

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Advances in Computer Vision Employing Deep Learning

Mubarak Shah

Center for Research in computer Vision
University of Central Florida
shah@crcv.ucf.edu
https://www.crcv.ucf.edu/





Classical Computer Vision Vs Deep Learning

Classical Computer Vision

- Solve Problem in a Modular Fashion
- Encode Expert Knowledge into constraints
- Convert constraints into Objective Function
- Optimize Objective Function

Deep Learning

- End-to-End
- Real Learning
- Motivated by human neural networks
- Requires Labeled/Annotated Data
- Need to Define Loss Function
- Requires Massively Parallel Computations
- Use simple Stochastic Gradient Descend
- Excellent Results



Contents

- Person RE-ID
- Multiple Object Tracking
- Drone Video Analysis
- Anomaly Detection
- Action Localization
- Video Object Segmentation
- Building Damage Assessment
- Face Retrieval
- Geo-spatial Localization
- Grounded Visual Q&A



Human Semantic Parsing for Person Re-identification

Mahdi M. Kalayeh, Emrah Basaran, Muhittin G¨okmen, Mustafa E. Kamasak Mubarak Shah

CVPR 2018

http://crcv.ucf.edu/people/phd_students/mahdi/papers/CVPR18.pdf

https://www.youtube.com/watch?v=nN0U0i7gBhE&list=PLd3hlSJsX_Ilqyw513wCDl3zS35YYGsBu&index=3



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Person Re-Identification

- Retrieving the images of a person from a large gallery
- Querry and gallery images are captured by different cameras
- A cross-camera data association problem

CUHK03

- **1,360** identities captured by **6** cameras.
  - Each identity is viewed by **2** disjoint cameras.
- **13,164** person images
  - Each identity has on average **4.8** images in each viewpoint.
- The dataset partitions:
  - Training: **1160** persons
  - Validation: **100** persons
  - Test: **100** persons



Person Re-Identification



Challenges

- Illumination conditions
- Observable human body parts
- Perceived posture of the person
- **Background clutter**
- **Occlusion**

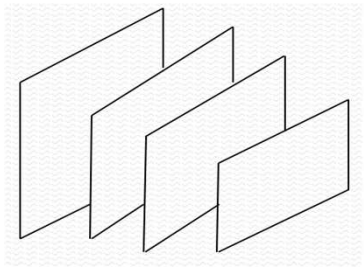


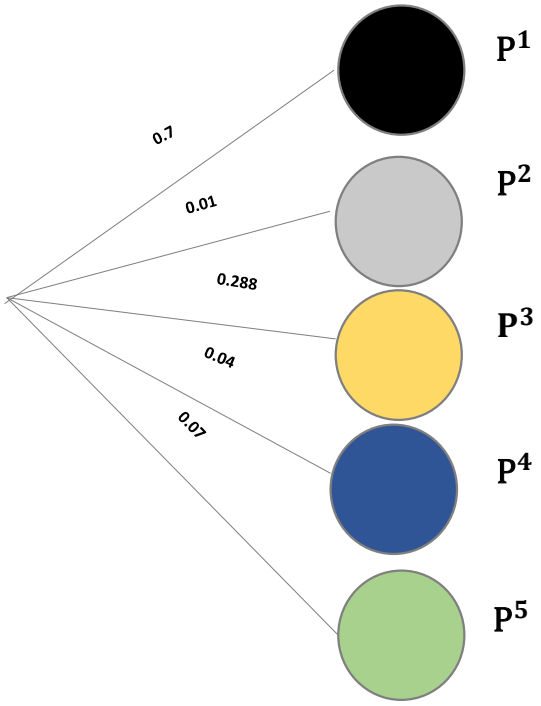




Person Classification Network



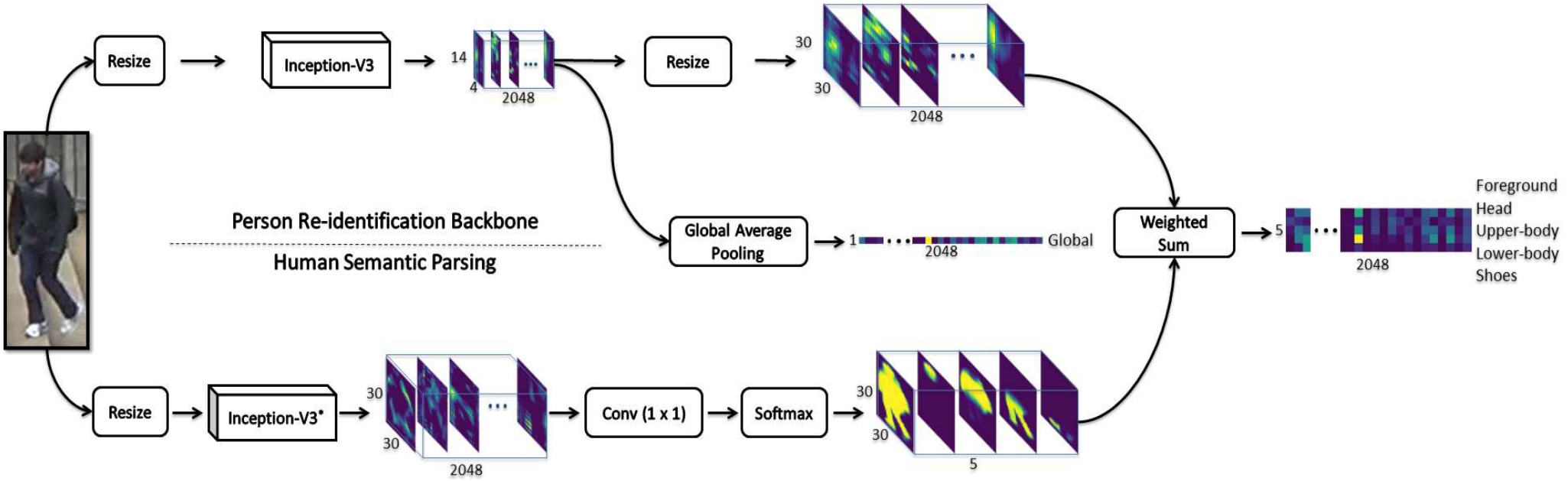




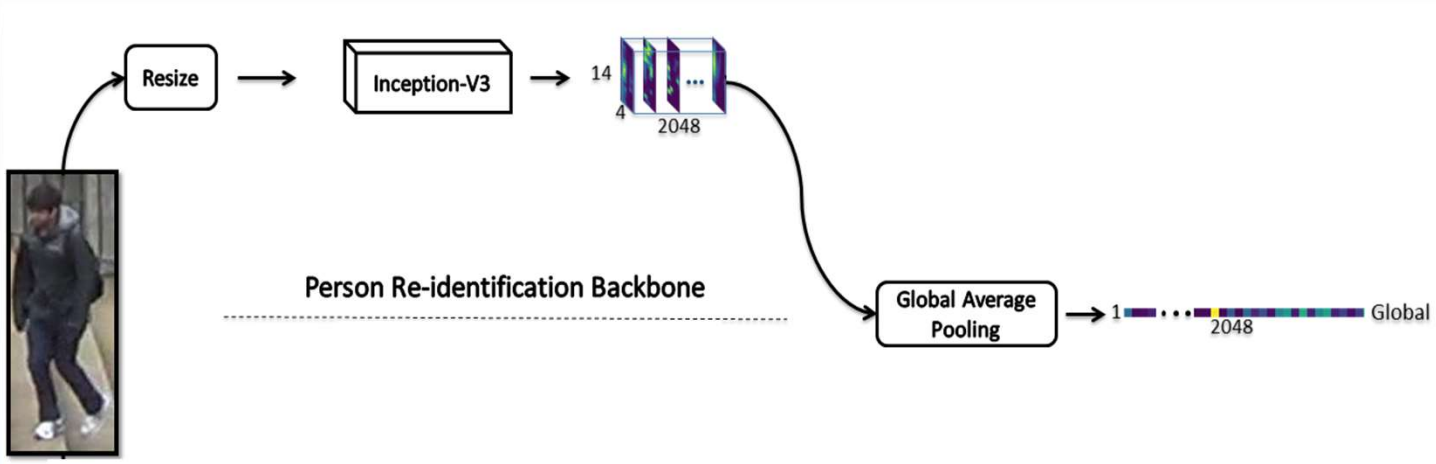
Proposed Approach

- Employ Human **S**emantic **P**arsing
- Train on aggregation of multiple datasets To
- Use Features of Trained Network to solve Re-Identification Problem

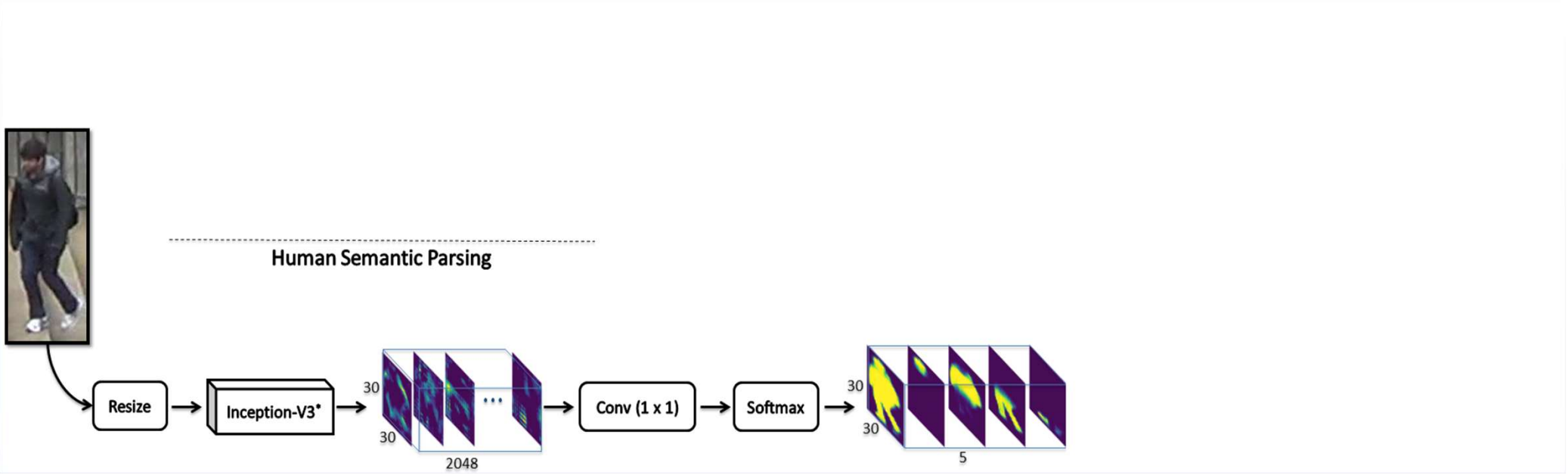
SPReID: Human Semantic Parsing for Person Re-identification



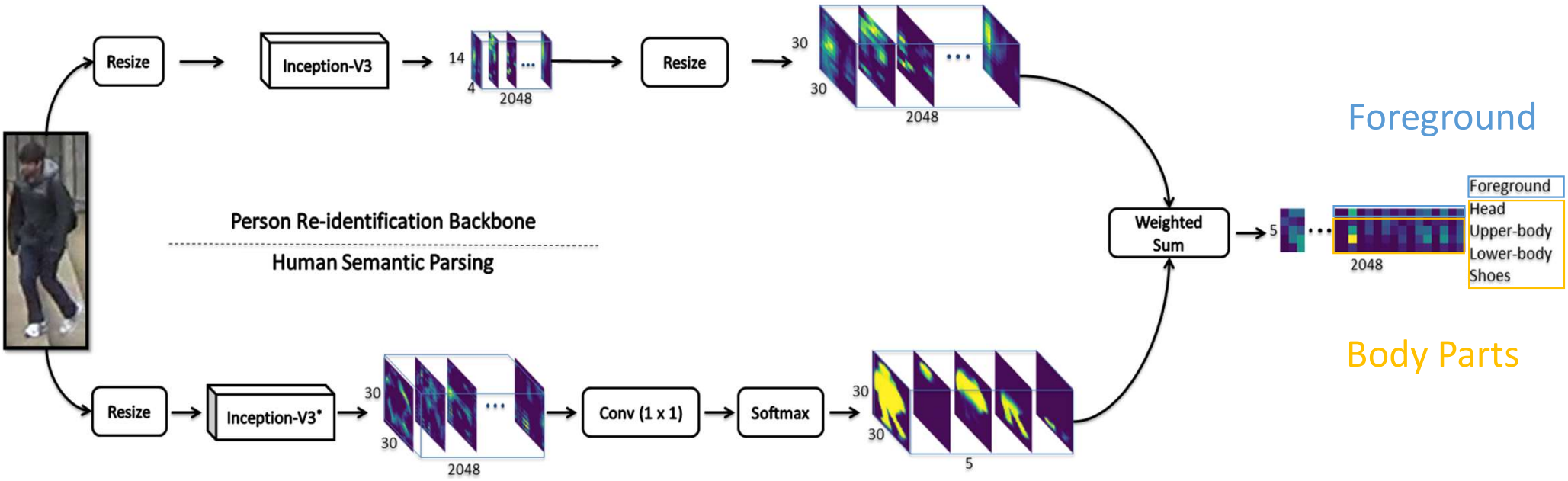
SPReID - Global Representation



SPReID - Human Semantic Parsing

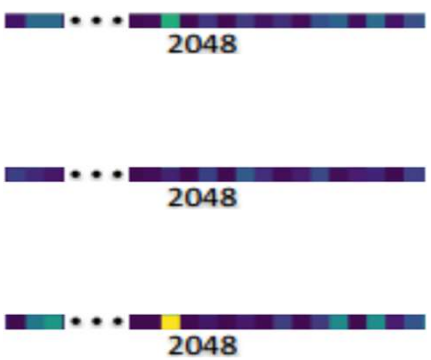


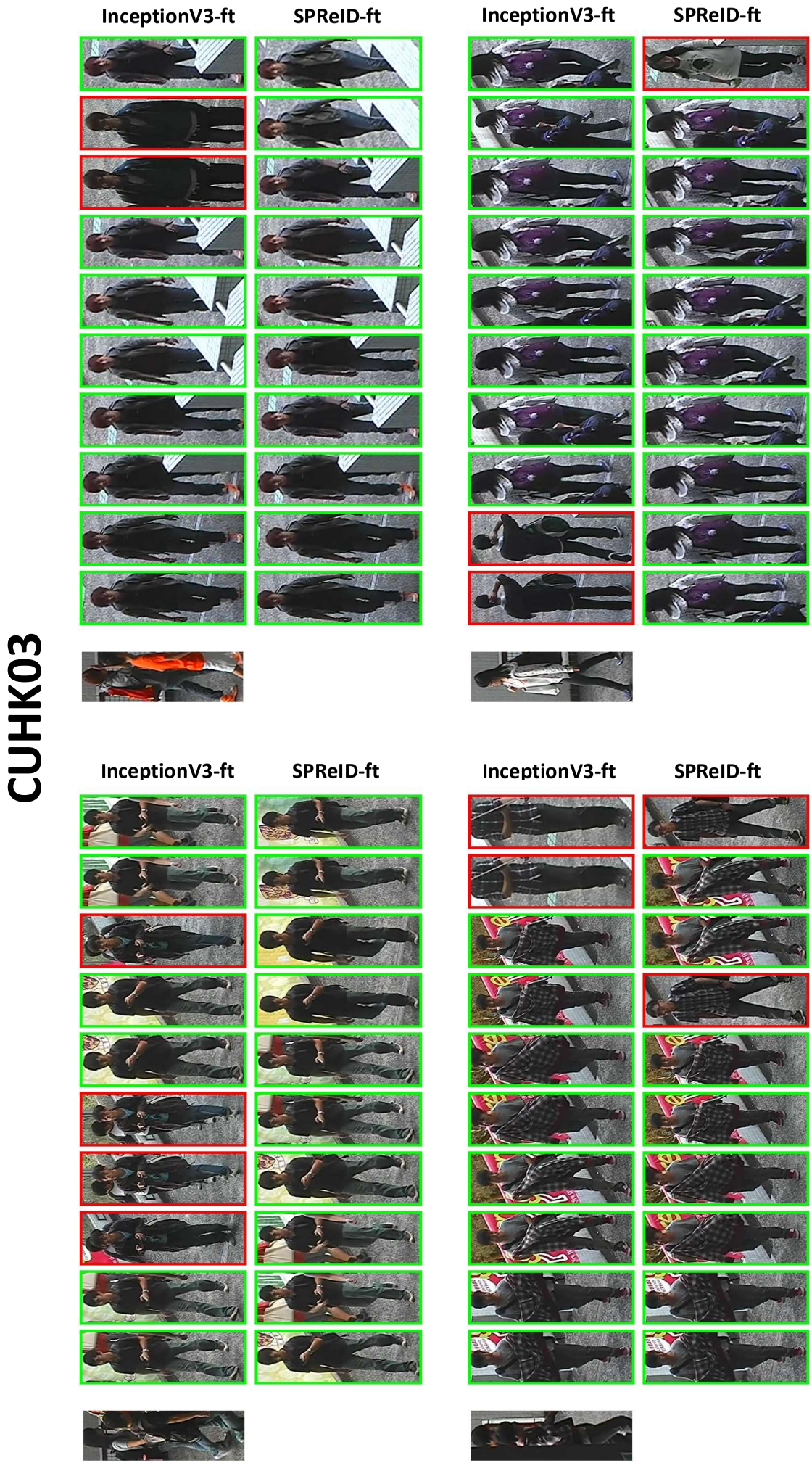
SPReID - Human Semantic Parsing

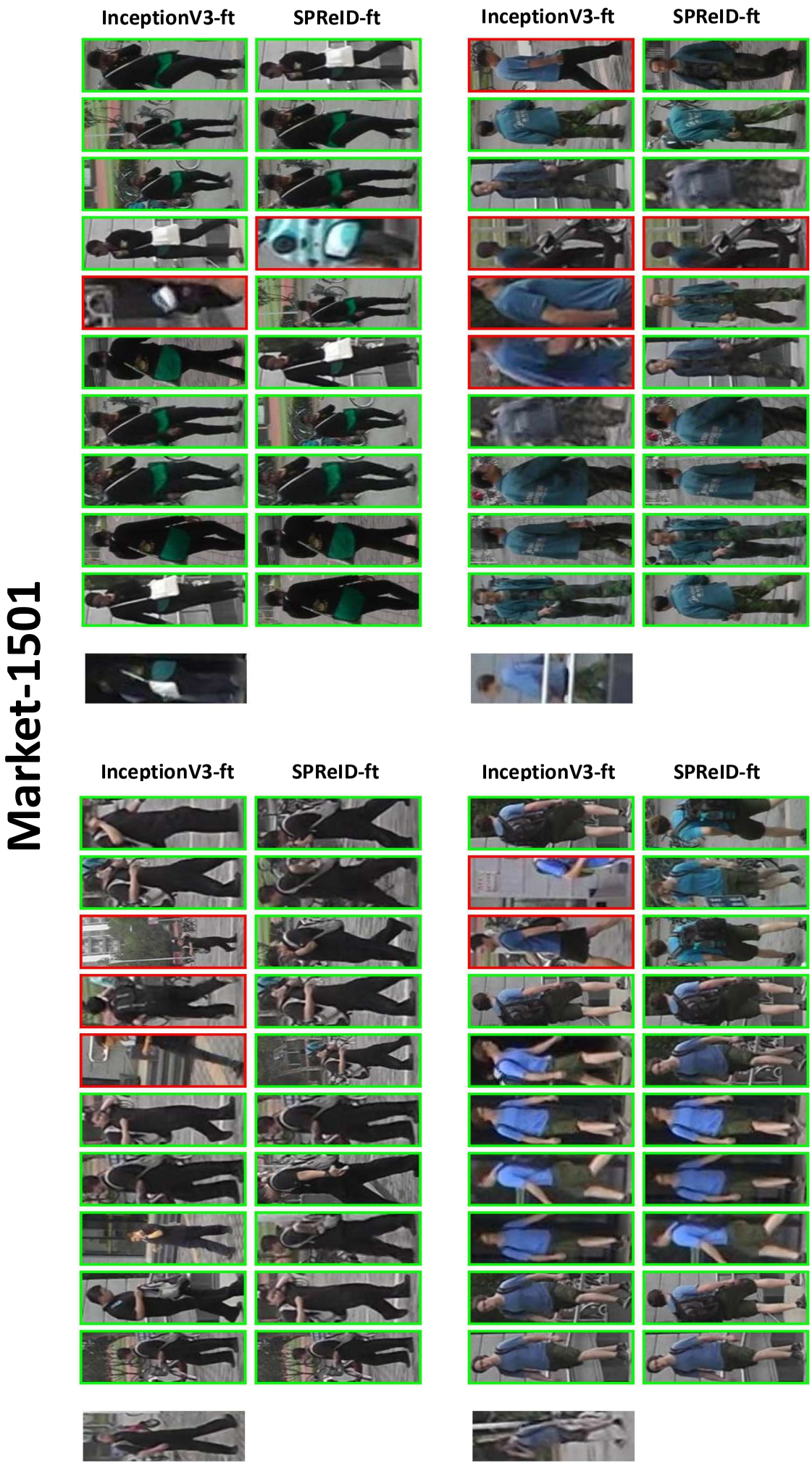


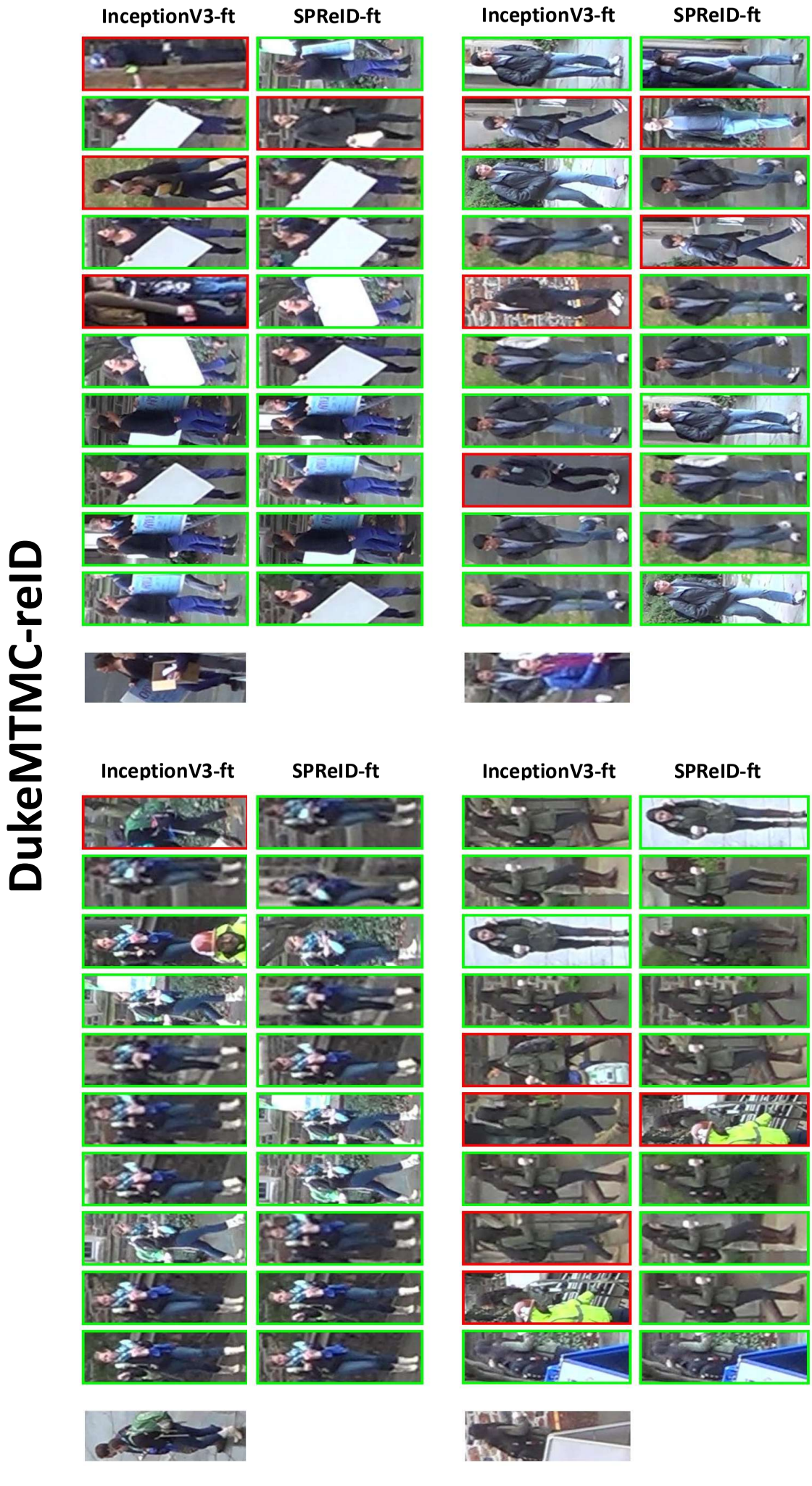
SPReID Representation

- Concatenation of
  - Global representation
  - Foreground representation
  - Body Part representation




CUHK03
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft


Market-1501
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft


DukeMTMC-reID
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft
InceptionV3-ft
SPReID-ft

Results – CUHK03

State-of-the-art

Ours - **without** semantic parsing

Ours - **with** semantic parsing

Method	mAP (%)	rank-1	rank-5	rank-10
FT-JSTL+DGD	-	75.3	-	-
SSM	-	76.6	94.6	98.0
Spindle	-	88.5	97.8	98.6
DPAR	-	85.4	97.6	99.4
Chen et. Al.	82.8	86.7	-	-
HydraPlus	**-**	**91.8**	**98.4**	**99.1**
Inception-V3^{ft}	-	88.73	97.82	98.94
ResNet-152^{ft}	-	90.38	98.71	99.46
combined	-	92.81	98.9	99.35
SPReID^{ft}	-	93.89	98.76	99.51
+ re-ranking	-	96.31	99.25	99.71
ResNet-152^{ft} + SPReID^{ft}	-	94.28	99.04	99.56
+ re-ranking	**-**	**96.22**	**99.34**	**99.7**

Person Re-Identification

- Human Semantic Parsing for Person Re-Identification (CVPR2018**)**
- **Deep Constrained Dominant Sets for Person Re-Identification (ICCV2019)**







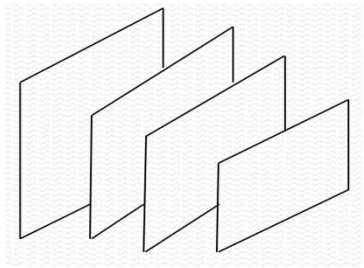
Deep Constrained Dominant Sets for Person Re-Identification

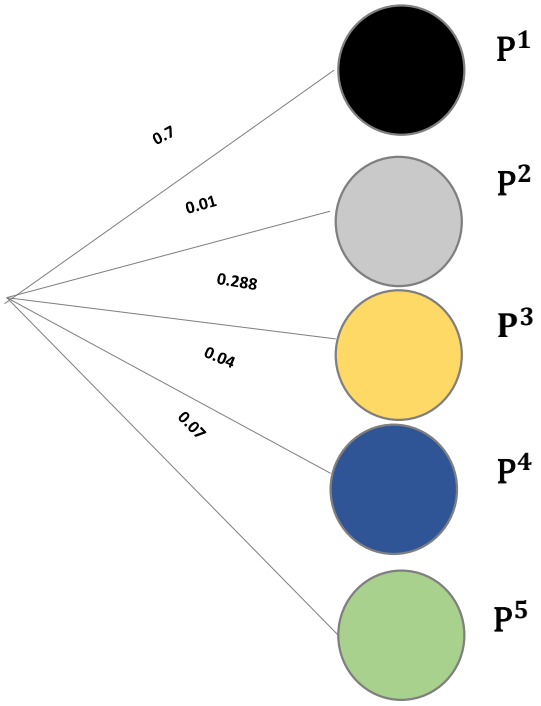
Alemu Leulseged Tesfaye **Marcello Pelillo** **Mubarak Shah**

https://www.youtube.com/watch?v=R8vXWtwwAbk&list=PLd3hlSJsX_IkSnnrMtzsMHl1q6vimipvp&index=2

