaly Detection Dataset

Anomaly	Number of videos
Burglary	100
Fighting	50
Road Accidents	150
Robbery	150
Shooting	50
Shoplifting	50
Stealing	100
Abuse	50
Arrest	50
Arson	50
Assault	50
Explosion	50
Vandalism	50
Normal	950

Example Normal Videos in the Dataset

Note: we fast play or trim some videos due to their long durations.


Frame No: = 001
JMC
Secure
Model: DS-2CE16D1T-IT5
Desc: Actual Video Footage, Day / Night Time - recorded on Hikvision HD-TVI DVR


Frame No: = 001


Frame No: = 001


Frame No: = 001
CAM 6

Example Anomalous Videos in the Dataset

Abuse

Arrest

Arson

Assault

Burglary

Stealing

Note: we fast play or trim some videos due to their long durations.

Explosion

Robbery

Arrest

Fighting

Road Accident

Shooting

Shoplifting

Vandalism

Weakly labeled Crime Detection Framework

Ranking

$$f(\mathcal{V}_a) > f(\mathcal{V}_n),$$

MIL Ranking

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

Loss Function

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

Weakly labeled Crime Detection Framework

Ranking

$$f(\mathcal{V}_a) > f(\mathcal{V}_n),$$

MIL Ranking

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

Loss Function

+1 0

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

$$+\lambda_1 \overbrace{\sum_{i}^{(n-1)} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f(\mathcal{V}_a^i)}^{②},$$

Weakly labeled Crime Detection Framework

Ranking

$$f(\mathcal{V}_a) > f(\mathcal{V}_n),$$

MIL Ranking

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

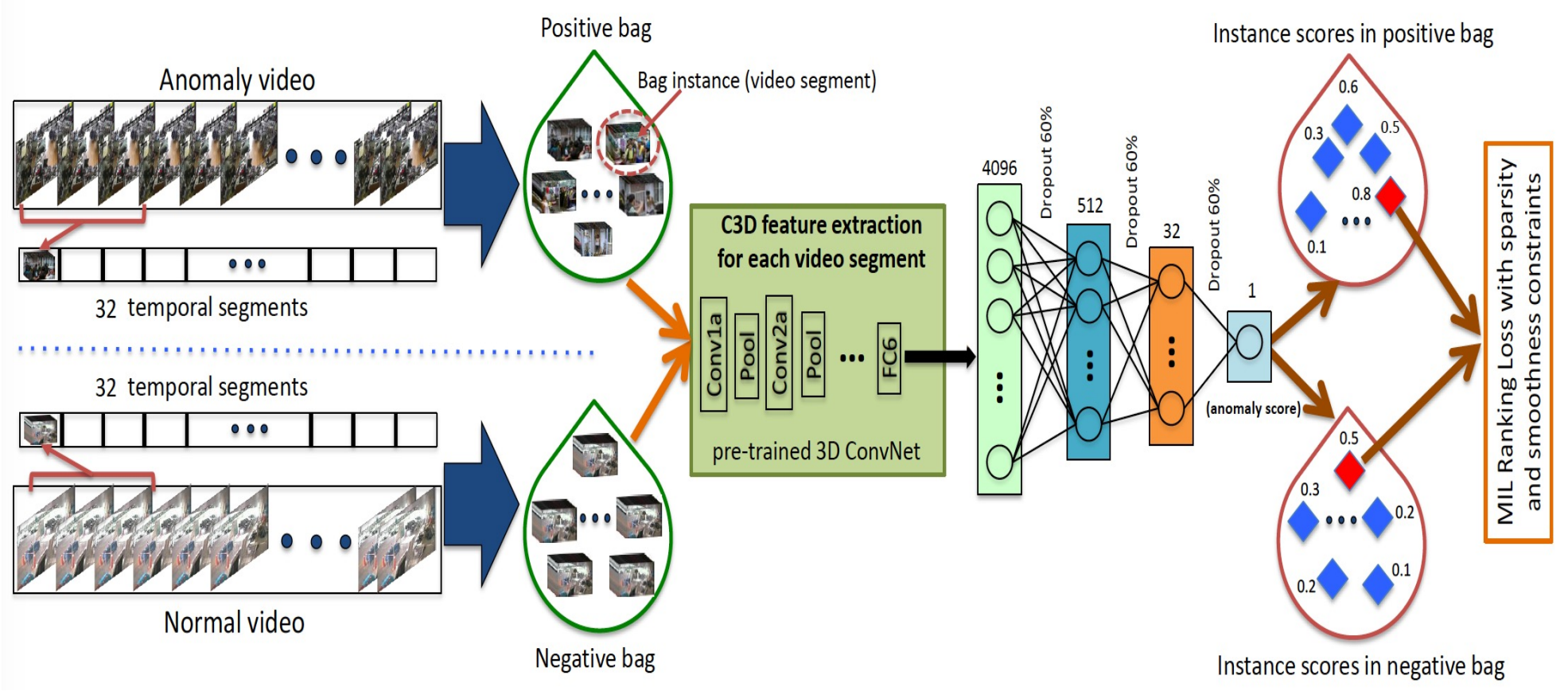
Loss Function

0 1

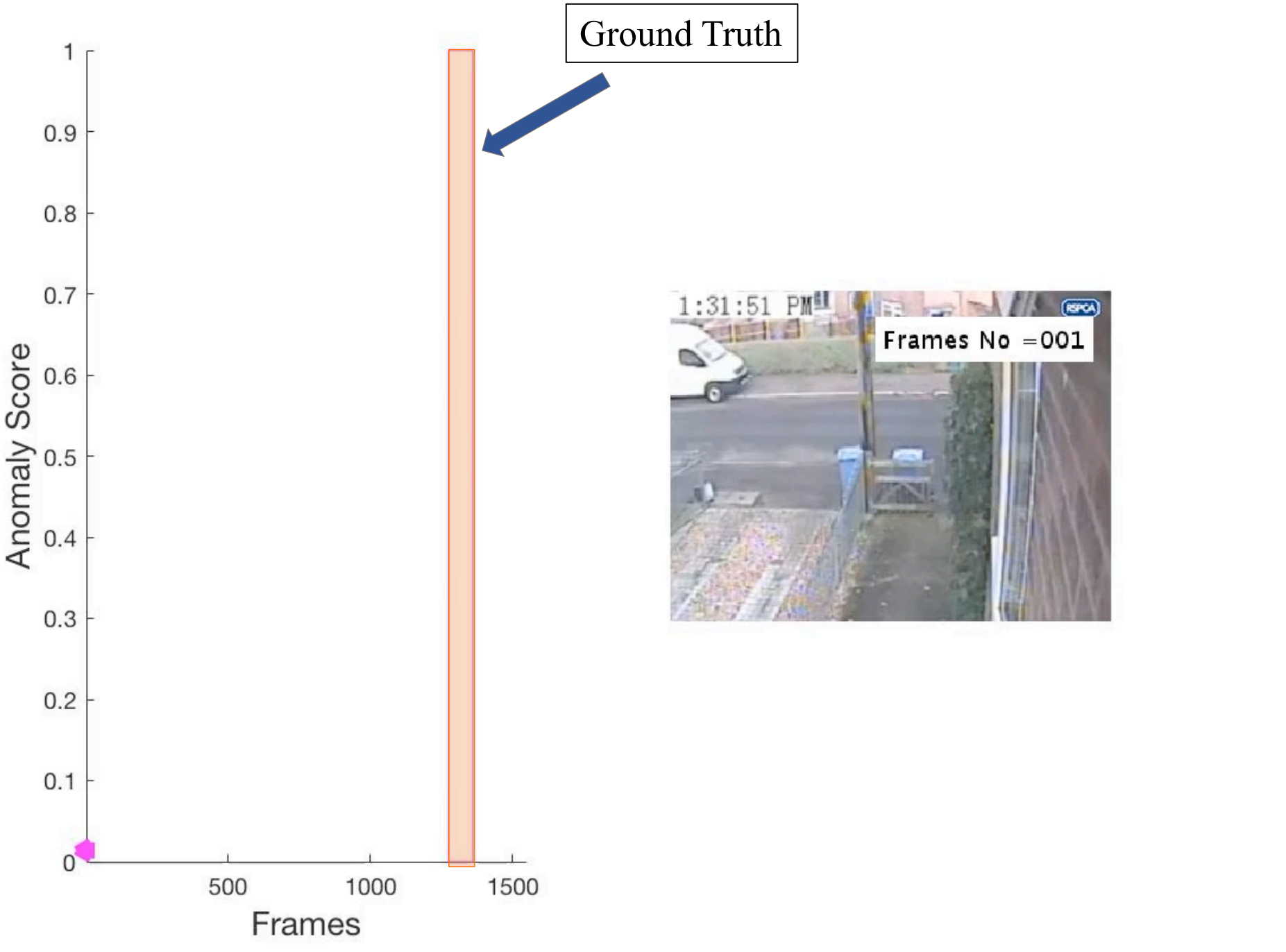
$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

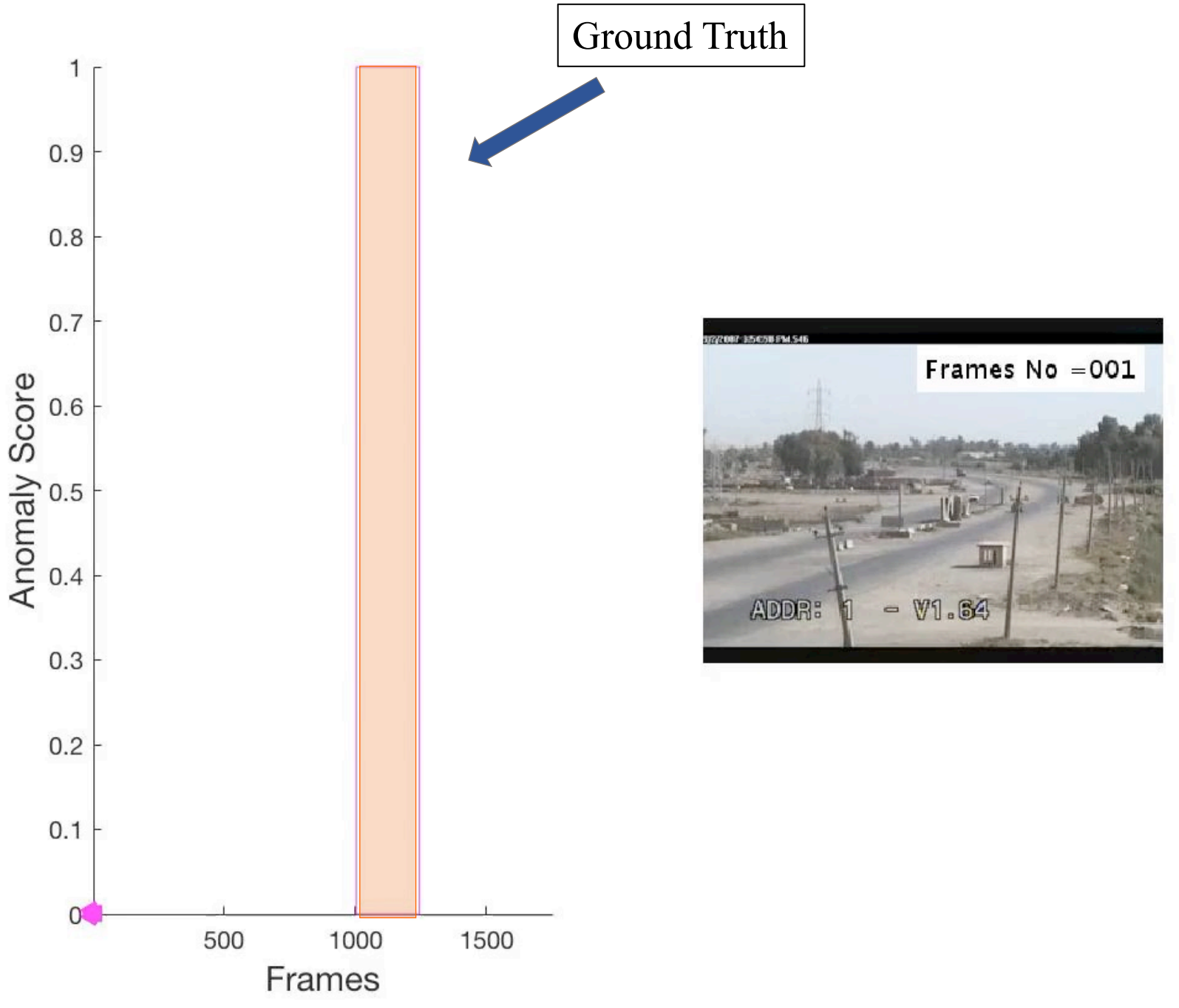
$$+\lambda_1 \overbrace{\sum_{i}^{(n-1)} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f(\mathcal{V}_a^i)}^{②},$$

Weakly labeled Crime Detection Framework

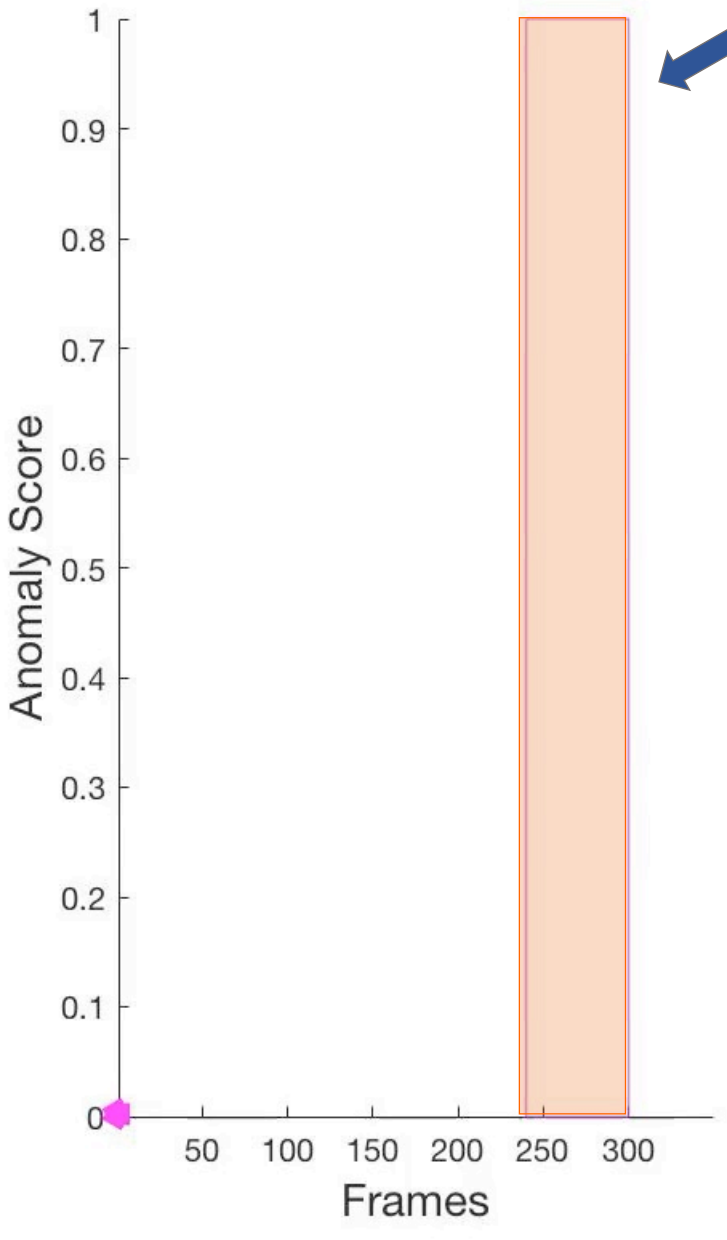


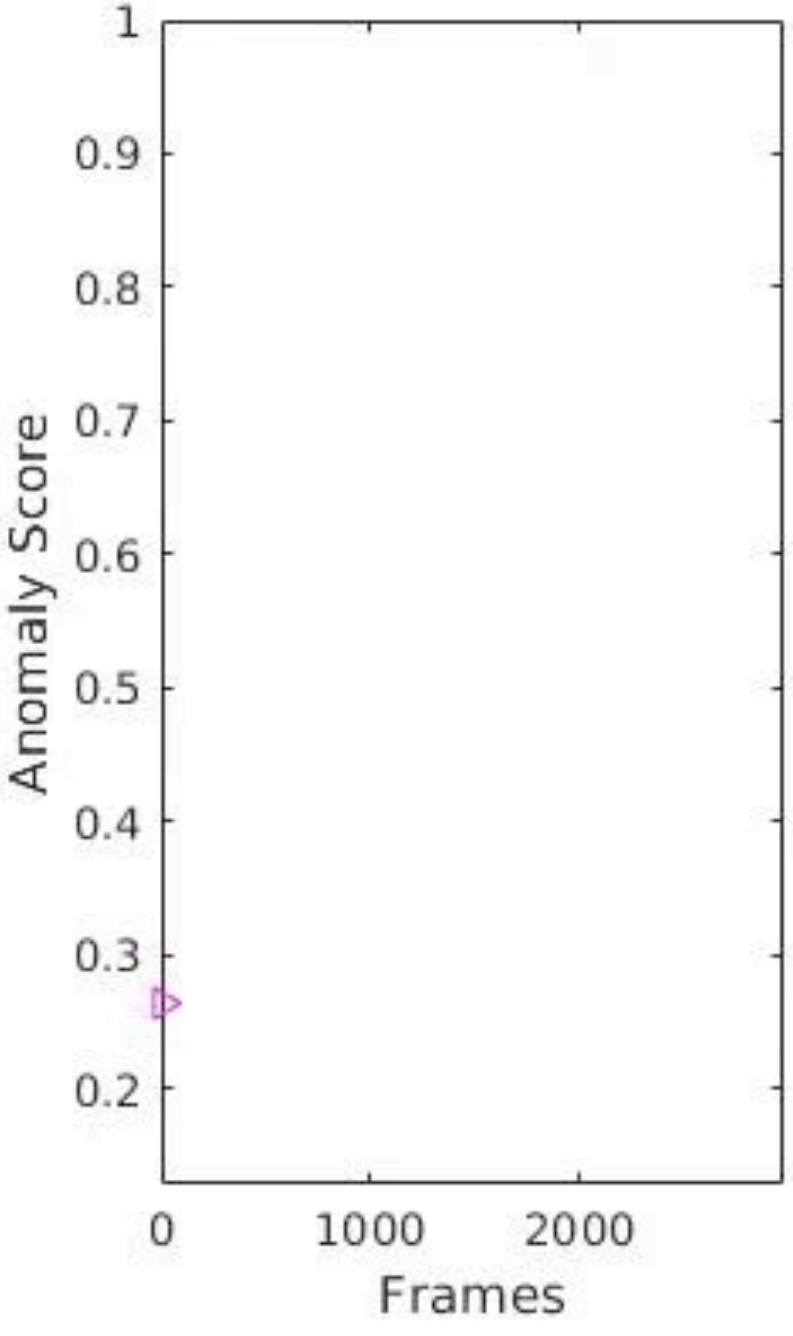
Anomaly Detection Examples of Our Method


Ground Truth
Anomaly Score
Frames
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
500
1000
1500
1:31:51 PM
Frames No =001


Ground Truth
Anomaly Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
500
1000
1500
Frames
Frames No =001
ADDR: 1 - V1.64

