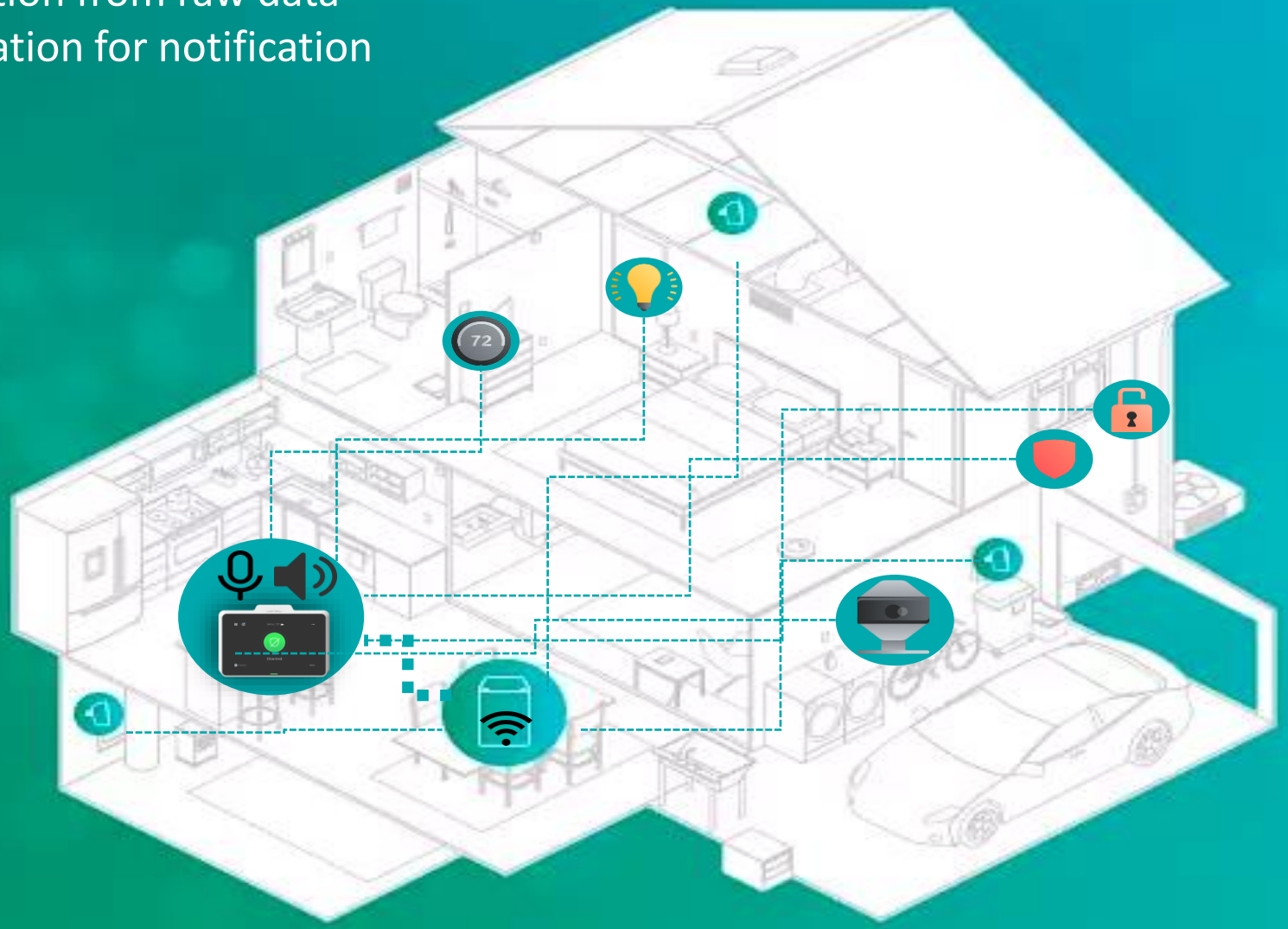
Video Activity Recognition with Limited Data for Smart Home Applications

Hongcheng Wang
Director of Machine Learning, Comcast AI & Discovery
September 2020

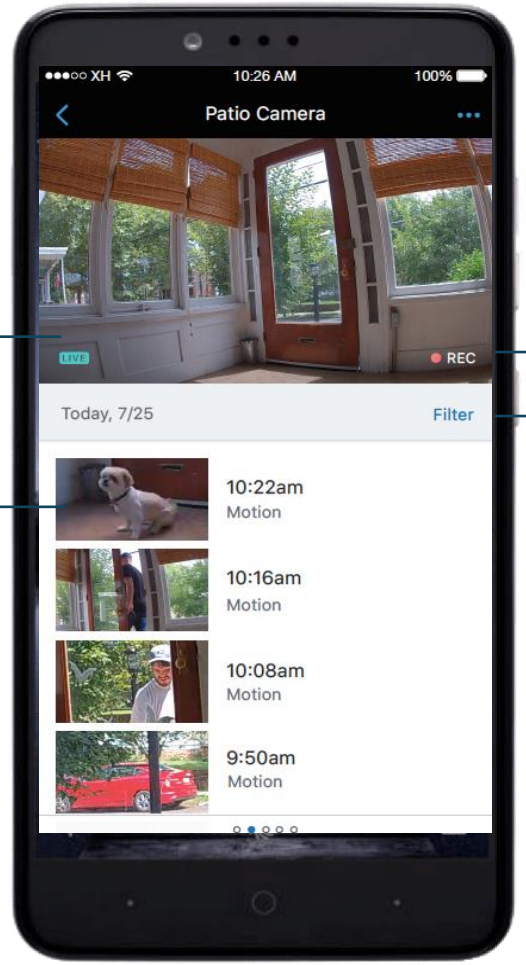
AI FOR CONNECTIVE LIVING

- Event generation from raw data
- Event aggregation for notification
- ML platform



XFINITY HOME APP

XFinity Home App is managing all the IoT devices and their events



Live Video

Motion Events with Thumbnails

24/7 Recording

Object Filter

Door/window events

TOO MANY CAMERA EVENTS

~175

Average Number of Daily Motion Events Per Camera

~7000

Events for 4 Cameras over 10 Days

~4

Hours to review all video events

COMCAST

NOT EASY FOR OUR CUSTOMERS TO FIND AN EVENT

COMCAST

UNIMPORTANT VIDEOS

- 73% of camera owners have a camera facing a front door or driveway. (Source: Google Survey, April 2019,)

- Common complaints about unimportant events on outdoor cameras (Source: Employee Survey, Sept 2018):
  - *"My cats, blowing trees and cars driving past my house cause too many triggers. If I can eliminate (most of) those I'd be golden!"*
  - *"I'd like to filter out stray cats and birds which I tend to see a lot in my feed."*
  - *"I have too many cars driving down the street unrelated to stopping at my house."*

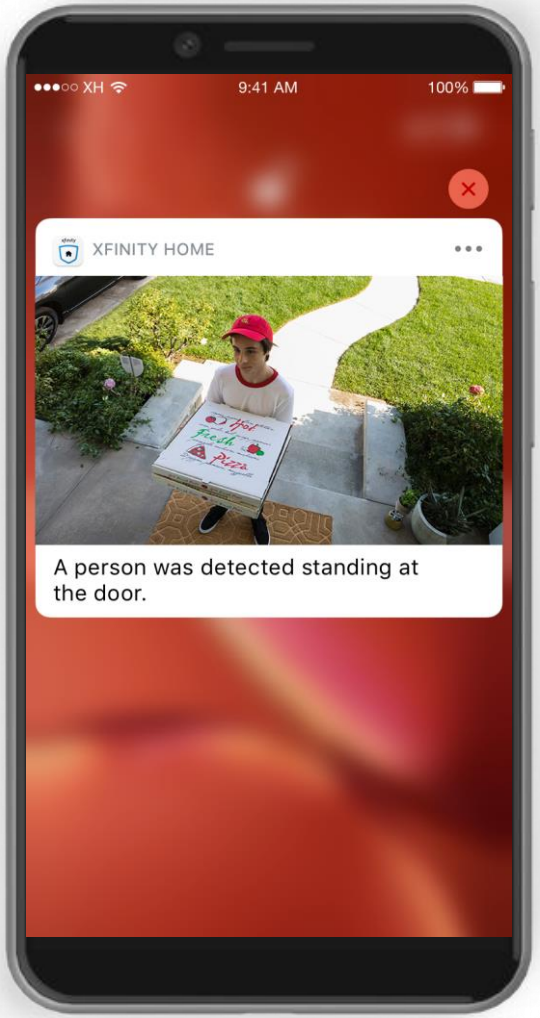
- TPX Research found that Notification of a Motion Event is a Must-Have, but users want to avoid excessive notifications. (https://research.xfinity.com/publication/xh-cvr-and-audio-handling-concept/, https://research.xfinity.com/publication/cvr-roadmap-prioritization-kano-categorization-model/)

GOAL: DELIVER RELEVANT ALERTS TO ENHANCE CUSTOMER EXPERIENCE

Develop computer vision and machine learning algorithms that can identify top 3 – 5 alert worthy videos each day.