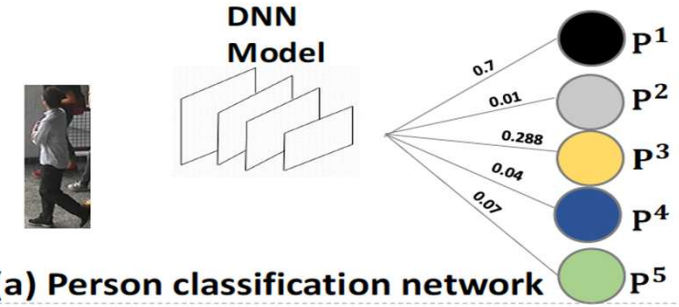
Person Classification Network



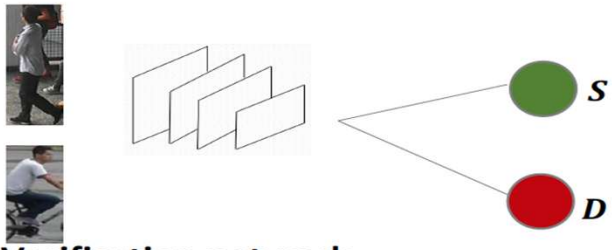




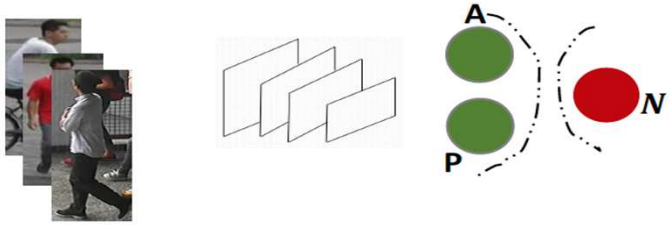
DNN Based Person Re-Id Methods



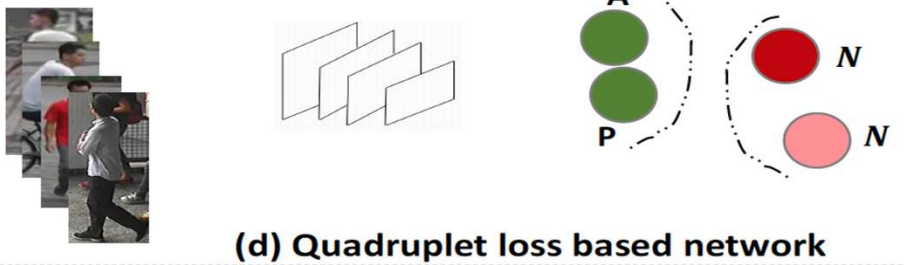
(a) Person classification network



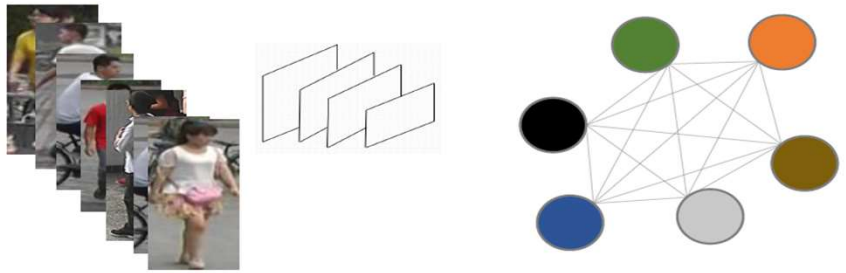
(b) Verification network



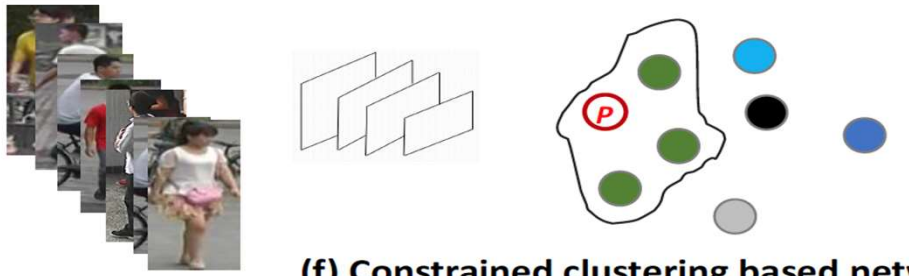
(c) Triplet loss based network



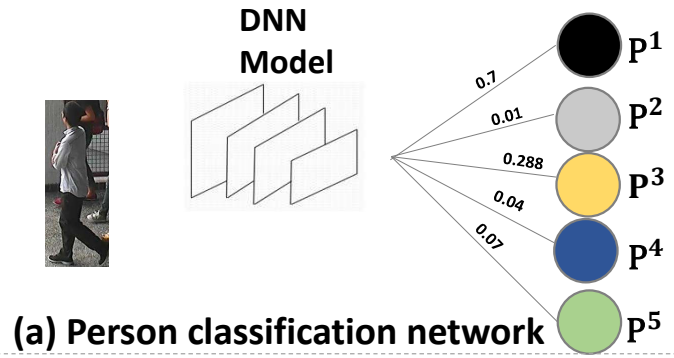
(d) Quadruplet loss based network



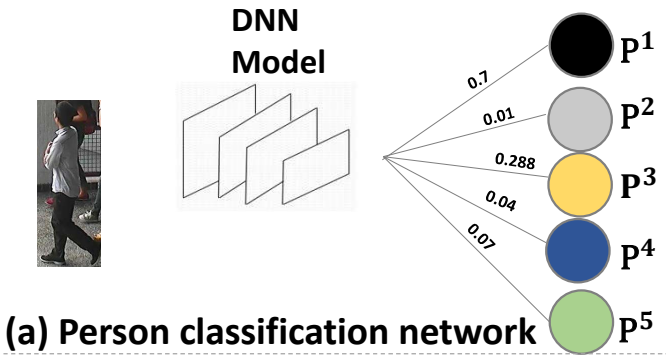
(e) Conventional diffusion based network



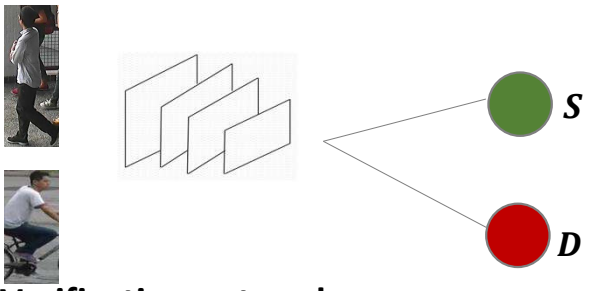
(f) Constrained clustering based network (DCDS)



(a) Person classification network

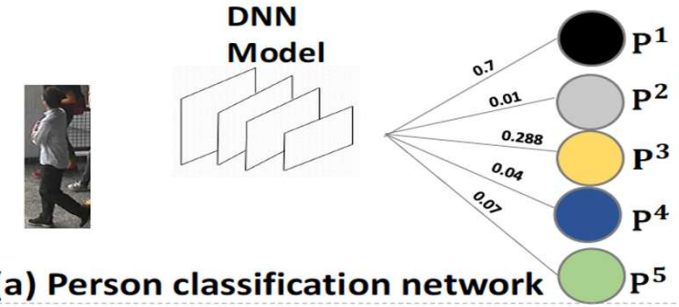


(a) Person classification network

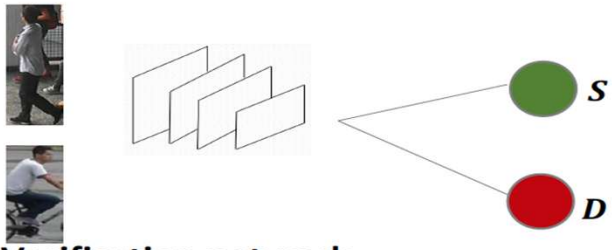


(b) Verification network

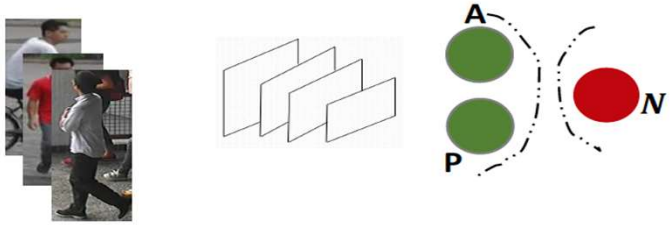
DNN Based Person Re-Id Methods



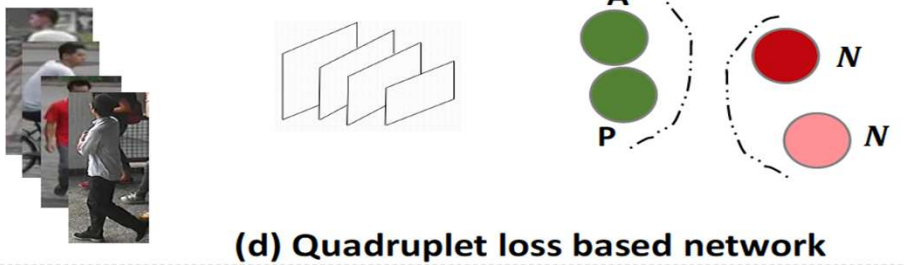
(a) Person classification network



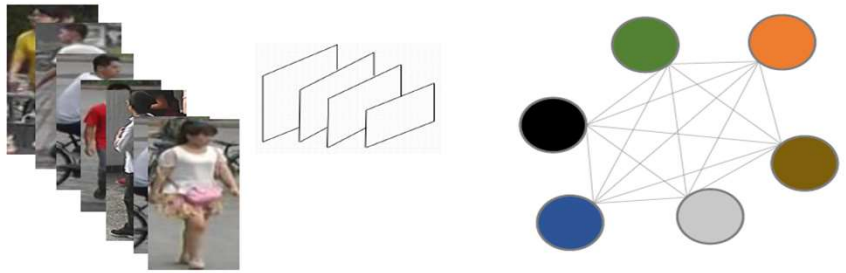
(b) Verification network



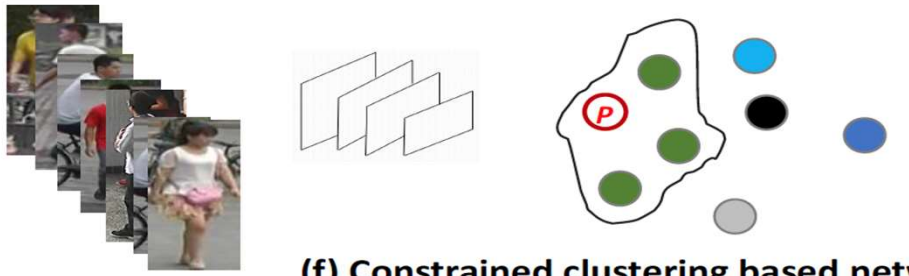
(c) Triplet loss based network



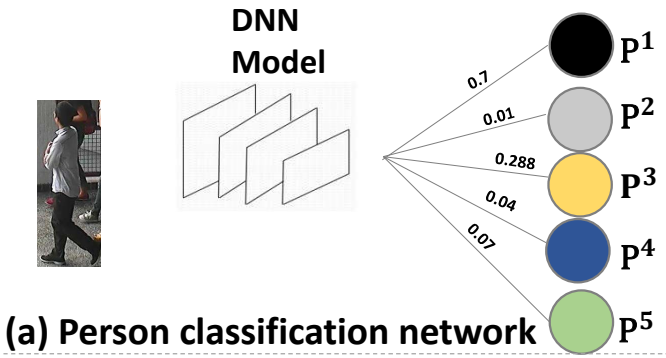
(d) Quadruplet loss based network



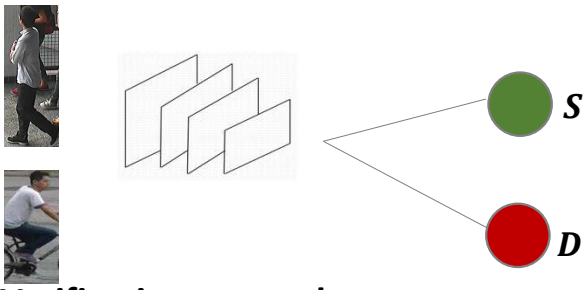
(e) Conventional diffusion based network



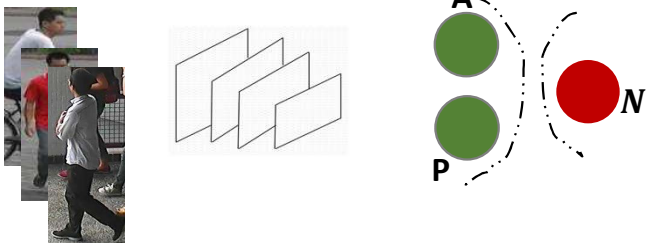
(f) Constrained clustering based network (DCDS)



(a) Person classification network



(b) Verification network



(c) Triplet loss based network

Triplet Loss Based Network

Positive



Query



Negative



Triplet Loss Based Network

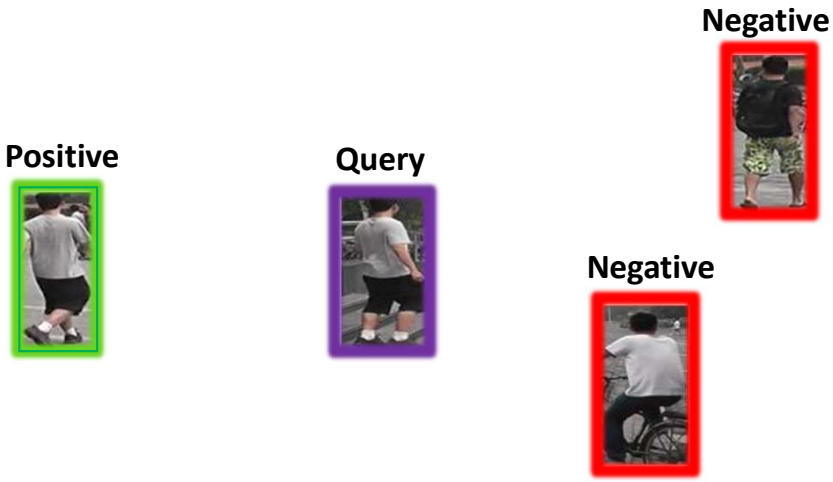
Positive **Query**



Negative



Quadruplet Loss Based Network



Quadruplet Loss Based Network

Positive



Query



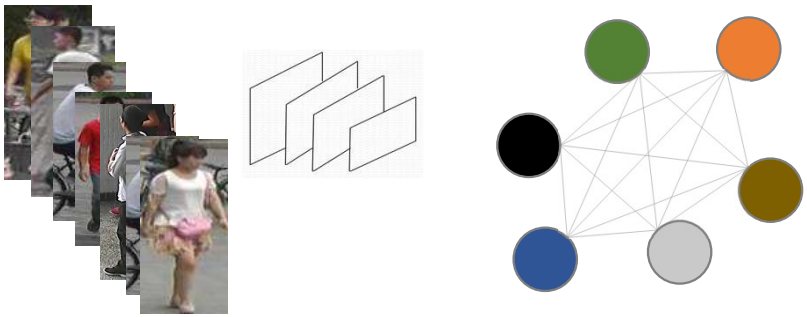
Negative



Negative



Diffusion Based Network

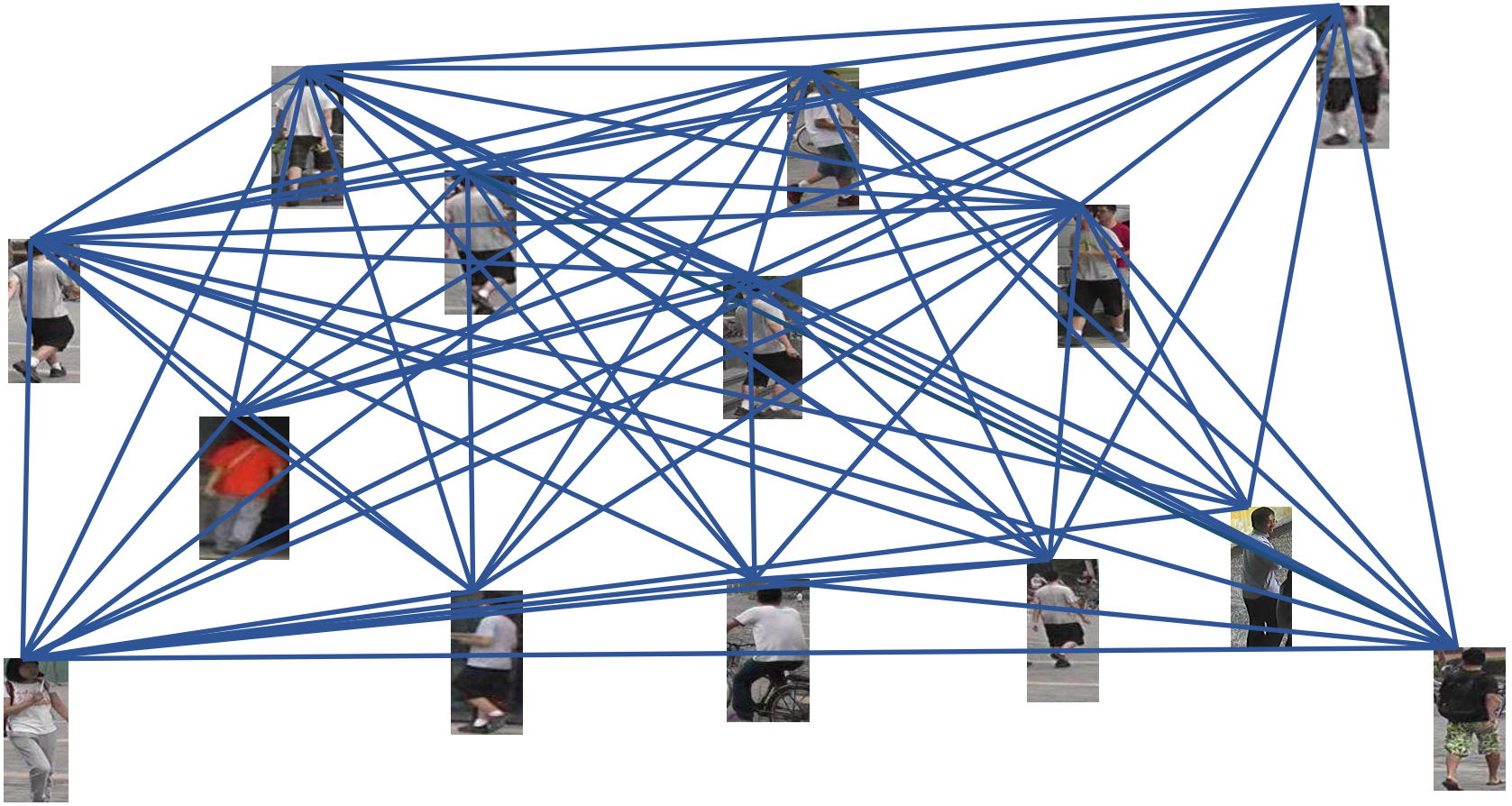


(e) Conventional diffusion based network

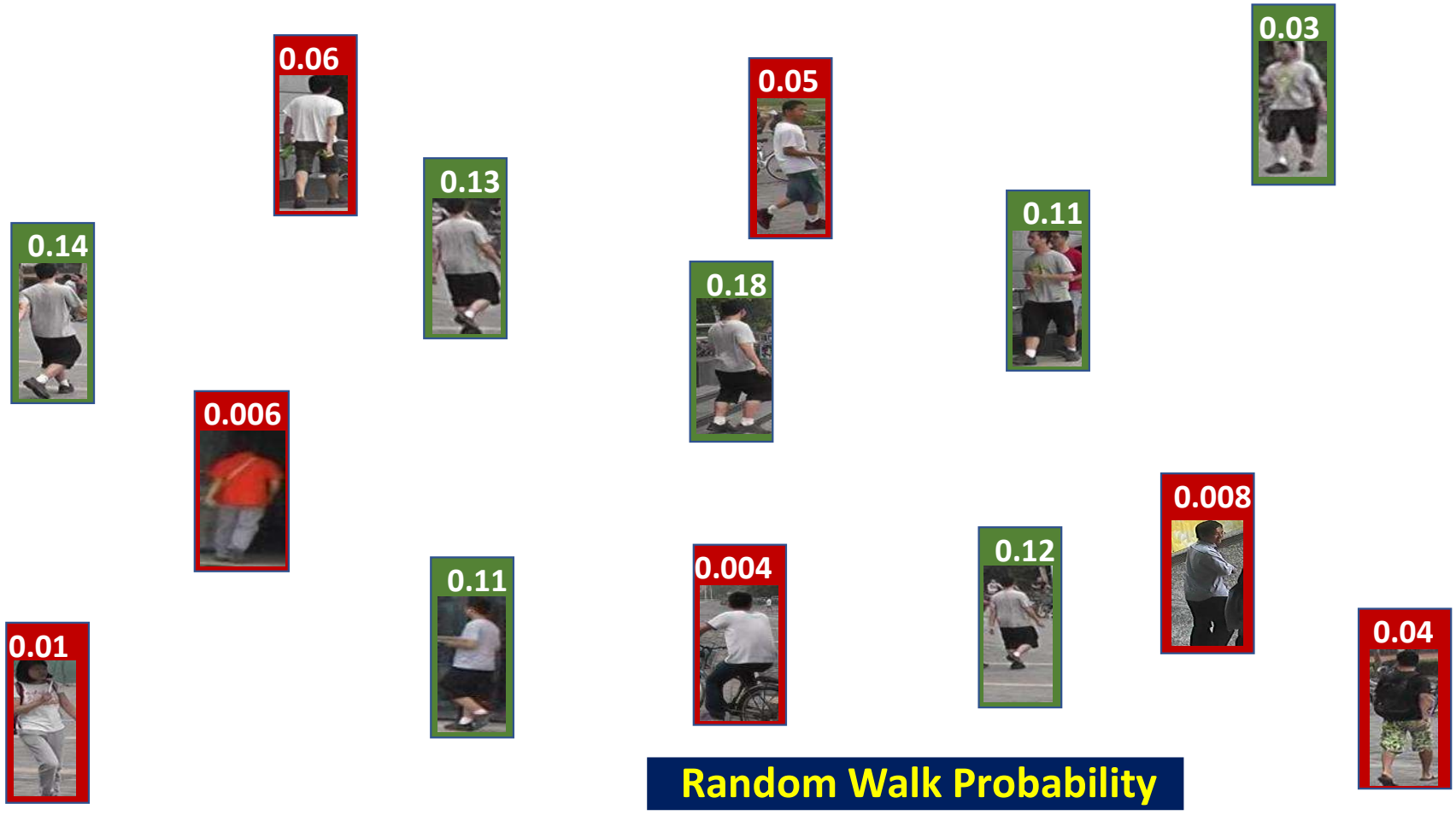
Diffusion Based Network



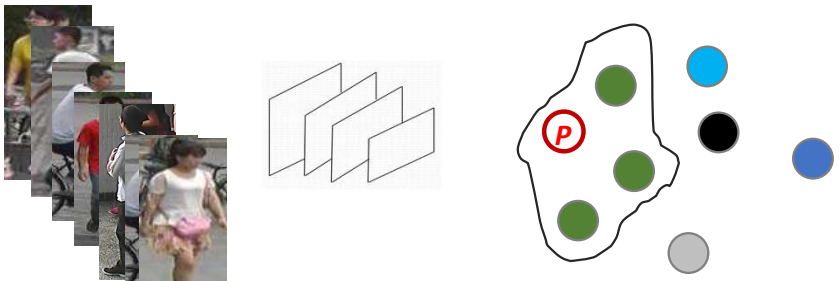
Diffusion Based Network



Diffusion Based Network



DCDS Based Network



(f) Constrained clustering based network (DCDS)

DCDS Based Network








Query (Constraint)

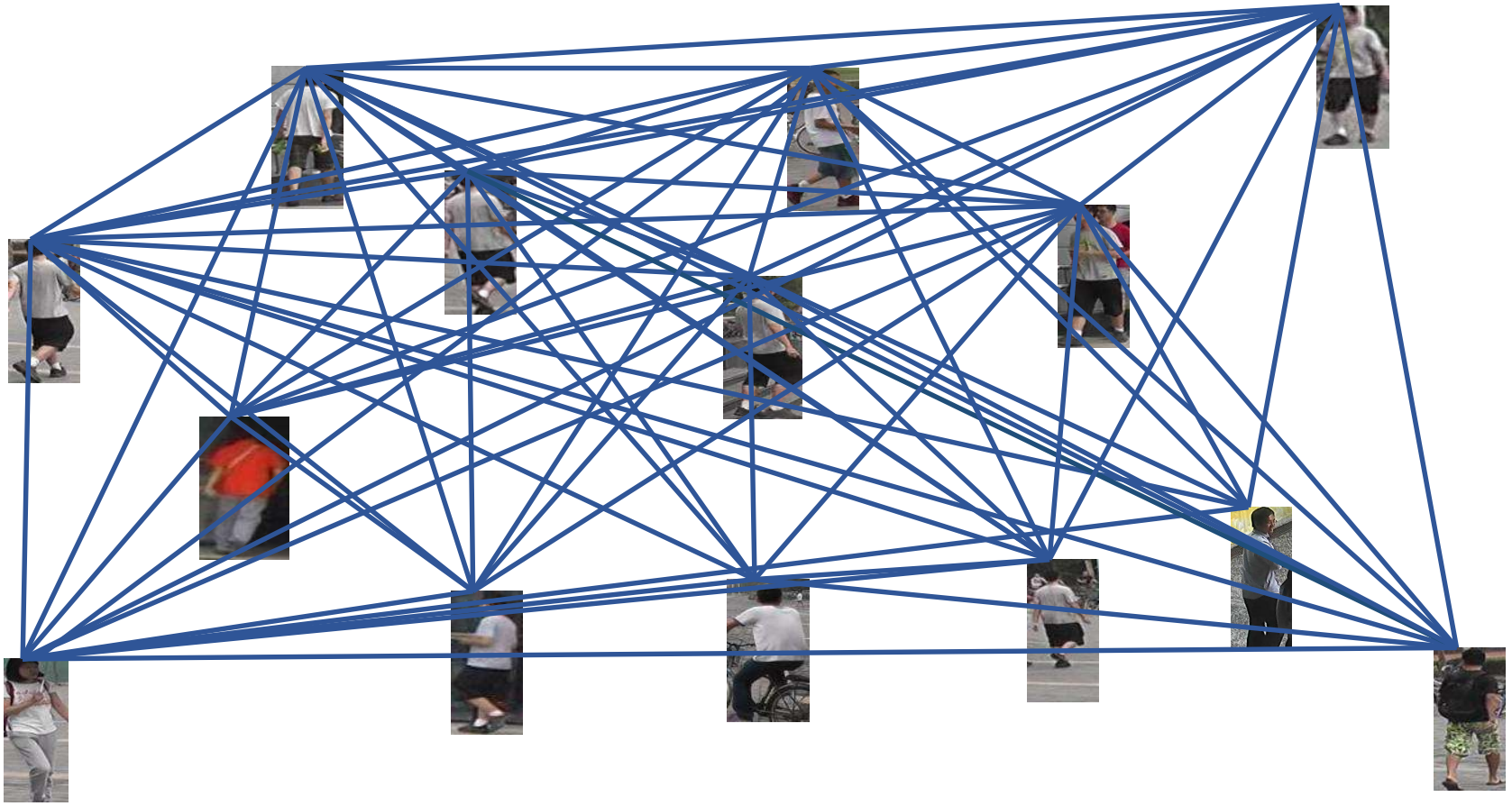









Diffusion Based Network



DCDS Based Network



Results

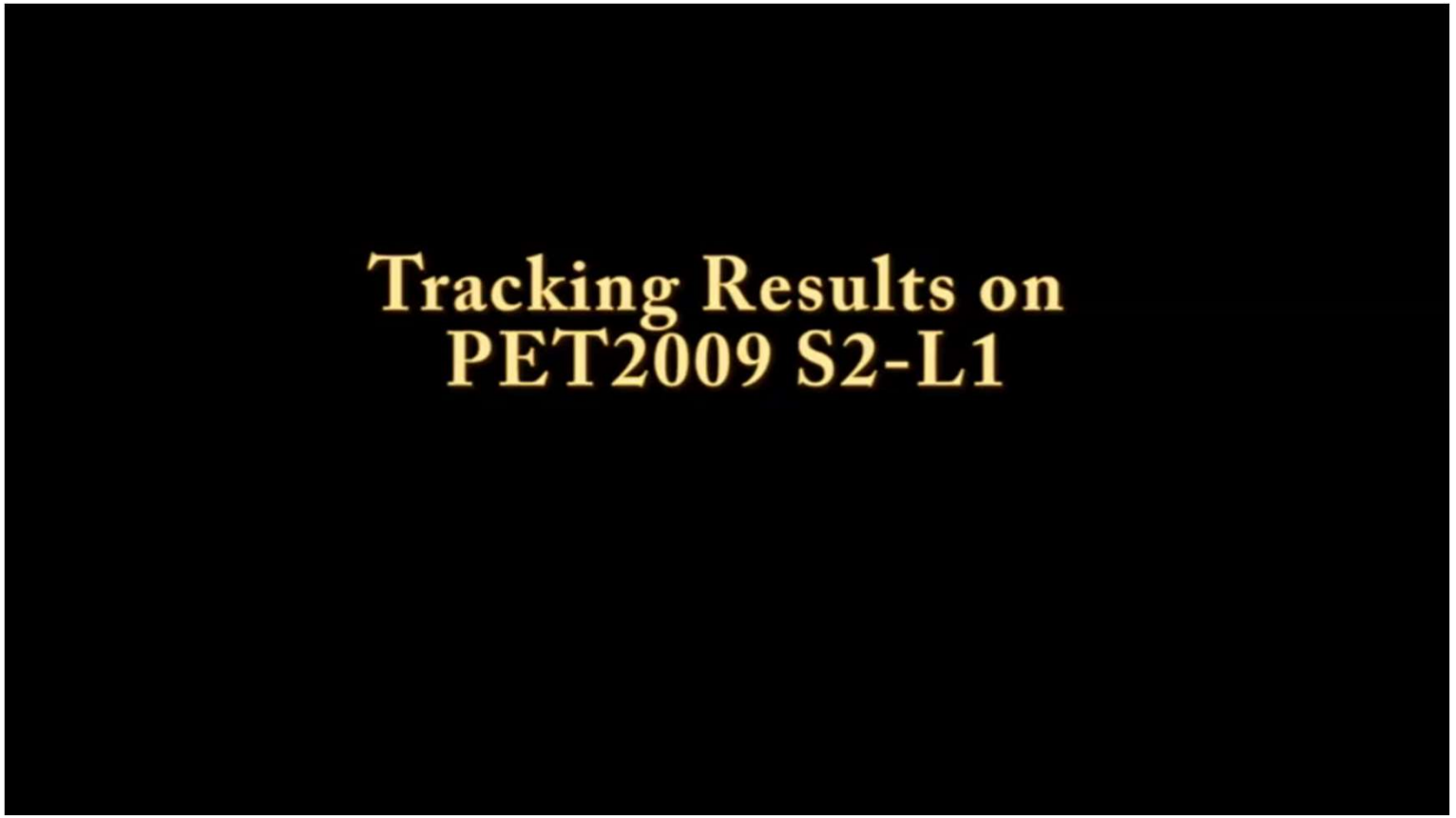
On CUHK03 Dataset

Methods	rank-1	rank-5
SGGNN [21] ECCV18	95.3	99.1
DKPM [22] CVPR18	91.1	98.3
DGSRW [20] CVPR18	94.9	98.7
GCSL [7] CVPR18	90.2	98.5
MLFN [6] CVPR18	89.2	-
CPC [28] CVPR18	88.1	-
PA [23] ECCV18	88.0	97.6
HSP [14] CVPR18	94.3	99.0
Ours	**95.8**	**99.1**