Ground Truth

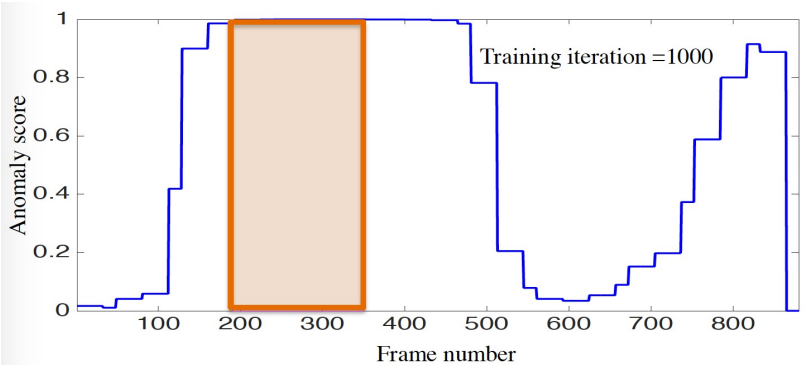


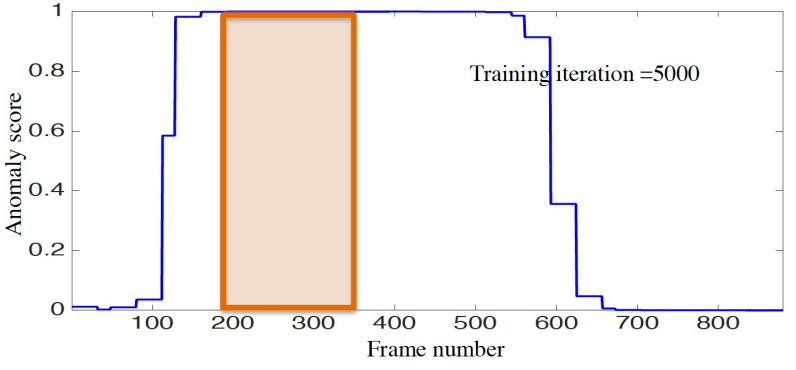

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Anomaly Score
0
1000
2000
Frames

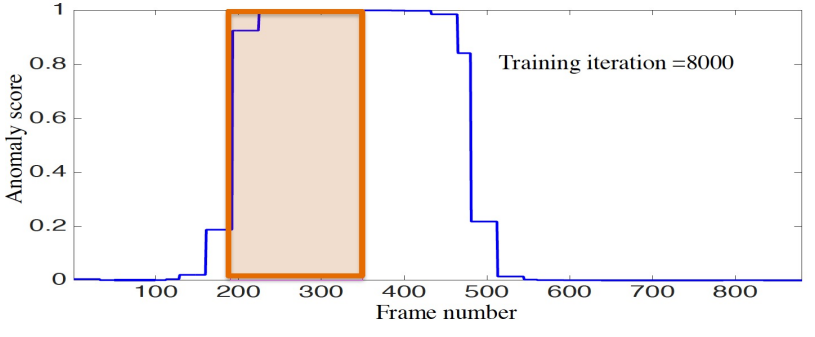

Frame No: = 002
大街上強搶小孩，嚇死人！
在中國大陸真要非常小心哦！

Analysis

- Evolution of score on training video over iterations







Anomaly Detection Results

Method	AUC
Binary classifier	50.0
Hasan *et al.* [19]	50.6
Lu *et al.* [28]	65.51
Proposed w/o constraints	74.44
Proposed w constraints	**75.41**

Anomaly Detection Results

Method	[18]	[28]	**Proposed**
False alarm rate	27.2	3.1	**1.9**

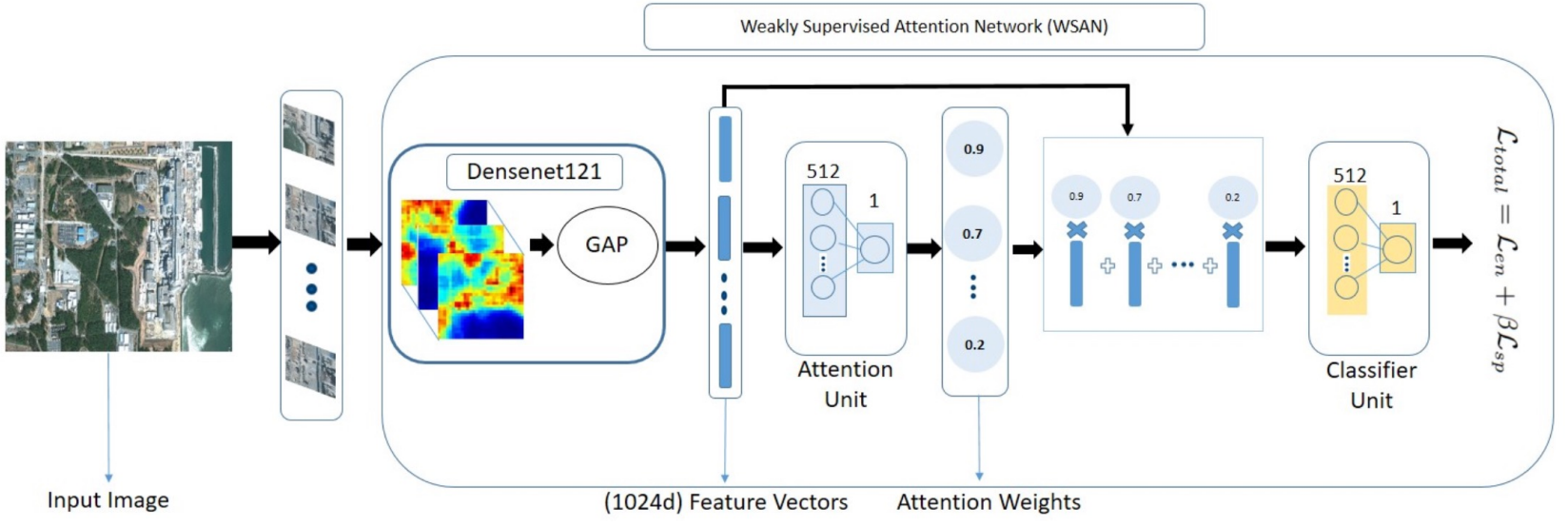
Destruction from Sky: Destruction Detection in Satellite Imagery

Muhammad Usman Ali, Waqas Sultani, Mohsen Ali

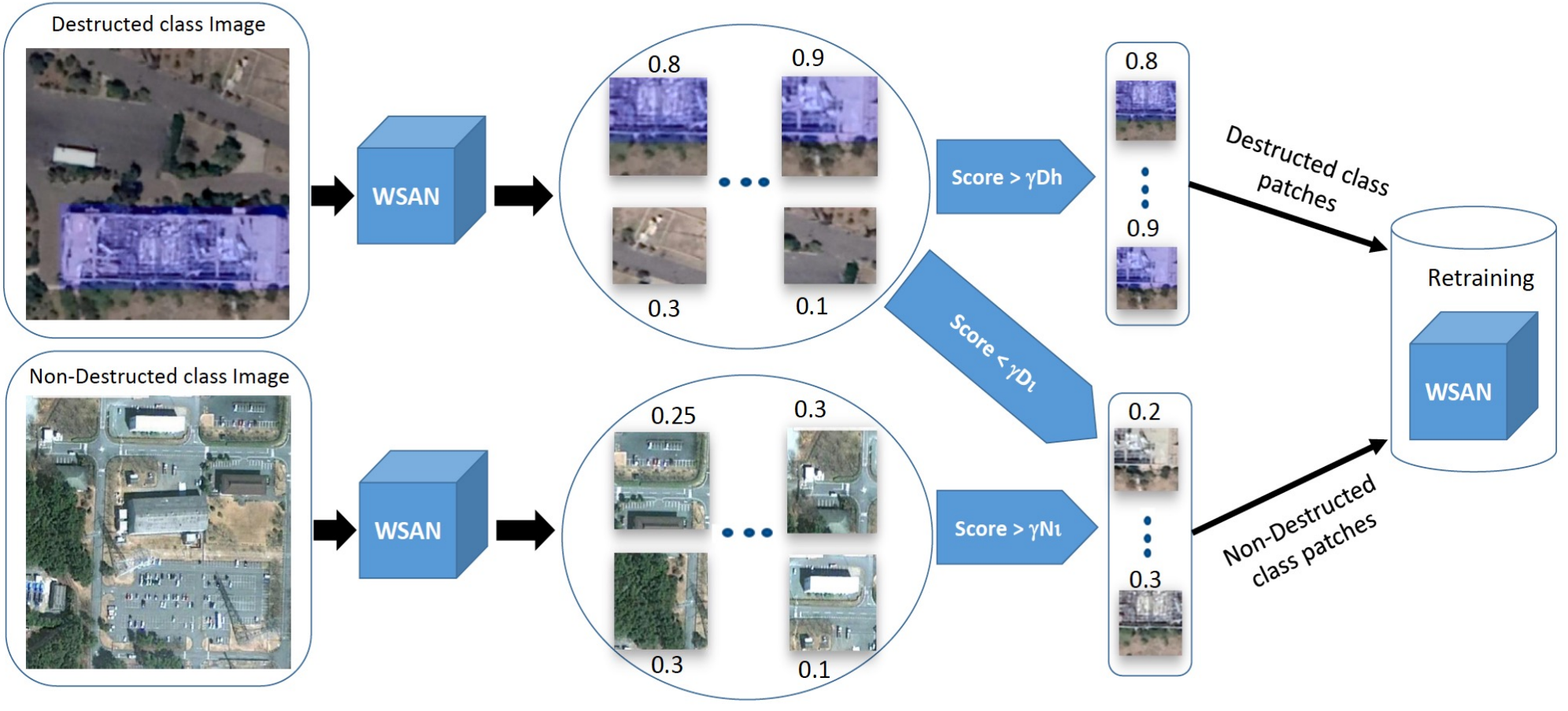
ISPRS, 2020

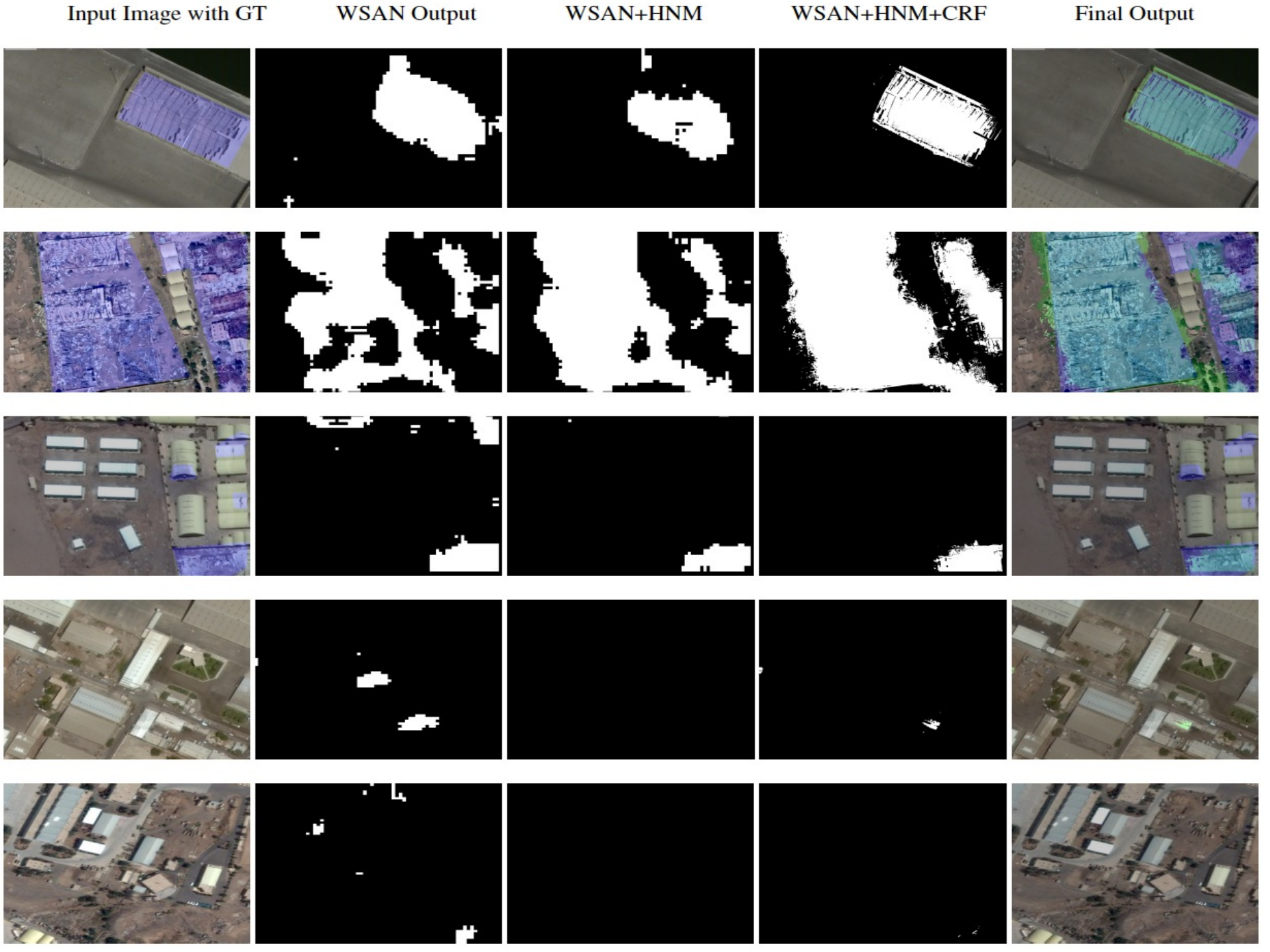
a. Aden (Yemen) b. Sanaa (Yemen) c. Ofunato (Japan)

Weakly labeled Destruction Detection Framework



Hard Negative Mining




Input Image with GT
WSAN Output
WSAN+HNM
WSAN+HNM+CRF
Final Output

Qualitative Results

Quantitative Results

Patch level results

Metrics	WSAN	WSAN+HNM	WSAN+HNM+CRF
Accuracy	85%	89%	**91%**
Precision	0.35	0.45	**0.51**
Recall	0.58	0.55	**0.64**
F1-Score	0.43	0.50	**0.61**

Pixel level results

Metrics	WSAN	WSAN+HNM	WSAN+HNM+CRF
Accuracy	92%	**94%**	93%
Precision	0.41	**0.54**	0.49
Recall	**0.58**	0.49	**0.58**
F1-Score	0.48	0.51	**0.53**
IOU Score	0.31	0.35	**0.36**

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

1. Human Action Recognition Across Datasets

Waqas Sultani, Imran Saleemi

IEEE Conference on Computer Vision and Pattern Recognition
CVPR 2014

Problem

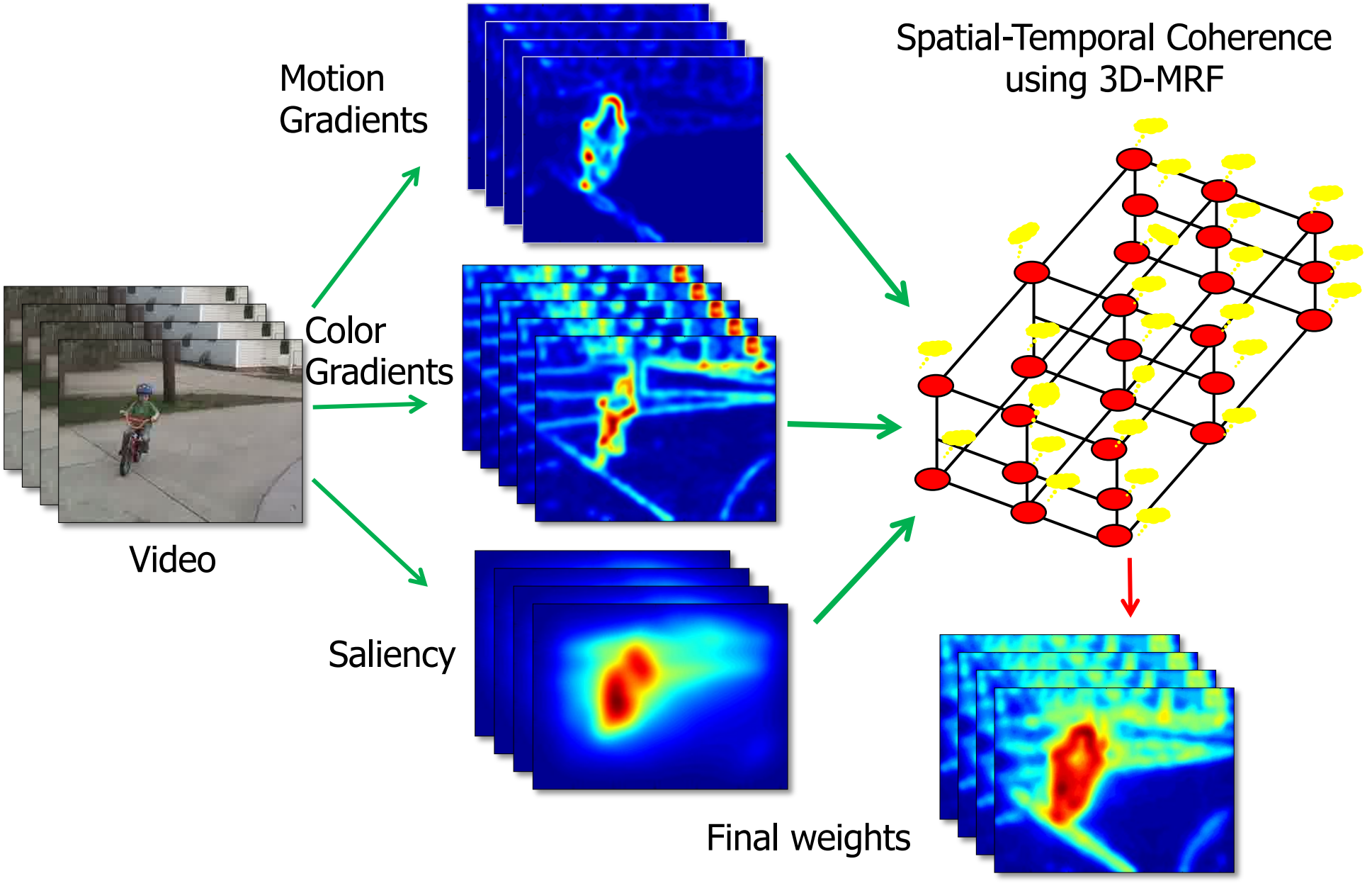
- Recognition Accuracy drops across the datasets!