Challenges:

- Limited data for training



COMCAST

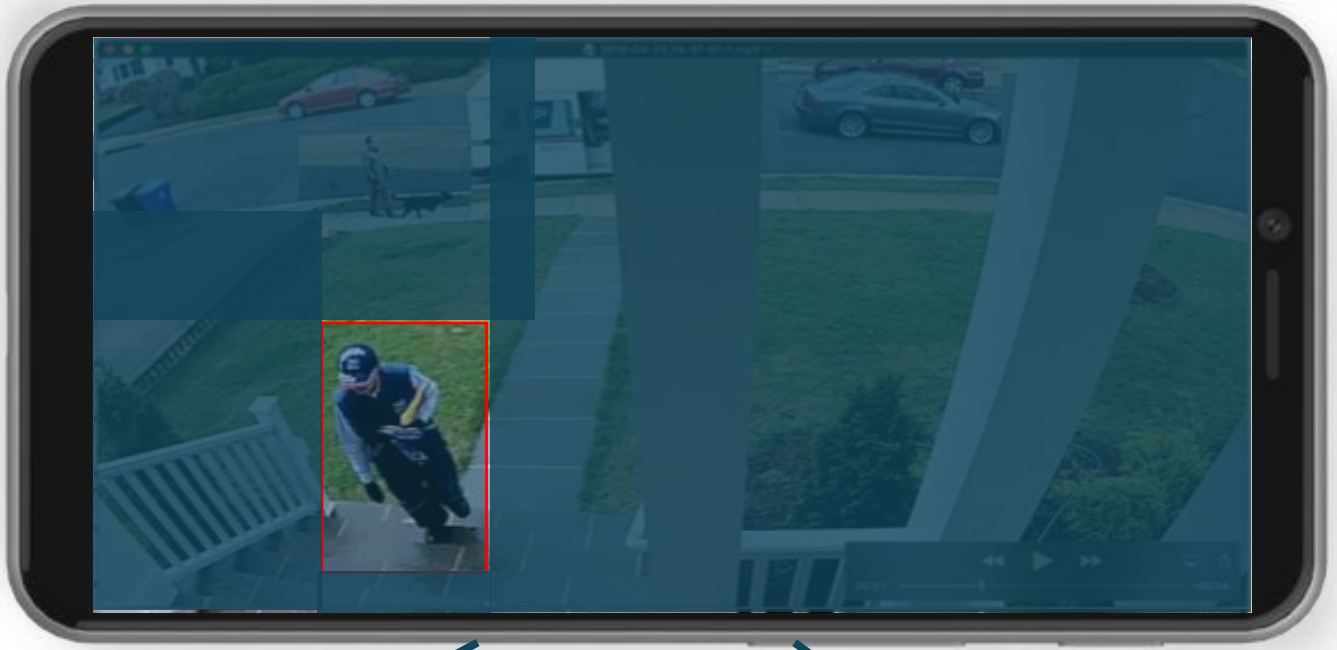
WHAT EVENTS ARE MORE RELEVANT/SIGNIFICANT?

- **What: object/event relevancy**
- **Where: spatial relevancy**
- **When: temporal relevancy**

COMCAST

1. OBJECT/EVENT RELEVANCY

- **What: object/event relevancy**
- **Where: spatial relevancy**
- **When: temporal relevancy**



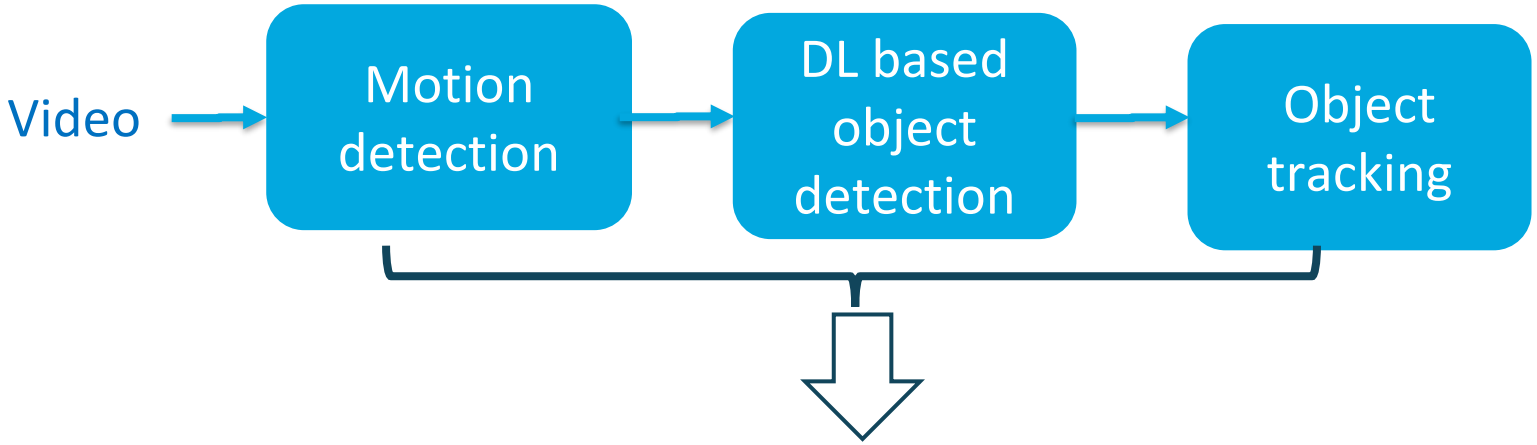
COMCAST

Object Relevancy

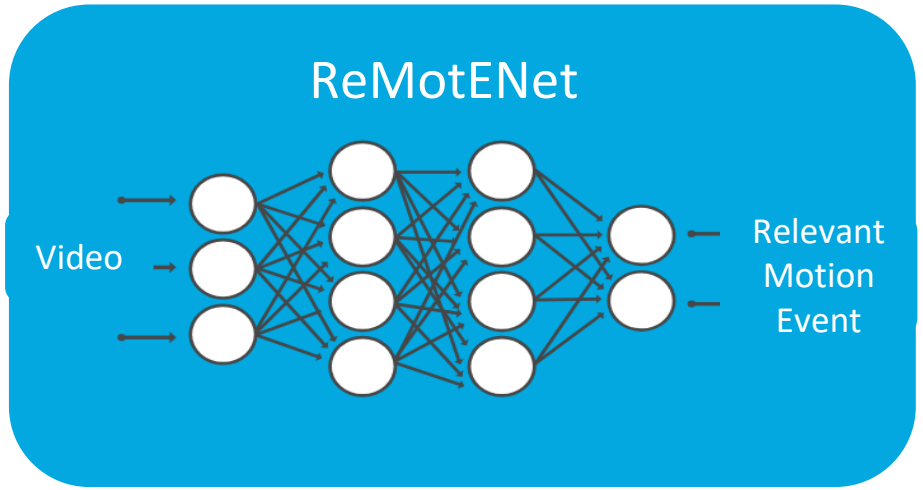
COMCAST

RELEVANT MOTION EVENT NETWORK (REMOTE-NET)

Prior Work:



Our Work:

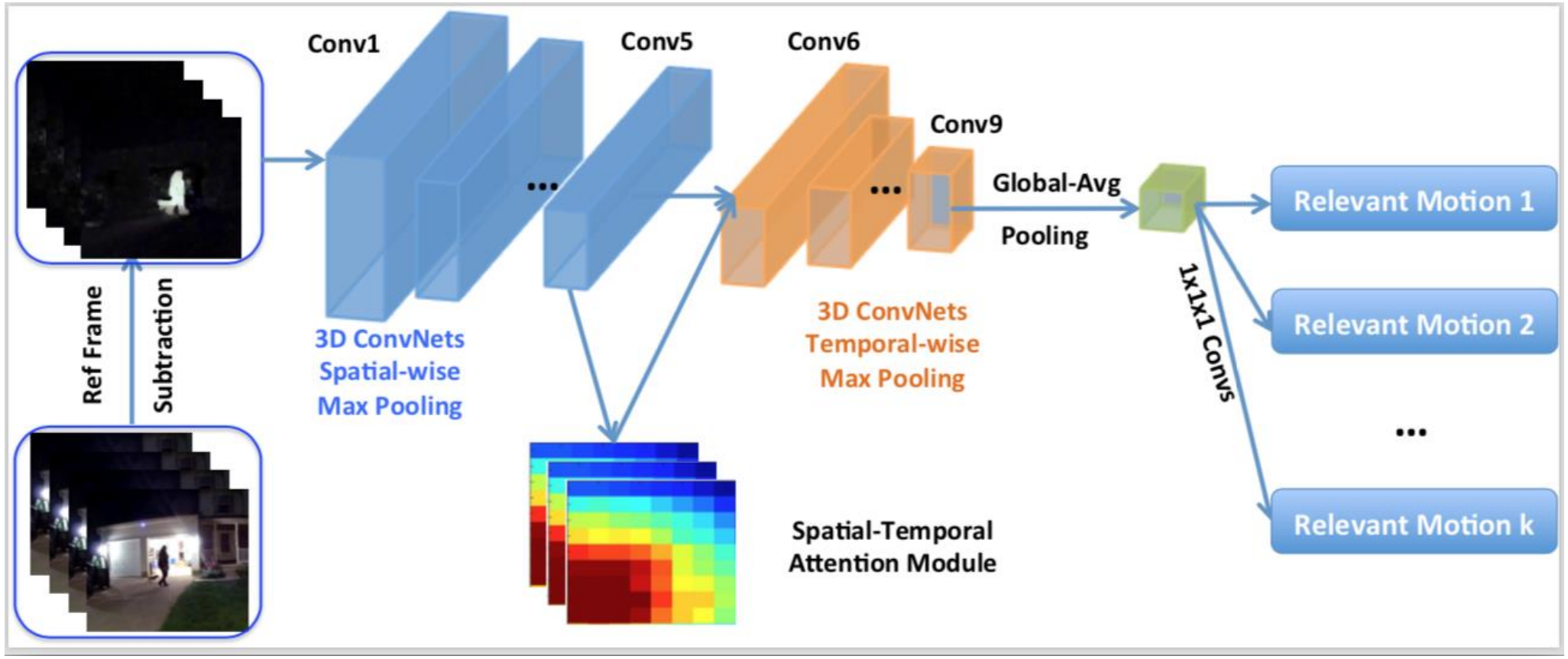


Advantages:

1. End-to-end network
2. Efficient

1. ReMotENet: Efficient Relevant Motion Event Detection for Large-scale Home Surveillance Videos, published at IEEE WACV'2018
2. Relevant Motion Detection in Video, US Patent App. 20,190/244,366, 2019

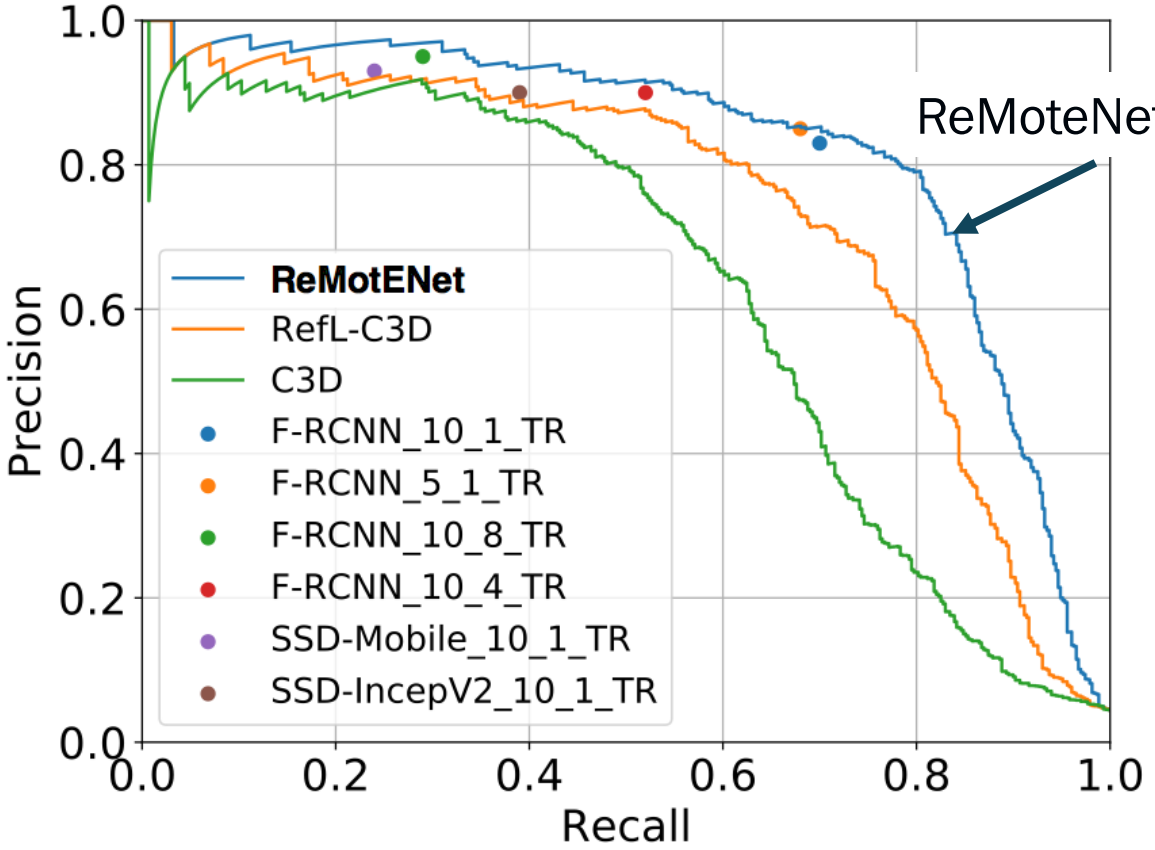
RELEVANT MOTION EVENT DETECTION (REMOTENET)



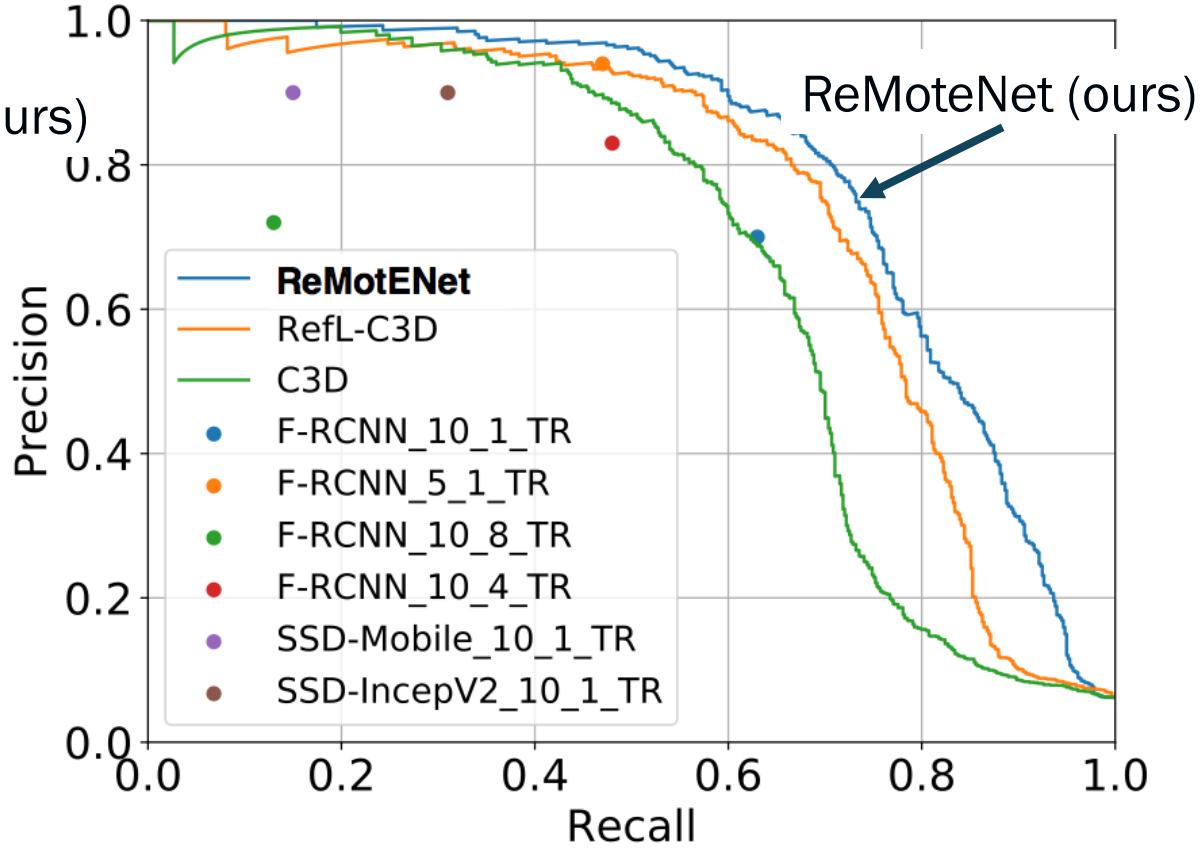
ReMotENet: Efficient Relevant Motion Event Detection for Large-scale Home Surveillance Videos, IEEE WACV 2018