Deep affinity network for multiple object tracking

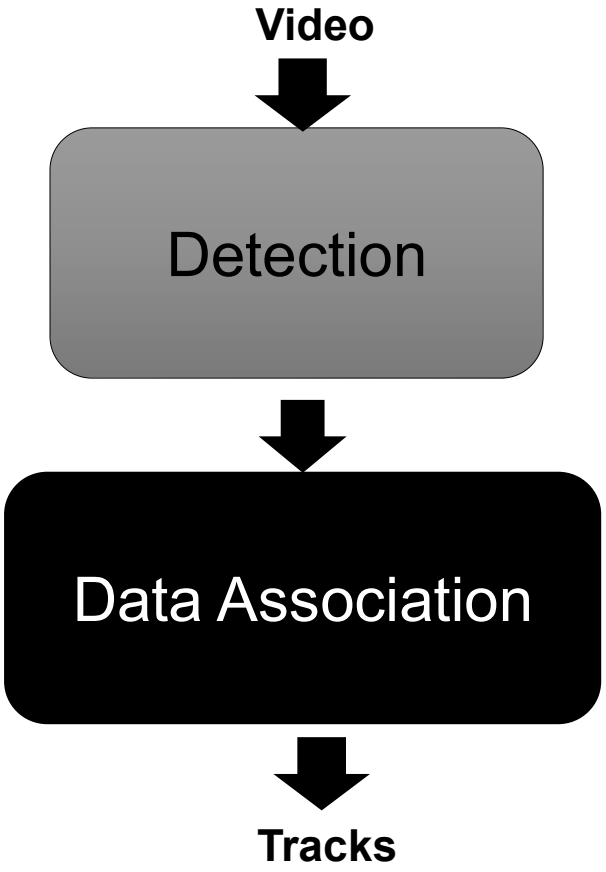
Sun, S., Akhtar, N., Song, H., Mian, A. and Shah, M.
IEEE PAMI 2018



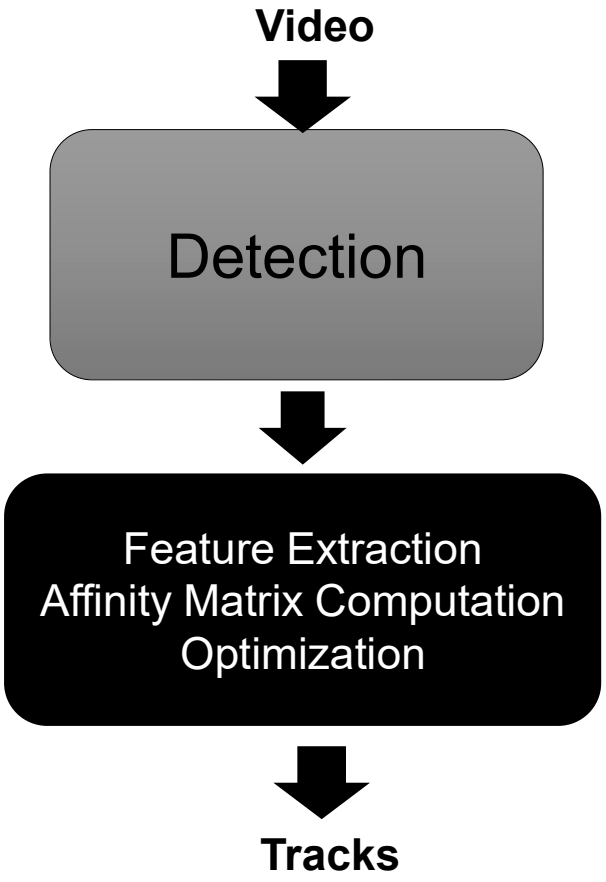

Tracking Results on
PET2009 S2-L1



Object Tracking

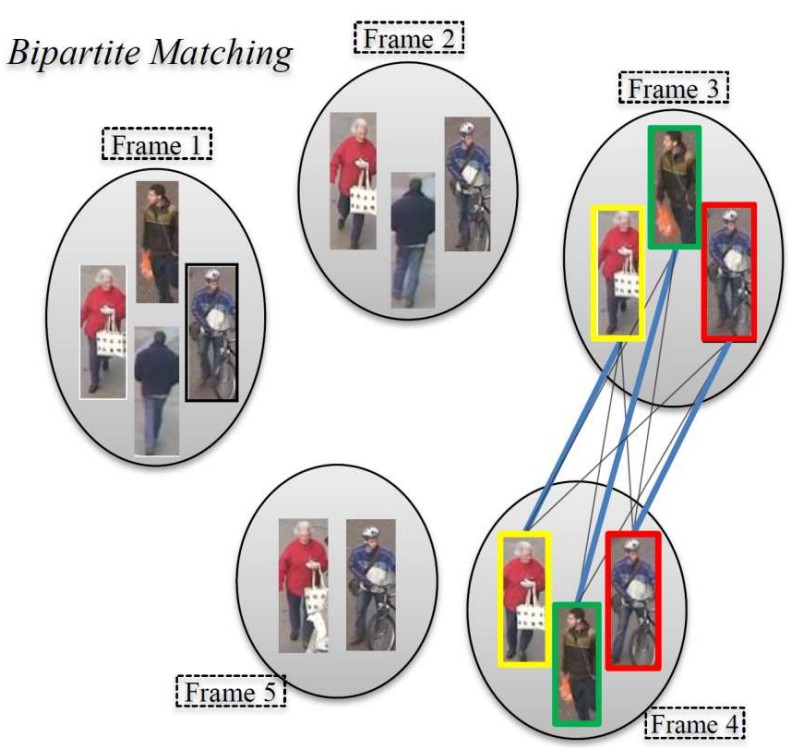


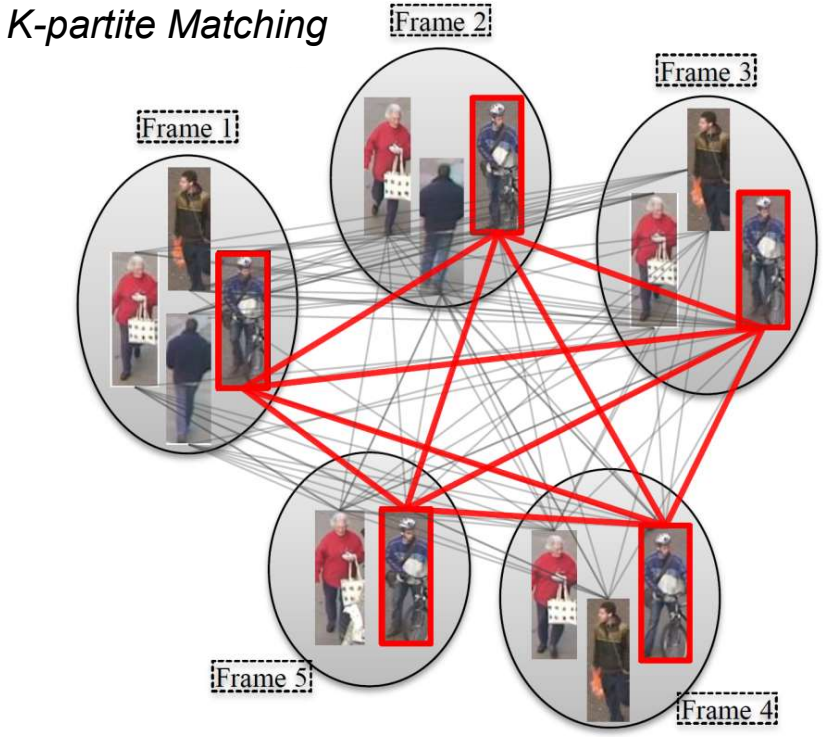
Object Tracking





Data Association

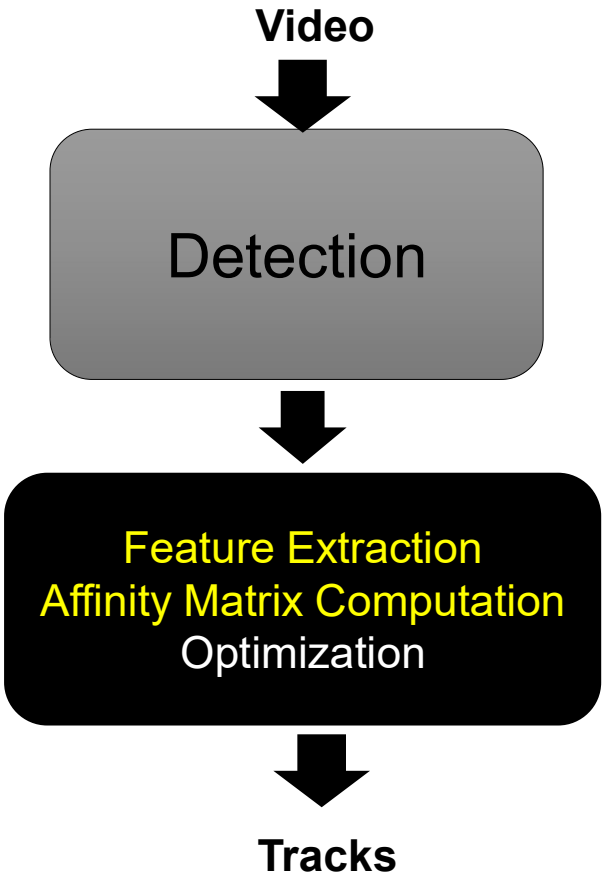




Can we End-to-End learn Affinity Matrix?



Object Tracking

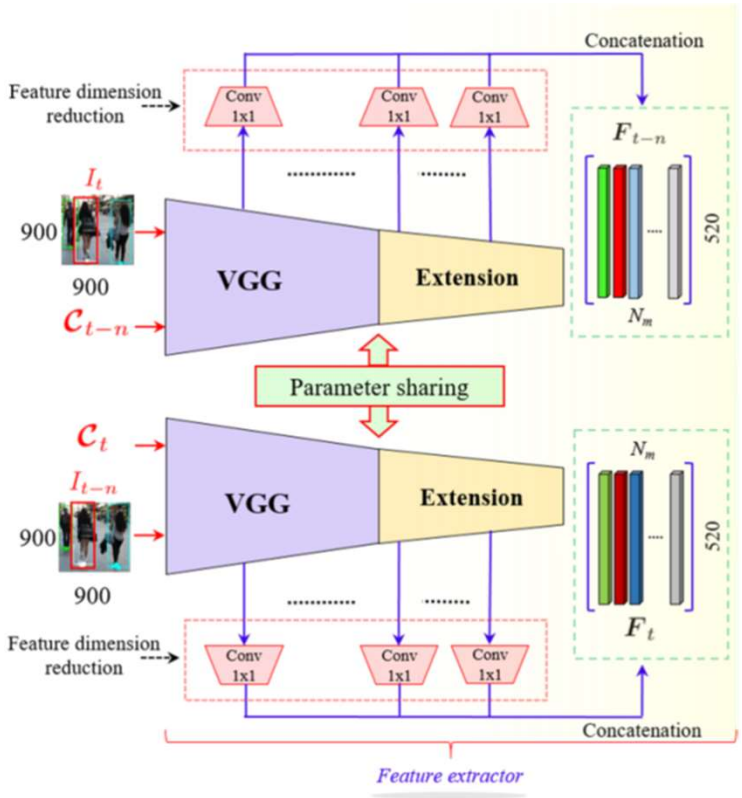


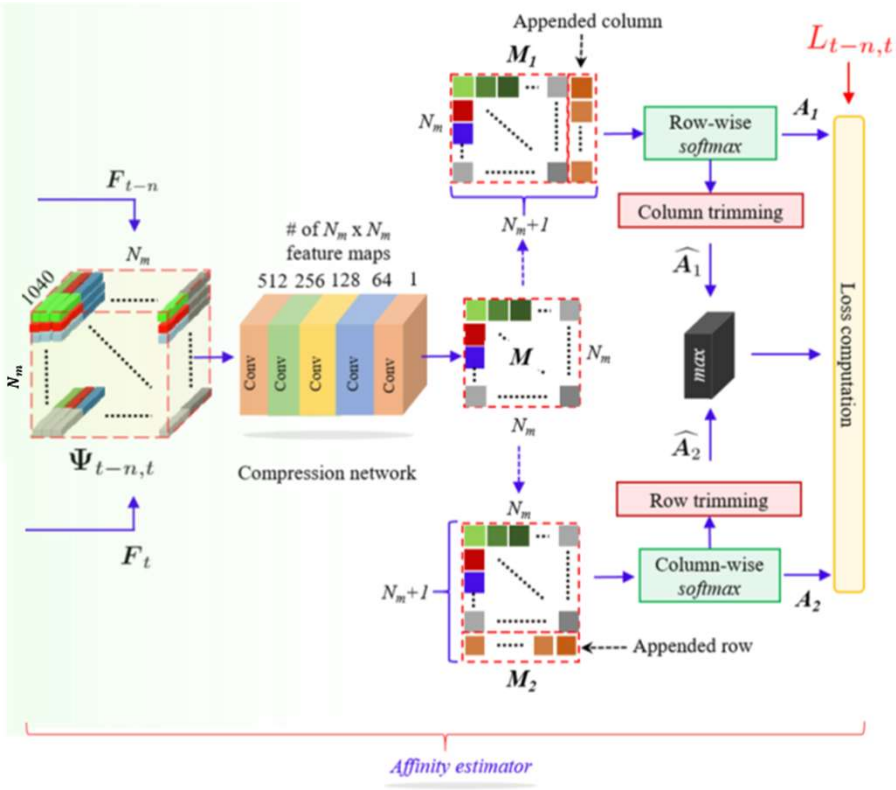
Deep Affinity Network (**DAN**)

- Learns compact features of objects at several levels of abstraction
- Performs exhaustive pairing permutations of the features
- Accounts for multiple objects appearing and disappearing.



DAN Architecture Simplified

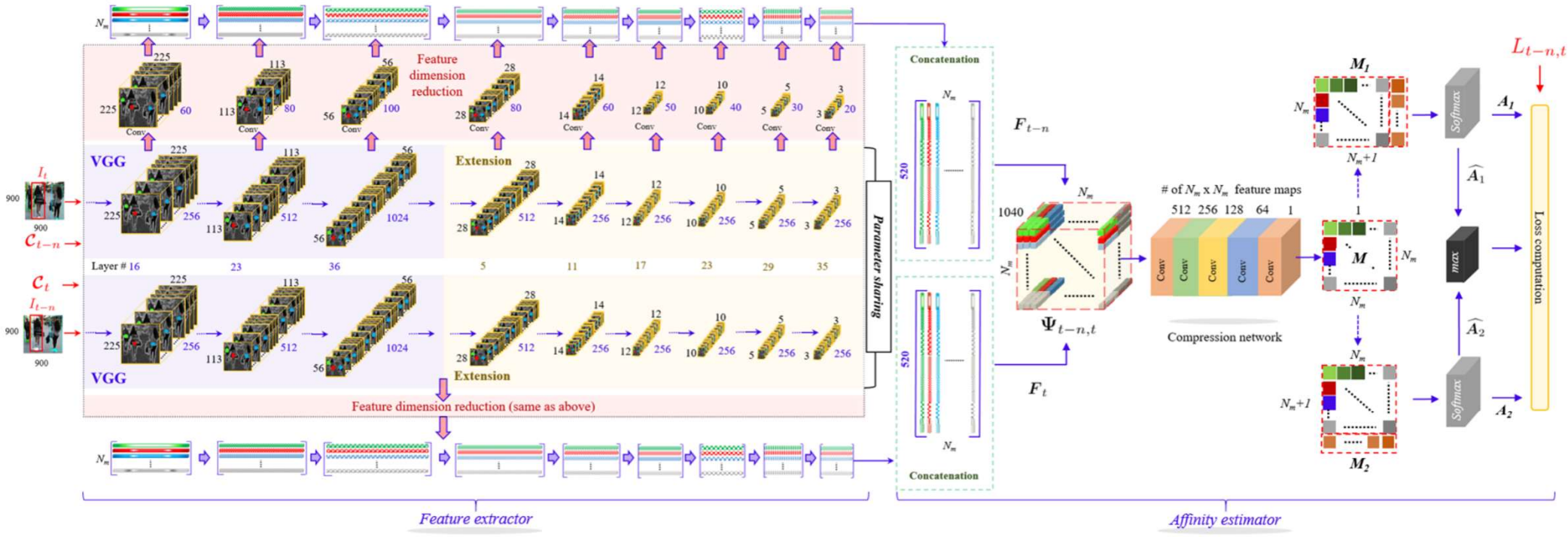




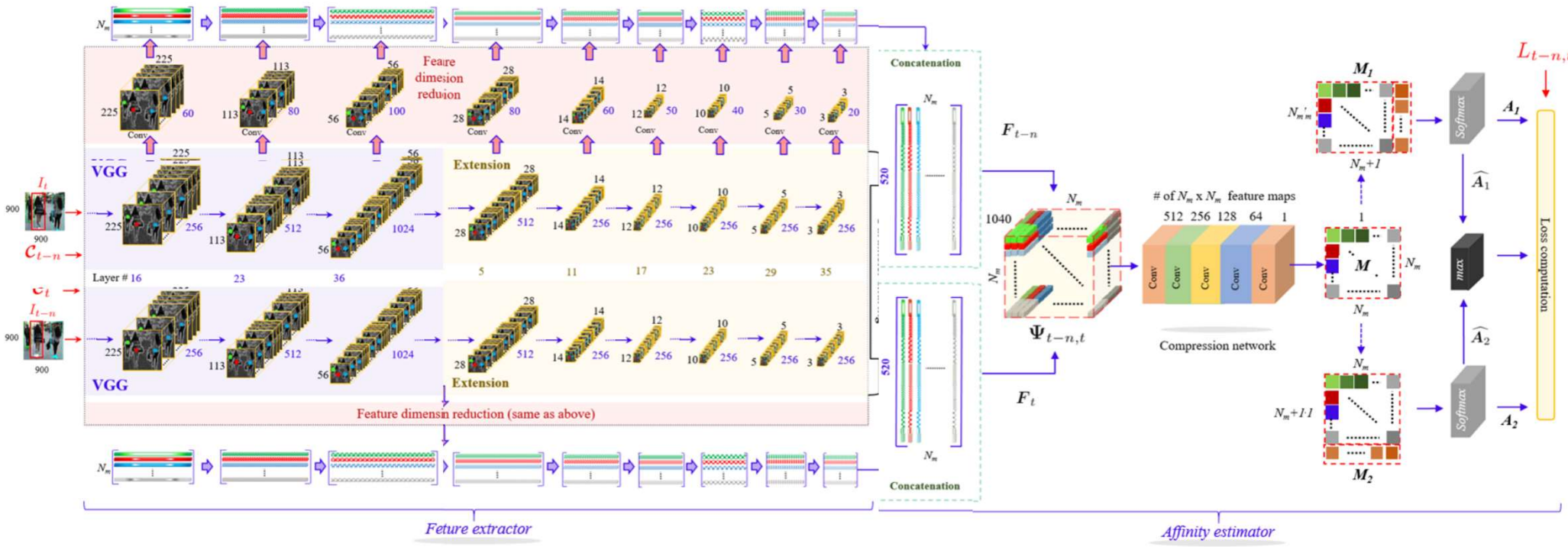
Pairs of video frames and object centers are passed through two shared-weights streams of convolution layers

Affinity Estimator encodes affinities between the objects using their extracted features

DAN Architecture



DAN Architecture



Network Losses

1) Forward-direction loss encourages correct identity association

$$\mathcal{L}_f(\boldsymbol{L}_1, \boldsymbol{A}_1) = \frac{\sum_{\text{coeff}} (\boldsymbol{L}_1 \odot (-\log \boldsymbol{A}_1))}{\sum_{\text{coeff}} (\boldsymbol{L}_1)}$$

2) Backward-direction loss ensures correct associations

$$\mathcal{L}_b(\boldsymbol{L}_2, \boldsymbol{A}_2) = \frac{\sum_{\text{coeff}} (\boldsymbol{L}_2 \odot (-\log \boldsymbol{A}_2))}{\sum_{\text{coeff}} (\boldsymbol{L}_2)}$$

3) Consistency loss for rebuffing any inconsistency

$$\mathcal{L}_c(\widehat{\boldsymbol{A}_1}, \widehat{\boldsymbol{A}_2}) = ||\widehat{\boldsymbol{A}_1} - \widehat{\boldsymbol{A}_2}||_1$$

4) Assemble loss that suppresses non-maximum

$$\mathcal{L}_a(\boldsymbol{L}_3, \widehat{\boldsymbol{A}_1}, \widehat{\boldsymbol{A}_2}) = \frac{\sum_{\text{coeff}} \left(\boldsymbol{L}_3 \odot \left(-\log(\max(\widehat{\boldsymbol{A}_1}, \widehat{\boldsymbol{A}_2}))\right)\right)}{\sum_{\text{coeff}} (\boldsymbol{L}_3)}$$

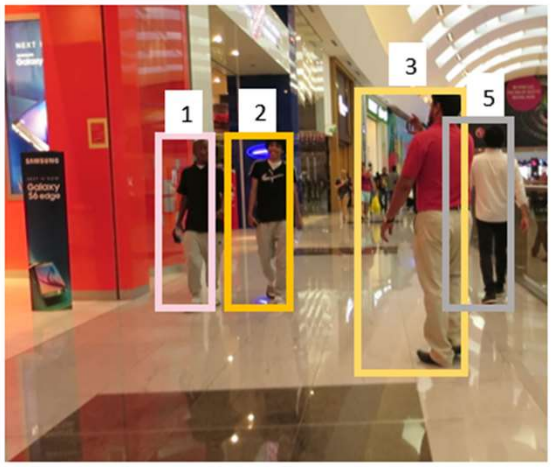


Data Association Matrix



(a) Frame 1





(b) Frame 30

Added to make N_m=5 (max # of tracked objects)

$L'_{1,30} =$

	5	3	2	1	x
3	0	1	0	0	0
1	0	0	0	1	0
2	0	0	1	0	0
x	0	0	0	0	0
4	0	0	0	0	0

(c) Matrix with dummy boxes

entered the frame

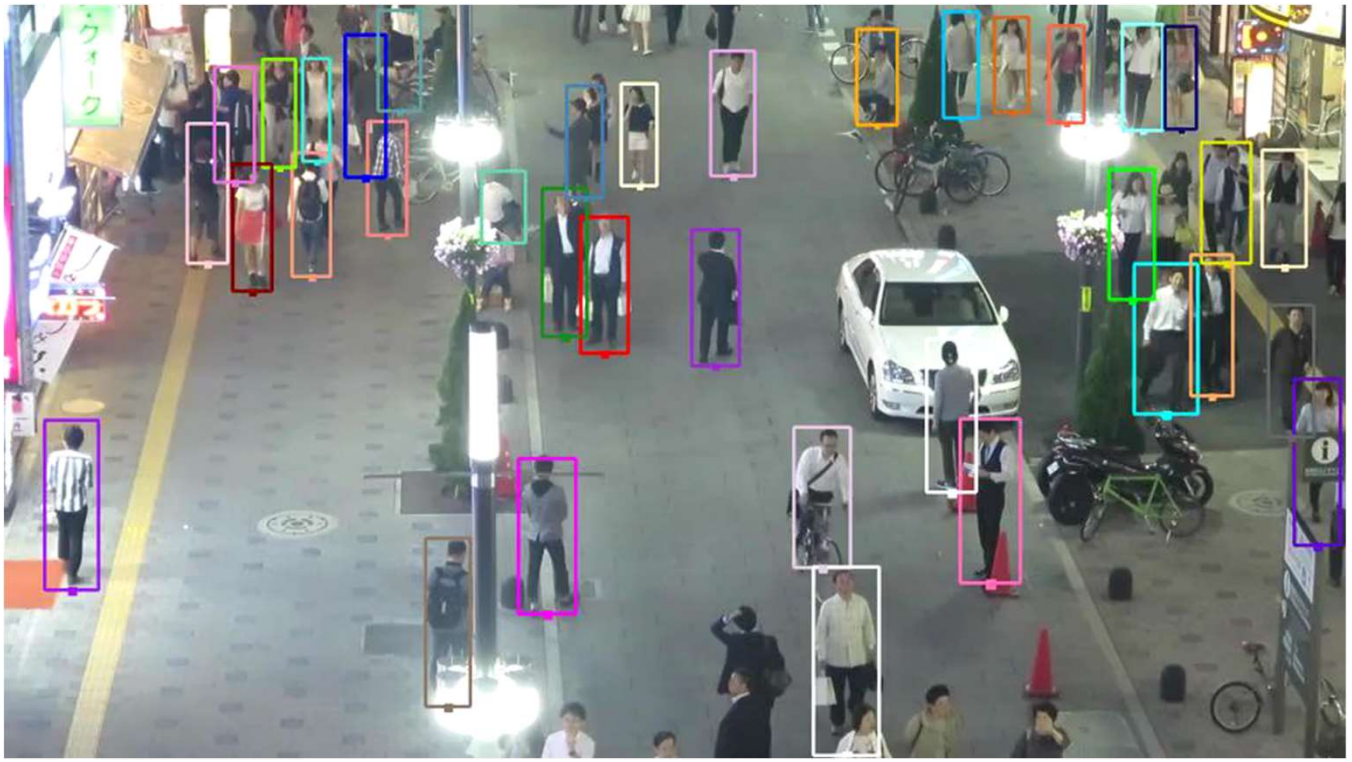
$L_{1,30} =$

	5	3	2	1	x	⊗
3	0	1	0	0	0	0
1	0	0	0	1	0	0
2	0	0	1	0	0	0
x	0	0	0	0	0	0
4	0	0	0	0	0	1
⊗	1	0	0	0	0	0

left the frame

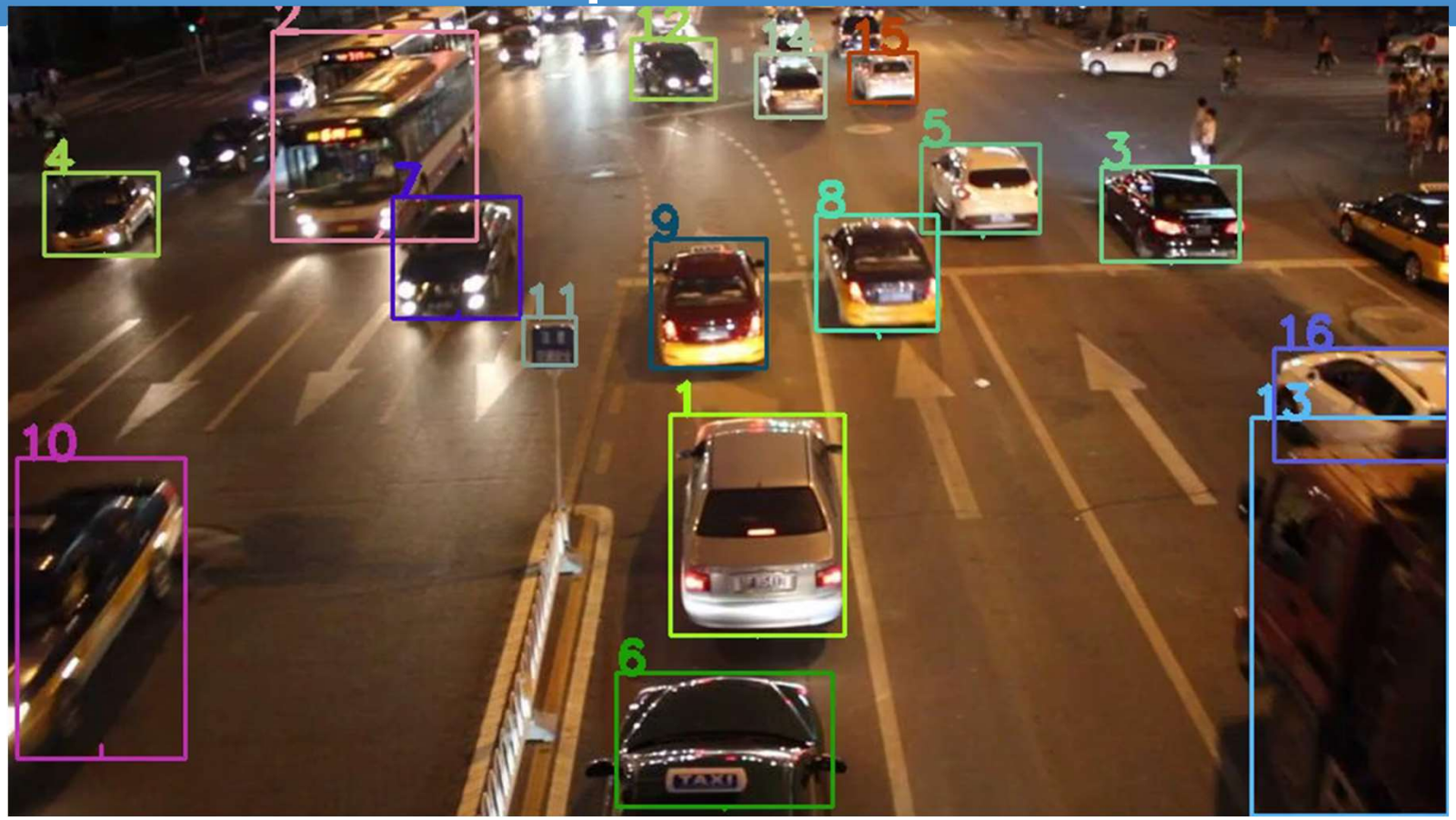
(d) Matrix with Un-Identified (UI) targets

MOT: Sample Video Results





MOT: Sample Video Results



Summary

- End-to-End learning of Affinity Matrix
- Learns compact features of objects at several levels of abstraction
- Performs exhaustive pairing permutations of the features
- Accounts for multiple objects appearing and disappearing.