Cross dataset recognition★

Training	Testing	Accuracy (avg)
UCF50	UCF50	70 %
UCF50	HMDB51	55.7 %
Olympic Sports	Olympic Sports	71.8 %
Olympic Sports	UCF50	16.67 %

★ Training and Testing is done on similar actions across the datasets

Obtain score of each pixel being the foreground using

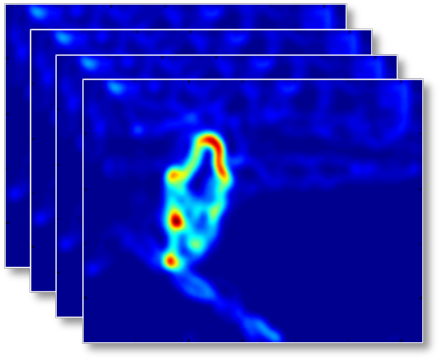


Foreground Focused Representation

- Motion Gradients

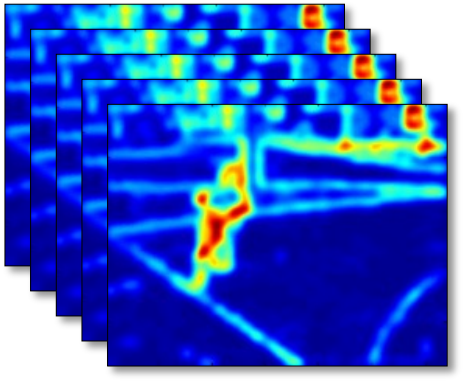
$$f_m(x,y) = \left\| \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \right\|_F * g$$

$$= \sqrt{u_x^2 + u_y^2 + v_x^2 + v_y^2} * g,$$



- Color Gradients

$$f_c(x,y) = \sqrt{L_x^2 + L_y^2 + a_x^2 + a_y^2 + b_x^2 + b_y^2} * g,$$



Coherence of Foreground Confidence using 3DMRF

- Initial aggregate of confidence map

$$\hat{f}_a = log\left(f_m\left(f_c + f_s\right) + 1\right)$$

- The score is max-normalized for each frame of a video
- The quality of labeling is given by:

$$E(\omega) = \sum_{\psi_p \in \mathcal{V}} D_p(\omega_p) + \sum_{(p,q) \in \mathcal{V}} V(\omega_p - \omega_q).$$

$$D_p(\omega_p) = (\hat{f}_a(p) - \omega_p)^2$$

$$V\left(\omega_p - \omega_q\right) = min((\omega_p - \omega_q)^2, \kappa)$$

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

2. Automatic Action Annotations using Weakly Labeled Videos

Waqas Sultani, Mubarak Shah

Journal of Computer Vision and Image Understanding (CVIU), 2017

Spatiotemporal Bounding box annotation of action in a video is difficult

Requires:

- Many human annotators
- Hundred of hours
- Expensive annotation interfaces
- Prone to error

Diving

Video level annotation of action in a video is easy to obtain

Requires:

- Few human annotators
- Less time
- Simple annotation interfaces
- Less prone to error

Diving

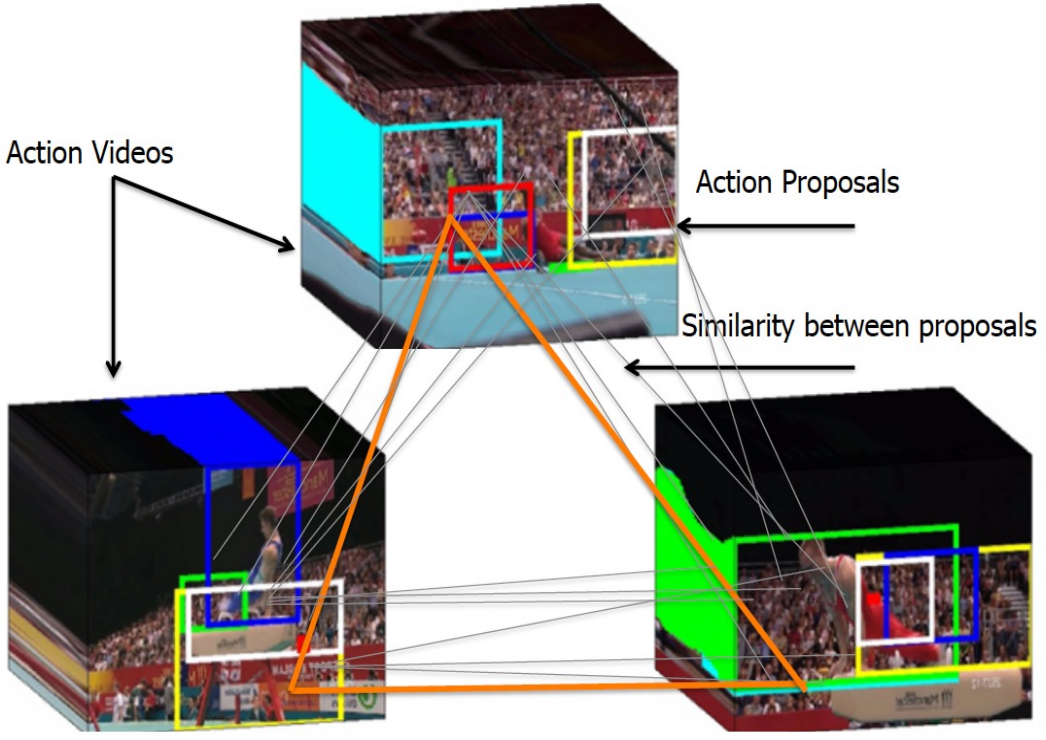
Goal

- Automatic Spatio-Temporal Annotations using video level labels only.

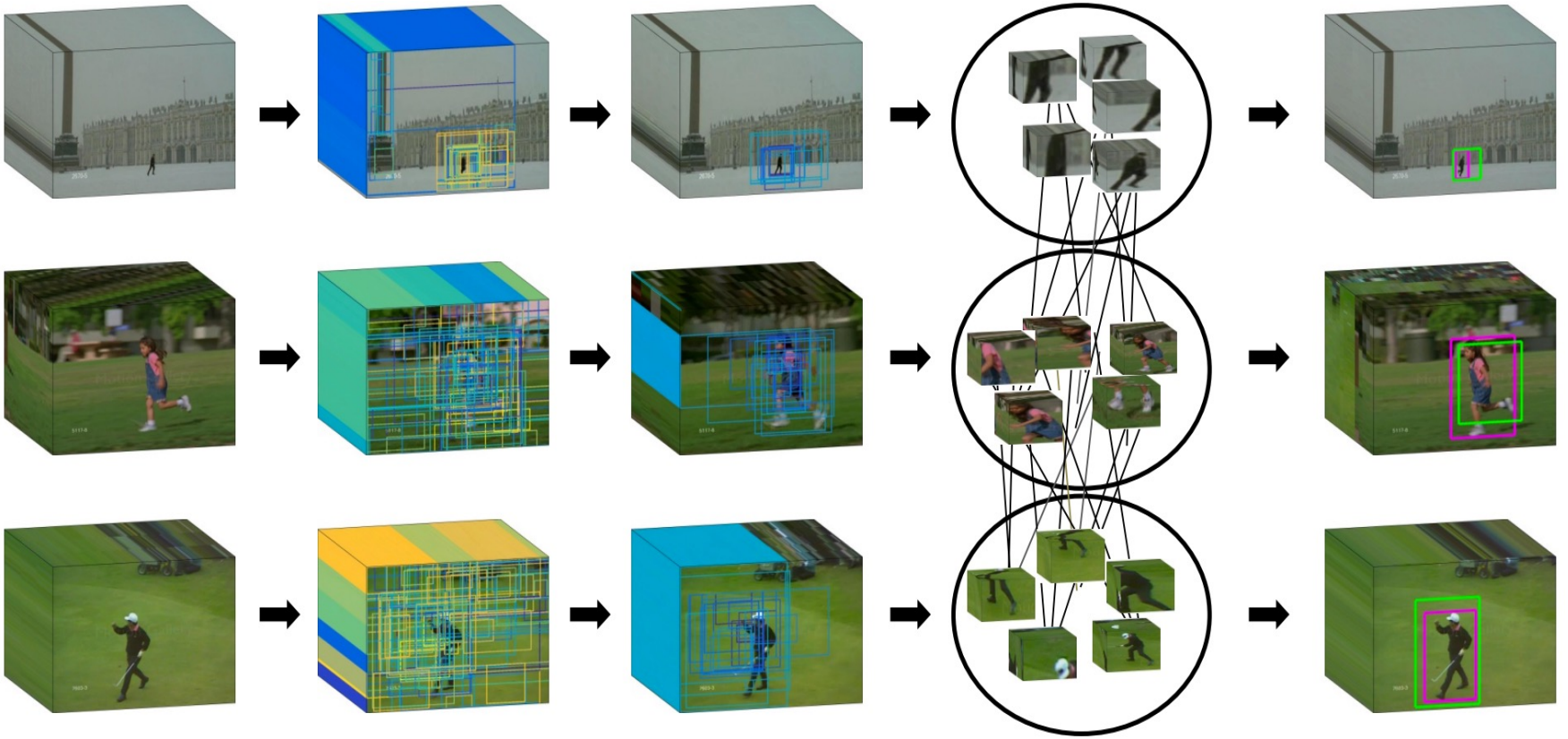
- Action Detection using automatic annotations

Key Idea

- Exploit Similarities between spatio-temporal volumes in multiple videos of the same action



Proposed Approach



Multiple videos | Action Proposals | Top Action Proposals | GMCP graph | Action Annotations

GMCP Graph

- Each proposal (node) in video is connected to all proposals in other videos
- No two proposals within the same are connected

$\Omega_{p_i} =$ Initial Action score

$\mathbf{\Pi_{ij}} =$ Shape similarity

$\mathbf{\Gamma_{ij}} =$ Fine-grained Similarity

$\mathbf{\Theta_{ij}} =$ Global Similarity

$$\sum_{i=1}^{i=N} \sum_{j=1, j\neq i}^{j=N} \left(\alpha \Omega_{p_i} + (\mathbf{\Theta_{ij}} + \mathbf{\Gamma_{ij}} + \mathbf{\Pi_{ij}}) \eta_{\mathbf{pi}} \times \eta_{\mathbf{pj}} \right)$$

Quantitative Results

Annotation Quality

- MABO (Mean Average Bounding box)
- CorrLoc (Correct Localization)

Quantitative Results

Method	UCF-Sports	Sub-JHMDB	THUMOS13
Co-segmentation[95]	53.15(23.3)	49.37(24.5)	9.56(5.48)
Co-segmentation[17]	42.25 (19.2)	47.78 (20.3)	21.41(12.4)
Negative Mining[59]	48.49(21.8)	61.39(22.3)	14.39(10.2)
CRANE [69]	61.18(27.9)	64.56(22.8)	14.17(10.0)
Ours	**83.83(34.6)**	**89.56(32.4)**	**41.69(19.1)**

Limitation

- What if we do not have multiple videos of the same action?