REMOTENET VERSUS "OBJECT DETECTION + TRACKING"

Person Event

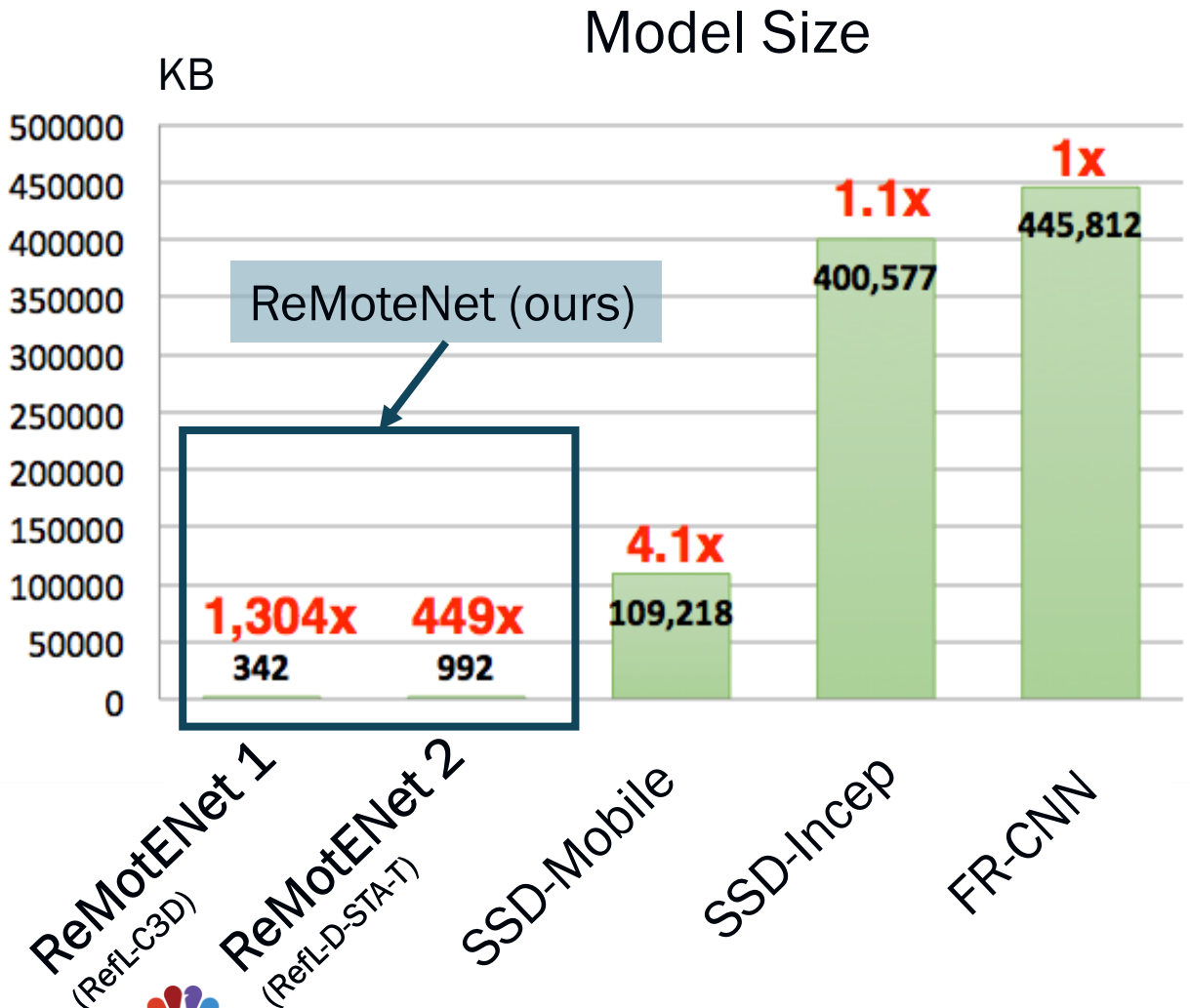


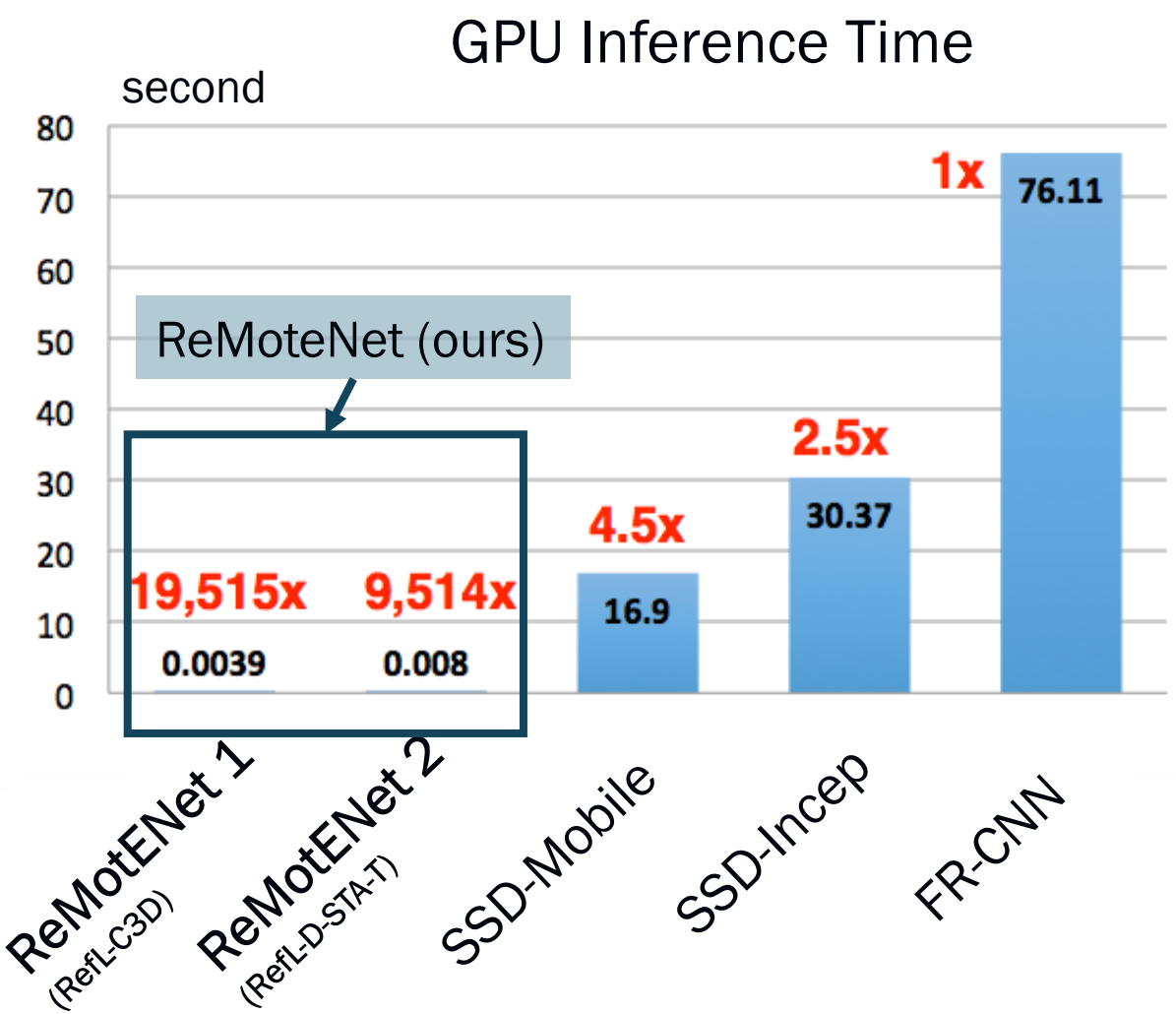
Vehicle Event



COMCAST

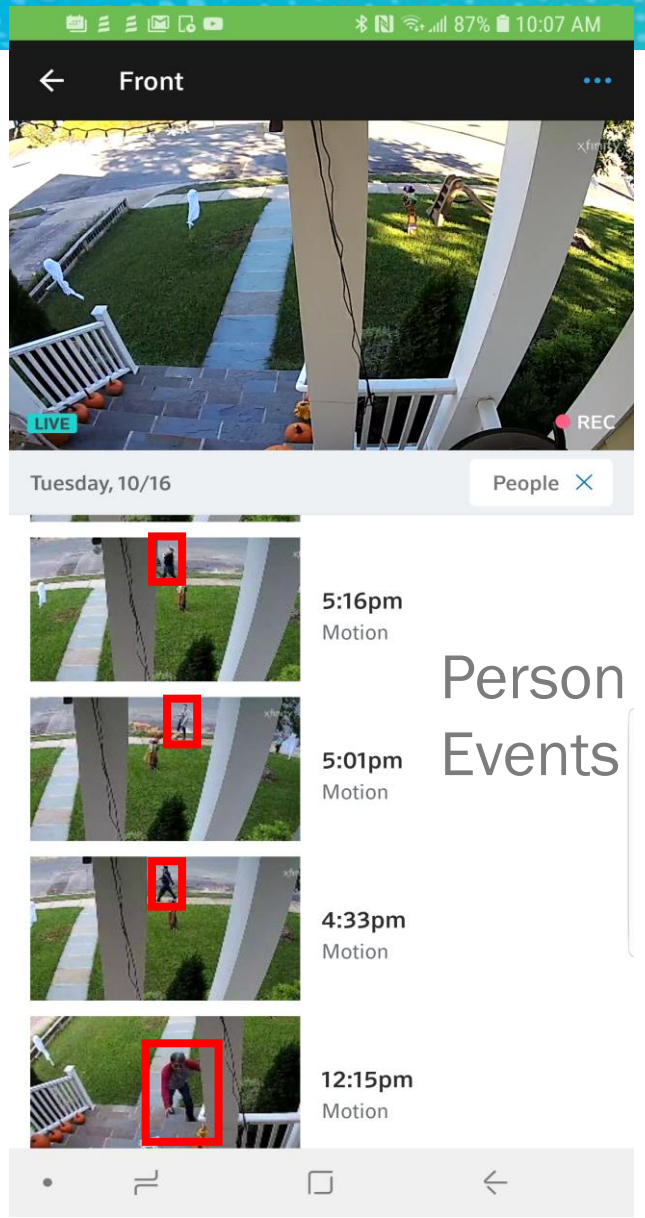
MODEL SIZE AND GPU COMPARISON



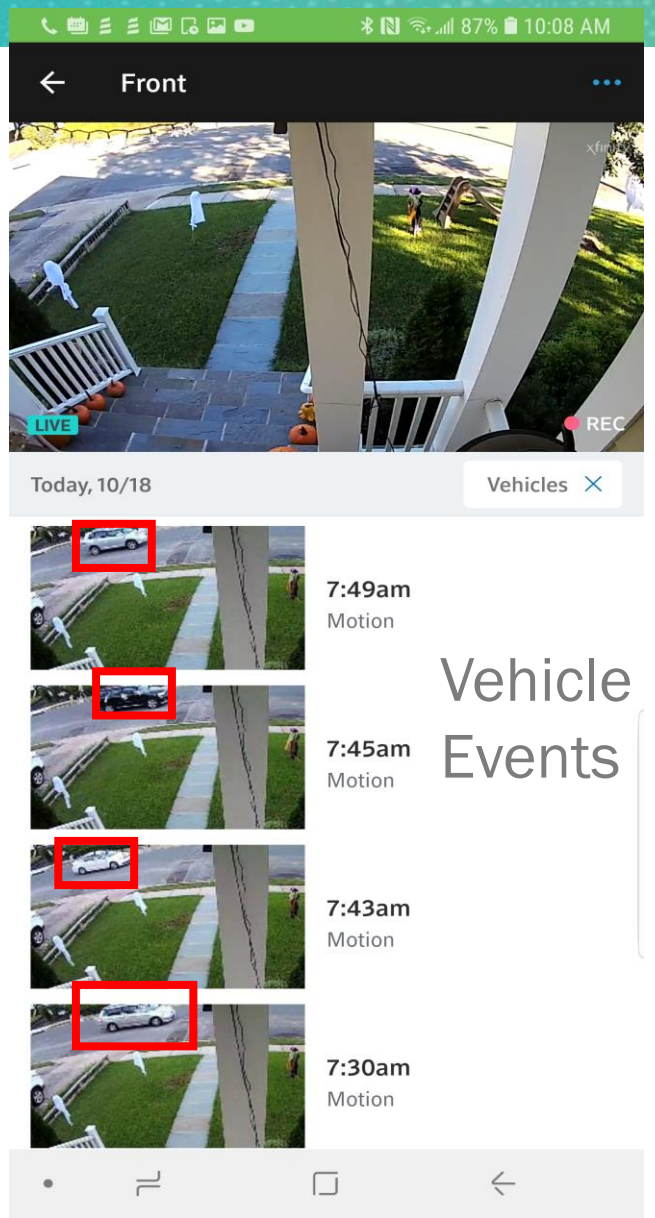


COMCAST

RESULTS



Person Events