Deep affinity network for multiple object tracking

Sun, S., Akhtar, N., Song, H., Mian, A. and Shah, M.
IEEE PAMI



2-D Tracking to 3-D Tracking

- Most tracking methods are for 2D tracking
- 3D MOT has applications in self-driving cars and robot navigation
- Lidar captures 3D information in the form of point-clouds
- 3D point-cloud preserves structural information and avoid perspective distortion
- Lidar is becoming widely available in automobiles, iPhone, iPAD



PC-DAN: Point Cloud based Deep Affinity Network for 3D Multi-Object Tracking

Aakash Kumar*, Jyoti Kini*, Ajmal Mian†, Mubarak Shah*
*Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA
†University of Western Australia, Crawley, WA, Australia

Challenge Winner in 3D tracking

2nd Workshop on Visual Perception for Navigation in Human Environments at CVPR 2021

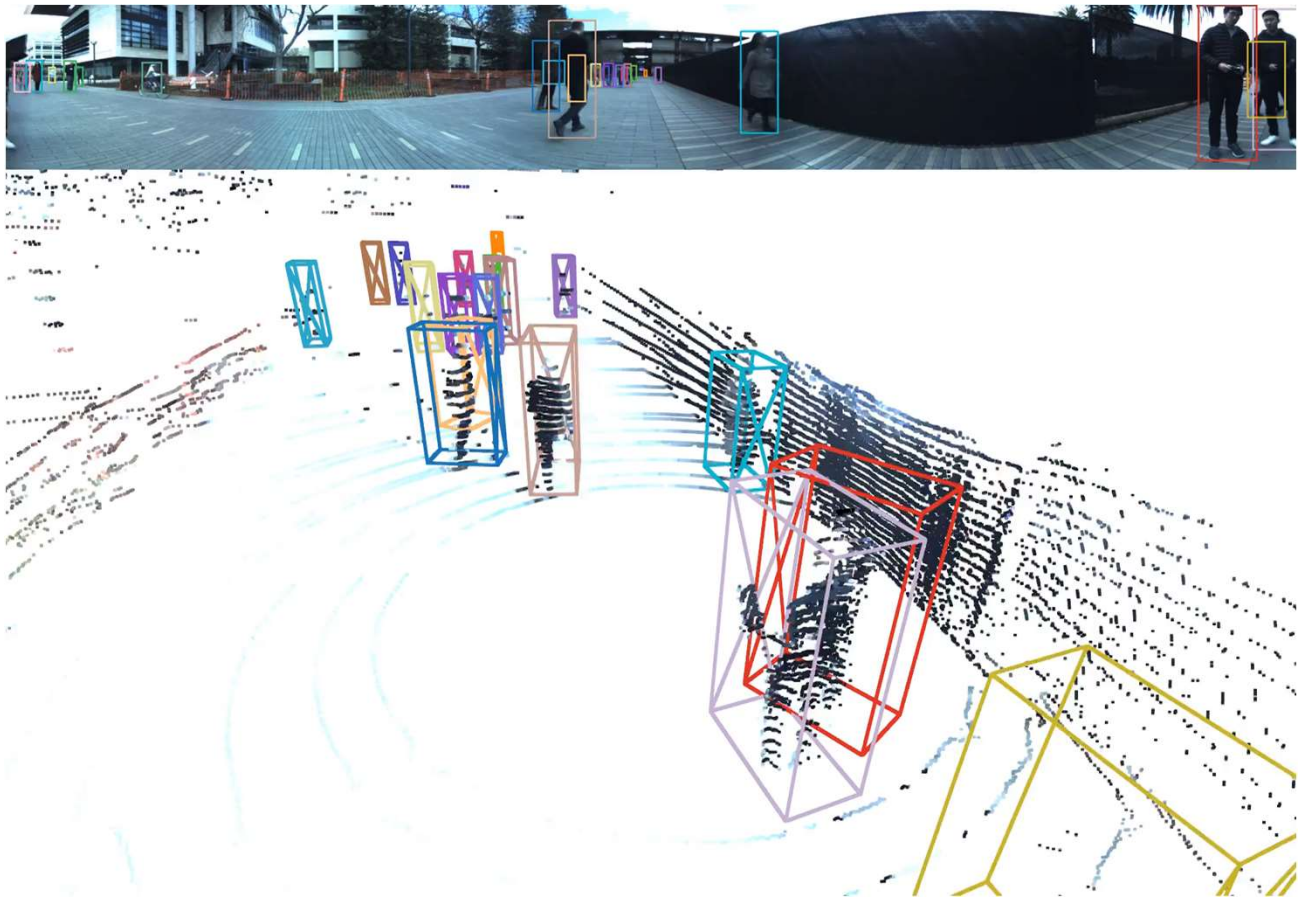
https://www.youtube.com/watch?v=2v7vlsY2RM8&list=PLd3hlSJsX_IlnEEKNg4bSUOIuVyHEhosy&index=4



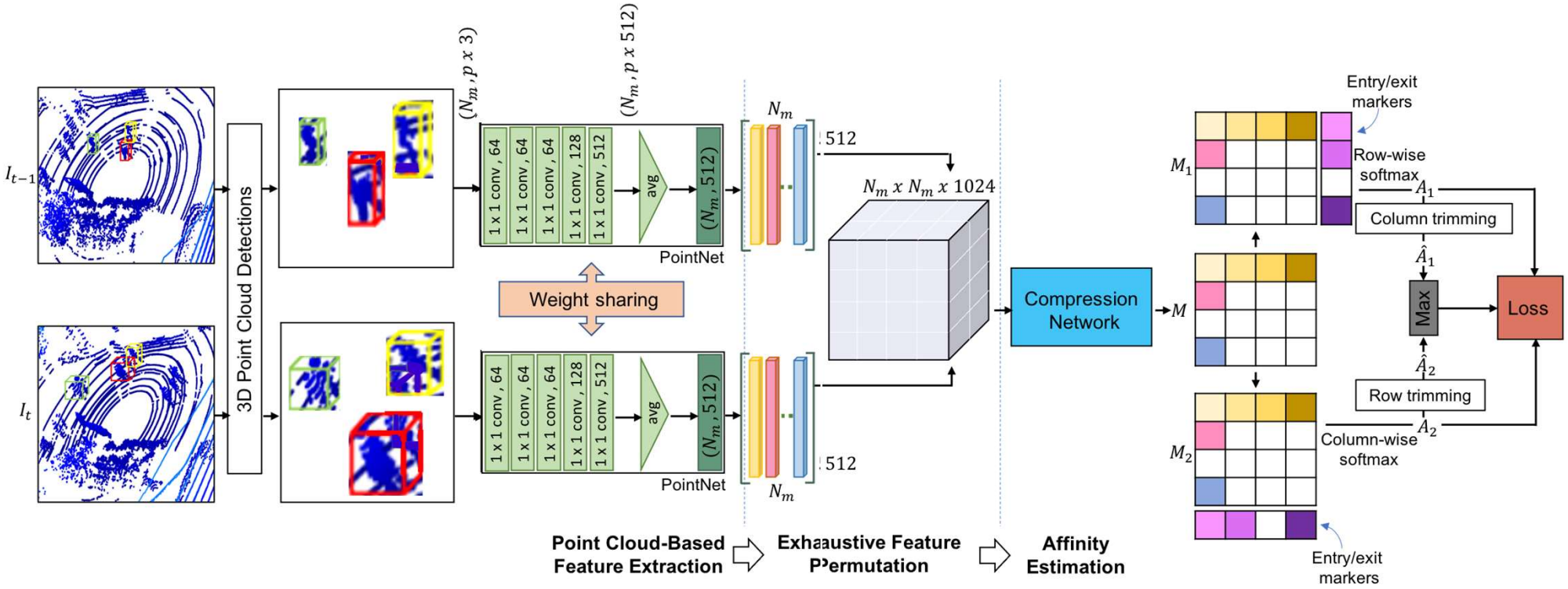
Sample Indoor Data: JRDB



Sample Outdoor Data: JRDB

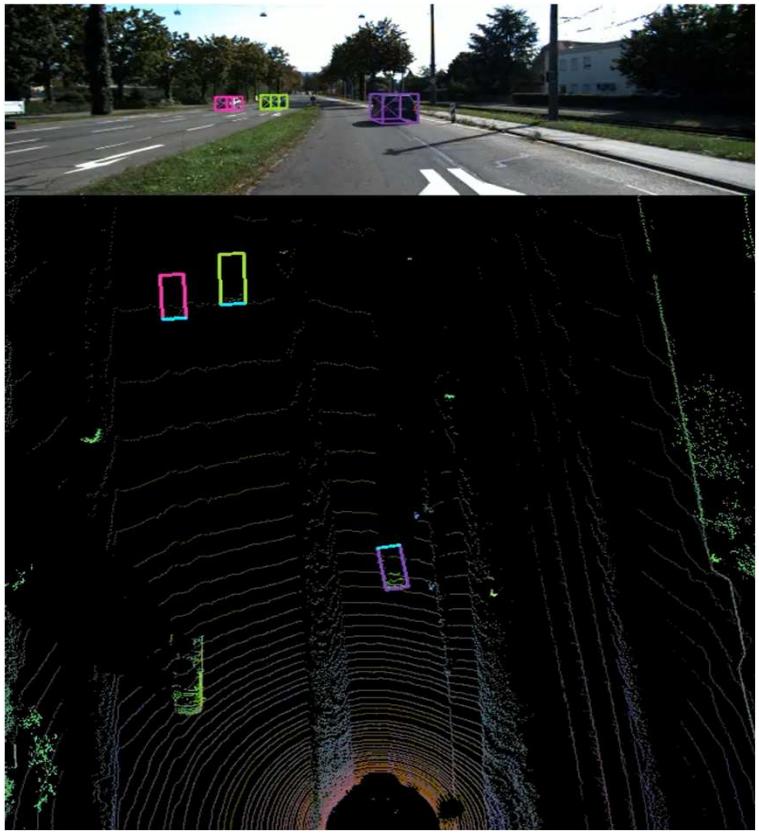


Methodology PC-DAN

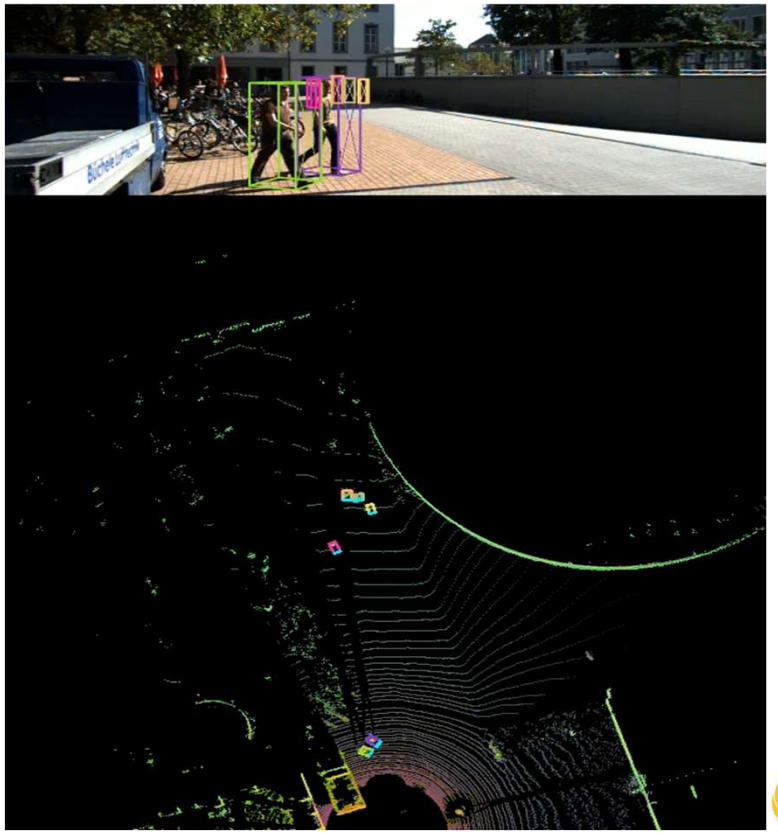


Qualitative Results

Cars



Pedestrians



PC-DAN: Point Cloud based Deep Affinity Network for 3D Multi-Object Tracking

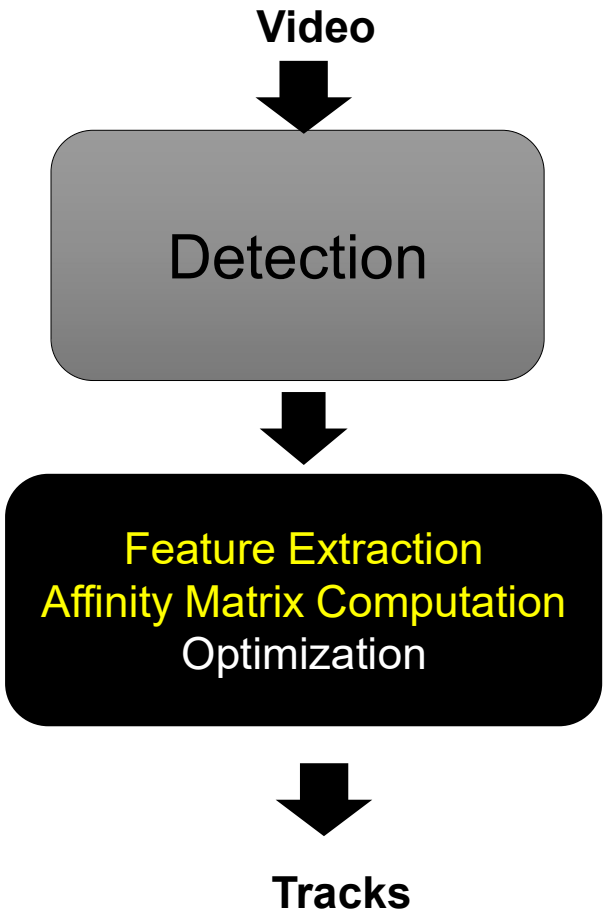
Aakash Kumar*, Jyoti Kini*, Ajmal Mian†, Mubarak Shah*
*Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA
†University of Western Australia, Crawley, WA, Australia

Challenge Winner in 3D tracking

2nd Workshop on Visual Perception for Navigation in Human Environments at CVPR 2021

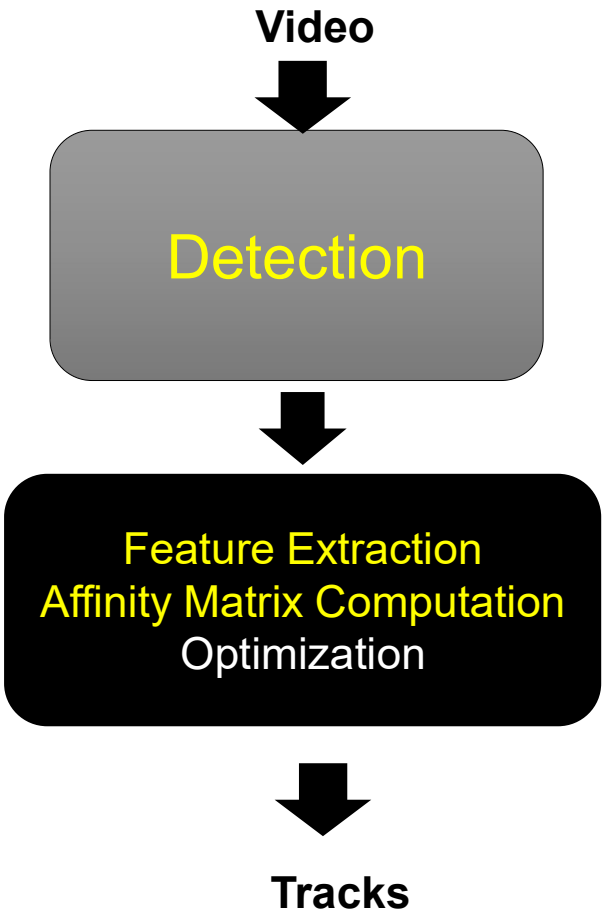


Object Tracking





Object Tracking







Simultaneous Detection and Tracking with Motion Modelling for Multiple Object Tracking

Naveed Akhtar, XiangYu Song, HuanSheng Song,

Ajmal Mian, Mubarak Shah

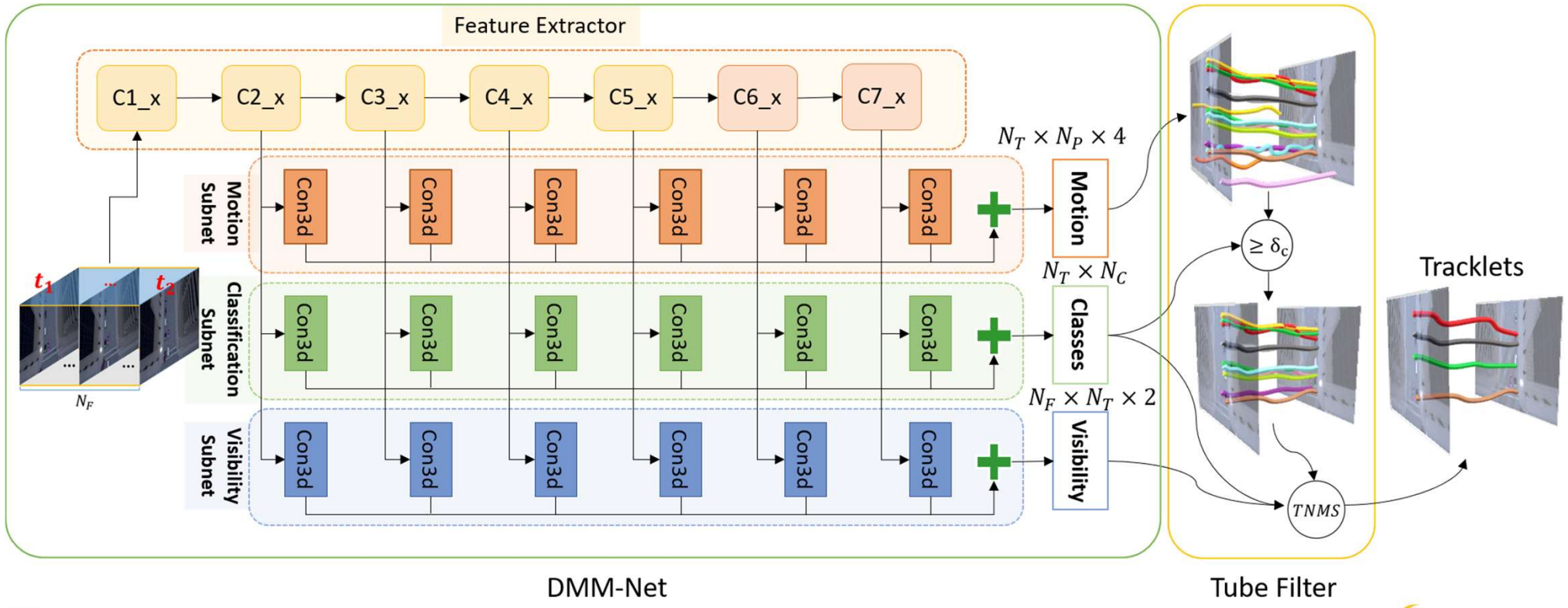
ECCV 2020





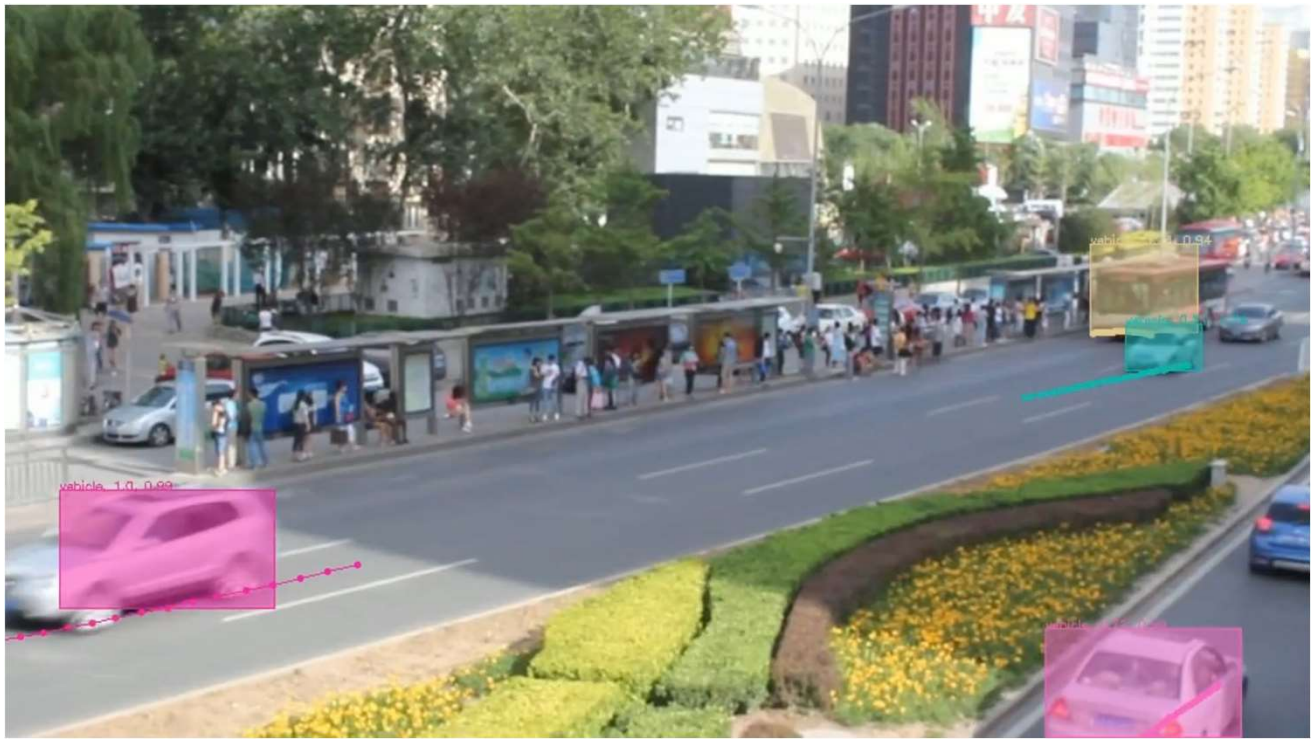


Deep Motion Modeling Network



Experiment

DMMN result in UA-DETRAC



Experiment

Here are the demo videos of DMMN result on Omni-MOT

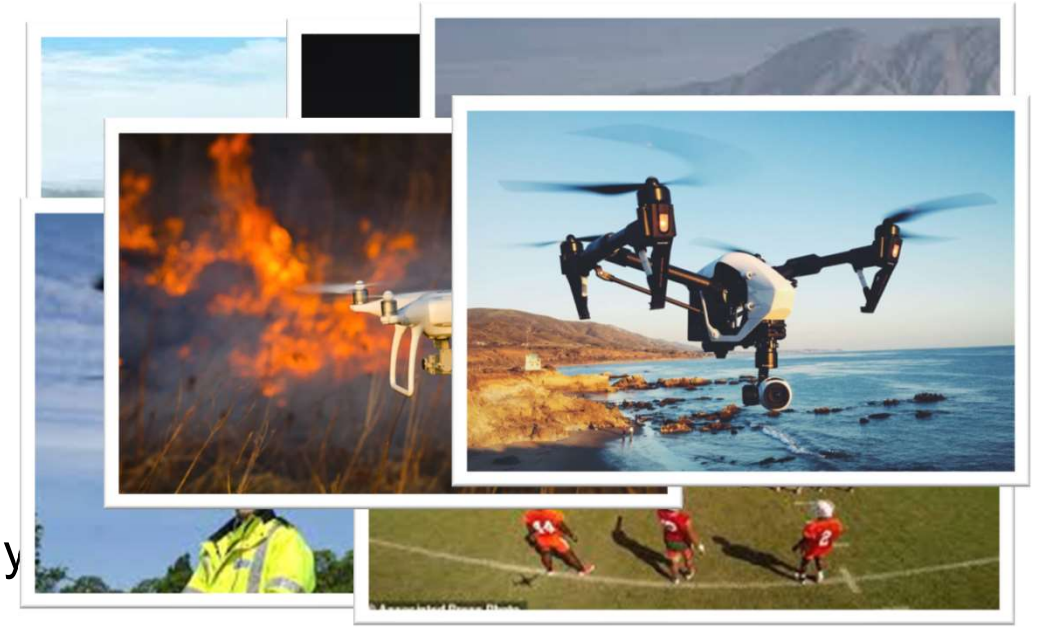




Drone Videos

Drones are ubiquitous and actively being used in:

- Agriculture
- Entertainment
- Military
- Surveillance
- Sports
- Wildfire fighting
- Cinematography





Cluster-Net: Detecting Small Objects in Large Scenes by Exploiting Spatiotemporal Information