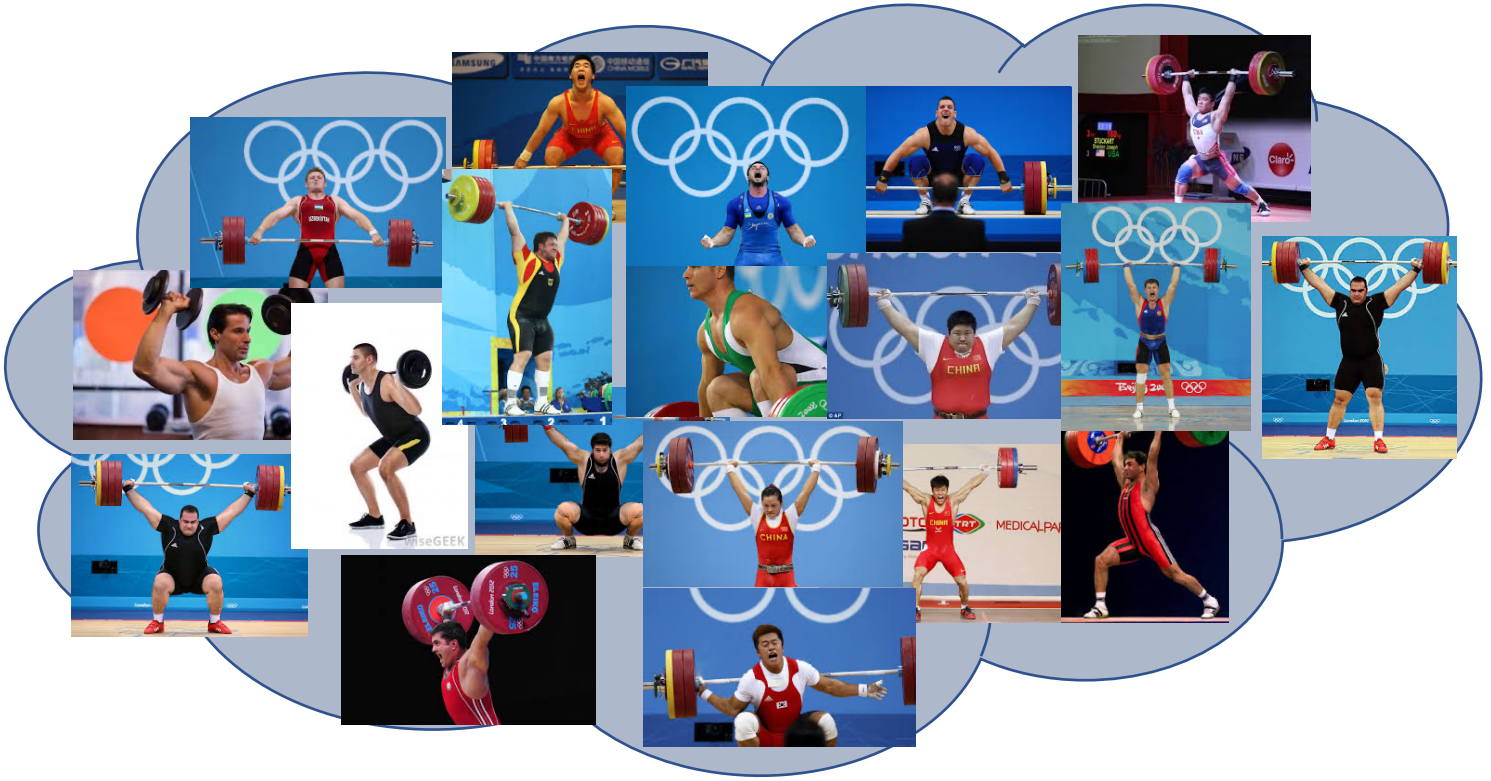
UCF
CENTER FOR RESEARCH IN COMPUTER VISION

3. Video Action Localization Using Web Images

Waqas Sultani, Mubarak Shah

IEEE Conference on Computer Vision and Pattern Recognition
CVPR 2016

Action images capture key poses



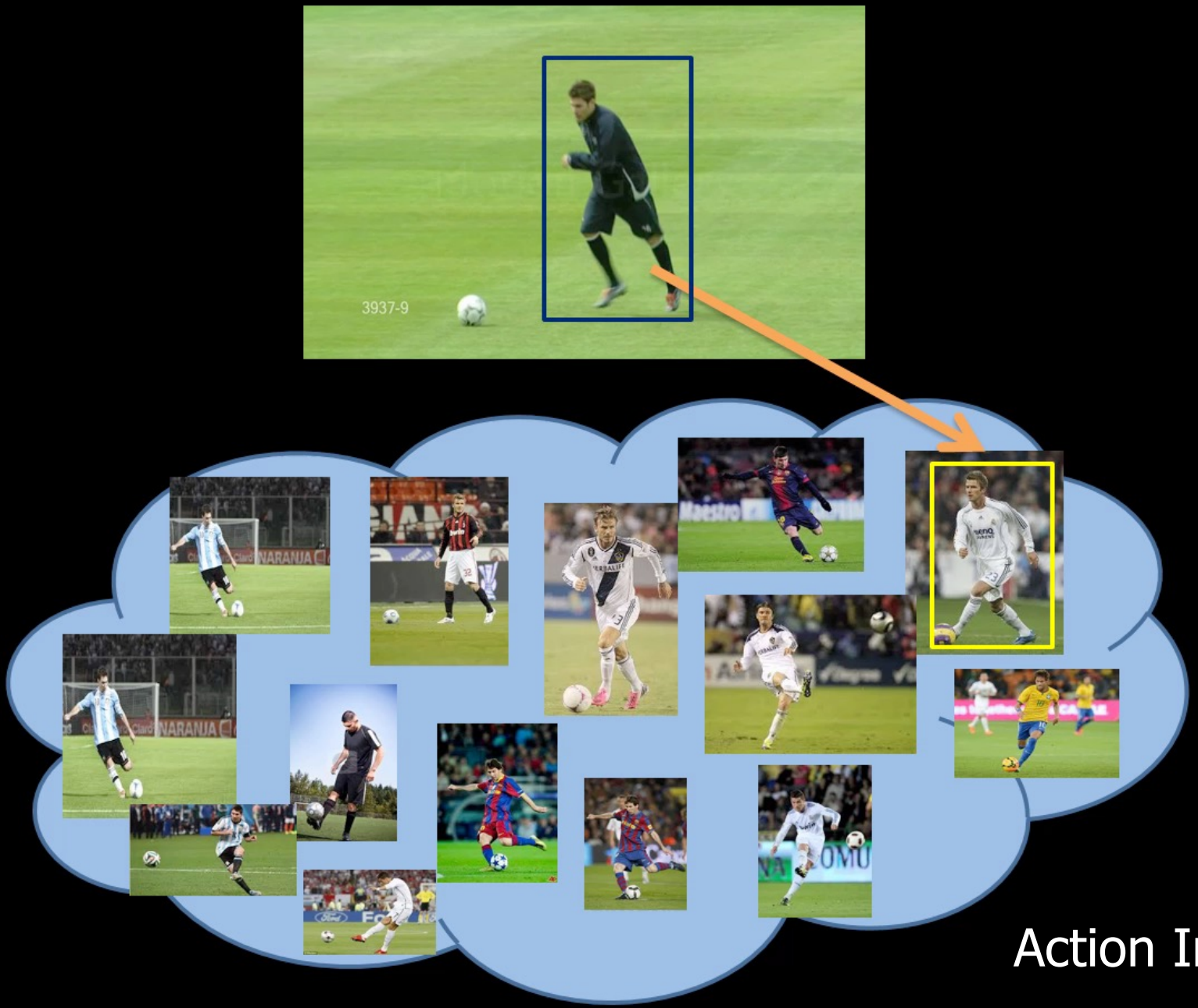
Weight Lifting
Web Images

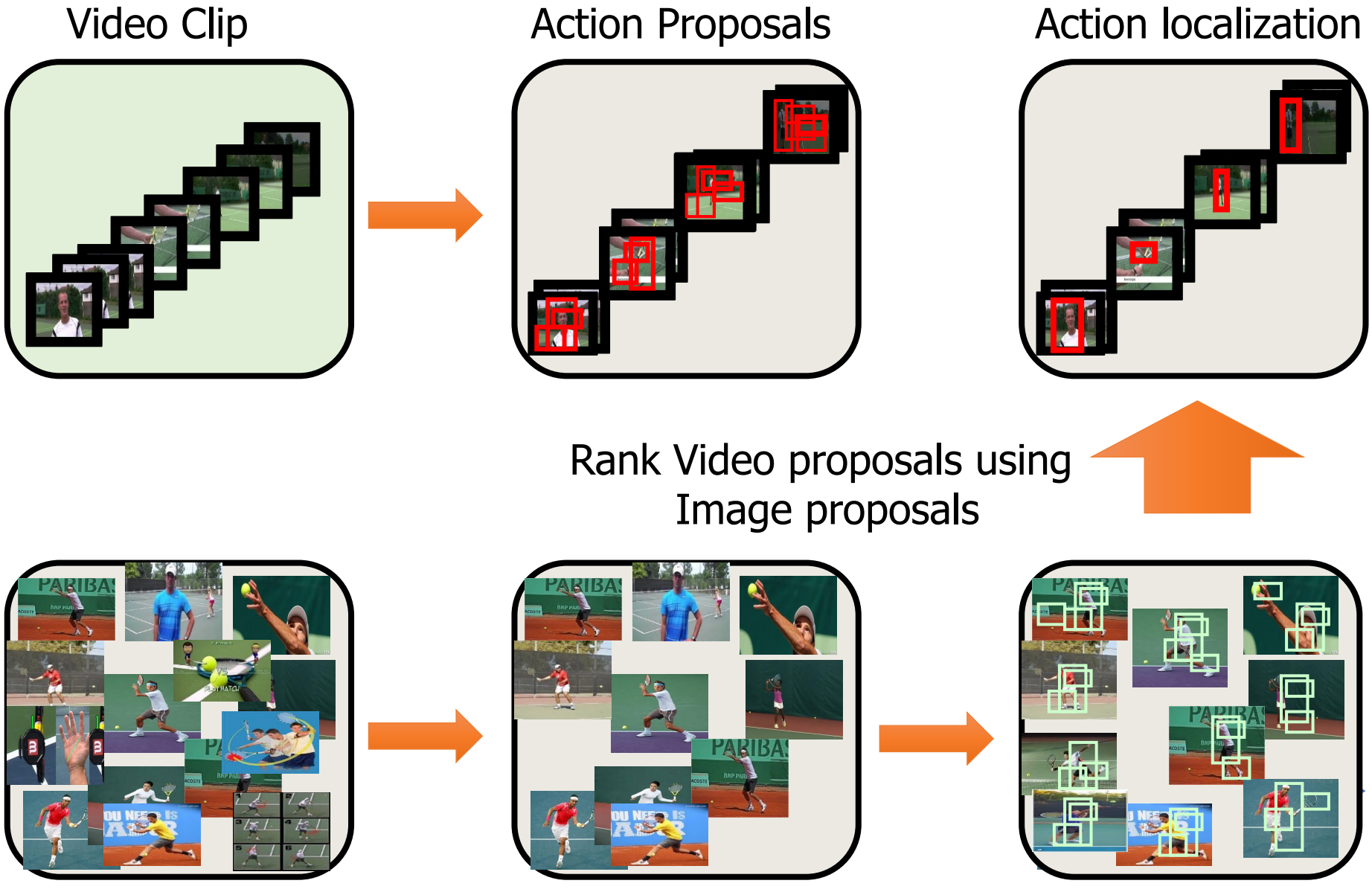
Action images capture key poses

Swing Side-Angle
Web Images

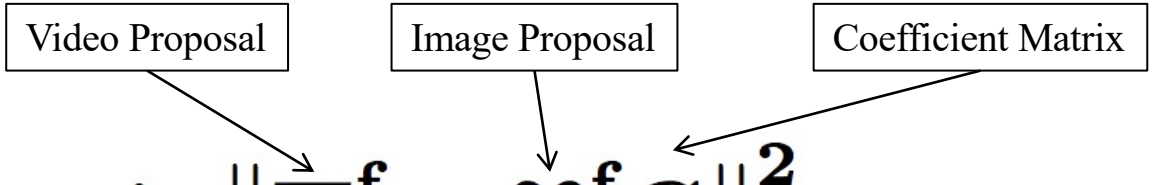
Key idea

Locations containing key poses in video frames can localize an action.


Action Images


Video Clip
Action Proposals
Action localization
Rank Video proposals using
Image proposals
Web Images
Remove Noisy Images
Action Proposal in Images

Rank Video Proposals using Image Action Proposals



$$\min_{\mathbf{C}} \left\| \mathbf{\Pi^f} - \mathbf{\Upsilon^f C} \right\|_{\mathbf{F}}^{\mathbf{2}}$$

Sparsity

$$\min_{\mathbf{C}} \left\| \mathbf{\Pi^f} - \mathbf{\Upsilon^f C} \right\|_{\mathbf{F}}^{\mathbf{2}} + \lambda_{\mathbf{2}} \left\| \mathbf{C} \right\|_{\mathbf{1}}$$

Consistency

$$\min_{\mathbf{C}} \left\| \mathbf{\Pi^f} - \mathbf{\Upsilon^f C} \right\|_{\mathbf{F}}^{\mathbf{2}} + \lambda_{\mathbf{2}} \left\| \mathbf{C} \right\|_{\mathbf{1}} + \lambda_{\mathbf{1}} \left\| \mathbf{C} - \bar{\mathbf{C}} \right\|_{\mathbf{F}}^{\mathbf{2}}$$

Experimental Results

Trimmed Datasets

- UCF-Sports
- THUMOS13

Un-Trimmed Datasets

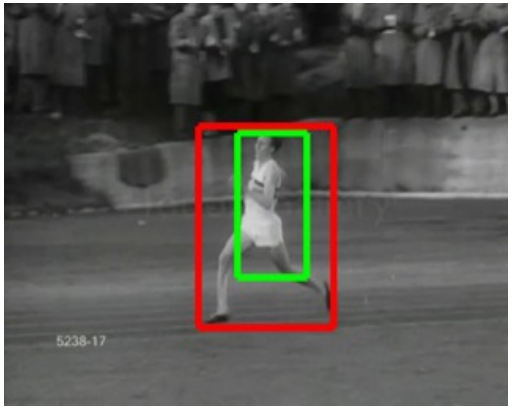
- THUMOS14

Trimmed Action Datasets

Proposed

Ground Truth

Running



Skateboarding

Horse Riding

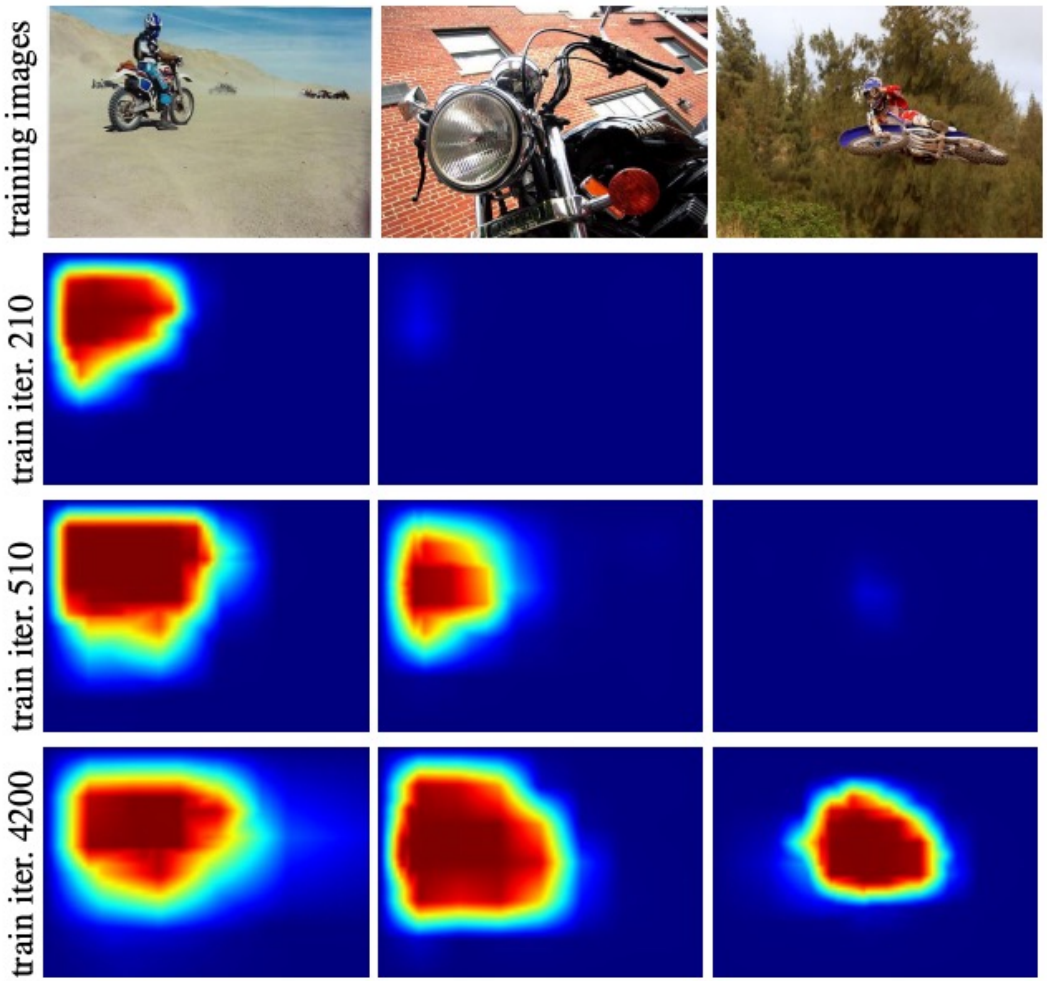
Quantitative Results

UCF Sports

Method	CRANE	Negative Mining	Ours
Localization Accuracy	65.41[1]	63.01[2]	**92.72**

[1] K.Tang et at., Discriminative Segment Annotation in Weakly Labeled Video. **CVPR2013.**

[2] P. Siva, et al., In defense of negative mining for annotating weakly labeled data. **ECCV, 2012.**



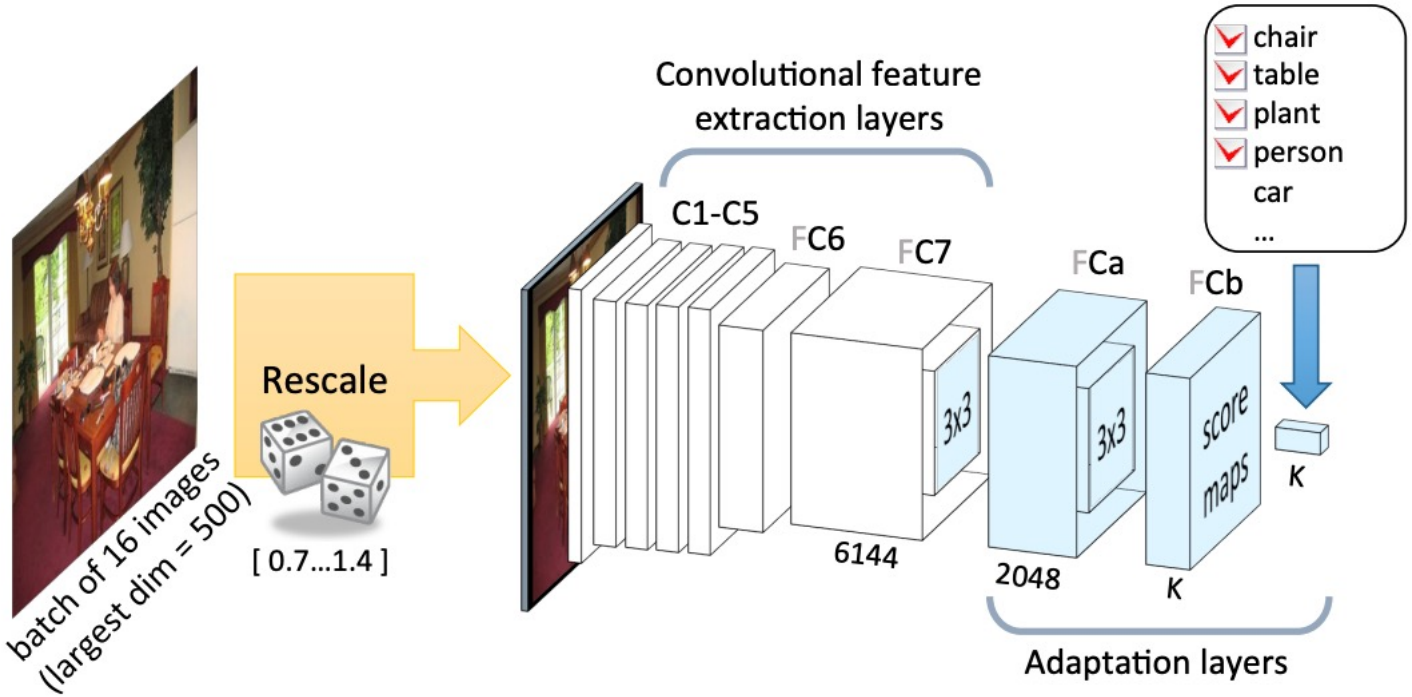
Is object localization for free? – Weakly-supervised learning with convolutional neural networks

Maxime Oquab*
INRIA Paris, France

Léon Bottou†
MSR, New York, USA

Ivan Laptev*
INRIA, Paris, France

Josef Sivic*
INRIA, Paris, France



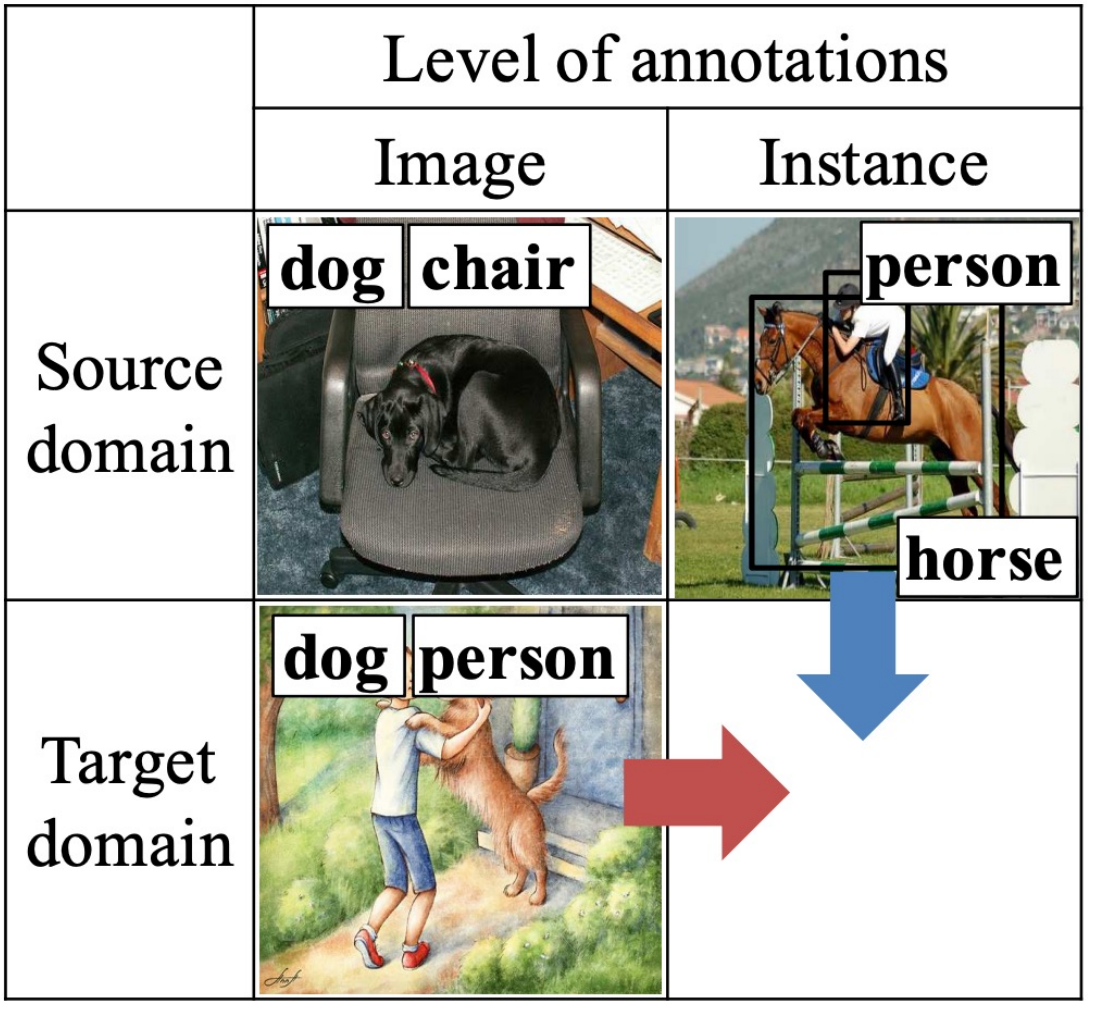
Is object localization for free? – Weakly-supervised learning with convolutional neural networks

Maxime Oquab*
INRIA Paris, France

Léon Bottou†
MSR, New York, USA

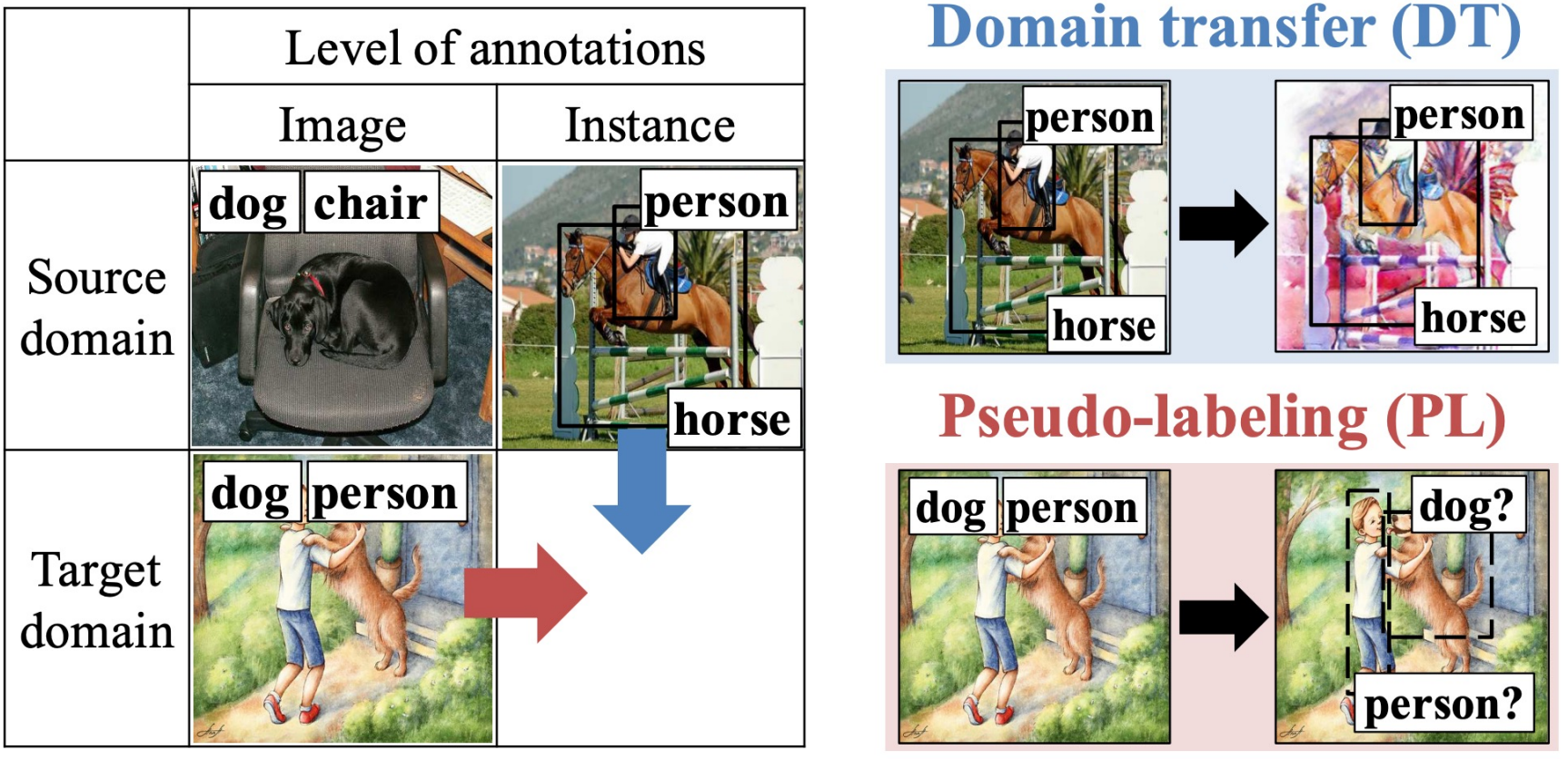
Ivan Laptev*
INRIA, Paris, France

Josef Sivic*
INRIA, Paris, France



Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation

Naoto Inoue Ryosuke Furuta Toshihiko Yamasaki Kiyoharu Aizawa
The University of Tokyo, Japan



Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation

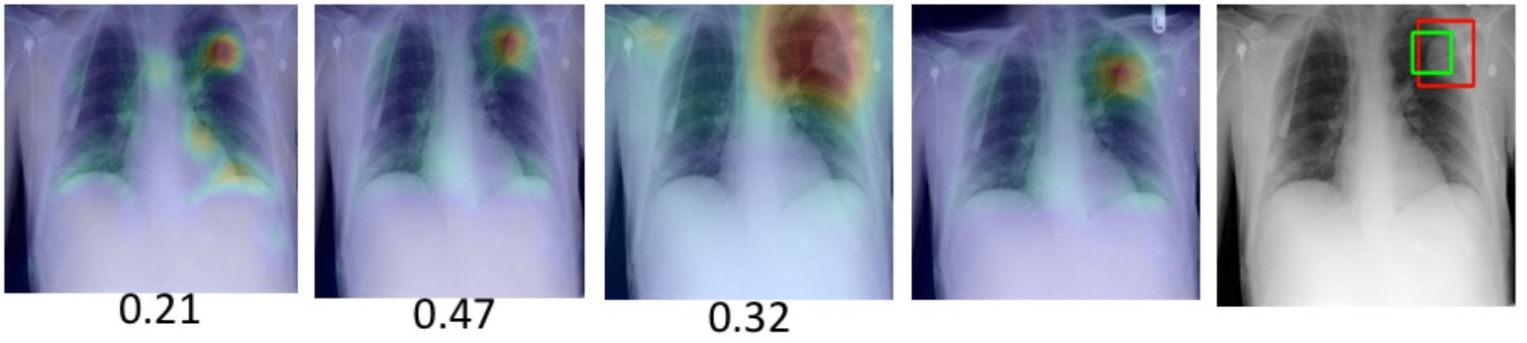
Naoto Inoue Ryosuke Furuta Toshihiko Yamasaki Kiyoharu Aizawa
The University of Tokyo, Japan

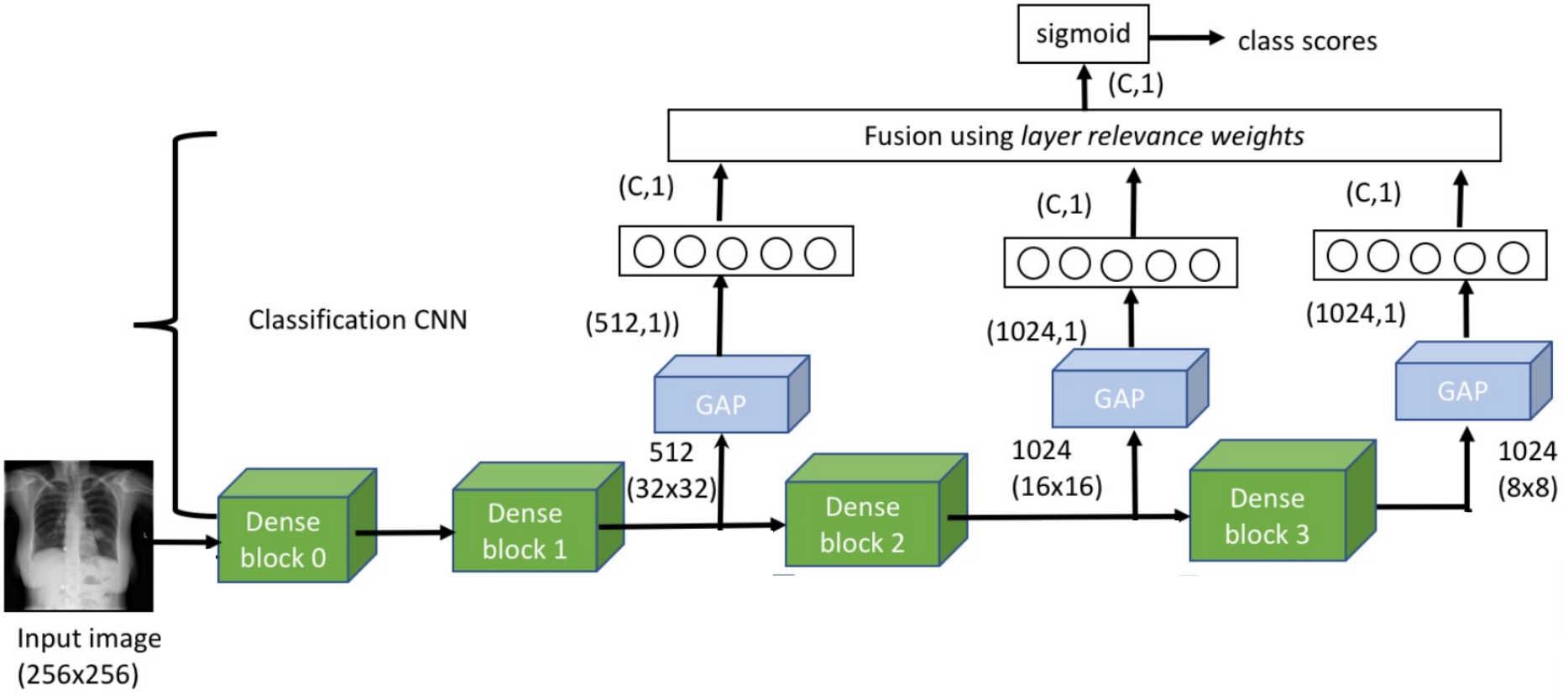
Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in X-ray images

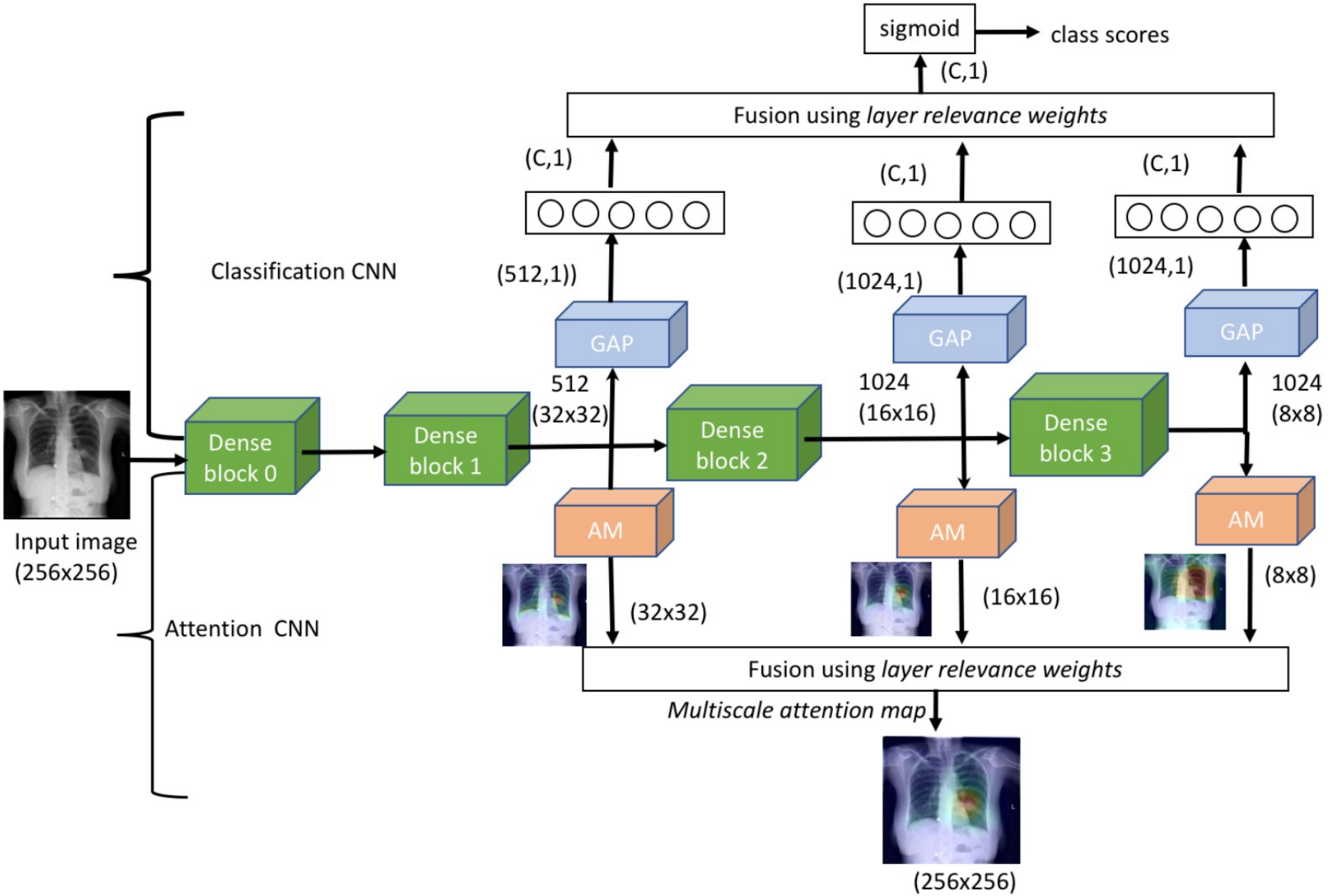
Suman Sedai, Dwarikanath Mahapatra, Zongyuan Ge,Rajib Chakravorty and Rahil Garnavi

IBM Research - Australia, Melbourne, VIC, Australia,
ssedai@au1.ibm.com

https://arxiv.org/pdf/1808.08280.pdf


0.21
0.47
0.32
Mass
0.35
0.45
0.20
Nodule
0.02
0.35
0.64
Cardiomegaly


sigmoid
class scores
(C,1)
Fusion using *layer relevance weights*
(C,1)
(C,1)
(C,1)
Classification CNN
(512,1))
(1024,1)
(1024,1)
GAP
GAP
GAP
512
(32x32)
1024
(16x16)
1024
(8x8)
Dense block 0
Dense block 1
Dense block 2
Dense block 3
Input image
(256x256)


sigmoid
class scores
(C,1)
Fusion using *layer relevance weights*
(C,1)
(C,1)
(C,1)
Classification CNN
(512,1))
(1024,1)
(1024,1)
GAP
GAP
GAP
512
(32x32)
1024
(16x16)
1024
(8x8)
Dense block 0
Dense block 1
Dense block 2
Dense block 3
Input image
(256x256)
AM
AM
AM
(32x32)
(16x16)
(8x8)
Attention CNN
Fusion using *layer relevance weights*
Multiscale attention map
(256x256)

Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis

Kangning Liu[1] KANGNING.LIU@NYU.EDU
Yiqiu Shen[1] YS1001@NYU.EDU
Nan Wu[1] NAN.WU@NYU.EDU
Jakub Chłędowski[4] JAKUB.CHLEDOWSKI@GMAIL.COM
Carlos Fernandez-Granda[*1,2] CFGRANDA@CIMS.NYU.EDU
Krzysztof J. Geras[*3,1] K.J.GERAS@NYU.EDU

[1] *NYU Center for Data Science*
[2] *Courant Institute of Mathematical Sciences, NYU*
[3] *NYU Grossman School of Medicine*
[4] *Jagiellonian University*

https://arxiv.org/pdf/2106.07049.pdf

Problem

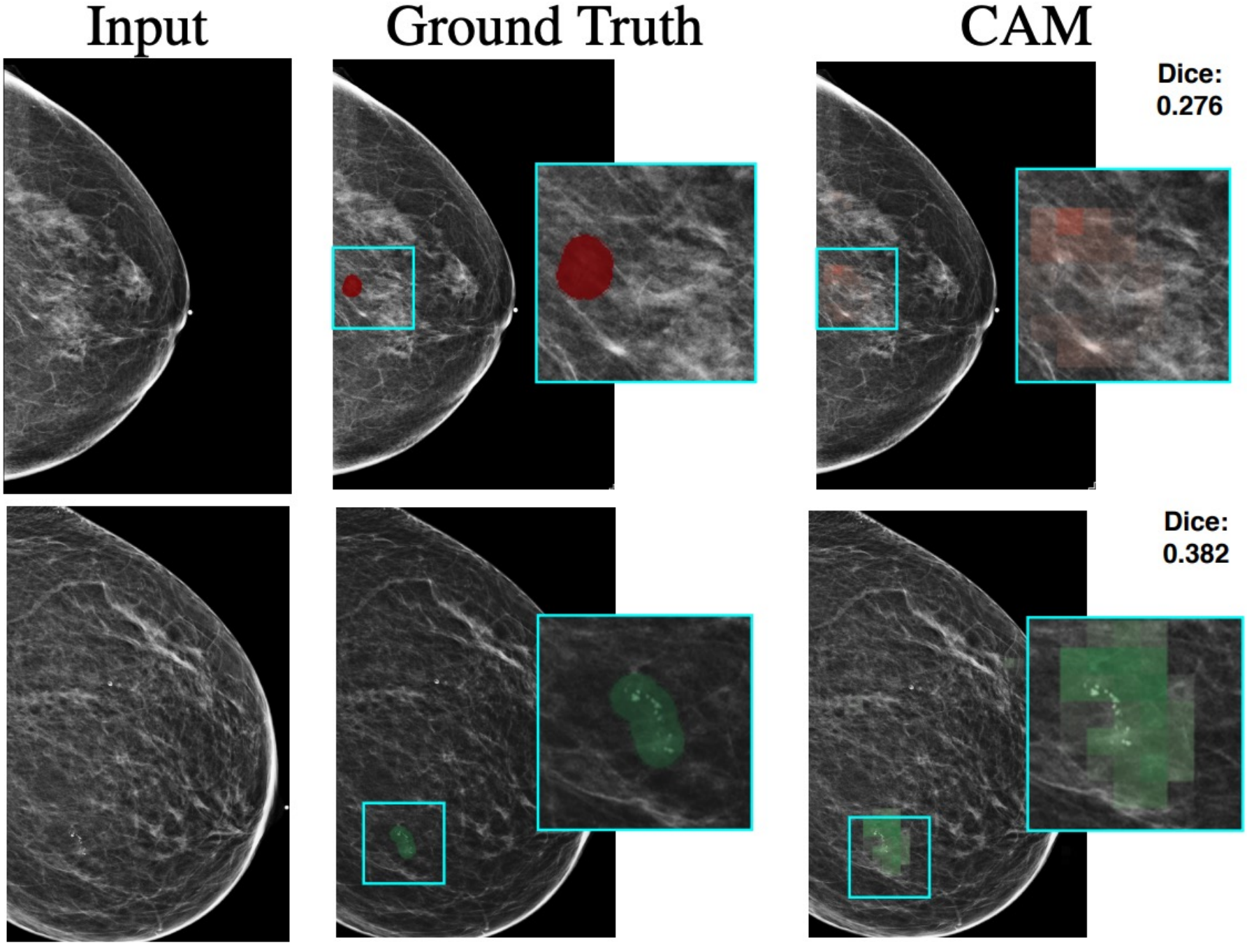
- In 2020, there were 2.3 million women diagnosed with **breast cancer** and
- 685 000 deaths **globally.**
- The risk of having breast cancer has increased in Pakistan, whereby **one in every 9 women in Pakistan** has a lifetime risk of being diagnosed with breast cancer
- High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis

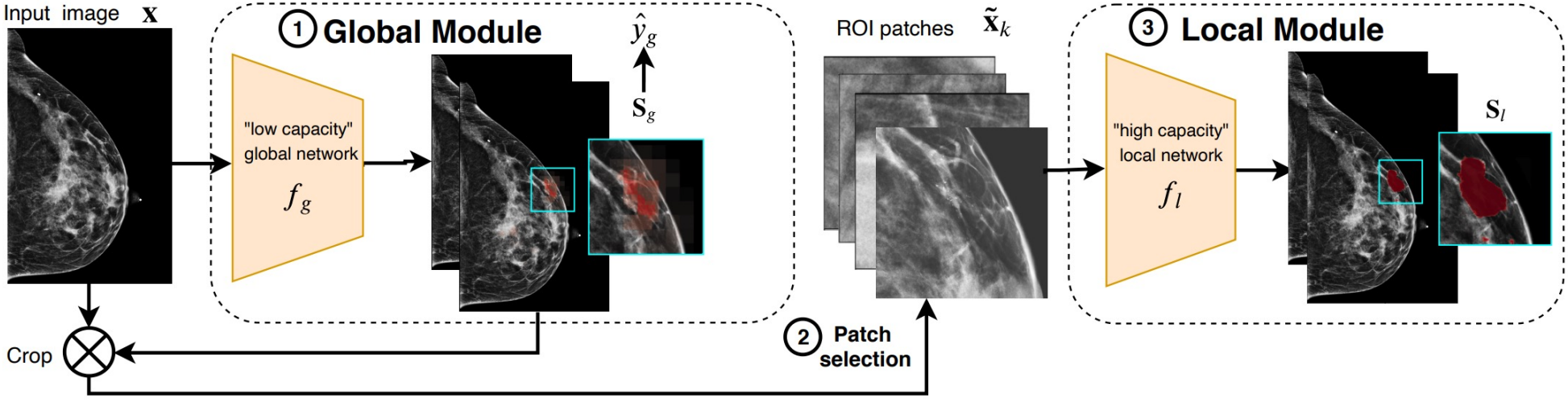
- High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis

- Unfortunately, acquiring such labels is labour-intensive and requires medical expertise. To overcome this difficulty, weakly supervised localization can be utilized

2 Stage Approach

- Generate Coarse Feature maps using CAM
- Generate Fine-Grained Feature maps using MIL


Input
Ground Truth
CAM
Dice:
0.276
Dice:
0.382


Input image $\mathbf{x}$
1 Global Module
"low capacity" global network
f_g
$\hat{y}_g$
$\mathbf{S}_g$
Crop
2 Patch selection
ROI patches $\tilde{\mathbf{x}}_k$
3 Local Module
"high capacity" local network
f_l
$\mathbf{S}_l$

Method

1. The image $\mathbf{x}$ is fed into the *global module*, a memory-efficient CNN denoted by f_g, to produce an image-level coarse saliency map $\mathbf{S}_g$ and an image-level class prediction $\hat{y}_g$.

2. We select M patches from $\mathbf{x}$ based on $\mathbf{S}_g$. To do that, we greedily select the patches for which the sum of the entries in $\mathbf{S}_g$ is the largest (see Algorithm 1 for a detailed description).