Vehicle Events

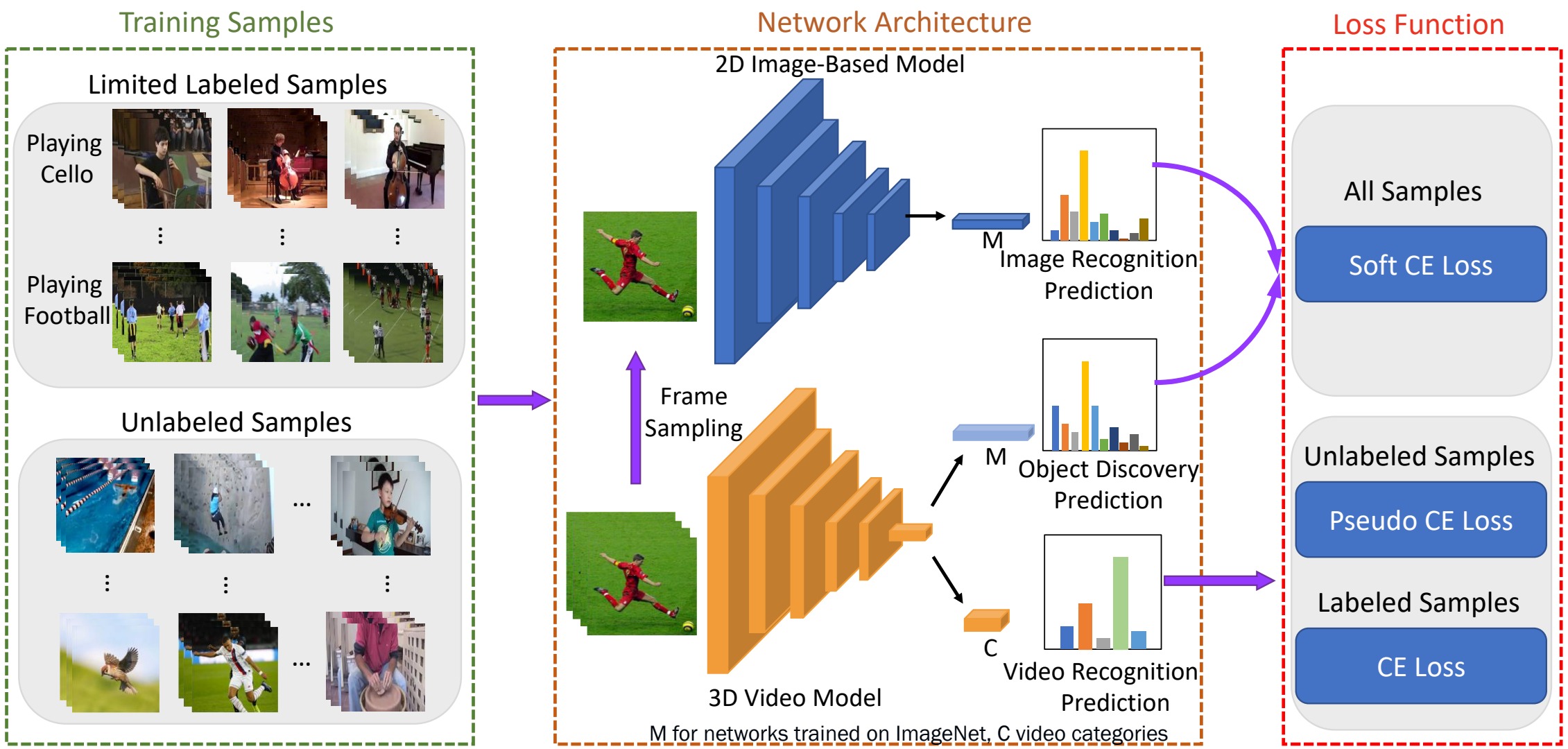
Event Relevancy

COMCAST

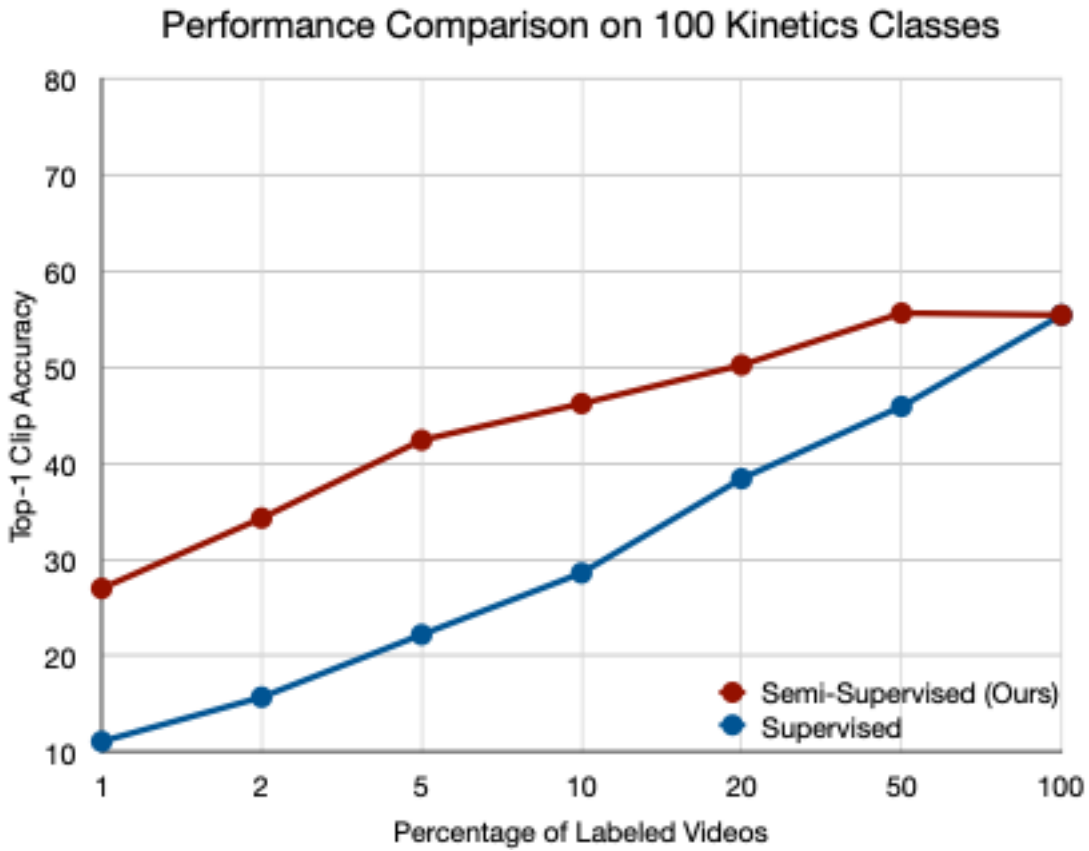
MOTIVATION

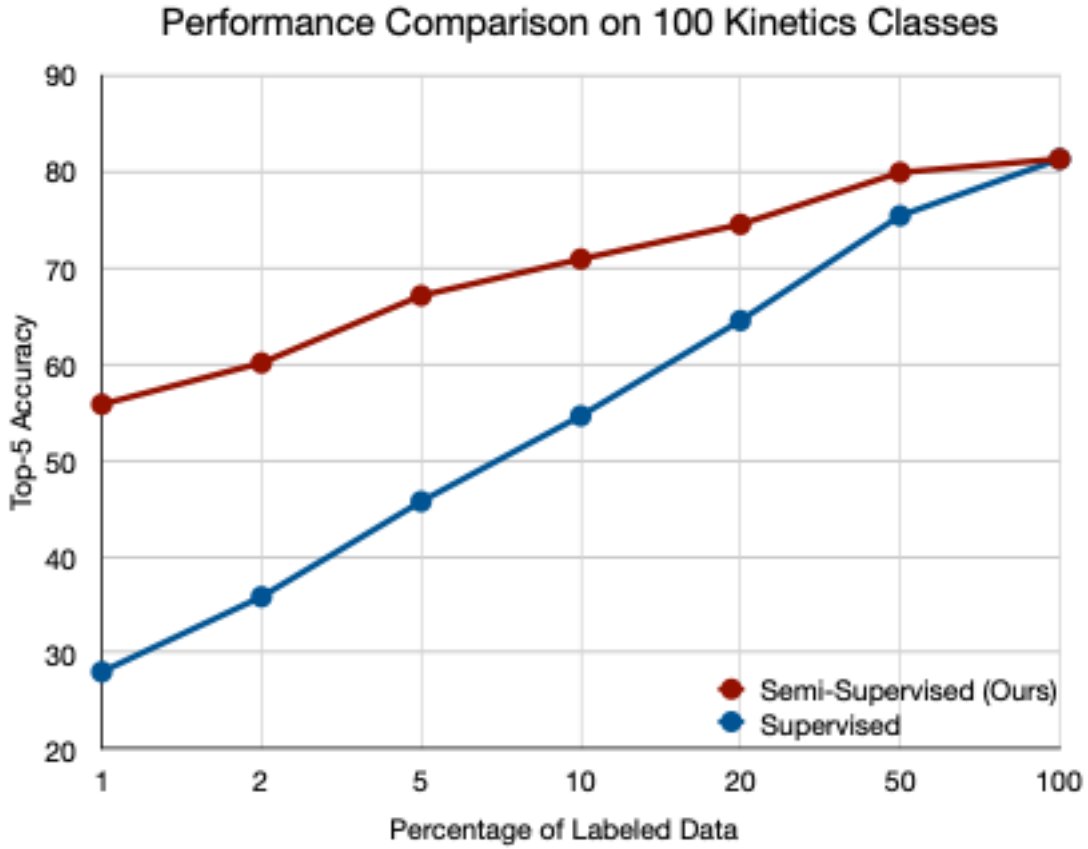
- Events of interest, such as package delivery, are very challenging to detect.
- Annotation of video data is also very time consuming.
- Also limited data are available to train a CNN model
- We propose a self-supervised model for video activity recognition

PROPOSED FRAMEWORK



PERFORMANCE ON KINETICS100 DATASET





With a small percentage of labeled examples, the 3D CNN trained by our proposed semi-supervised method significantly outperforms that trained in supervised setting.