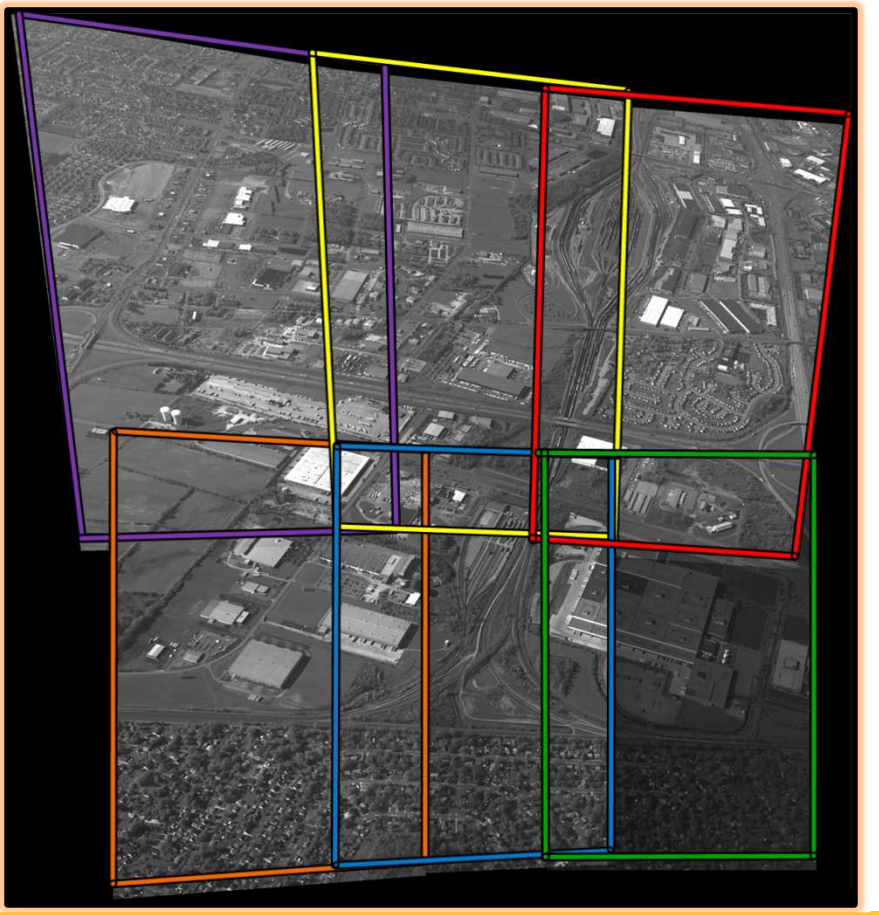
Rodney LaLonde, Dong Zhang and Mubarak Shah
CVPR-2018
https://www.youtube.com/watch?v=G6tobm5YZPg&list=PLd3hlSJsX_Ilqyw513wCDI3zS35YYGsBu&index=1



Wide Area Video Analytics



Six Camera mosaic





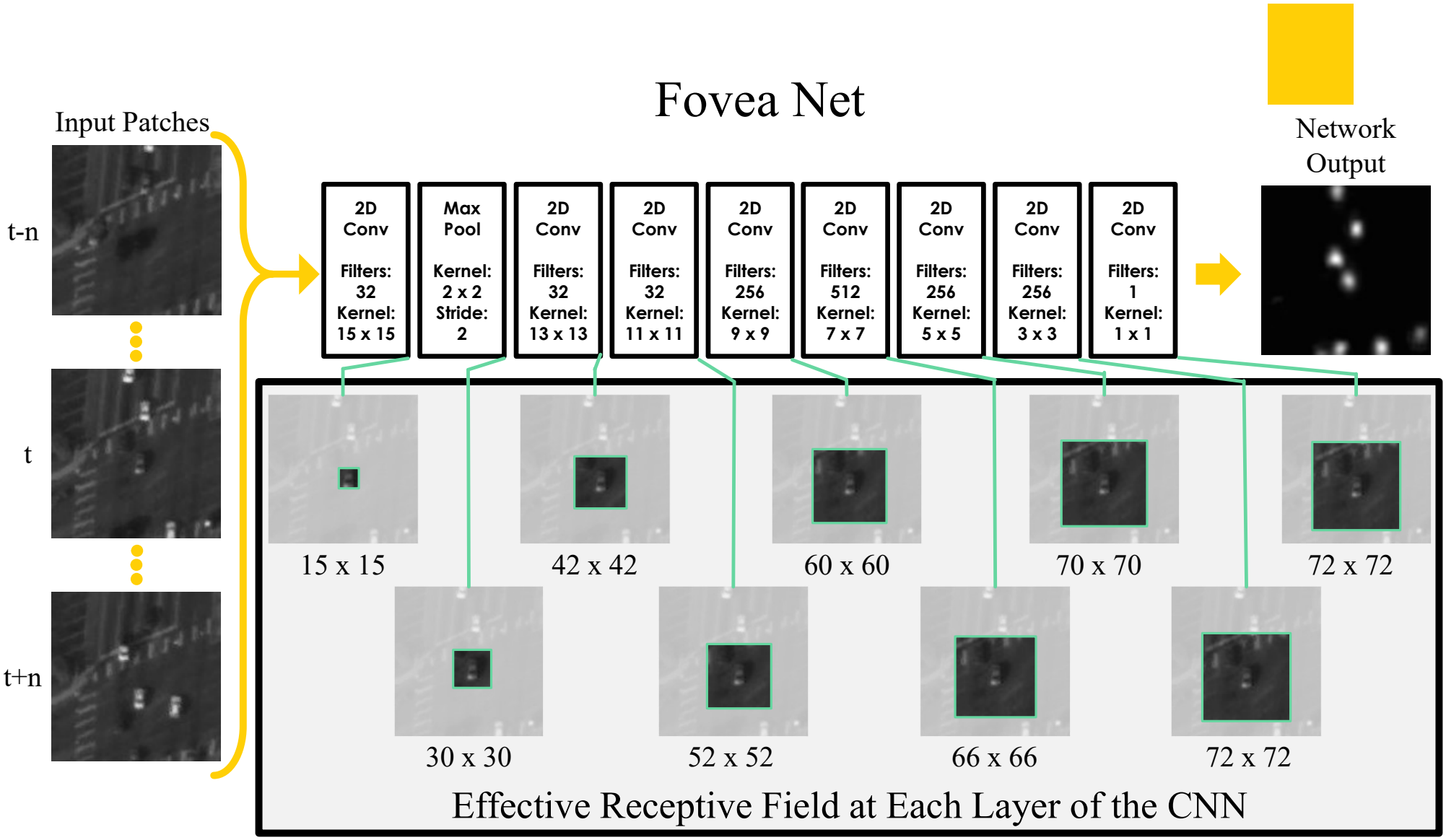
UCF
CENTER FOR RESEARCH IN
COMPUTER VISION

The Goal

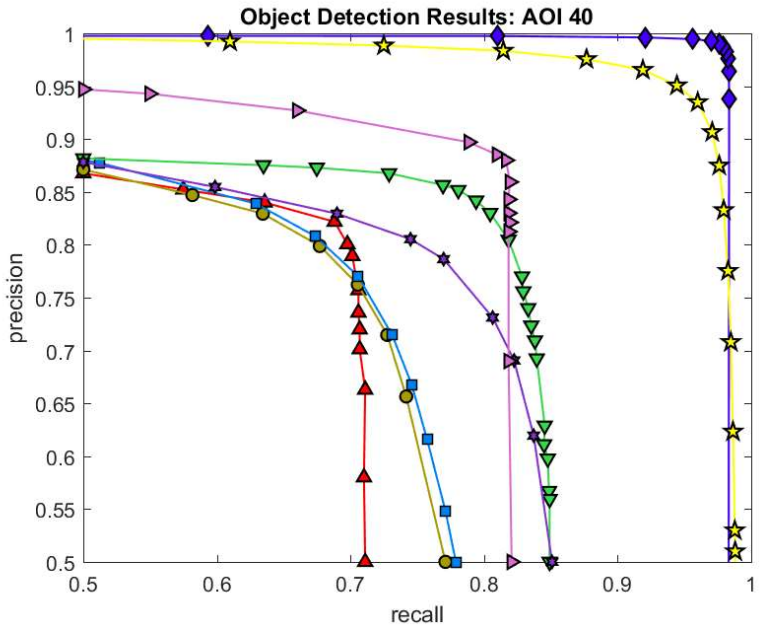
Multiple Vehicle Detection

Red dots are the ground truth (x,y) annotations.
(WPAFB 2009)




Fovea Net
Input Patches
t-n
t
t+n
2D Conv Filters: 32 Kernel: 15 x 15
Max Pool Kernel: 2 x 2 Stride: 2
2D Conv Filters: 32 Kernel: 13 x 13
2D Conv Filters: 32 Kernel: 11 x 11
2D Conv Filters: 256 Kernel: 9 x 9
2D Conv Filters: 512 Kernel: 7 x 7
2D Conv Filters: 256 Kernel: 5 x 5
2D Conv Filters: 256 Kernel: 3 x 3
2D Conv Filters: 1 Kernel: 1 x 1
Network Output
15 x 15
42 x 42
60 x 60
70 x 70
72 x 72
30 x 30
52 x 52
66 x 66
72 x 72
Effective Receptive Field at Each Layer of the CNN

Results AOI 40





Dogfight: Detecting Drones from Drones Videos

Muhammad Waseem Ashraf[1], Waqas Sultani[1], Mubarak Shah[2]

[1]Intelligent Machines Lab, Information Technology University, Pakistan
[2]Center for Research in Computer Vision, University of Central Florida

Accepted in CVPR 2021

http://im.itu.edu.pk/dogfight-detecting-drones-from-drones-videos/

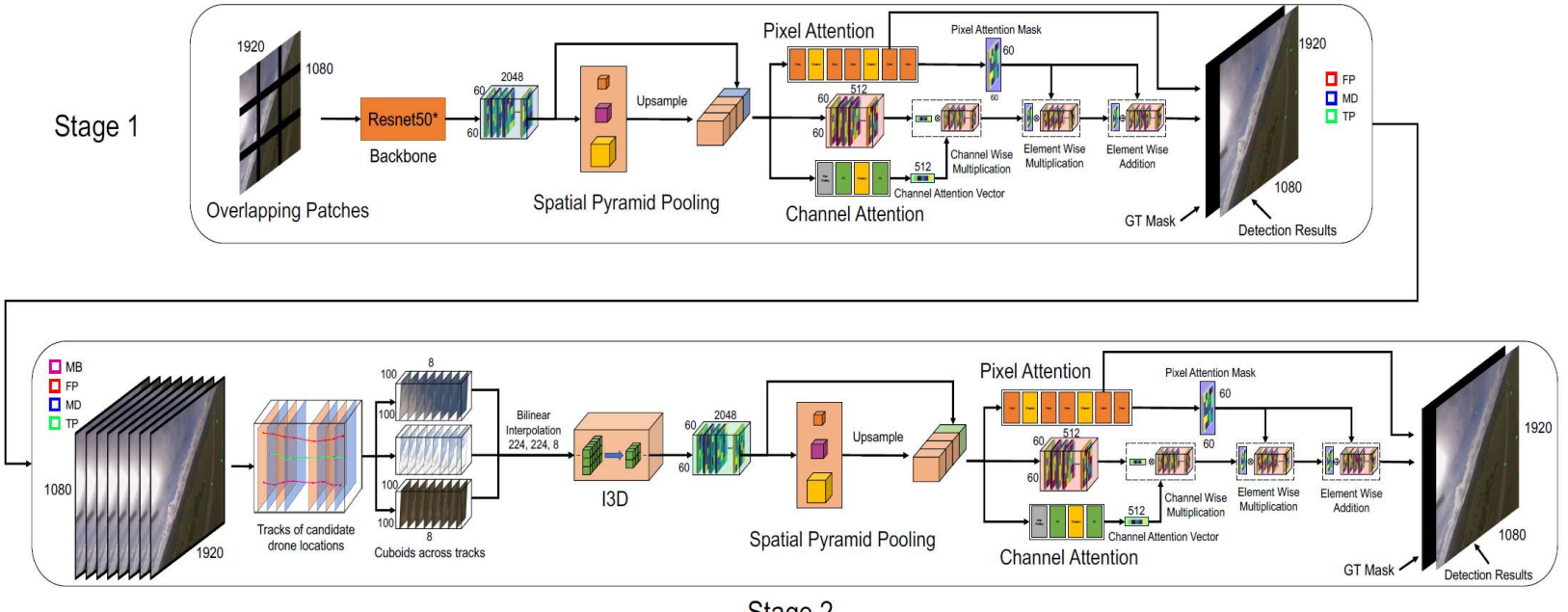








Proposed Approach








Ground Truth Proposed detections Bigger boxes are shown for better visualization



Ground Truth Proposed detections

Bigger boxes are shown for better visualization

Real-World Anomaly Detection in Surveillance videos

Waqas Sultani, Chen Chen, Mubarak Shah

CVPR-2018

https://www.youtube.com/watch?v=qS1iPiSW3RE&list=PLd3hlSJsX_Ilqyw513wCDI3zS35YYGsBu&index=2&t=5s



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Motivation

- Over 30 Millions cameras in US
- Over 4 Billions hours of videos per week
- Manual supervision is impossible
- Automatic Analysis is highly needed





Abuse



Arrest



Arson



Assault



Burglary



Stealing

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

Our Approach

- Learn anomalies by exploiting both normal and anomalous videos.
- Avoid annotating the anomalous clips in training videos.
- Learn anomaly through the deep multiple instance ranking
  - By leveraging weakly labeled training videos:
  - A video is normal or contains anomaly somewhere, but we do not know where.

Weakly labeled Crime Detection Framework

Ranking

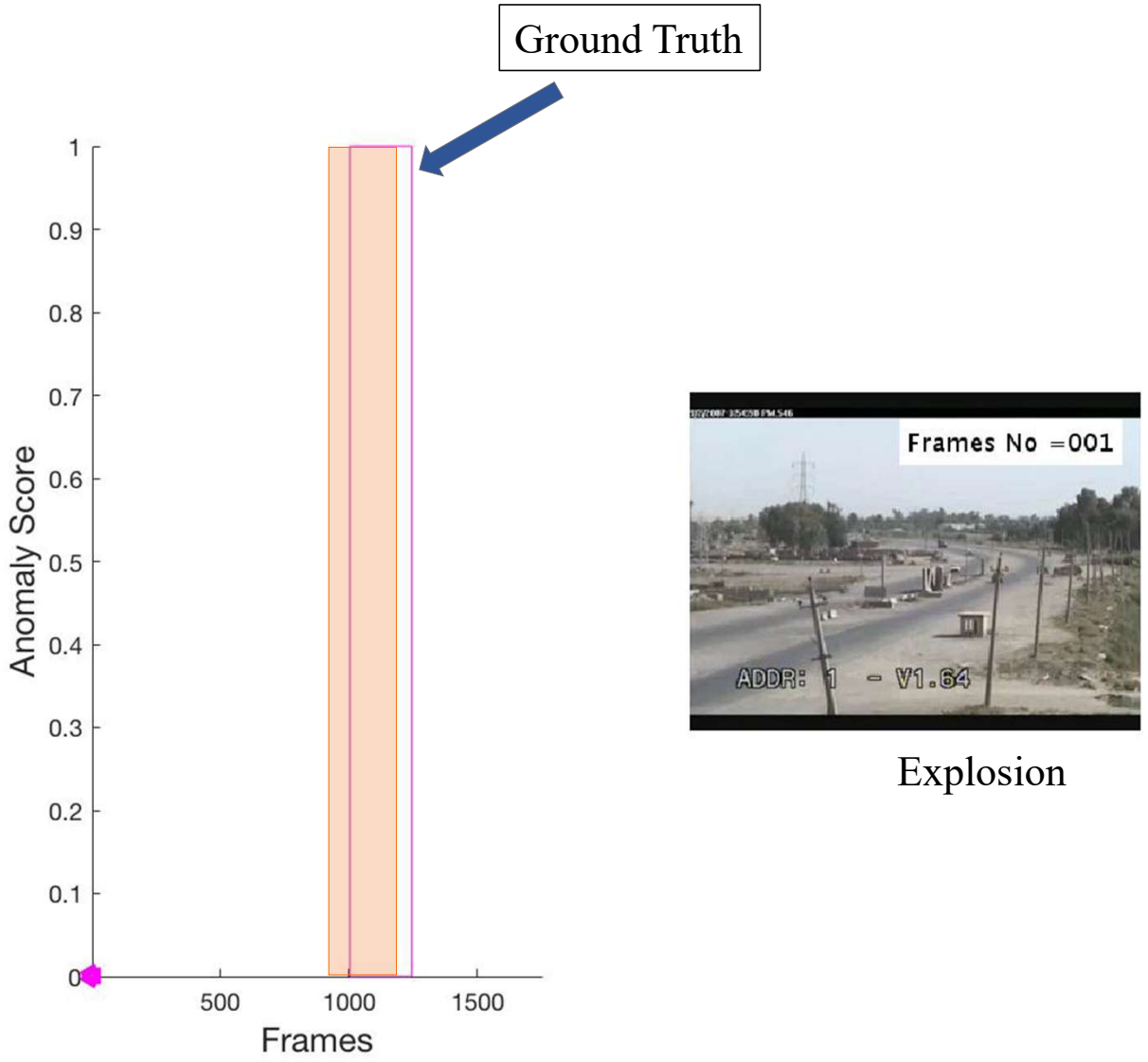
$$f(\mathcal{V}_a) > f(\mathcal{V}_n),$$

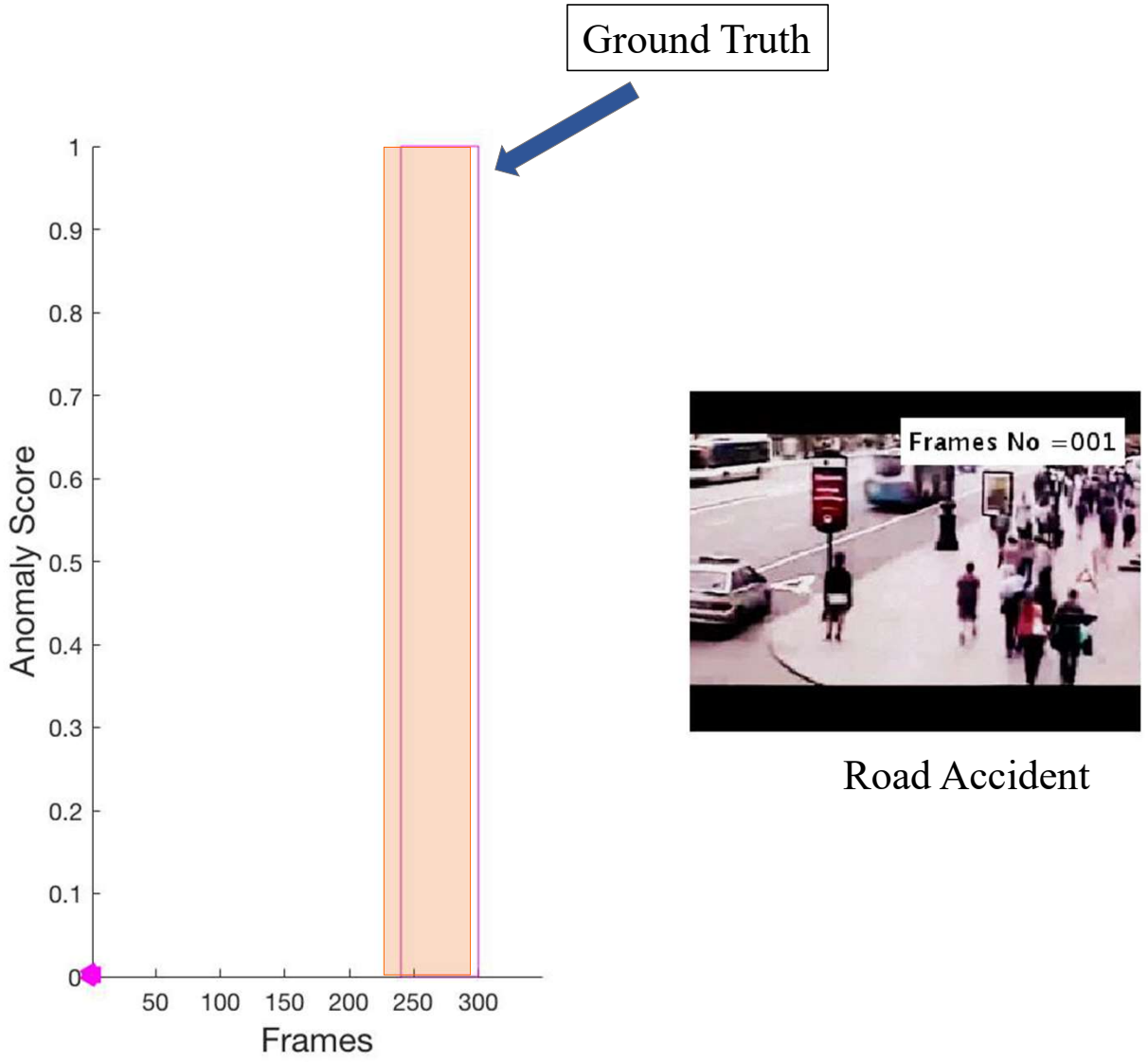
MIL Ranking

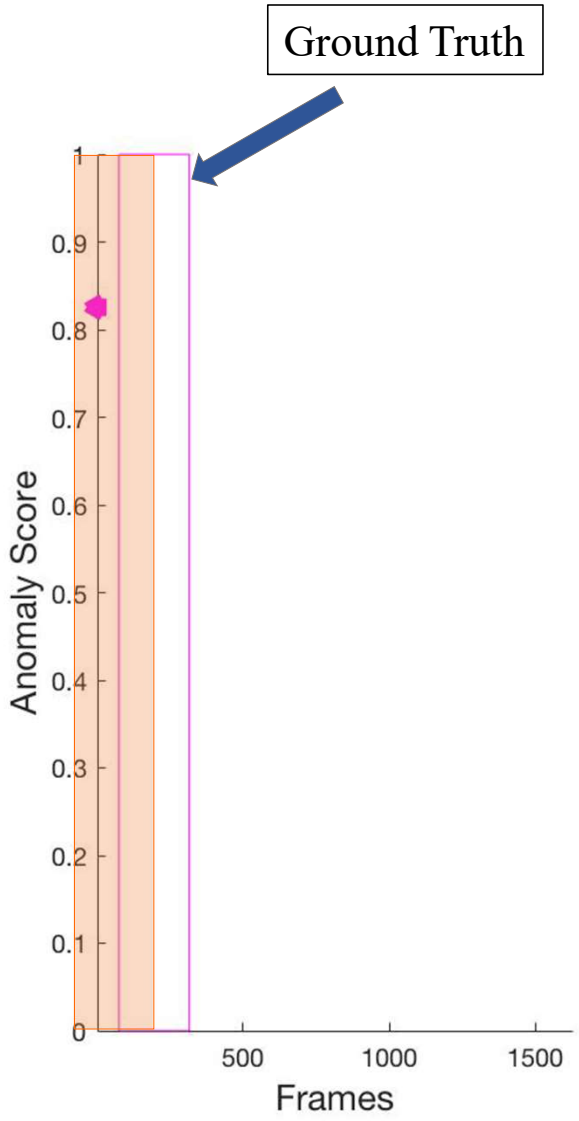
$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

Loss Function

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

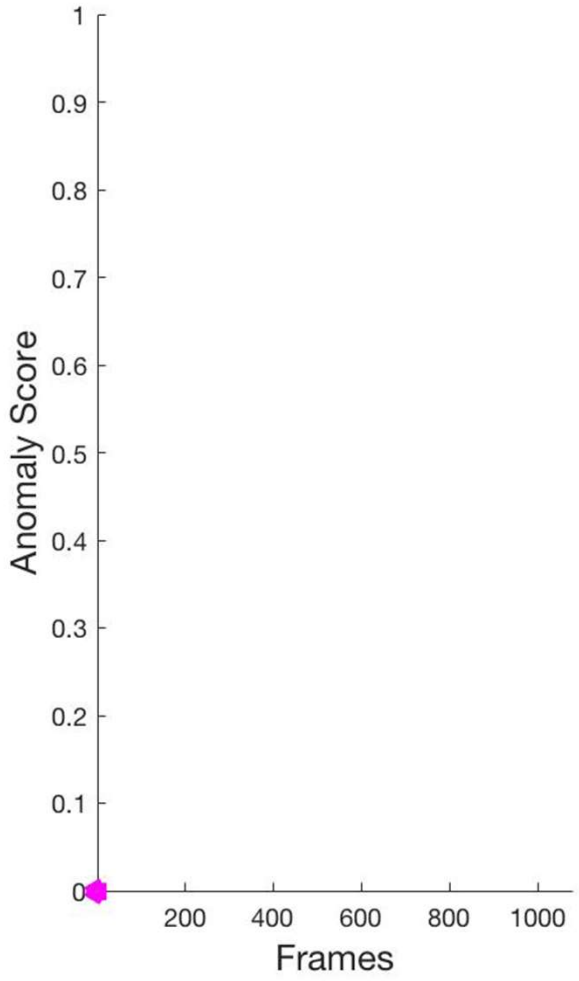

Ground Truth
Anomaly Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
500
1000
1500
Frames
Frames No =001
ADDR: 1 - V1.64
Explosion


Ground Truth
Anomaly Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300
Frames
Frames No =001
Road Accident


Ground Truth
Anomaly Score
Frames
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
500
1000
1500


Frames No =001

Shooting





Normal video

T-CNN for Action Detection in Videos

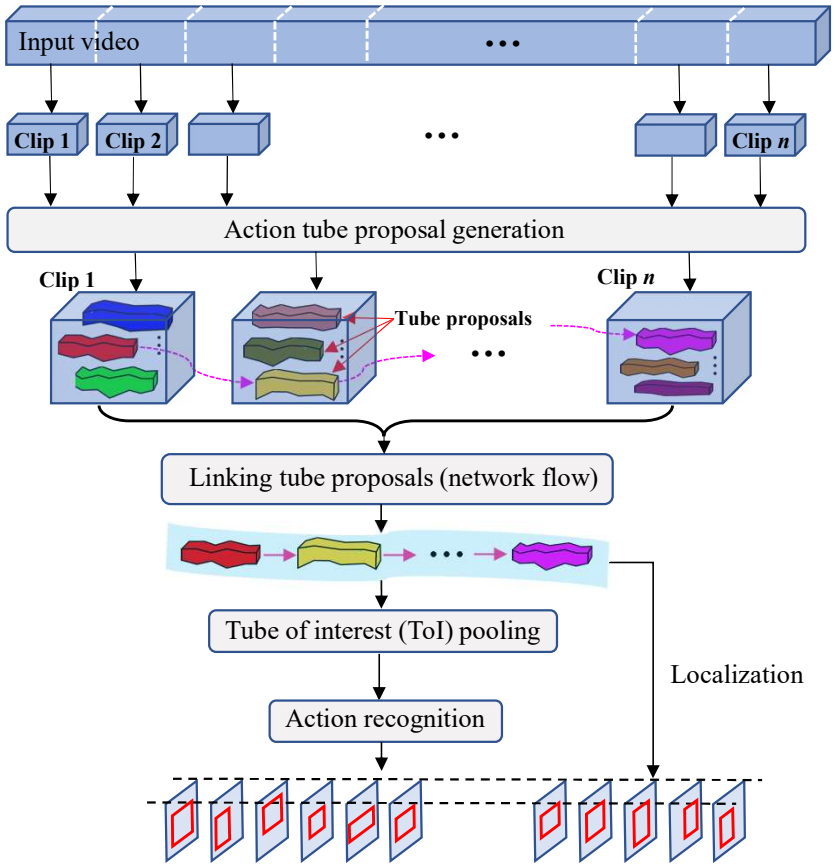
Rui Hou, Chen Chen, Mubarak Shah

ICCV-2017

http://crcv.ucf.edu/papers/iccv17/T-CNN-camera-ready.pdf

https://www.youtube.com/watch?v=3yBAANAMTyM&list=PLd3hlSJsX_IkYquVk8TANmuIgSVHz_Rg4&index=2

Overview of Tube-CNN



- Divide input video into equal length clips
- Generate tube action proposals
- Link tube proposals
- Tube of Interest Max-pooling
- Recognize and localize actions

Detection Results for UCF-Sports

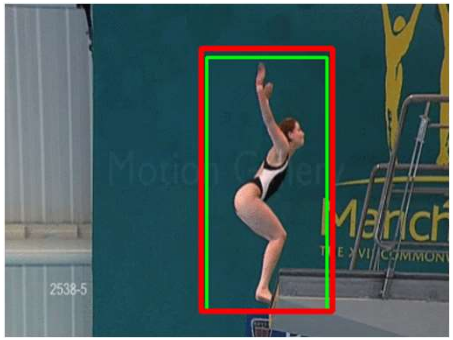


Horse Riding



Golf Swing

Red: Our detection
Green: Ground Truth



Diving



Running

Action Detection for J-HMDB



BrushHair



Claping



Picking

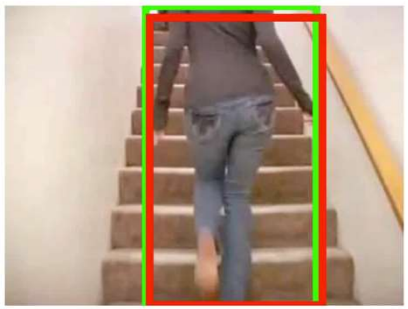
Red: Our detection
Green: Ground Truth



Kicking



Golf



Climbing Stairs

Convolutional Neural Networks (CNNs)

- Current gold-standard for most vision problems
- Achieve strong performance with
  - Large amounts of labeled data
  - Deeper network architectures
- In most video tasks, variants of 3D CNNs are state-of-the-art

Issues with CNNs

- CNNs do not explicitly model entities
- Max Pooling reduces spatial size and loses information