3. We feed the selected patches $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_M$ to the *local module* f_l, another CNN which produces a fine-grained saliency map associated with each patch. We then remap the patch-level saliency maps back to their location in the original input image. We denote the saliency map obtained through this procedure by $\mathbf{S}_l$.

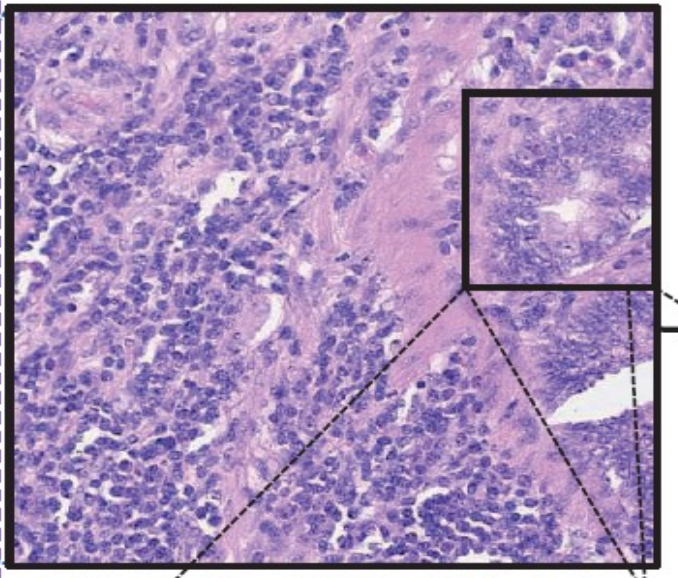
Multiplex Cellular Communities in Multi-Gigapixel Colorectal Cancer Histology Images for Tissue Phenotyping

Sajid Javed, Arif Mahmood, Naoufel Werghi, *Senior Member, IEEE*, Ksenija Benes, and Nasir Rajpoot, *Senior Member, IEEE*

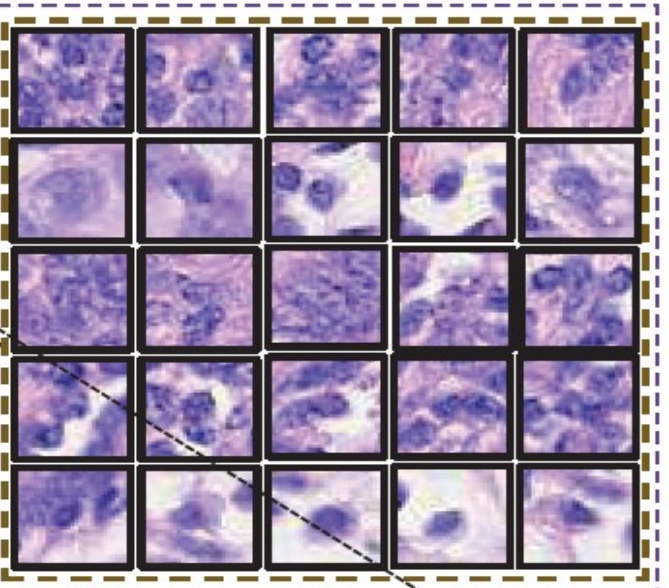
Colorectal cancer

- There are estimated **1.93 million new CRC cases diagnosed**, and 0.94 million CRC caused deaths in 2020 worldwide, representing 10% of the global cancer incidence (total 19.29 million new cases) and 9.4% of all cancer caused deaths (total 9.96 million deaths)

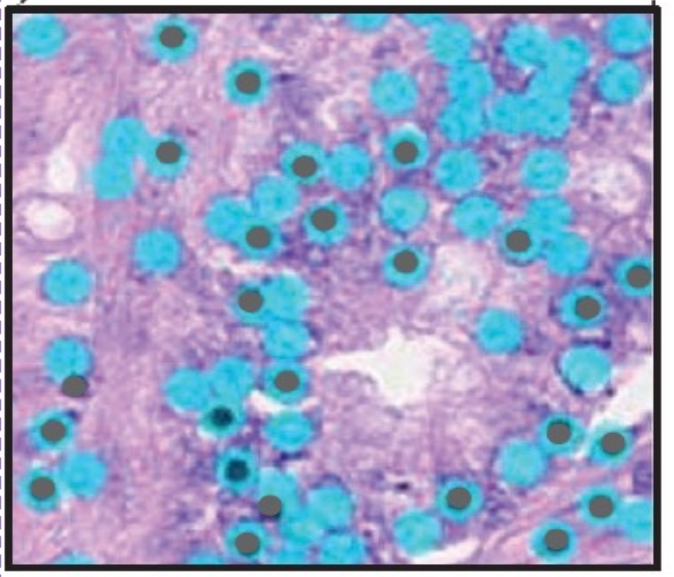
(a) Input CRC WSI



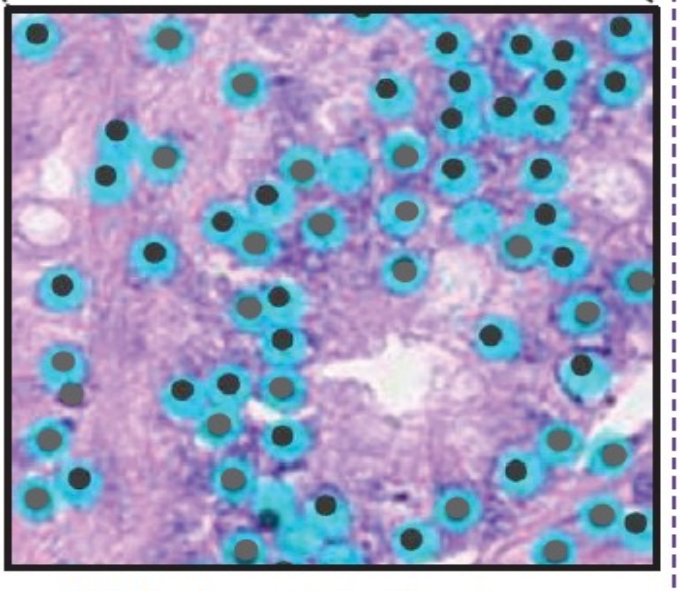
(b) Input Tile: @500 x 500 pixels at 20x magni□ca□on level



(c) Nuclear shape, texture and size varia□ons



(d) Nucleus detec□on using SC-CNN method



(e) Nucleus Detec□on by our proposed algorithm

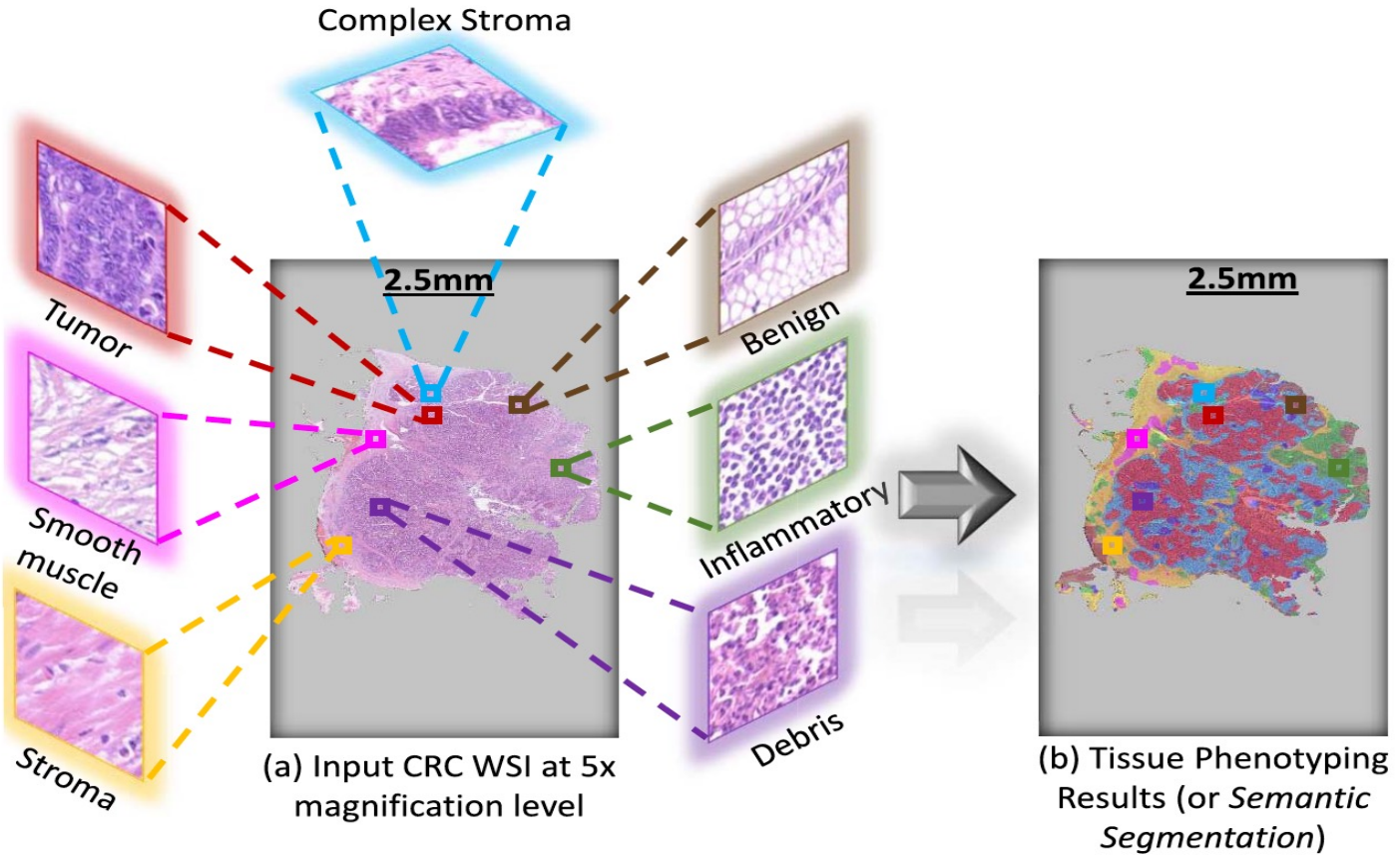


Fig. 1. (a) A sample multi-gigapixel Whole Slide Image (WSI) of human ColoRectal Cancer (CRC) tissue slide. The WSI is shown at 5× magnification level which contains 22.5K × 11.5K pixels. The same WSI is also captured at 10× magnification level containing 45.1K × 23K pixels, 20× magnification level containing 90K × 46K pixels, and 40× magnification level containing 180K × 92K pixels. The WSI contains different tissue components including Tumor, Stroma, Inflammatory, Debris, Complex stroma, Benign, and Smooth muscle, which are shown at 20× magnification level; (b) Results of our proposed Multiplex Cellular communities for Tissue Phenotyping (MCTP) algorithm for tissue phenotyping where seven distinct tissue classes in different colors are overlaid on the input WSI.

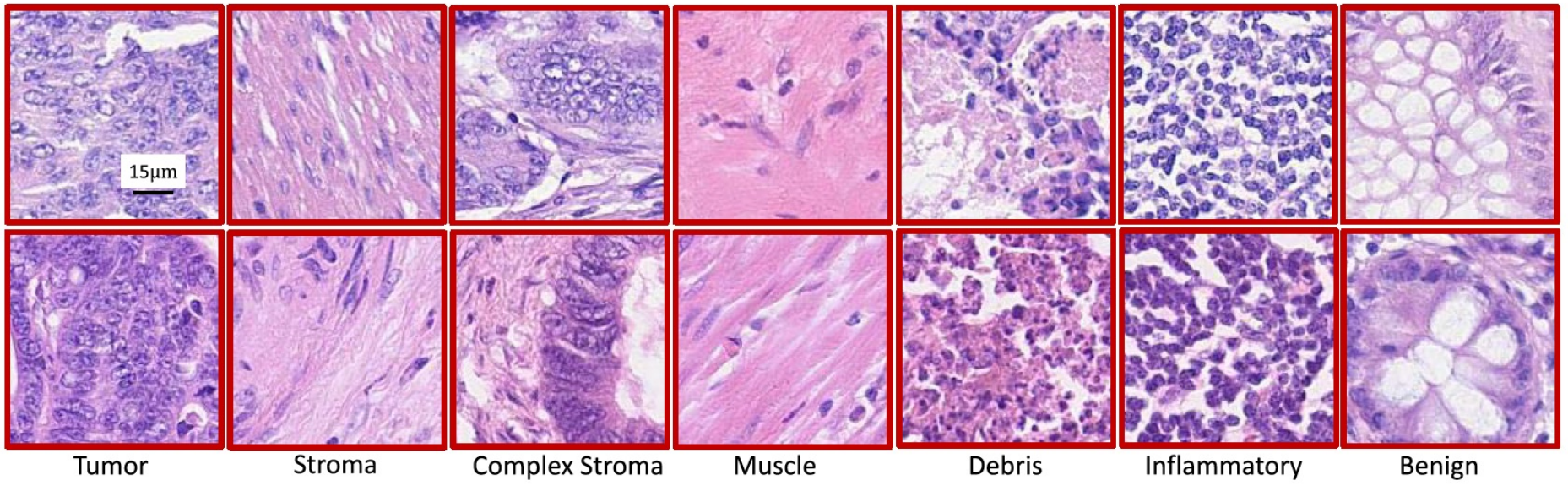


Fig. 4. Sample images for each of the seven distinct tissue phenotypes represented in CRC-TP dataset [33]. Each tissue image contains 150 × 150 pixels ($75\mu m \times 75\mu m, 0.5\mu m/px$) extracted at 20× magnification level.

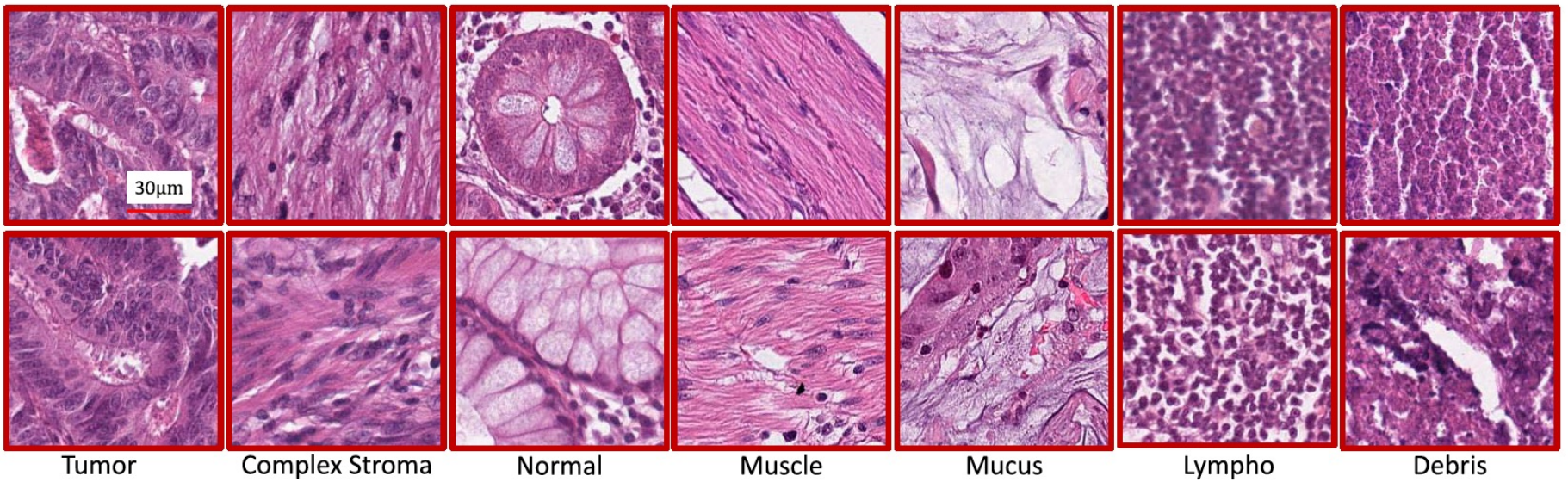


Fig. 5. Sample images for each of the seven tissue phenotypes represented in the CRCH dataset [35]. The tissue images in each phenotypes contain 224 × 224 pixels ($112\mu m \times 112\mu m, 0.5\mu m/px$) captured at 20× magnification level.