PERFORMANCE ON UCF101 AND HMDB51

UCF101

%Label	Supervised[13]		PL[29]		MT[44]		SD		MT+SD		S^4L [51]		Ours	
	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video
5	15.1	16.9	17.2	17.6	15.3	17.5	29.3	31.2	28.4	30.3	21.0	22.7	30.9	**32.4**
10	21.6	24.0	23.5	24.7	24.0	25.6	38.6	40.7	37.5	40.5	27.1	29.1	40.2	**42.0**
20	30.0	32.2	33.9	37.0	33.4	36.3	42.1	45.4	41.7	45.5	34.7	37.7	46.2	**48.7**
50	35.1	38.3	43.9	47.5	42.5	45.8	49.8	53.9	49.2	53.0	44.9	47.9	51.5	**54.3**

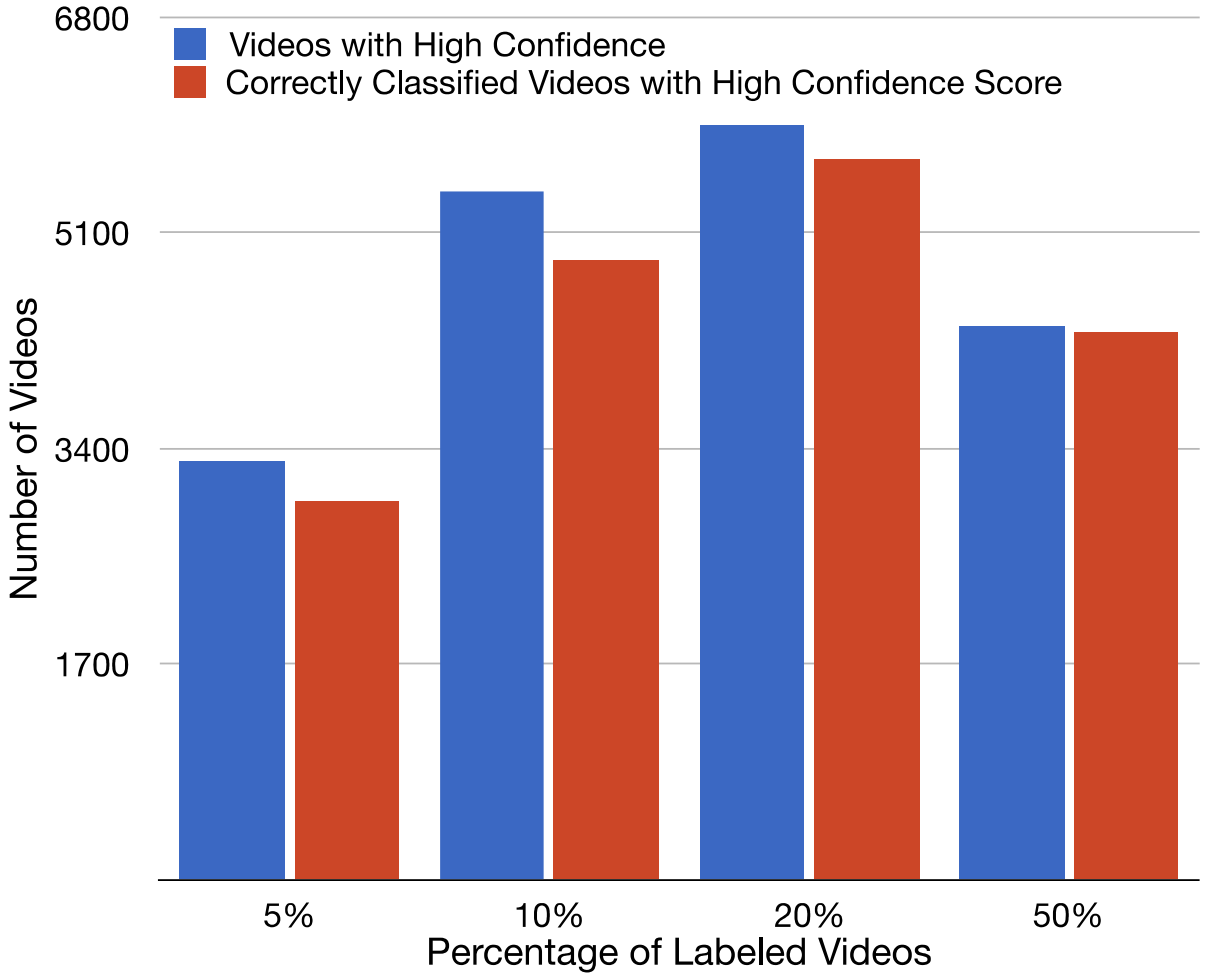
HMDB51

%Label	Supervised[13]		PL[29]		MT[44]		SD		MT+SD		S^4L [51]		Ours	
	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video
40	17.1	18.0	26.3	27.3	26.4	27.2	31.6	32.6	32.1	32.3	28.8	29.8	32.6	32.7
50	29.1	30.7	30.9	32.4	29.2	30.4	34.1	35.1	30.8	33.6	28.9	31.0	34.9	**36.2**
60	30.0	31.2	31.4	33.5	31.1	32.2	35.4	36.3	34.5	35.7	32.5	35.6	35.7	**37.0**

Our method significantly outperform the supervised baseline model on both UCF101 and HMDB51 datasets.

A PROMISING TOOL FOR DATA LABELING

- More than 90% of the videos receiving confident predictions ($P > 0.95$) are correctly classified by CNNs trained by our approach.
- Potential use of the resulting classifier for frameworks to expand datasets through active learning.

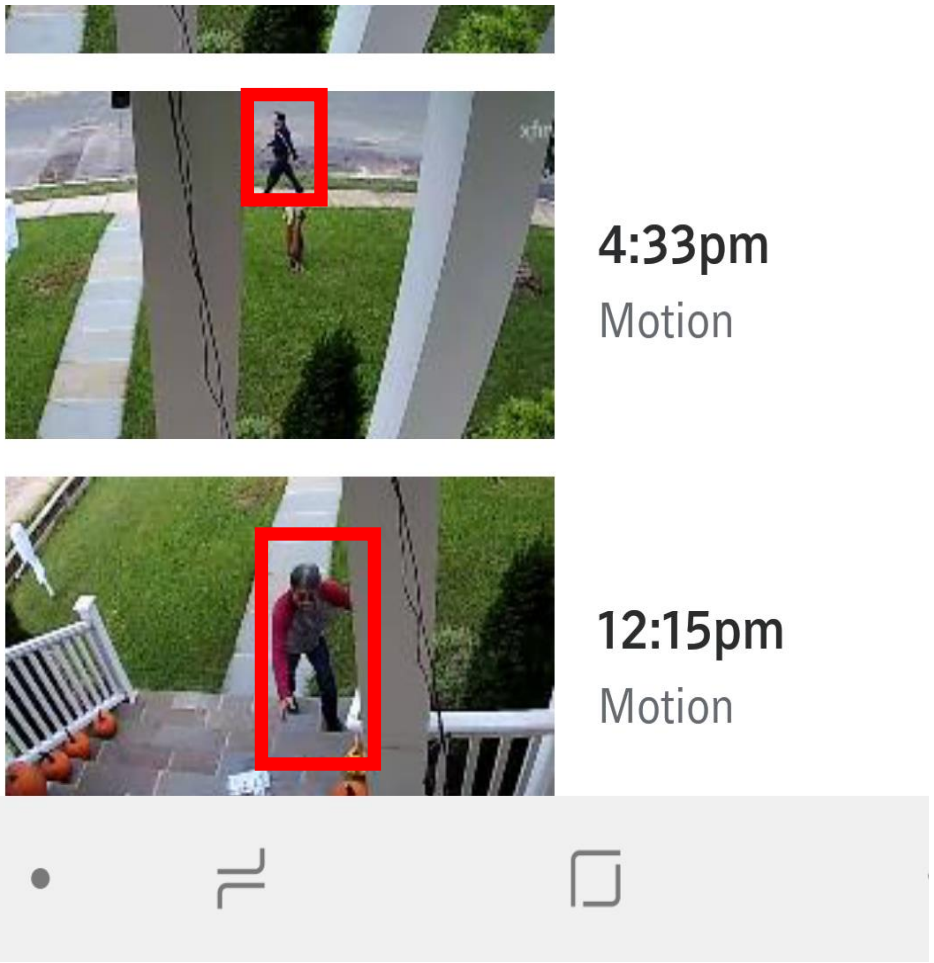


Spatial Relevancy

COMCAST

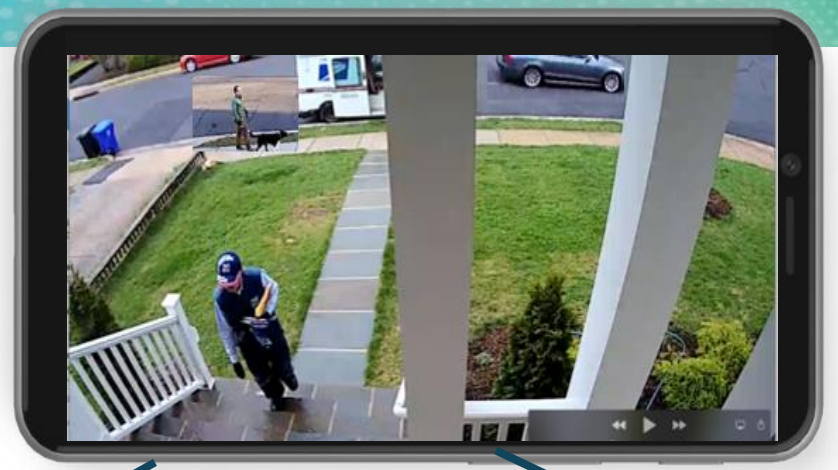
WHAT IS MISSING?