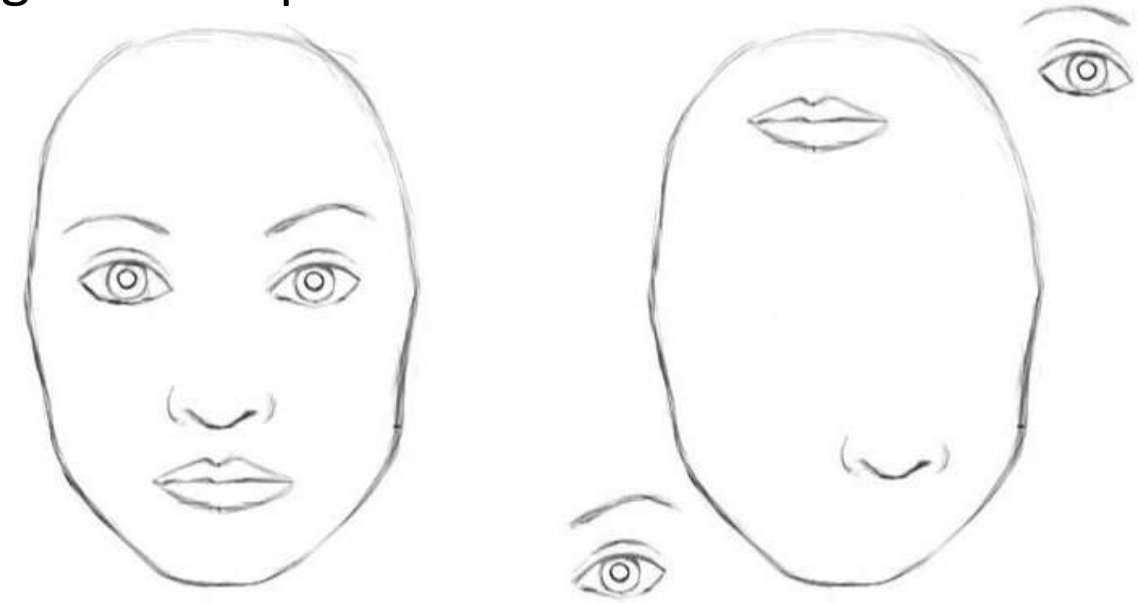


Image Source: Pechyonkin, M. (2017, November 03). Understanding Hinton's Capsule Networks. Part I: Intuition. from https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-networks-part-i-intuition-b4b559d1159b

Capsule Networks

https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Duarte_Defense.mp4

Capsule Networks

- Add **extra structure** to model entities
  - Entities modeled as a **group of neurons**
  - **Routing-by-agreement** to model **part-to-whole** relationships
- Inspiration from Inverse Graphics

Computer Graphics

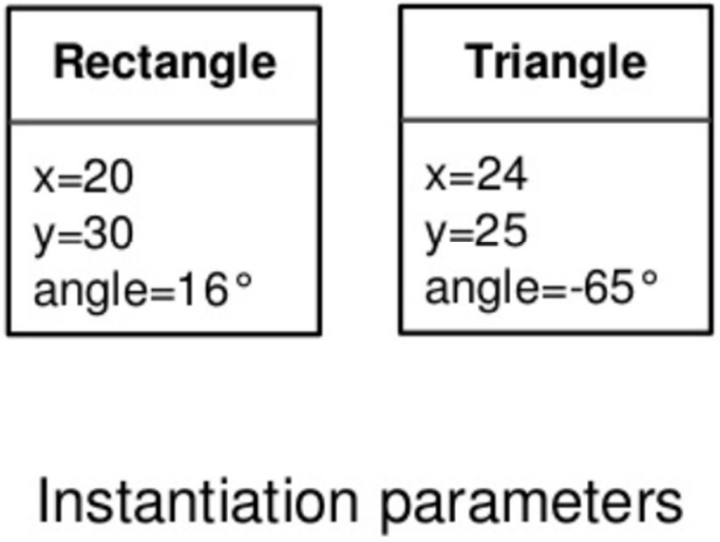


Image source: Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Computer Graphics

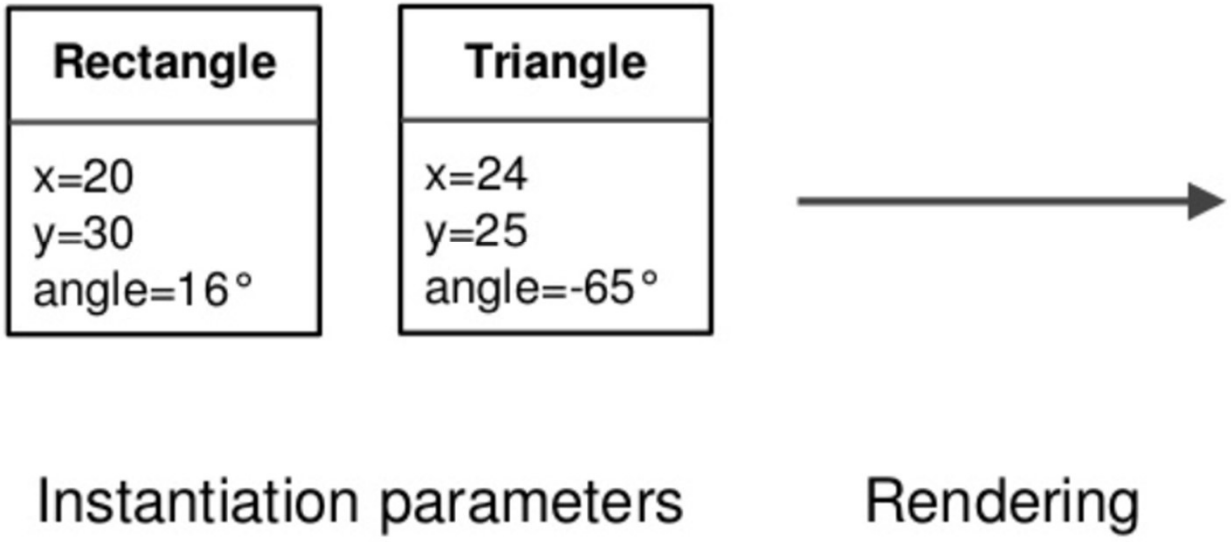


Image source: Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Computer Graphics

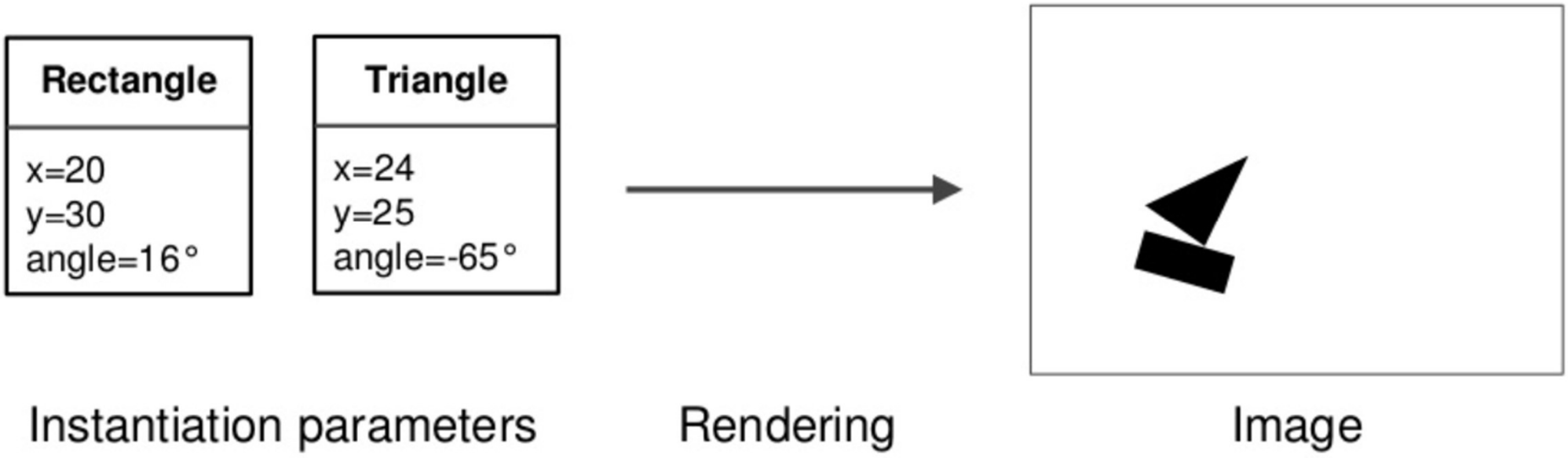
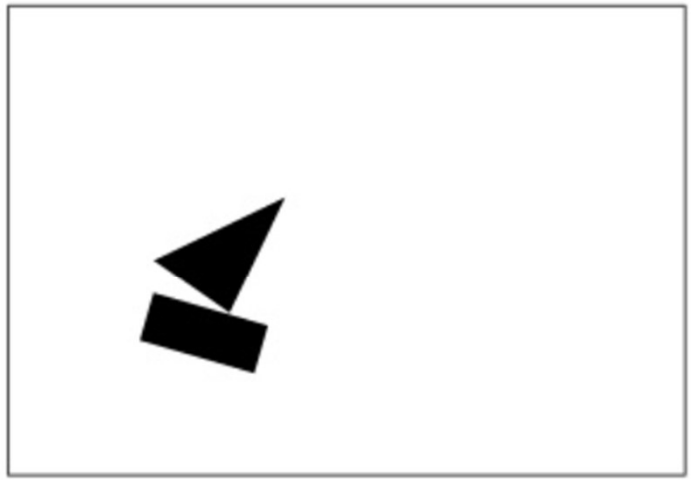


Image source: Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Inverse Graphics



Image

Image source: Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Inverse Graphics

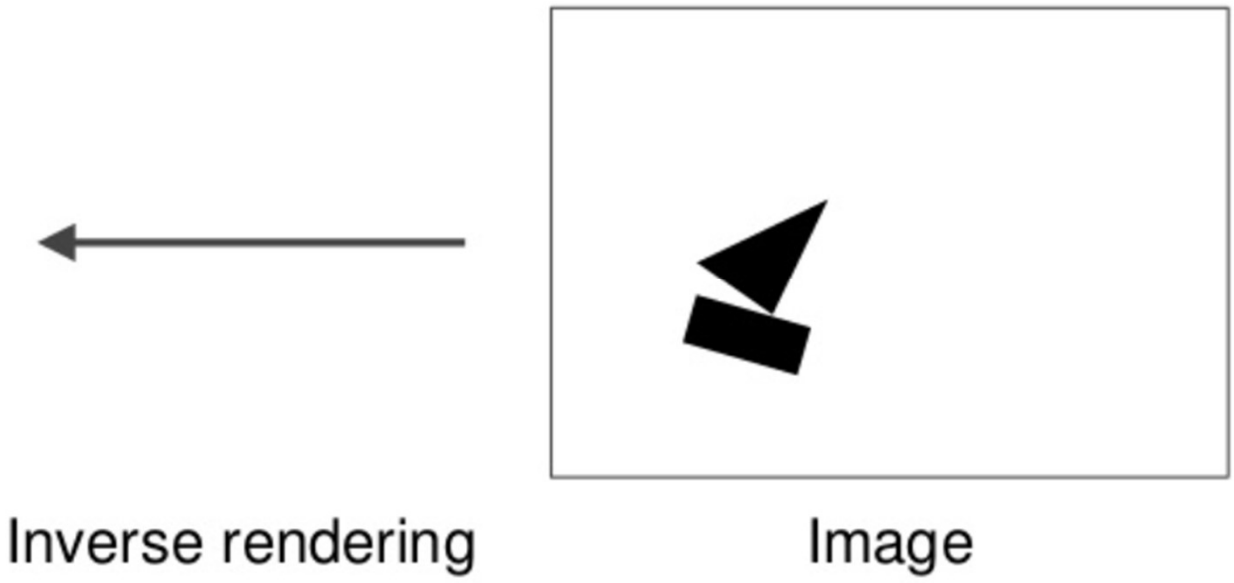


Image source: Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Inverse Graphics

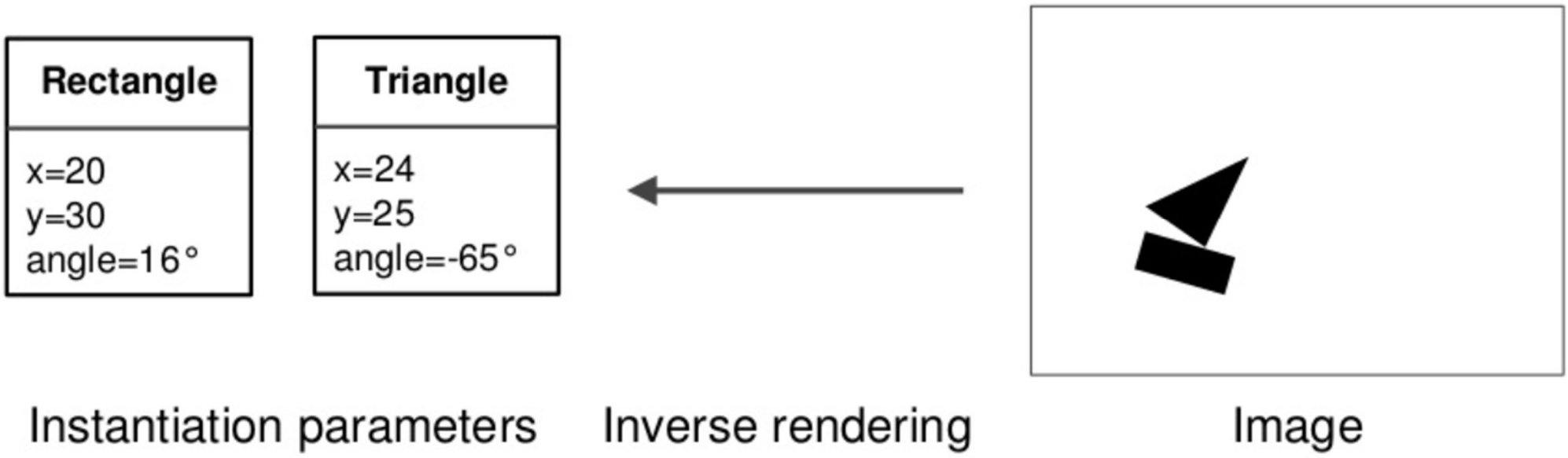
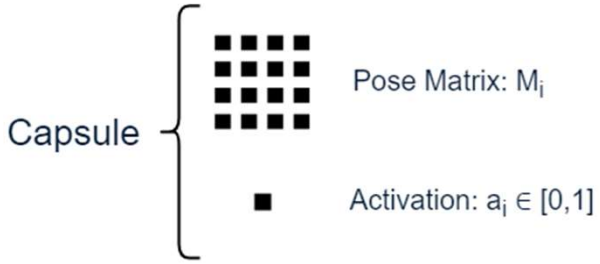


Image source: Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Capsules

- 4x4 pose matrix and an activation neuron

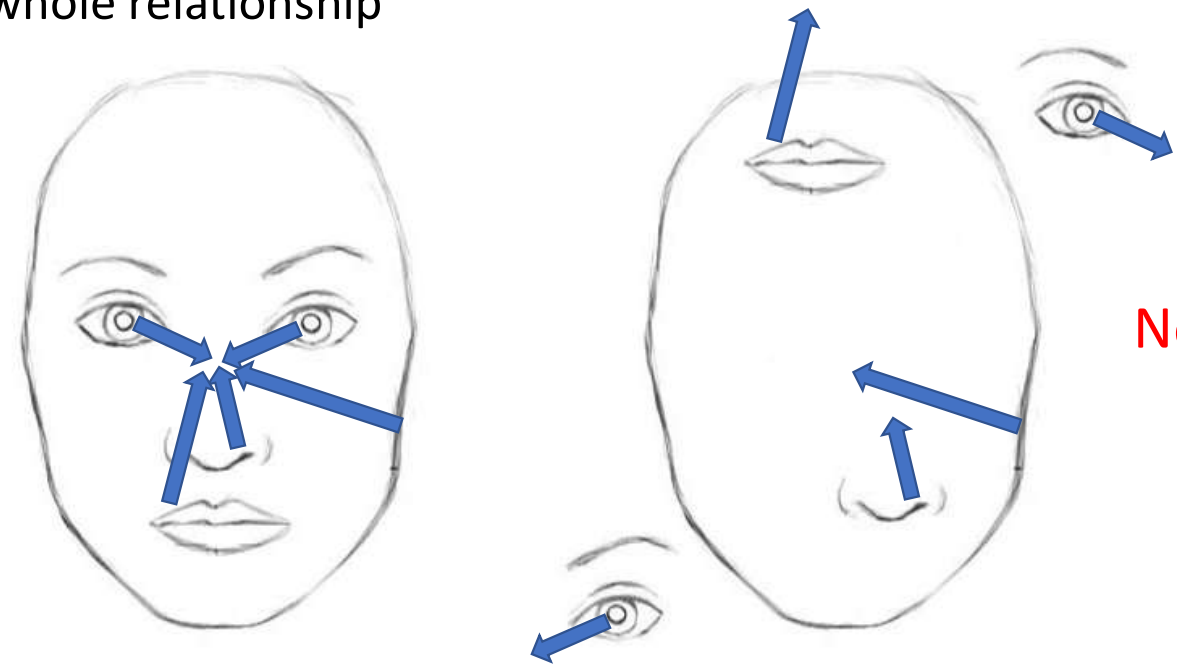


- Capsules in lower levels “vote” on capsules in higher level
- Vote is parts’ prediction of the whole
- Votes are grouped together using “Routing by Agreement”

Routing

- Transforms low-level capsules to higher level capsules
  - Parts-to-whole relationship

Agreement ✓

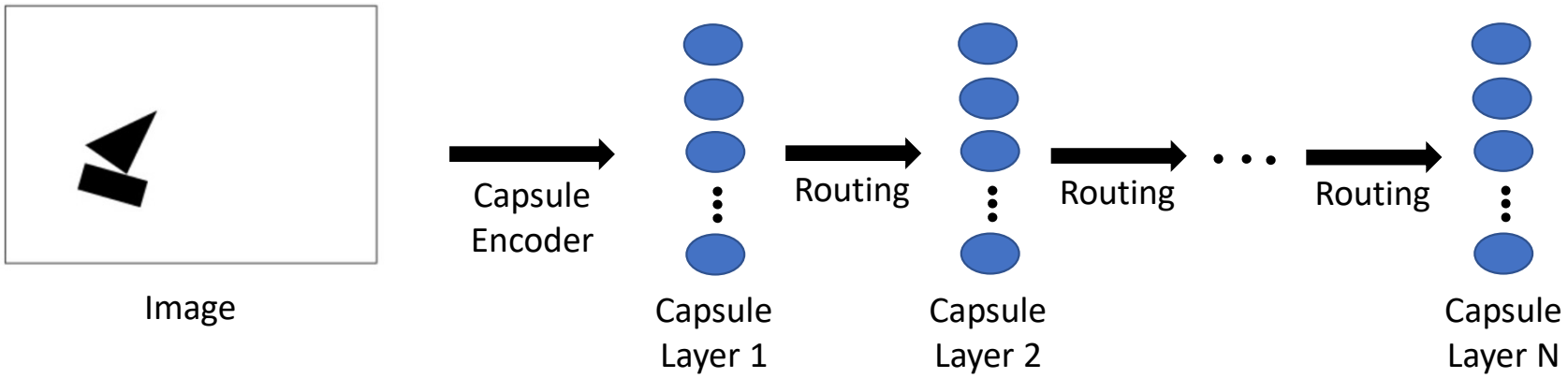


No Agreement ✘

Image Source: Pechyonkin, M. (2017, November 03). Understanding Hinton's Capsule Networks. Part I: Intuition. from https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-networks-part-i-intuition-b4b559d1159b

Routing

- Transforms low-level capsules to higher level capsules
  - Parts-to-whole relationship
- Capsule networks are composed of layers
  - Each layer routes its information to the next



Routing

- Transforms low-level capsules to higher level capsules
  - Parts-to-whole relationship
- Capsule networks are composed of layers
  - Each layer routes its information to the next
- EM Routing [1]

[1] Hinton, Geoffrey E., Sara Sabour, and Nicholas Frosst. "Matrix capsules with EM routing." ICLR (2018).

CapsuleVOS: Semi-Supervised Video Object Segmentation Using Capsule Routing

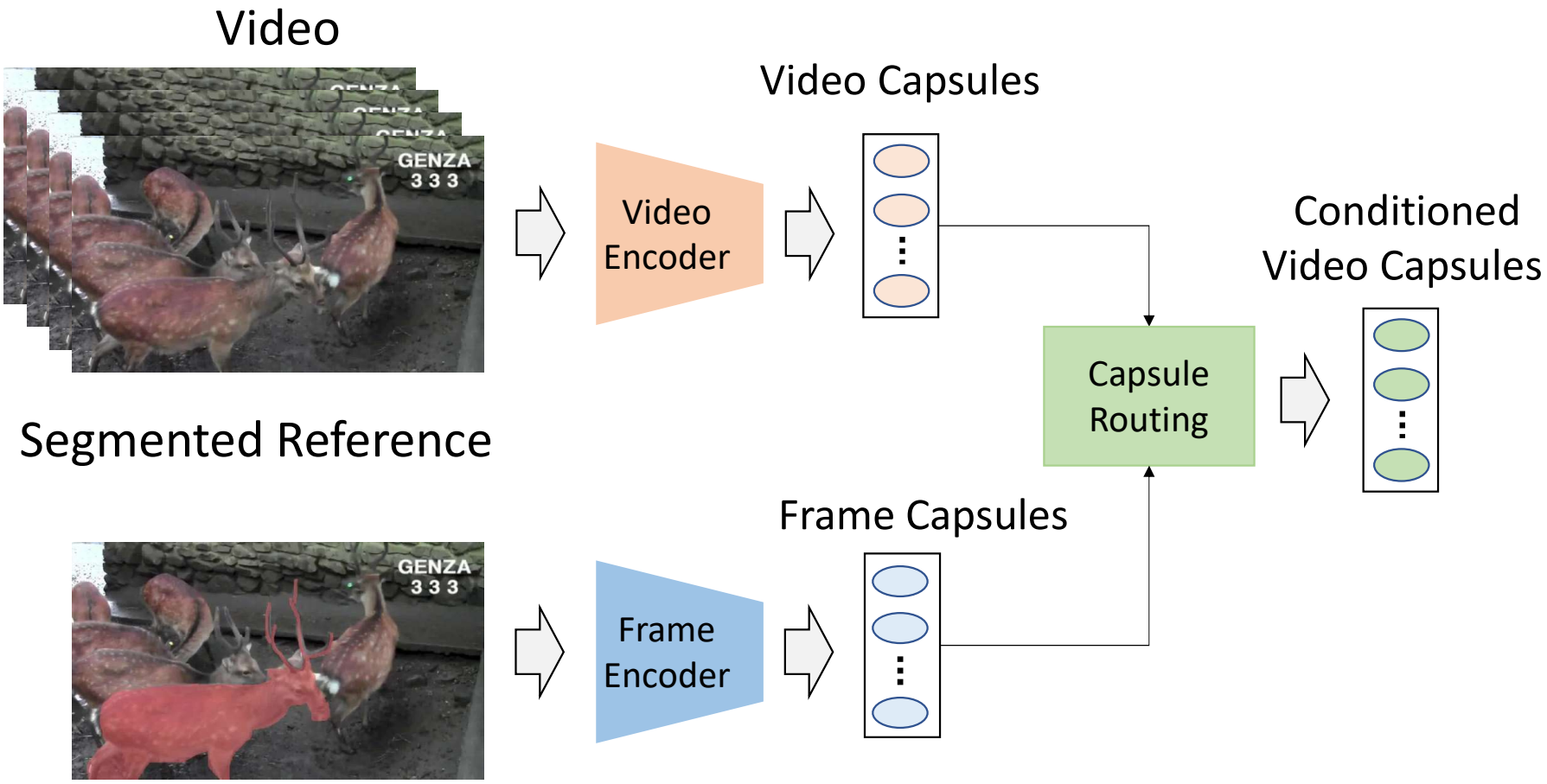
Kevin Duarte, Yogesh Rawat and Mubarak Shah
ICCV-2019

https://www.youtube.com/watch?v=YjjVatilGpA&list=PLd3h

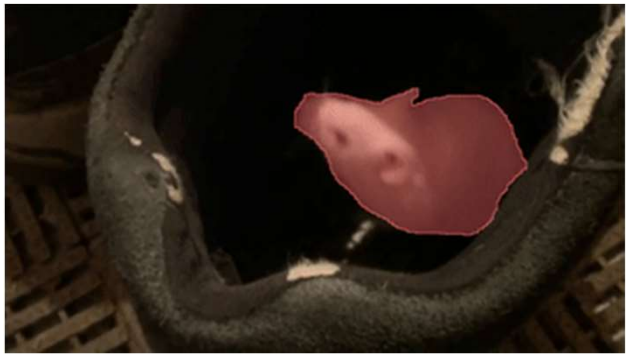


UCF
CENTER FOR RESEARCH IN COMPUTER VISION

VOS using Capsules



Qualitative Results – Single Object








Qualitative Results – Multiple Objects








Visual-textual Capsule Routing for Text-based Video Segmentation

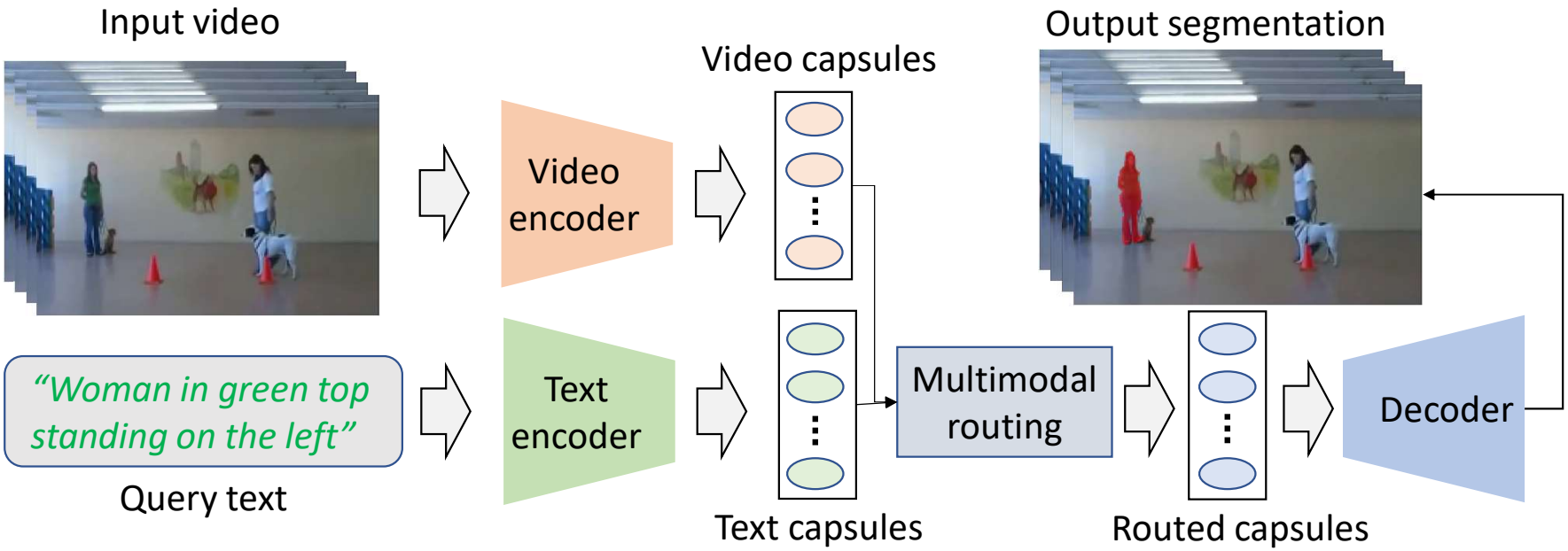
Bruce McIntosh, Kevin Duarte, Yogesh Rawat and Mubarak Shah
ICCV-2019

https://www.youtube.com/watch?v=YjjVatiIGpA&list=PLd3hlSJsX_IkSnnrMtzsMHI1q6vimipvp&index=3&t=3s



Overall Approach

- Create video and text capsules
- Textual Conditioned capsules
- Produce segmentations for all frames



Qualitative Results

"a person is walking backwards while being followed by a dog"
"a black dog is crawling"



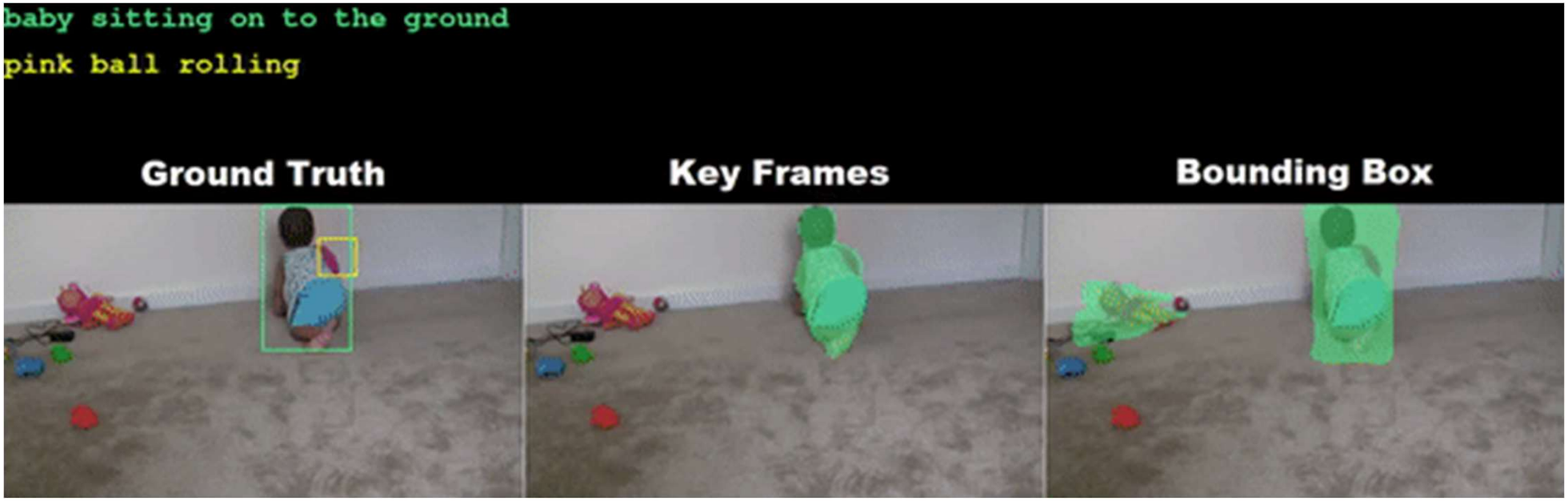
"baby crawling in corridor"
"the dog on the right is crawling"
"dog on the left crawling"



Qualitative Results



Qualitative Results



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

ICCV 2019
Seoul, Korea

Bridging the Domain Gap for Ground-to-Aerial Image Matching



Krishna Regmi

Krishna Regmi & Mubarak Shah
University of Central Florida

https://www.youtube.com/watch?v=gmAhQXCYCEQ&list=PLd3hlSJsX_IkSnnrMtzsMHl1q6vimipvp&index=1

“Where Am I?”