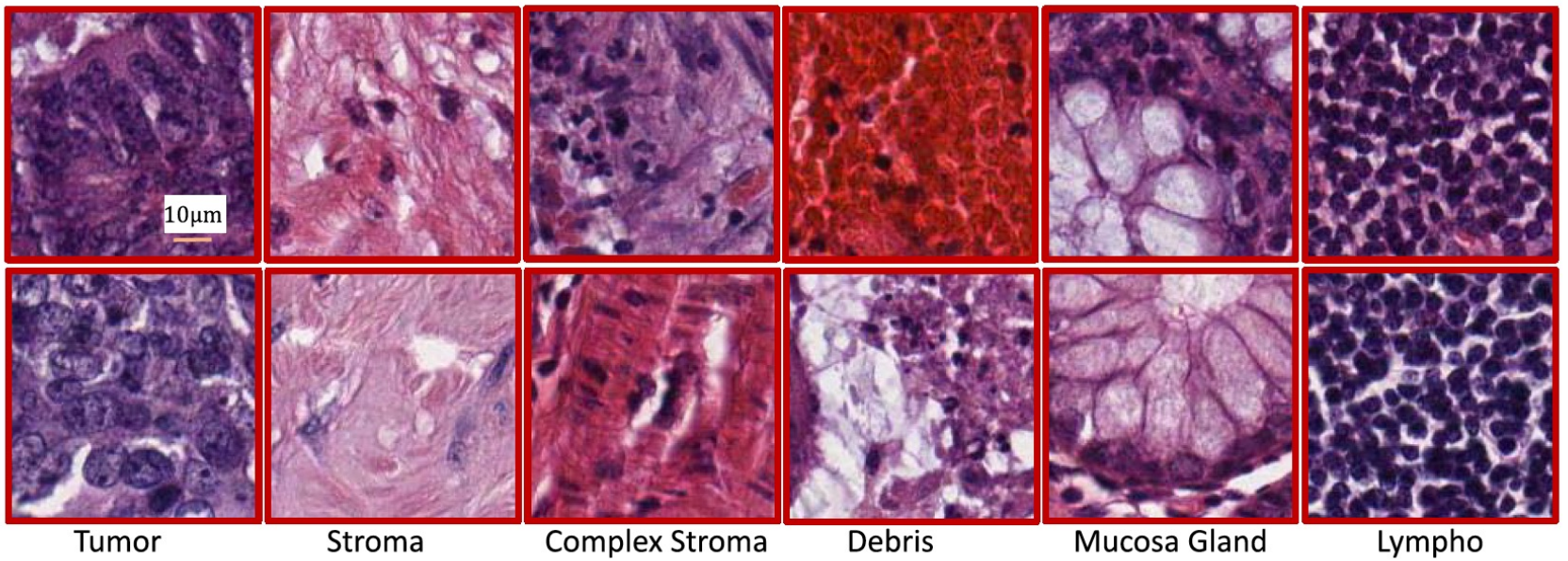
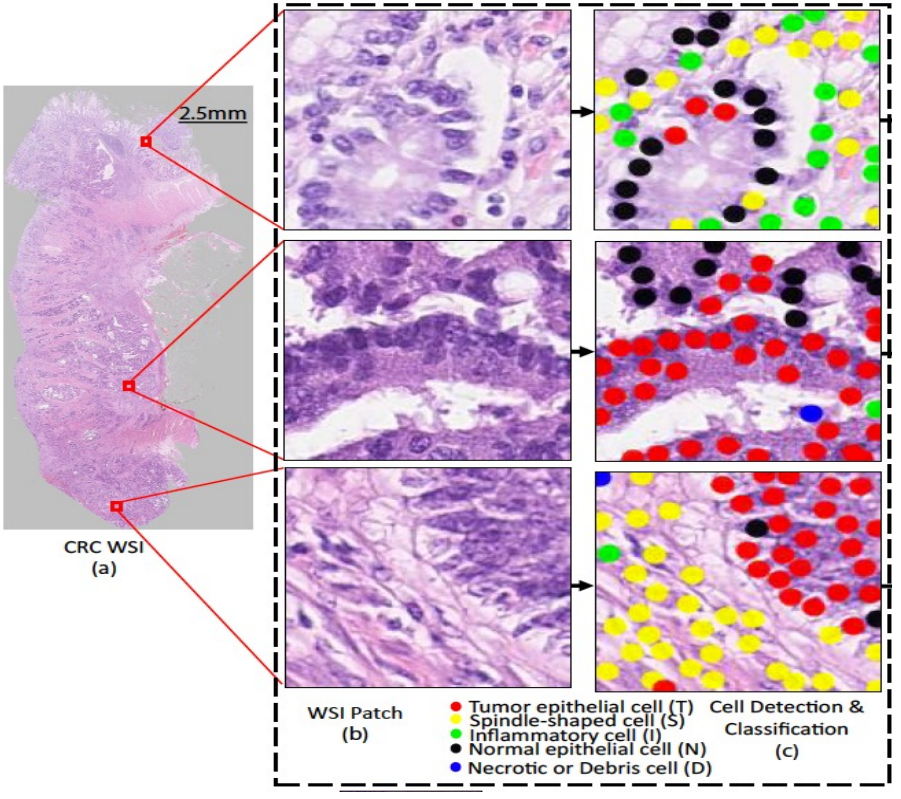
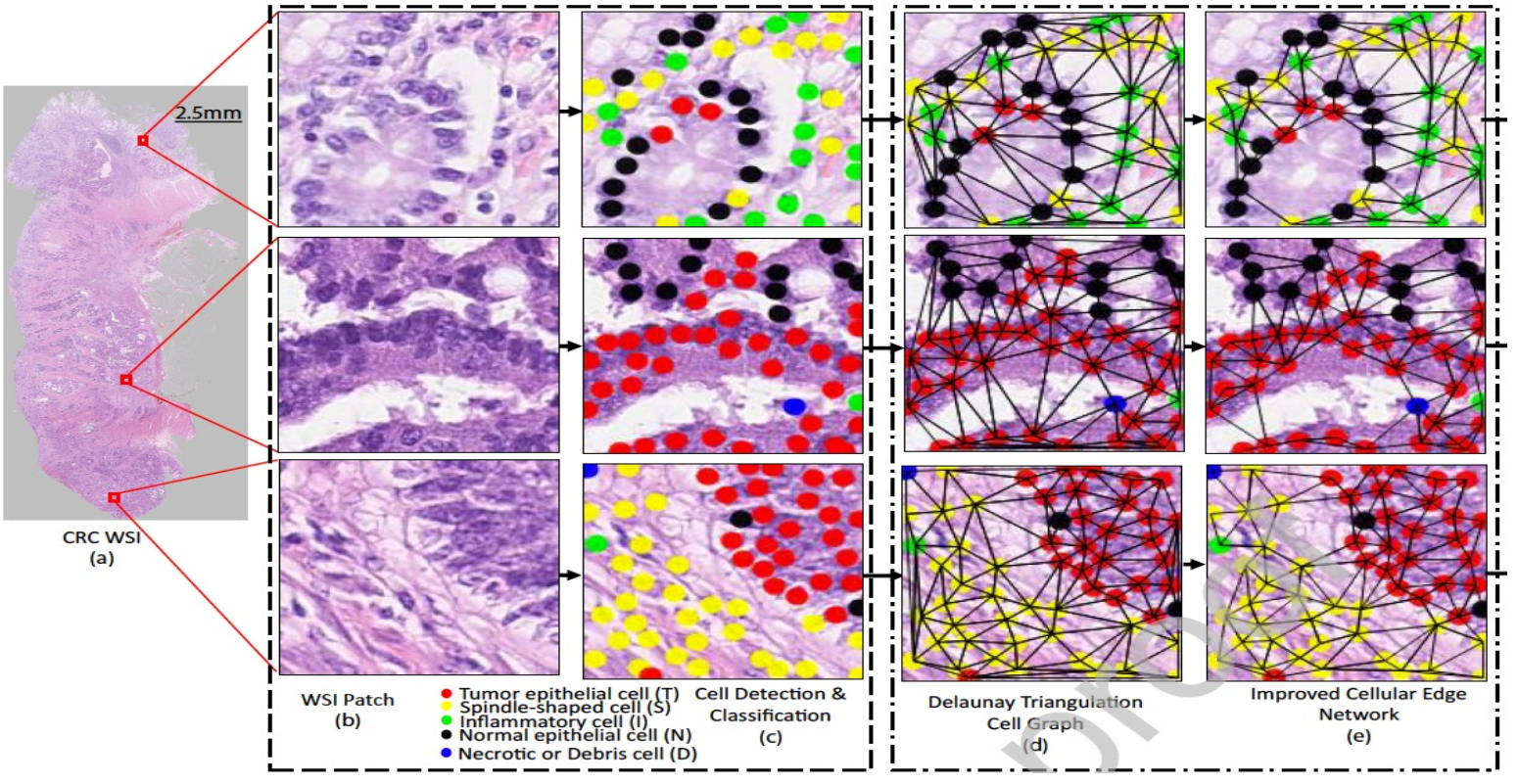
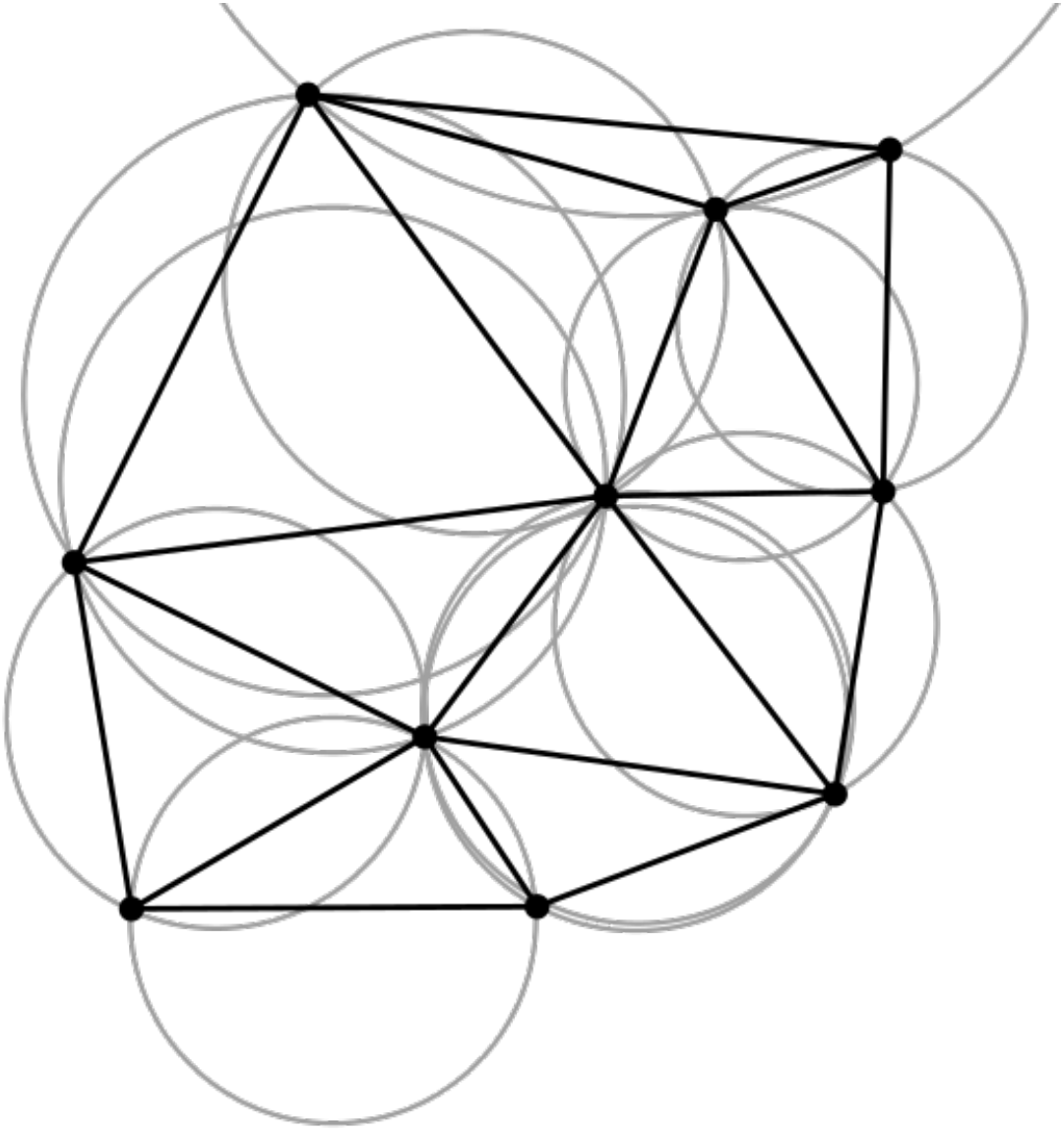


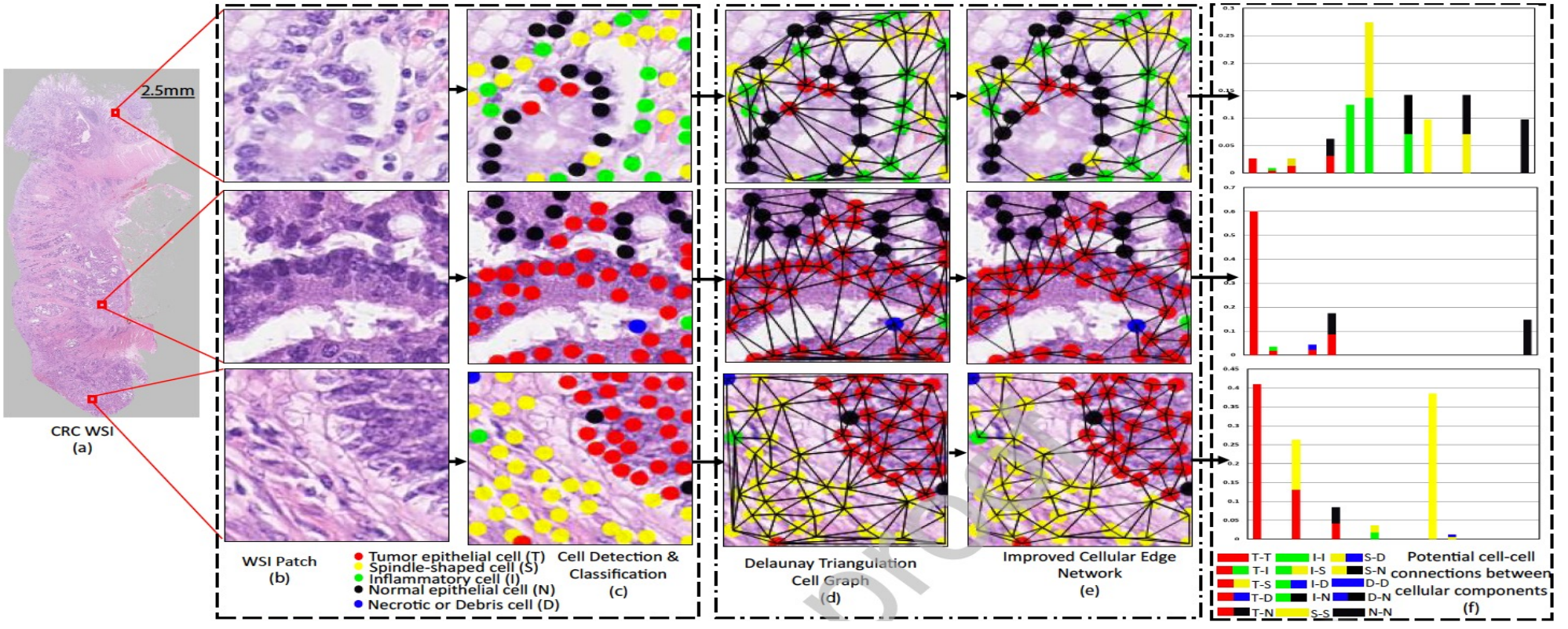
Fig. 6. Sample images for six distinct tissue phenotypes represented in CCH dataset [36]. Each tissue image contains 150 × 150 pixels ($75\mu m \times 75\mu m, 0.5\mu m/px$) extracted at 20× magnification level.

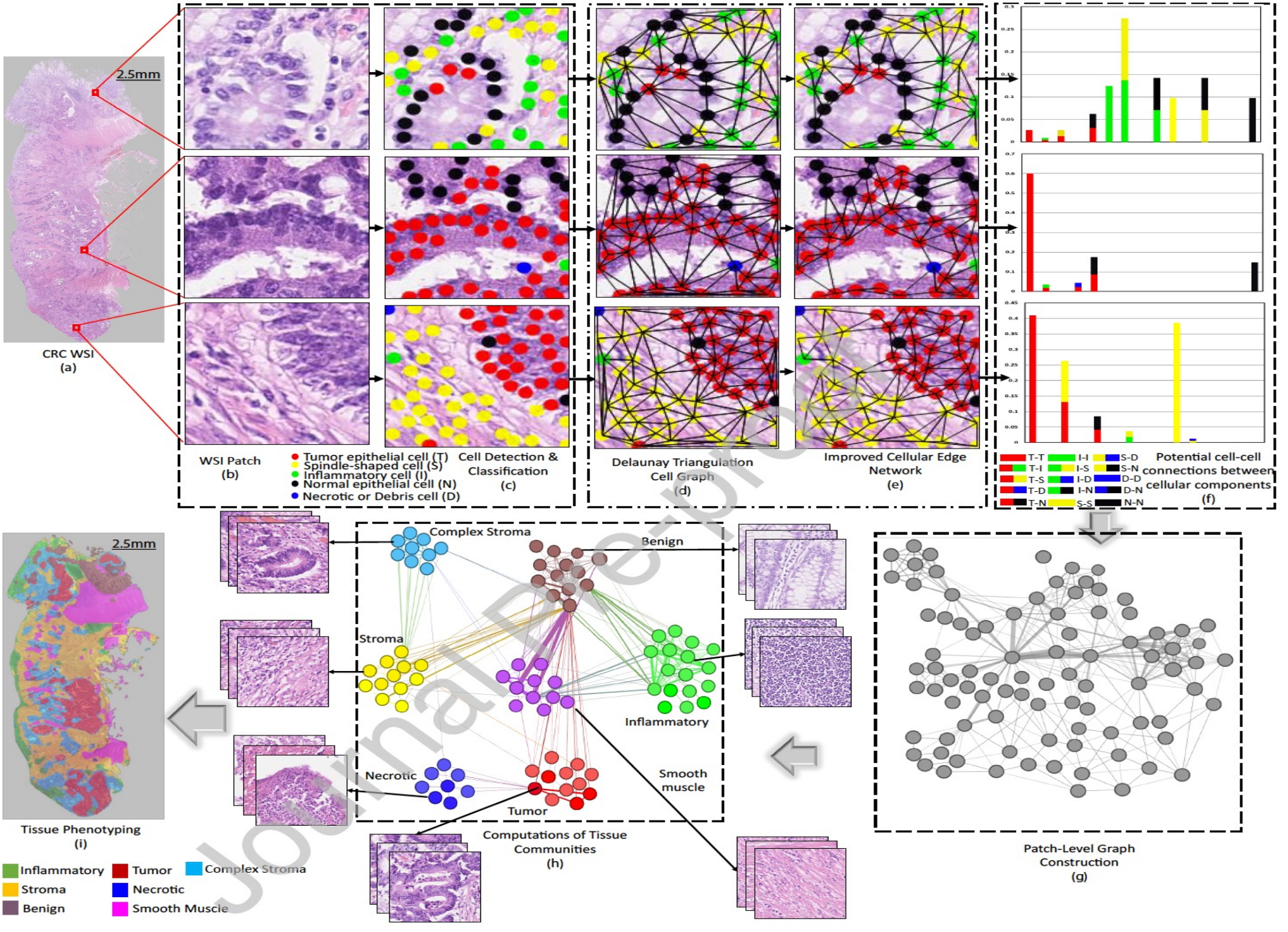

2.5mm
CRC WSI
(a)
WSI Patch
(b)
Tumor epithelial cell (T)
Spindle-shaped cell (S)
Inflammatory cell (I)
Normal epithelial cell (N)
Necrotic or Debris cell (D)
Cell Detection &
Classification
(c)

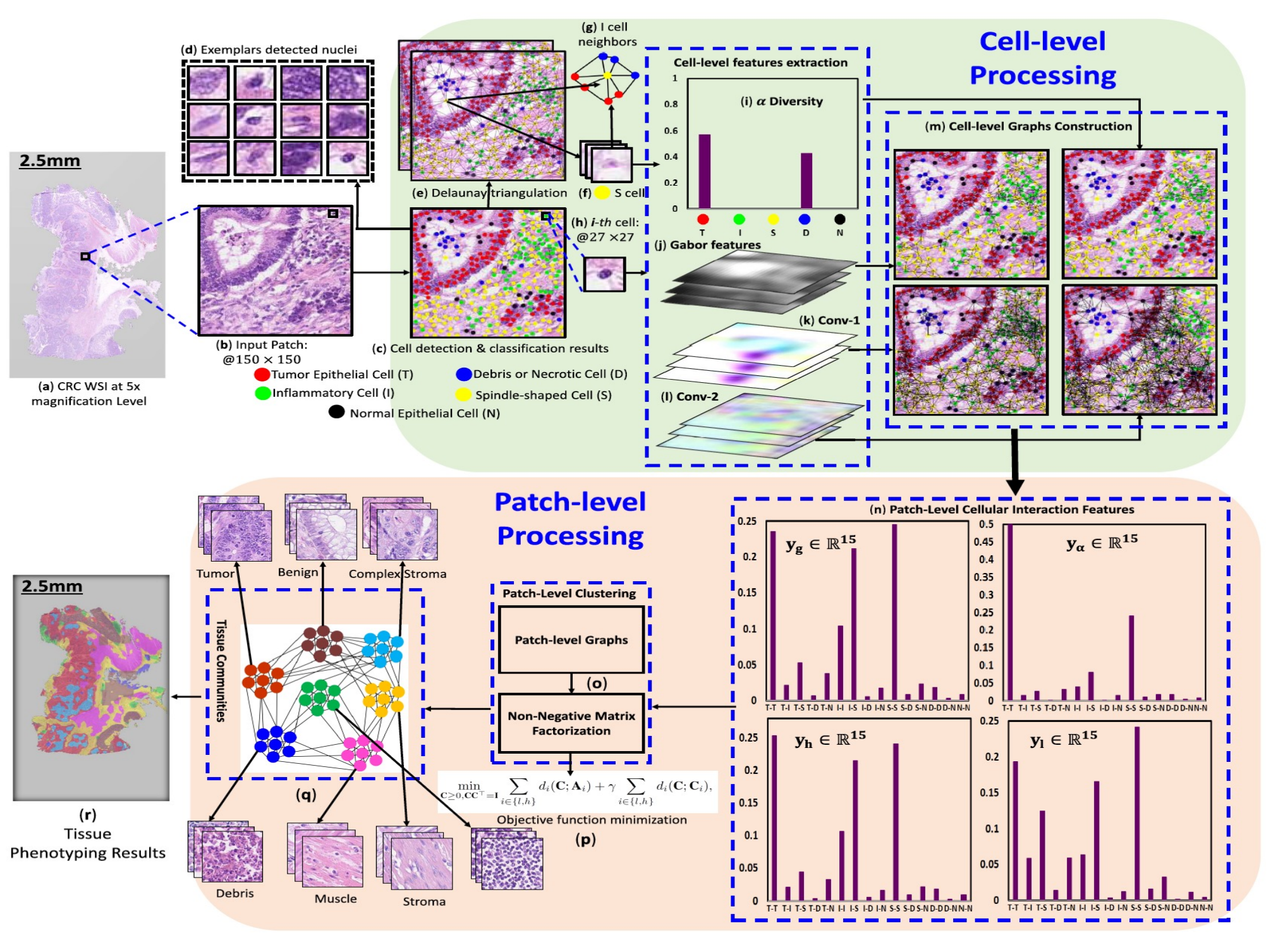

2.5mm
CRC WSI
(a)
WSI Patch
(b)
Tumor epithelial cell (T)
Spindle-shaped cell (S)
Inflammatory cell (I)
Normal epithelial cell (N)
Necrotic or Debris cell (D)
Cell Detection &
Classification
(c)
Delaunay Triangulation
Cell Graph
(d)
Improved Cellular Edge
Network
(e)