- Problem:

  Image Localization

  Input





Mere Visual Information(Images)

Location in Terms of λ (Lon.) and φ (Lat.)

“Where Am I?”

- Problem:

  Image Localization

  Input



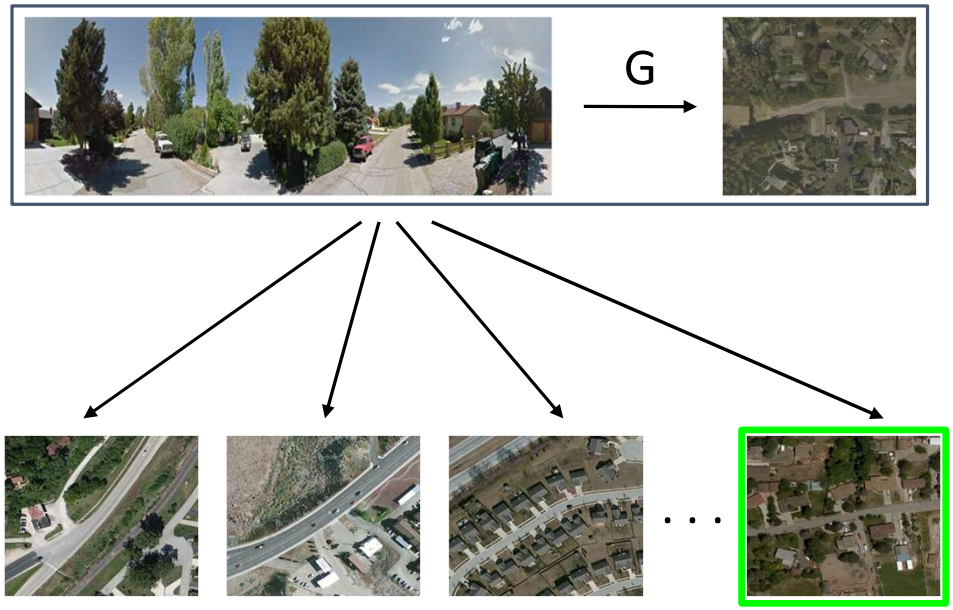


Mere Visual Information(Images)

Location in Terms of λ (Lon.) and φ (Lat.)

***φ*=40.4419, *λ*=-79.9986**

Cross-View Geo-Localization

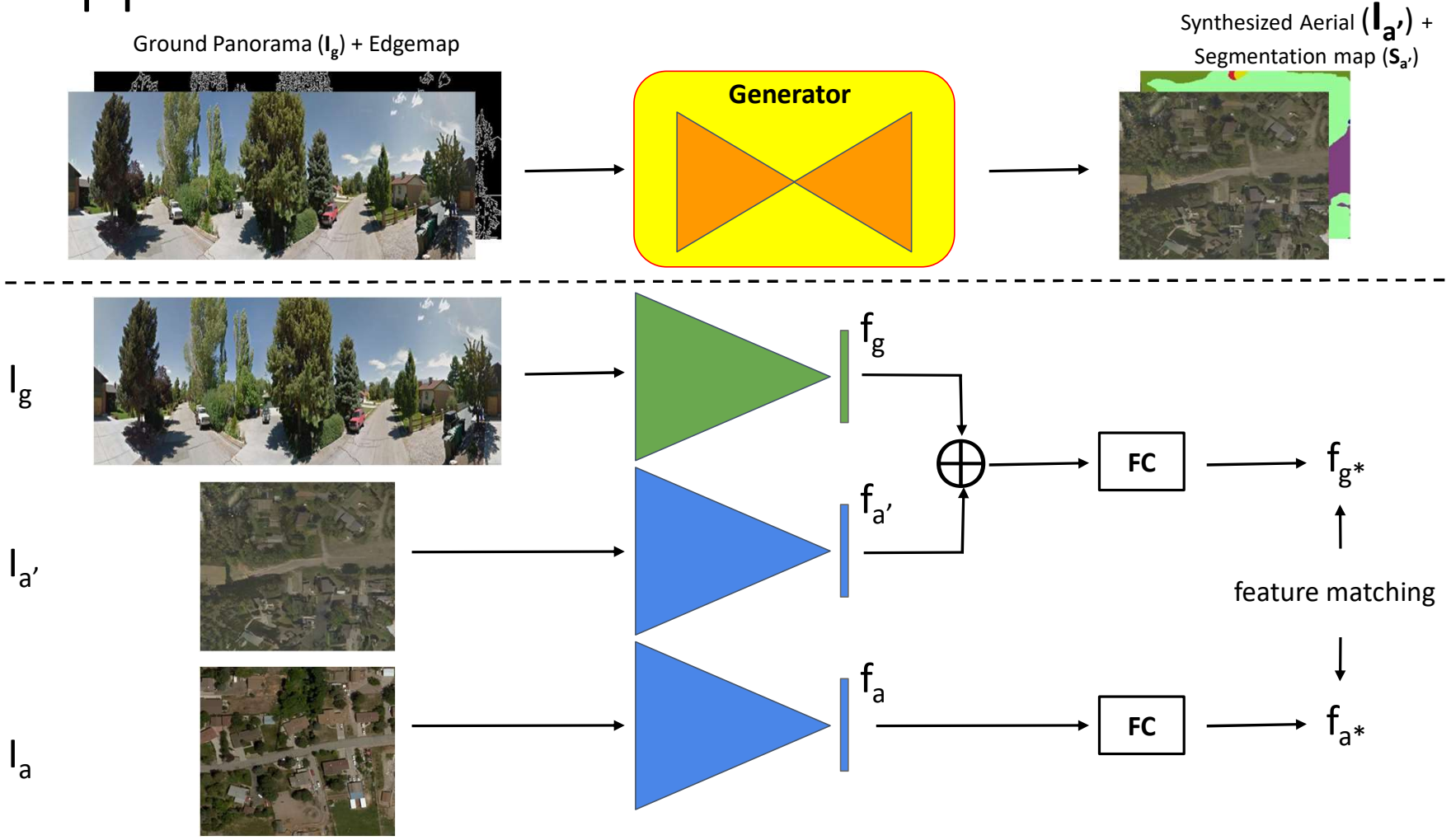


Our Approach

Cross-view image synthesis followed by

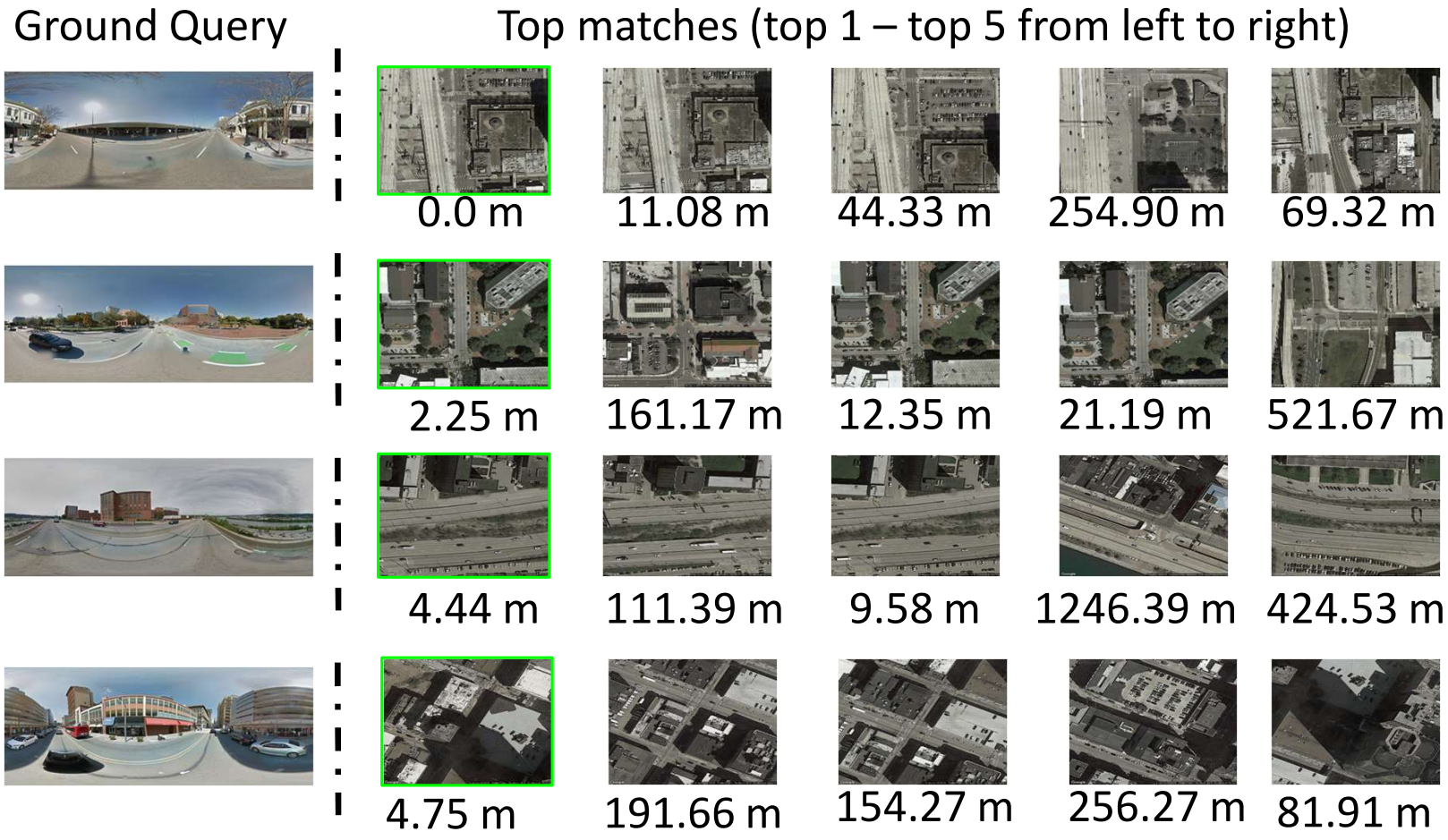
Joint Feature Learning and Feature Fusion

Our Approach



Center for Research in Computer Vision
UCF
UNIVERSITY OF CENTRAL FLORIDA

Cross-View Geo-localization



Video Geo-Localization Employing Geo-Temporal Feature Learning and GPS Trajectory Smoothing

Krishna Regmi and Mubarak Shah

ICCV-2021

https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Regmi_Defense.mp4



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Problem statement

- Given a query video clip,
  - Estimate the GPS for each frame in the video clip
  - Smooth the estimated GPS trajectory

Center for Research in Computer Vision
UNIVERSITY OF CENTRAL FLORIDA

Video Geo-Localization



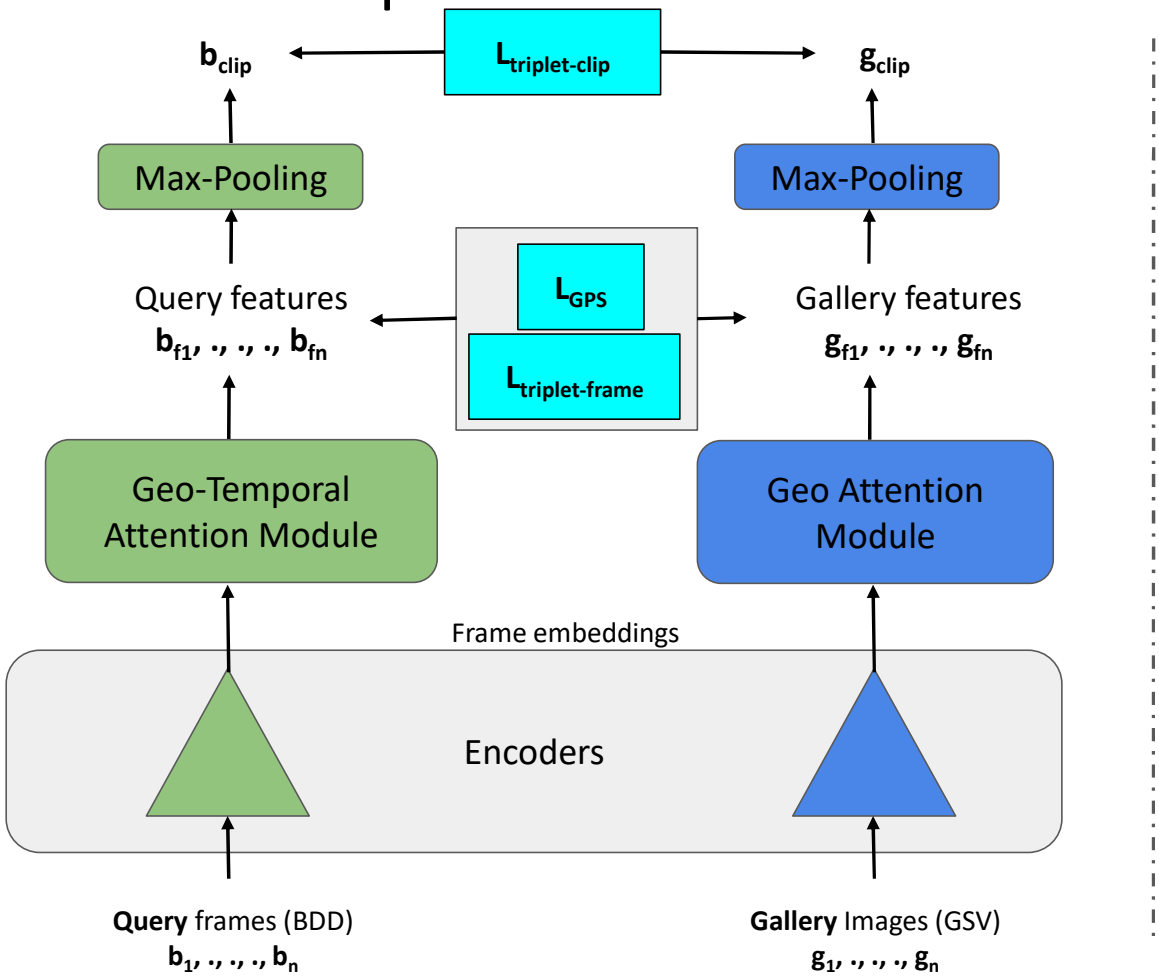


T = 3 T = 6 T = 9 T = 12 T = 15

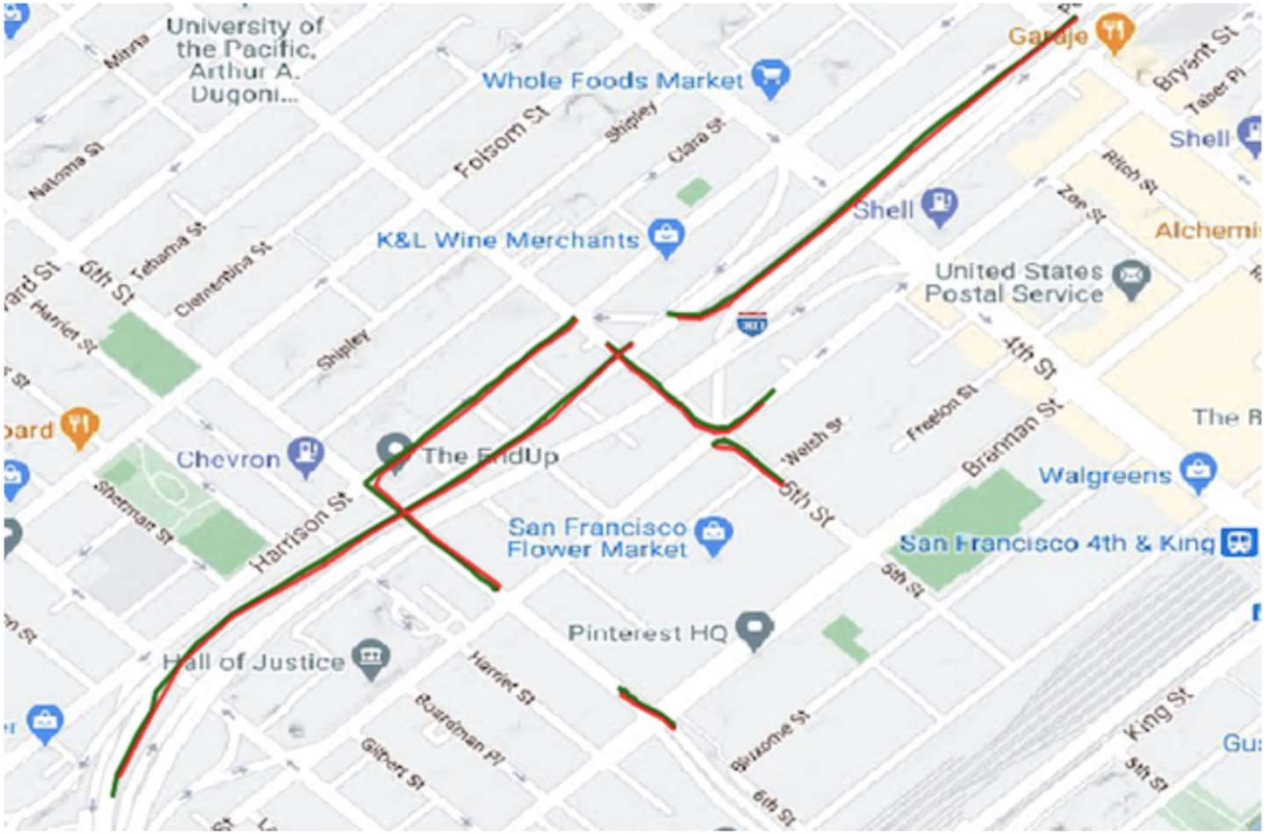


...

Geo-Temporal Feature Learning Network

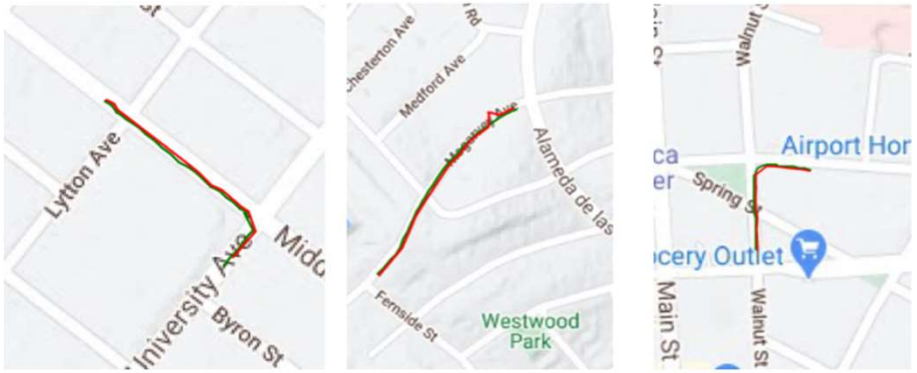


Qualitative Results

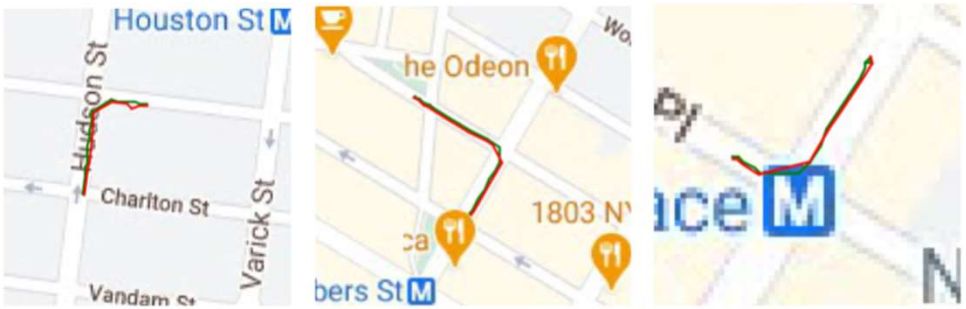


Green : Ground truth trajectory
Red : Estimated trajectory

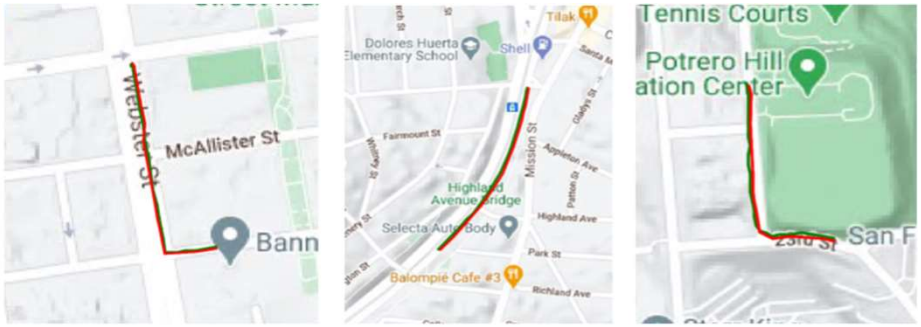
Additional visualizations



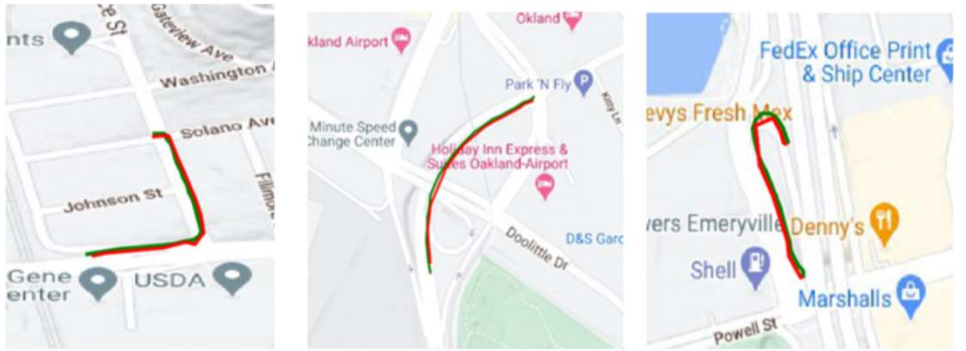
Bay Area



New York



San Francisco



Berkeley

Green : Ground truth trajectory
Red : Estimated trajectory

Quantitative Evaluation (Localization error in meters)

Methods	SF	Bay Area	Berkeley	NY
2D CNN	2516	4686	7020	1818
2D CNN + Smoothing	2290	3999.17	5425	1292
2D CNN^f	2091.66	4509.15	6687.61	1332.08
2D CNN^f + Smoothing	1710.54	4112.15	4565.46	1222.88
[45]	1742.39	4257.08	5031.35	1164.83
[45] + Smoothing*	1327.77	3253.06	3755.98	935.78
3D CNN	4247.09	6183.71	6677.93	1572
3D CNN + Smoothing	3848.83	5201.65	6503.96	1399.17
Proposed	300.47	524.28	424.79	493.43
Proposed + Smoothing	**128.94**	**206.32**	**161.51**	**285.41**

[45] E. Zemene et al. Large-scale image geo-localization using dominant sets. IEEE T-PAMI, 41(1):148–161, 2019

RescueNet: Joint Building Segmentation & Damage Assessment from Satellite Imagery

Rohit Gupta and Mubarak Shah

ICPR-2020
https://arxiv.org/pdf/2004.07312.pdf



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

xBD Dataset

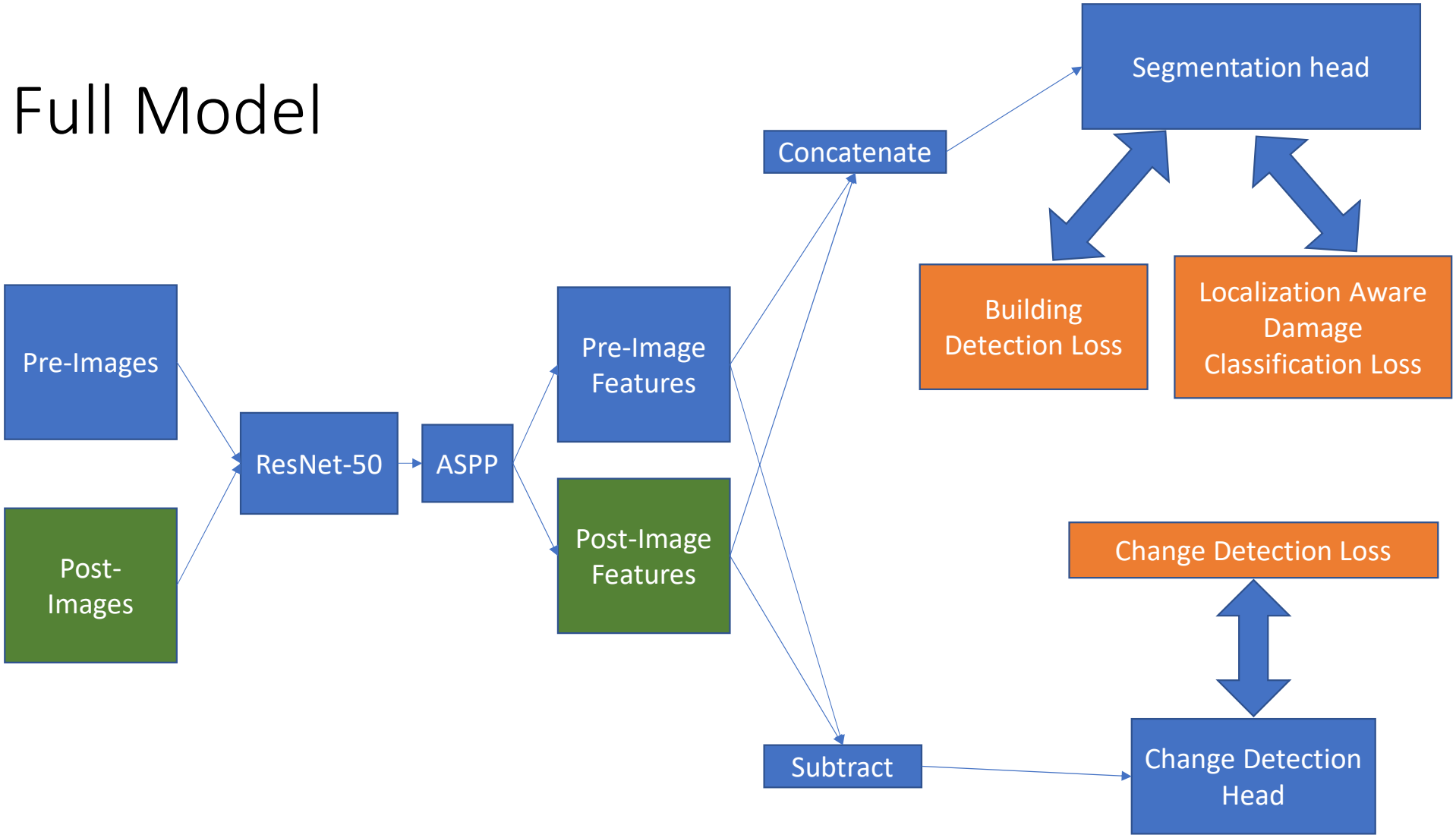
- Collected for Humanitarian Aid and Disaster Relief
- 19 disaster events (such as floods, volcanic eruptions, earthquakes, and hurricanes) spread across the world
- High-resolution (1024X1024) satellite imagery
- 4-classes damage labels provided for each building

xBD Dataset

- Collected for Humanitarian Aid and Disaster Relief
- 19 disaster events (such as floods, volcanic eruptions, earthquakes, and hurricanes) spread across the world
- High-resolution (1024X1024) satellite imagery
- 4-classes damage labels provided for each building