Delaunay triangulation




2.5mm
CRC WSI
(a)
WSI Patch
(b)
Tumor epithelial cell (T)
Spindle-shaped cell (S)
Inflammatory cell (I)
Normal epithelial cell (N)
Necrotic or Debris cell (D)
Cell Detection &
Classification
(c)
Delaunay Triangulation
Cell Graph
(d)
Improved Cellular Edge
Network
(e)
T-T
T-I
T-S
T-D
T-N
I-I
I-S
I-D
I-N
S-S
S-D
S-N
D-D
D-N
N-N
Potential cell-cell
connections between
cellular components
(f)




Cell-level Processing
2.5mm
(a) CRC WSI at 5x magnification Level
(b) Input Patch: @150 × 150
(c) Cell detection & classification results
Tumor Epithelial Cell (T)
Inflammatory Cell (I)
Normal Epithelial Cell (N)
Debris or Necrotic Cell (D)
Spindle-shaped Cell (S)
(d) Exemplars detected nuclei
(e) Delaunay triangulation
(f) S cell
(g) I cell neighbors
(h) i-th cell: @27 ×27
Cell-level features extraction
(i) α Diversity
T I S D N
(j) Gabor features
(k) Conv-1
(l) Conv-2
(m) Cell-level Graphs Construction
Patch-level Processing
Tumor
Benign
Complex Stroma
Tissue Communities
(q)
Patch-Level Clustering
Patch-level Graphs
(o)
Non-Negative Matrix Factorization
Objective function minimization
(p)
Debris
Muscle
Stroma
2.5mm
(r) Tissue Phenotyping Results
(n) Patch-Level Cellular Interaction Features
$\mathbf{y_g} \in \mathbb{R}^{15}$
$\mathbf{y_\alpha} \in \mathbb{R}^{15}$
$\mathbf{y_h} \in \mathbb{R}^{15}$
$\mathbf{y_l} \in \mathbb{R}^{15}$
T-T T-I T-S T-D T-N I-I I-S I-D I-N S-S S-D S-N D-D D-N N-N

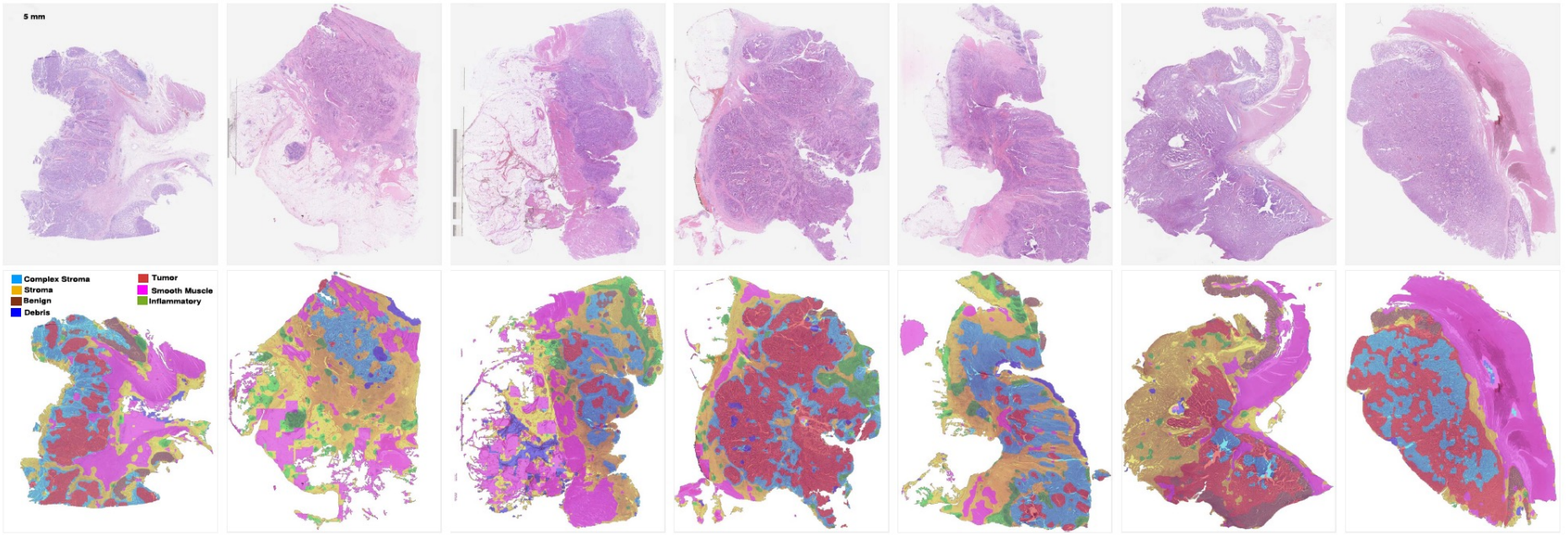


Fig. 7. Visual assessment: the results of the proposed MCTP-2 algorithm are overlaid on seven WSIs taken from CRC-TP dataset [33]. The color coded WSIs are manually inspected by expert pathologist (KB) and found to be biologically meaningful and in agreement with manual annotations.

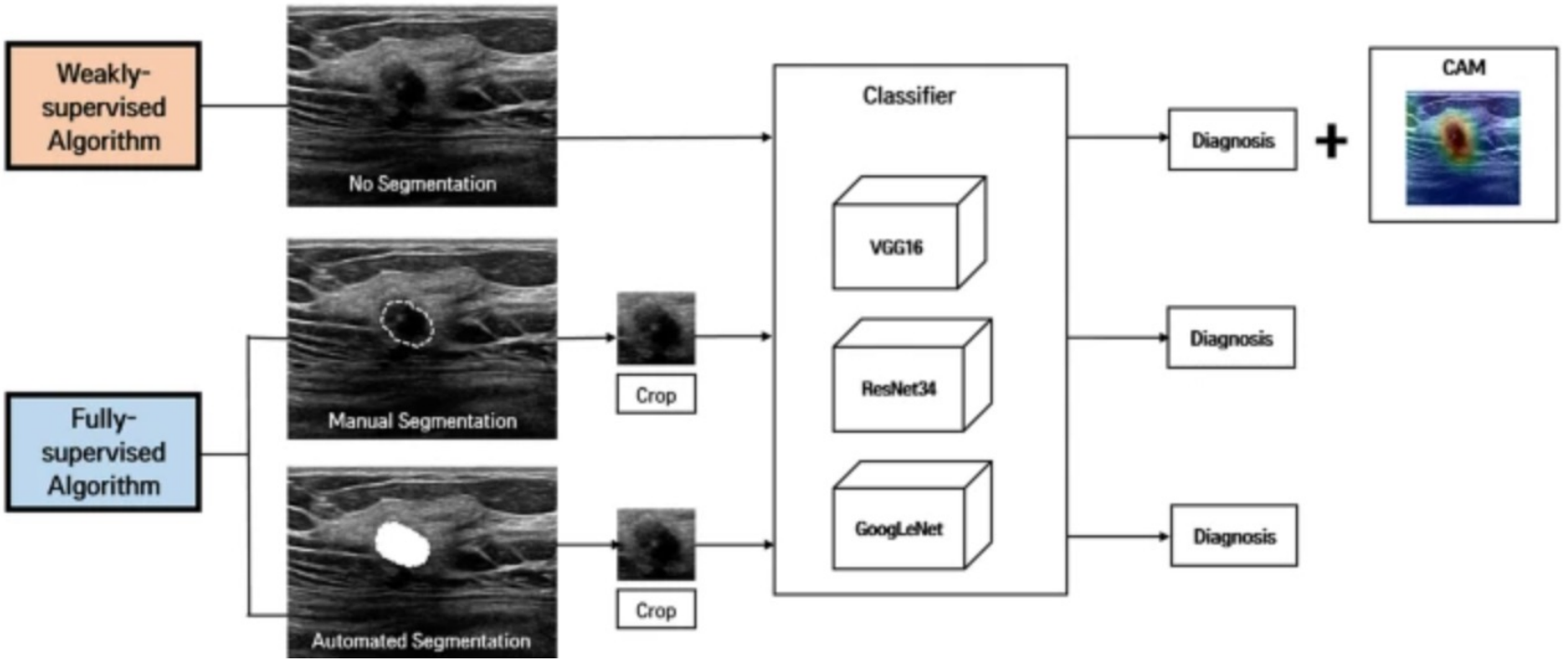
- Trained using 1000 unannotated US images (500 benign and 500 malignant masses).
- Two sets of 200 images (100 benign and 100 malignant masses) were used for internal and external validation sets.

nature › scientific reports › articles › article

Article | Open Access | Published: 21 December 2021

Weakly-supervised deep learning for ultrasound diagnosis of breast cancer

Jaeil Kim, Hye Jung Kim, Chanho Kim, Jin Hwa Lee, Keum Won Kim, Young Mi Park, Hye Won Kim, So Yeon Ki, You Me Kim & Won Hwa Kim



Article | Open Access | Published: 21 December 2021

Weakly-supervised deep learning for ultrasound diagnosis of breast cancer

Jaeil Kim, Hye Jung Kim, Chanho Kim, Jin Hwa Lee, Keum Won Kim, Young Mi Park, Hye Won Kim, So Yeon Ki, You Me Kim & Won Hwa Kim

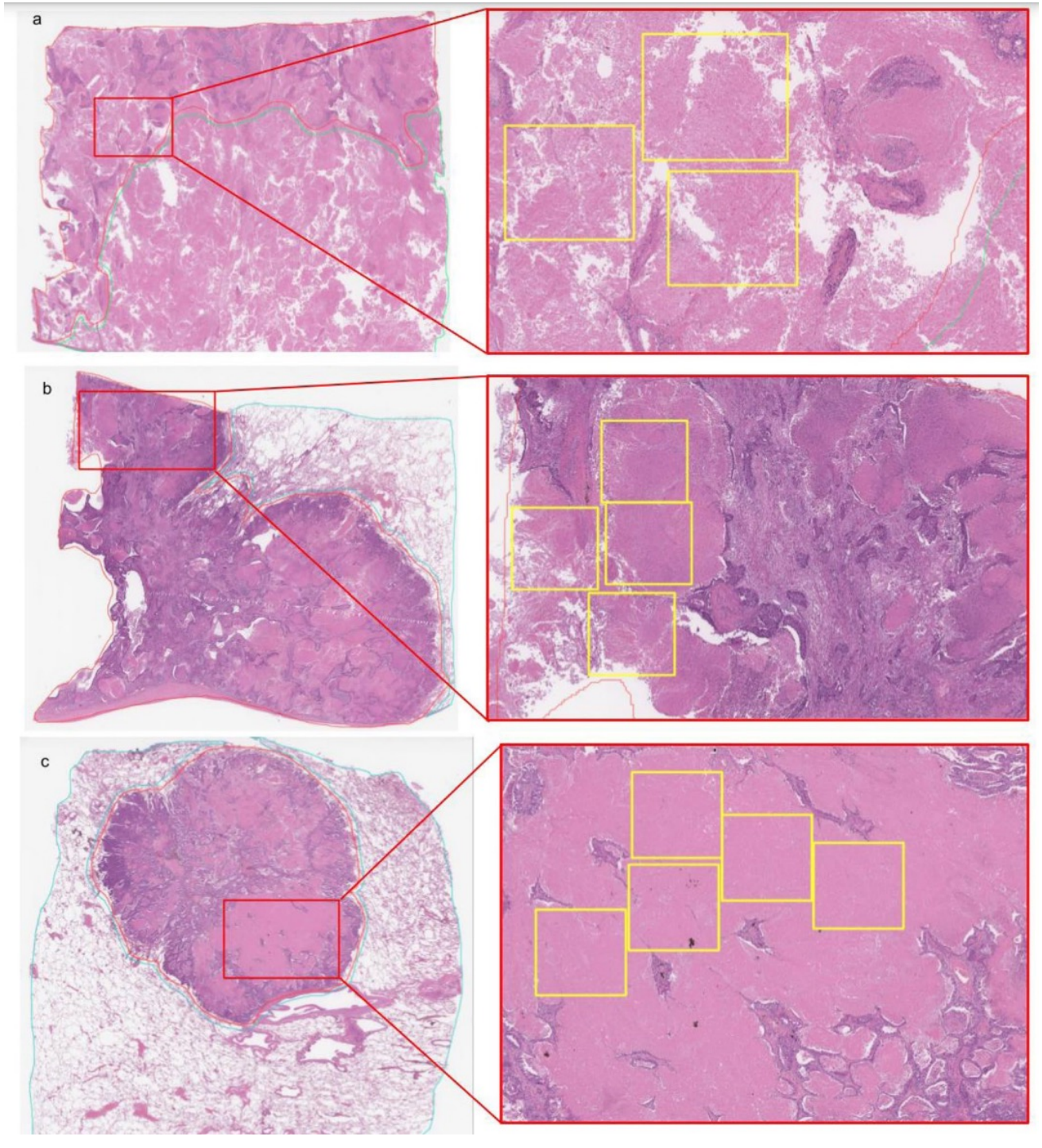
SCIENTIFIC REPORTS

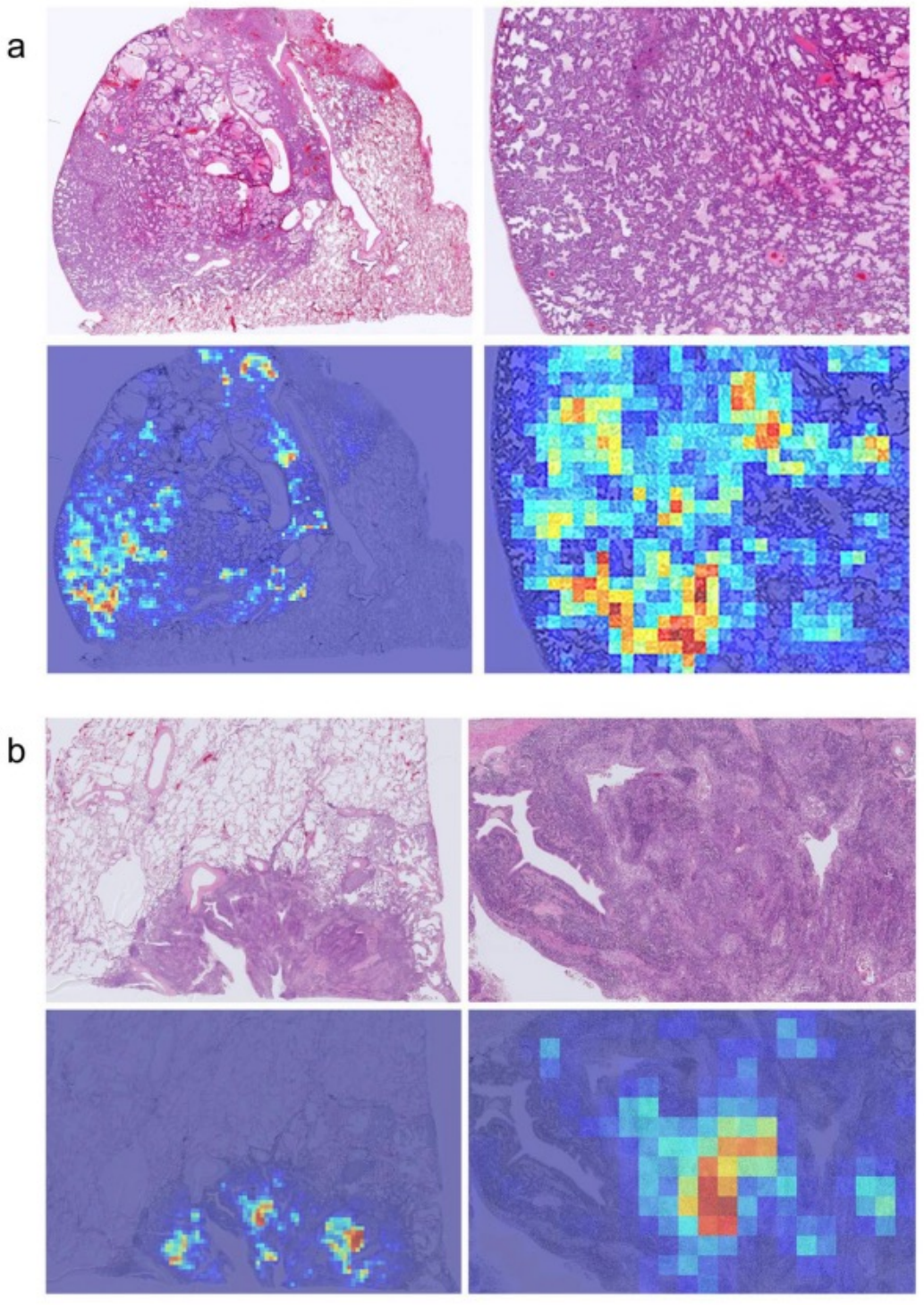
nature research

OPEN

Weakly-supervised learning for lung carcinoma classification using deep learning

Fahdi Kanavati[1,5], Gouji Toyokawa[2,5], Seiya Momosaki[3,5], Michael Rambeau[4], Yuka Kozuma[2], Fumihiro Shoji[2], Koji Yamazaki[2], Sadanori Takeo[2], Osamu Iizuka[4] & Masayuki Tsuneki[1,4] ✉


a
b
c


a
b