Data Split	Image Pairs
Tier1 Train	2,495
Tier3 Train	6,369
Tier1 Validation	304

Full Model


Pre-Images
Post-Images
ResNet-50
ASPP
Pre-Image Features
Post-Image Features
Concatenate
Subtract
Segmentation head
Building Detection Loss
Localization Aware Damage Classification Loss
Change Detection Loss
Change Detection Head

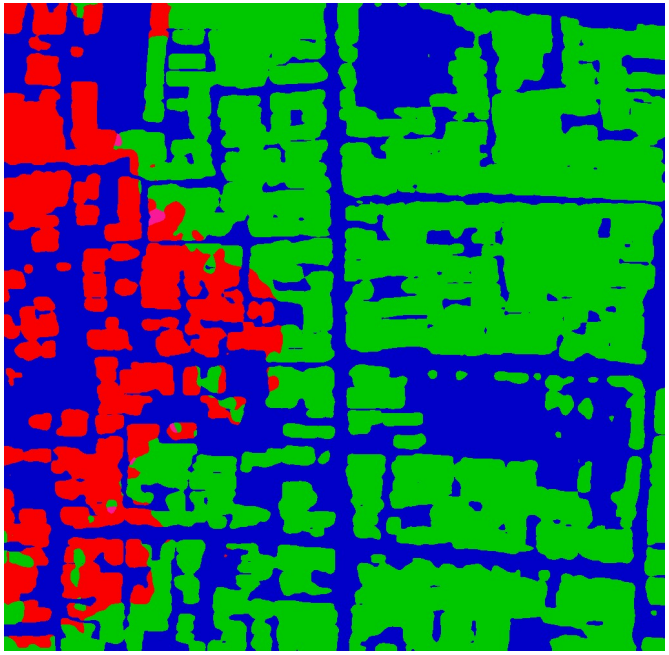
Qualitative Results

Blue: Areas outside buildings
Green: Undamaged Buildings
Red: Completely Damaged
Pink: Building with major damage
Orange: Building with minor damage

Ground Truth



Ours



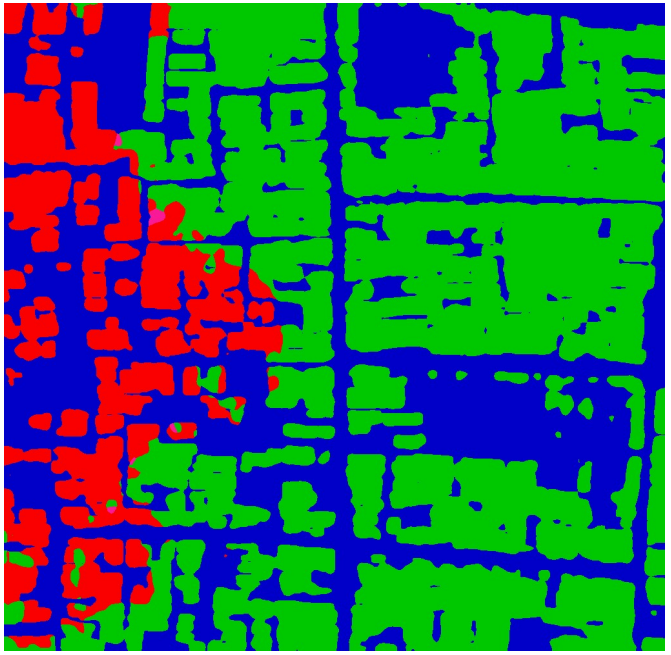
Qualitative Results

Blue: Areas outside buildings
Green: Undamaged Buildings
Red: Completely Damaged
Pink: Building with major damage
Orange: Building with minor damage

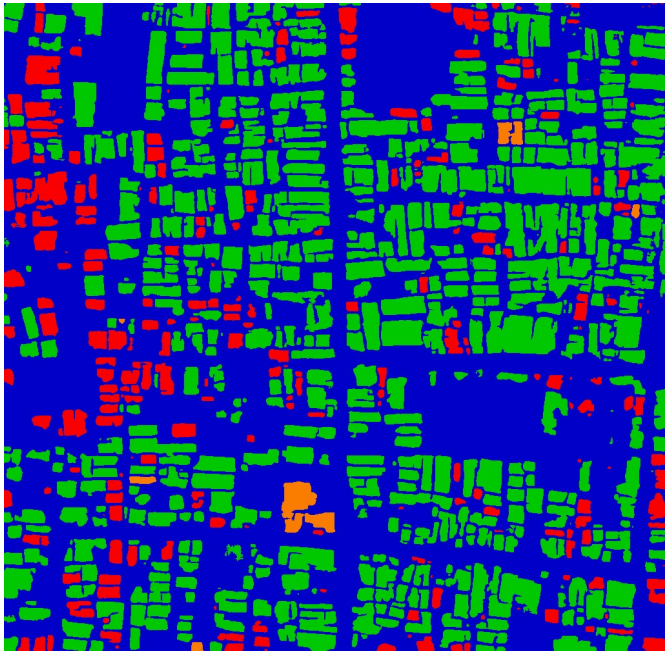
Ground Truth



Ours



Baseline [1]



[1] Gupta, Ritwik, et al. "xBD: A dataset for assessing building damage from satellite imagery." arXiv:1911.09296 (2019).







Face Image Retrieval with Attribute Manipulation

Alireza Zaeemzadeh[*,1], Shabnam Ghadar[2], Baldo Faieta[2], Zhe Lin[2], Nazanin Rahnavard[1], Mubarak Shah[1], Ratheesh Kalarot[2]

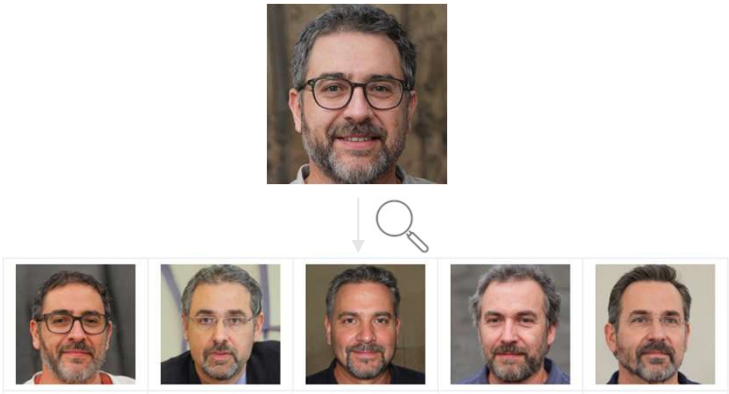
[1] University of Central Florida

[2] Adobe Inc.

[*] Work done as part of an internship at Adobe Inc.

Face Image Retrieval with Attribute Manipulation

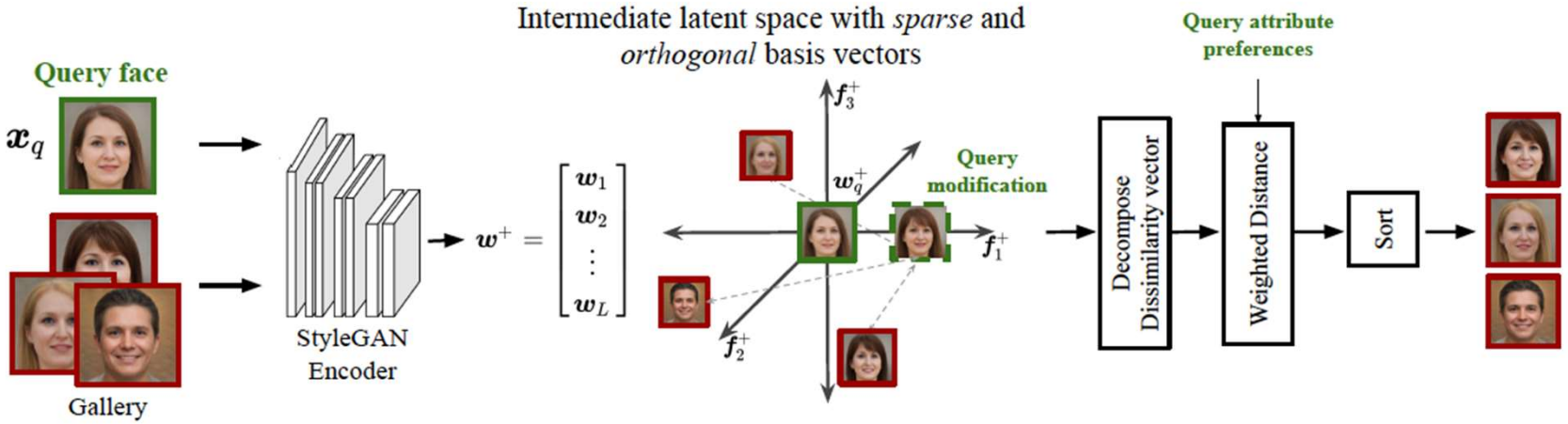
Given a query face image, find the most similar faces



User should be able to emphasize/change attributes

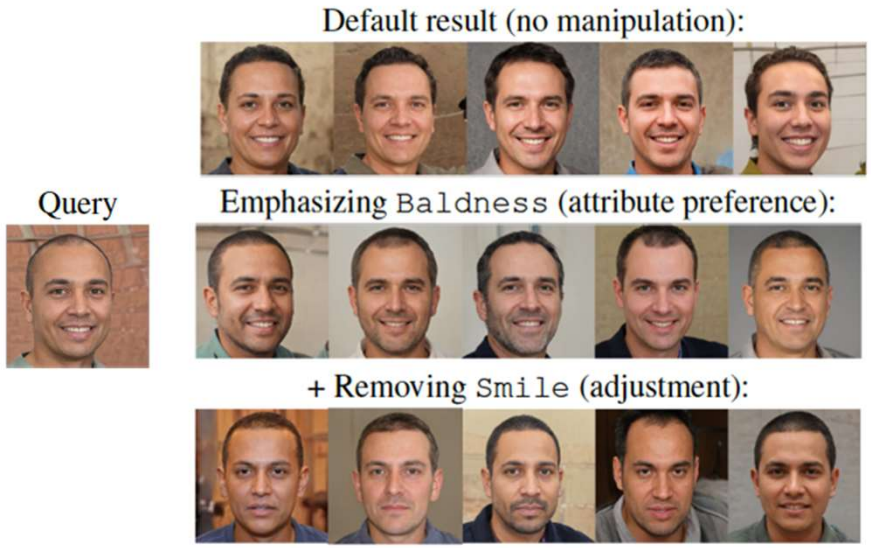
- Simultaneously
  - *Adjusts* the facial attributes and
  - *Assigns preference* to different attributes

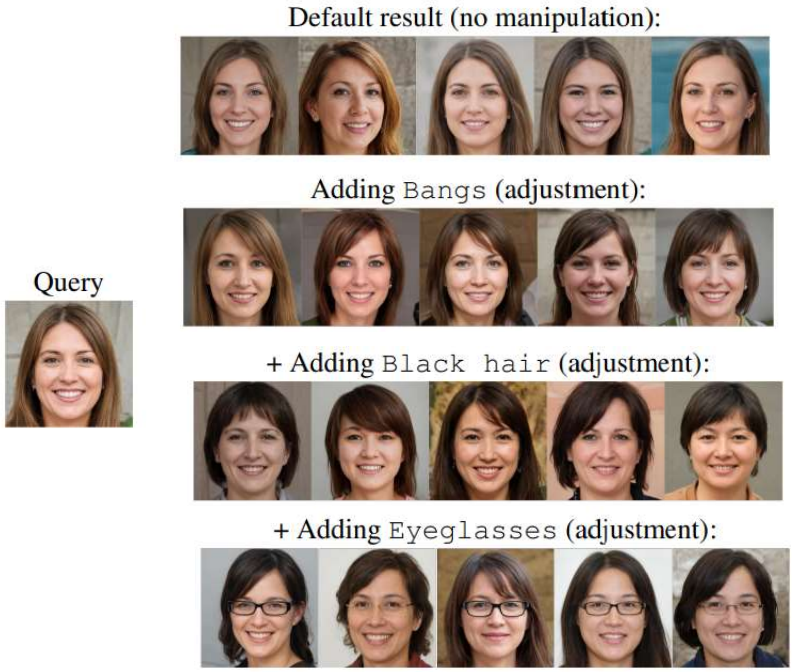
Proposed Method: Overview



- Dissimilarity can be easily decomposed, due to orthogonality.
- Weighted similarity can be defined to emphasize attributes

Qualitative Results: Retrieval





Found a Reason for me? Weakly-supervised Grounded Visual Question Answering using Capsules



Aisha Urooj Khan[1,2], Hilde Kuehne[1], Kevin Duarte[2], Chuang Gan[1], Niels Da Vitoria Lobo[2], Mubarak Shah[2]

[1] MIT-IBM Watson AI Lab, [2] University of Central Florida

CVPR 2021

MIT-IBM Watson AI Lab

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Visual Question Answering

Q: Are the black horses to the right of the vehicle on the road?

Yes





UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Grounded Visual Question Answering

Q: Are the black horses to the right of the vehicle on the road?

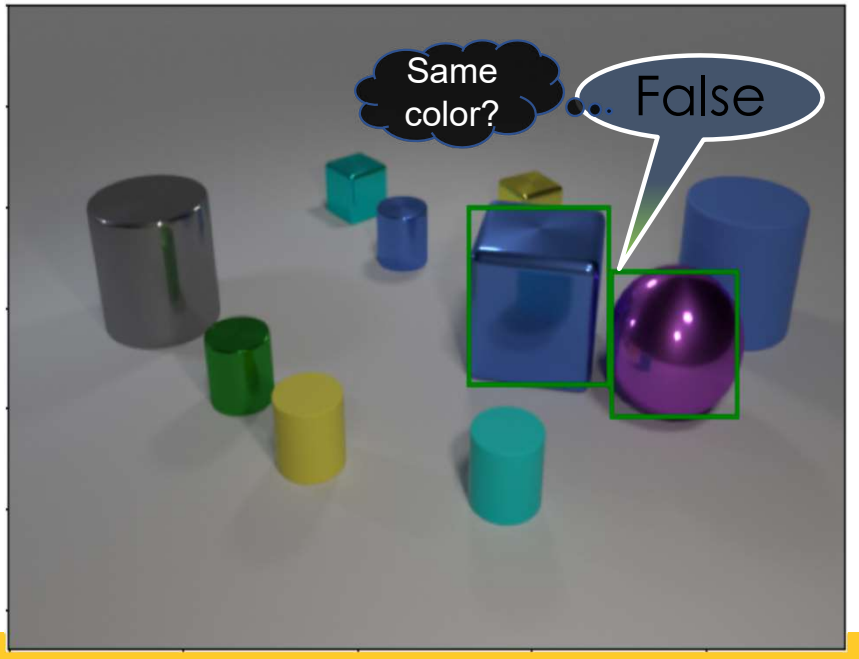
Yes



Grounded Visual Question Answering

There is a cube that is in front of the blue rubber cylinder; is it the same color as metal sphere?

False





Summary

- Person RE-ID
- Multiple Object Tracking
- Drone Video Analysis
- Anomaly Detection
- Action Localization
- Video Object Segmentation
- Building Damage Assessment
- Face Retrieval
- Geo-spatial Localization
- Grounded Visual Q&A



Classical Computer Vision Vs Deep Learning

Classical Computer Vision

- Solve Problem in a Modular Fashion
- Encode Expert Knowledge into constraints
- Convert constraints into Objective Function
- Optimize Objective Function

Deep Learning

- End-to-End
- Real Learning
- Motivated by human neural networks
- Requires Labeled/Annotated Data
- Need to Define Loss Function
- Requires Massively Parallel Computations
- Use simple Stochastic Gradient Descend
- Excellent Results



Classical Computer Vision Vs Deep Learning

Classical Computer Vision

- Solve Problem in a Modular Fashion
- Encode Expert Knowledge into constraints
- Convert constraints into Objective Function
- Optimize Objective Function

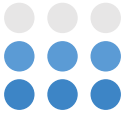
Deep Learning

- End-to-End
- Real Learning
- Motivated by human neural networks
- Requires Labeled/Annotated Data
- Need to Define Loss Function
- Requires Massively Parallel Computations
- Use simple Stochastic Gradient Descend
- Excellent Results



Supervised Learning

- Supervised learning relies on large labeled datasets
- Constructing large labeled datasets
  - expensive
  - time-consuming



Labeled Data

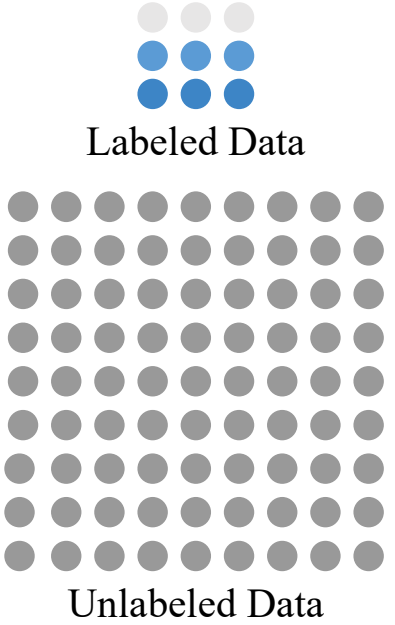
Learning with Less Labels

- Semi-supervised Learning
- Few Shot Learning
- Self-Supervised Learning



Semi-Supervised Learning

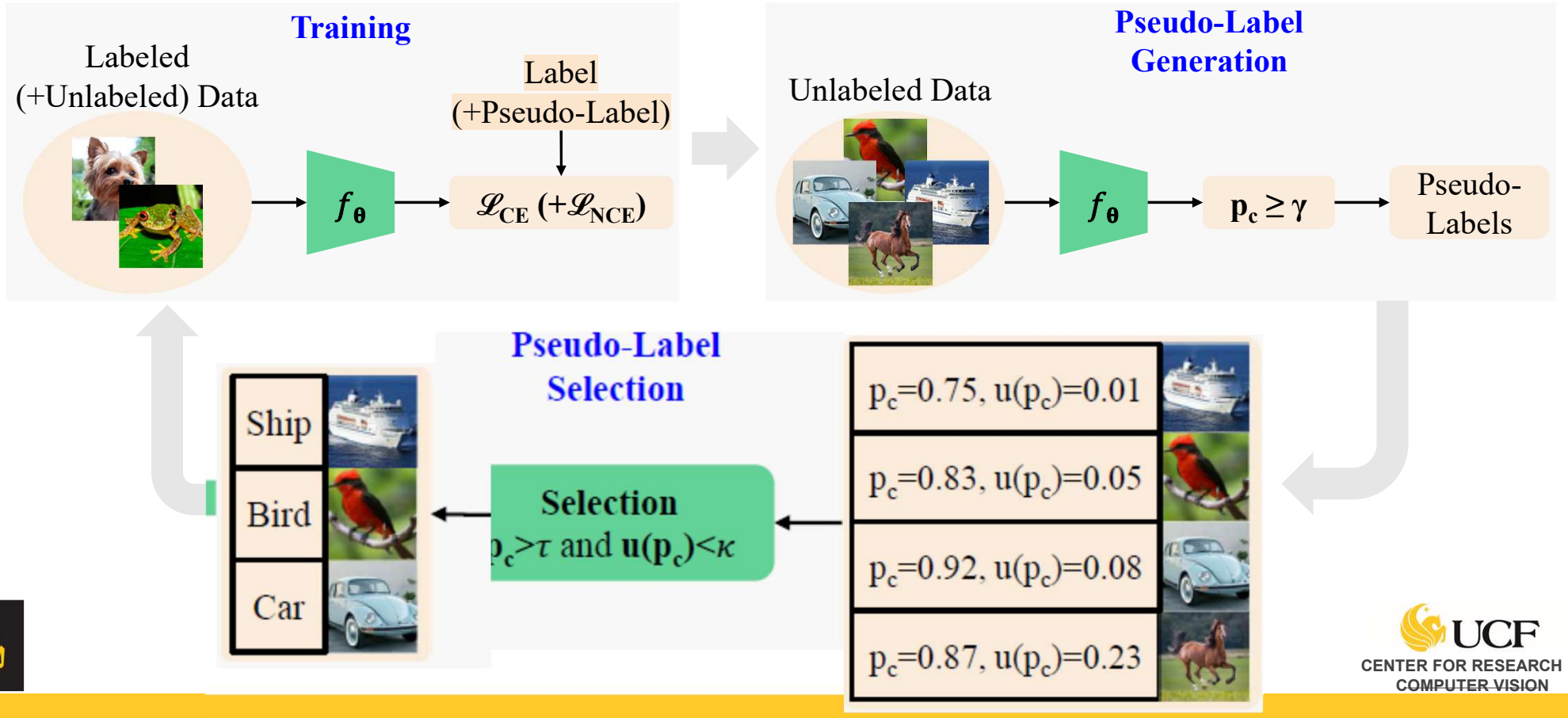
- SSL leverages
  - a small amount of labeled data
  - a large amount of unlabeled data concurrently



Rizve, Mamshad Nayeem; Duarte, Kevin; Rawat, Yogesh Singh; Shah, Mubarak
In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning
Ninth International Conference on Learning Representations (ICLR), 2021



UPS Method


Training
Labeled (+Unlabeled) Data
Label (+Pseudo-Label)
f_θ
$\mathcal{L}_{CE}$ (+$\mathcal{L}_{NCE}$)
Pseudo-Label Generation
Unlabeled Data
f_θ
$p_c \geq \gamma$
Pseudo-Labels
Pseudo-Label Selection
Selection
$p_c > \tau$ and $u(p_c) < \kappa$
Ship
Bird
Car
p_c=0.75, $u(p_c)$=0.01
p_c=0.83, $u(p_c)$=0.05
p_c=0.92, $u(p_c)$=0.08
p_c=0.87, $u(p_c)$=0.23

Results (UCF-101 and Pascal VOC2007)

Accuracy (%) on the UCF-101 test set:

Method	20% labeled	50% labeled
Supervised	33.5	45.6
MT*	36.3	45.8
PL*	37.0	47.5
S4L*	37.7	47.9
UPS	**39.4**	**50.2**

mAP scores on the Pascal VOC2007 test set:

Method	10% labeled	20% labeled
Supervised	18.36 ± 0.65	28.84 ± 1.68
PL	27.44 ± 0.55	34.84 ± 1.88
MixMatch	29.57 ± 0.78	37.02 ± 0.97
MT	32.55 ± 1.48	39.62 ± 1.66
UPS	$\mathbf{34.22 \pm 0.79}$	$\mathbf{40.34 \pm 0.08}$



Few-Shot Learning

- Few-shot learning (FSL)
  - trains a model on a set of **base** classes
  - adapts the model to **novel** classes with only a few samples
- Few-shot learning is challenging
  - limited data and limited annotation

Self-Supervised Learning

- Self-Supervised Learning
  - Trains a model on a set of **unlabelled data**
  - Employs pre-text tasks
  - Finetunes on downstream supervised task

TCLR: Temporal Contrastive Learning for Video Representation I Dave, R Gupta, MN Rizve, M Shah
arXiv preprint arXiv:2101.07974

Classical Computer Vision Vs Deep Learning

Classical Computer Vision

- Solve Problem in a Modular Fashion
- Encode Expert Knowledge into constraints
- Convert constraints into Objective Function
- Optimize Objective Function

Deep Learning

- End-to-End
- Real Learning
- Motivated by human neural networks
- Requires Labeled/Annotated Data
- Need to Define Loss Function
- Requires Massively Parallel Computations
- Use simple Stochastic Gradient Descend
- Excellent Results



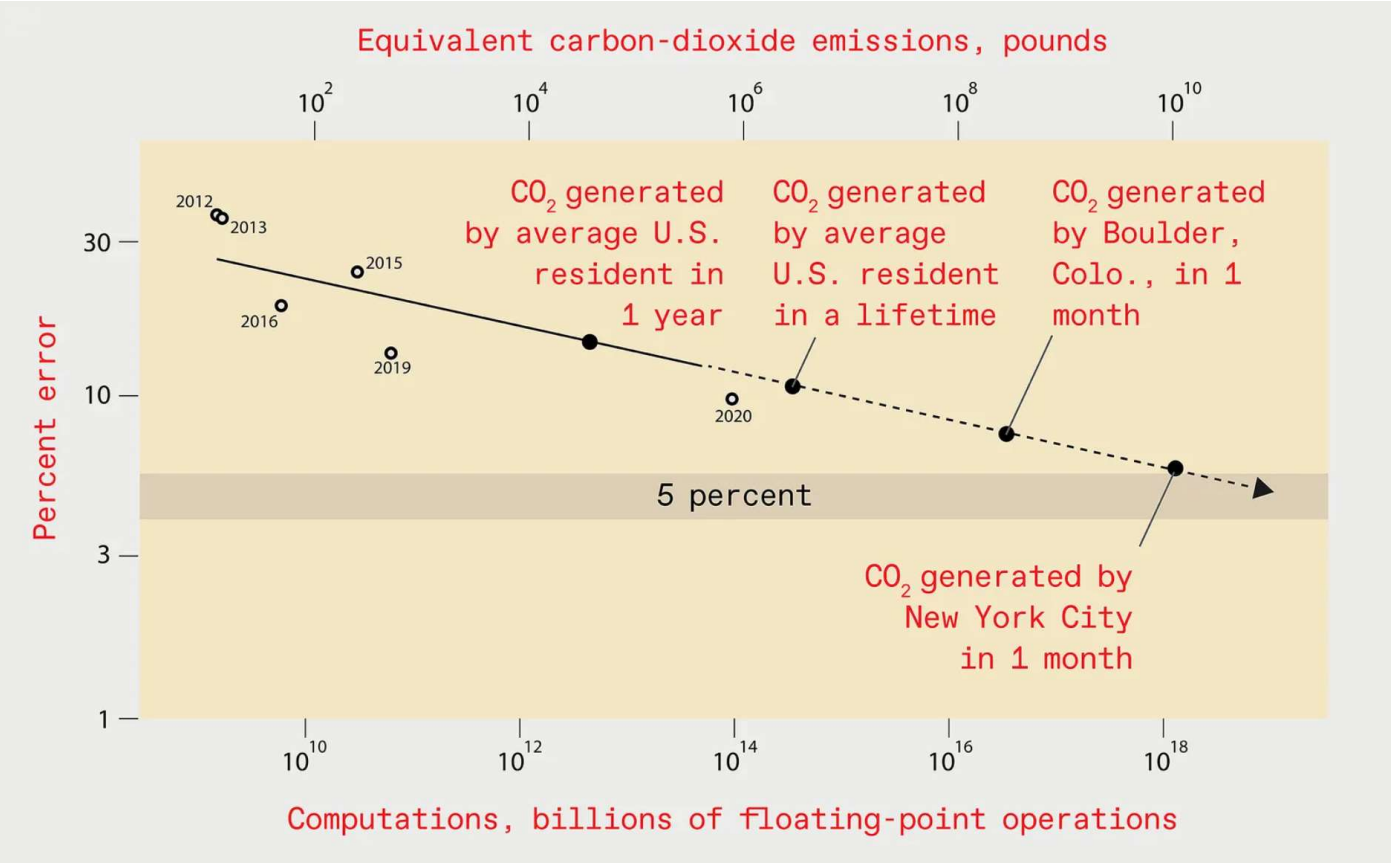
Deep Learning's Diminishing Returns*

- Cost of improvement is becoming unsustainable
  - DeepMind's system to play Go cost $35 Million
  - OpenAI's GPT3 cost $4 Million
  - Enormous Computing resources and energy required to train an ImageNet system with 5% error



*Neil C. Thompson *et. al*., IEEE Spectrum, October 2021

ImageNet Error



Deep Learning's Diminishing Returns*

- Solution maybe different hardware frameworks
  - Analog
  - Neuromorphic
  - Optical
  - Quantum



* Neil C. Thompson *et. al*., IEEE Spectrum, October 2021

Thank you!





UCF CENTER FOR RESEARCH IN COMPUTER VISION

Advances in Computer Vision Employing Deep Learning

Mubarak Shah

Center for Research in computer Vision
University of Central Florida
shah@crcv.ucf.edu
https://www.crcv.ucf.edu/