- Not all events are equal
- Not all regions are equal



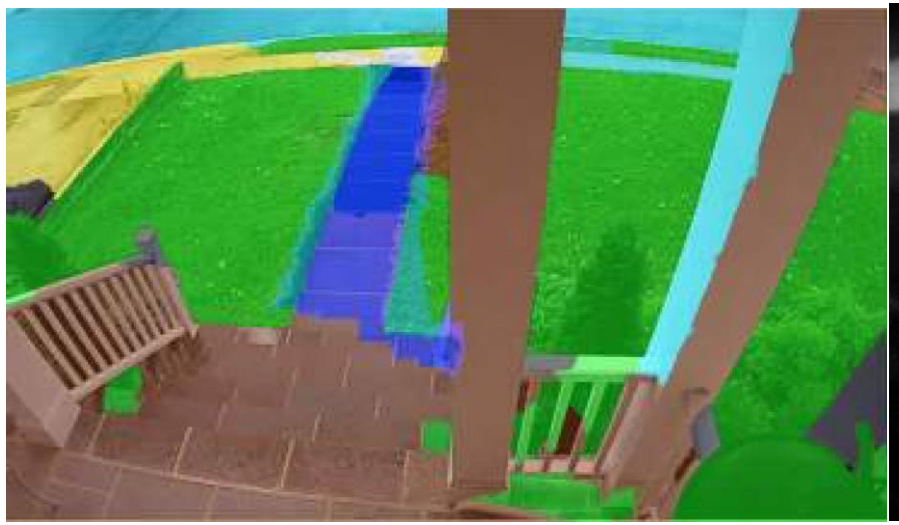
2. SPATIAL RELEVANCY

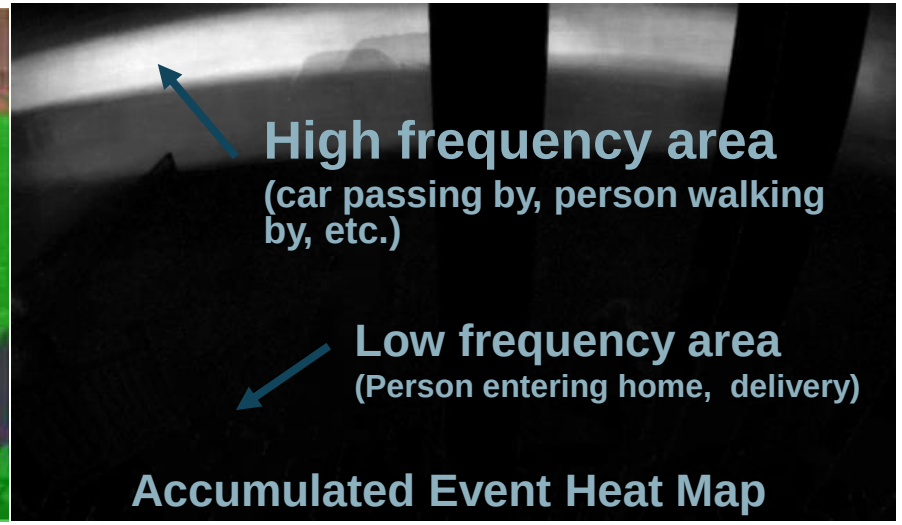
- What: object/event relevancy
- **Where: spatial relevancy**
- When: temporal relevancy



Semantic Regions

Region of Interest

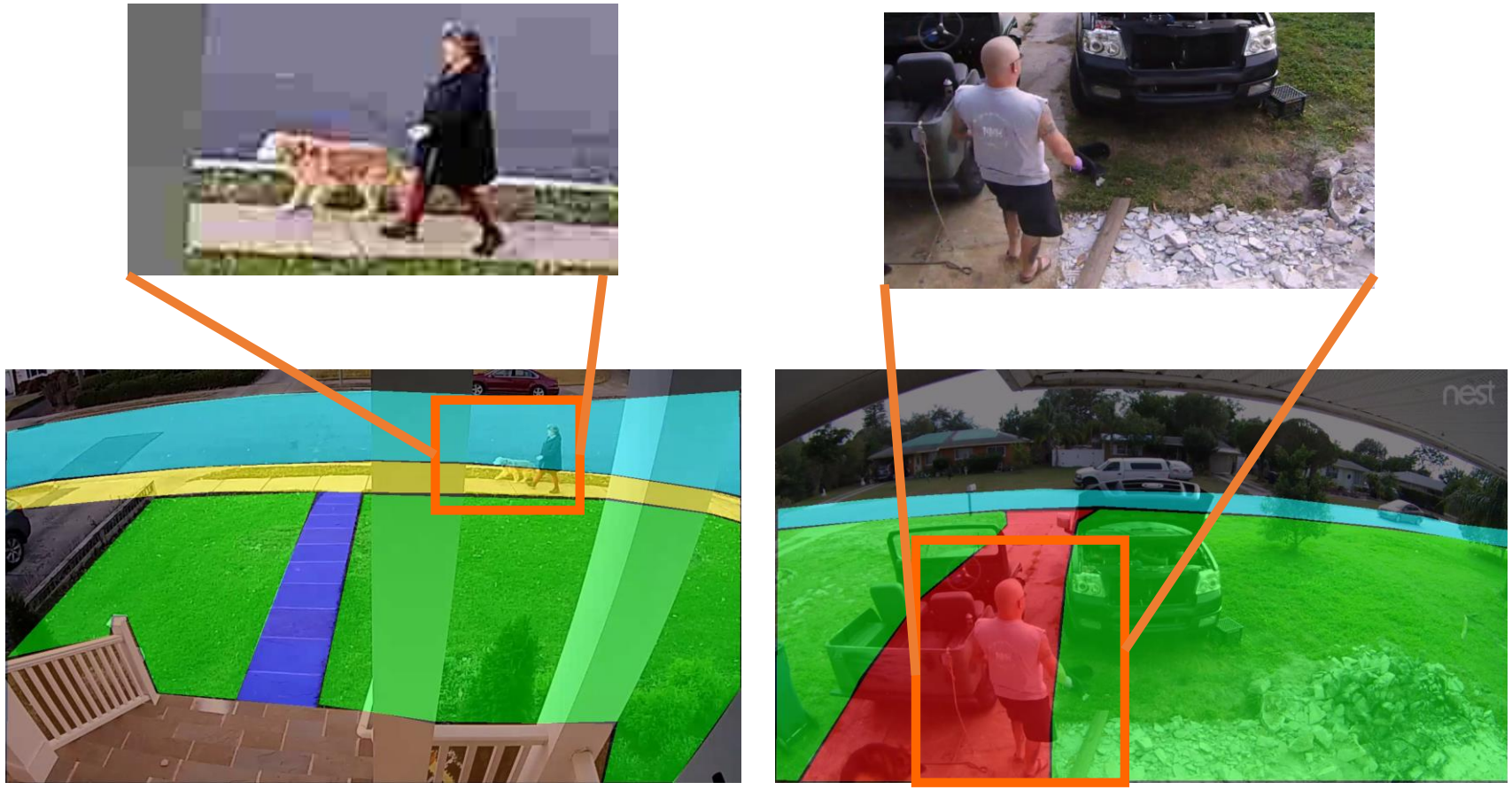




1) Patent pending: Methods and Systems for Determining Object Activity within ROI, 2019
2) Relevant Motion Event Detection with Scene Layout-Induced Video Representation, IEEE ICCV'2019

COMCAST

DETECTION TRIPLET <AGENT, ACTION, PLACE>

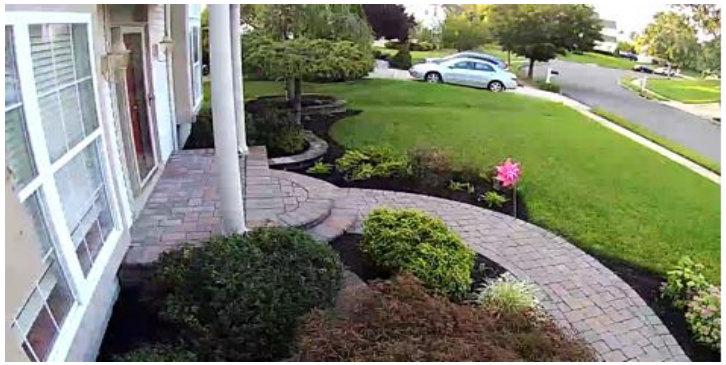


<pet, move along, sidewalk>

<person, move along, sidewalk>

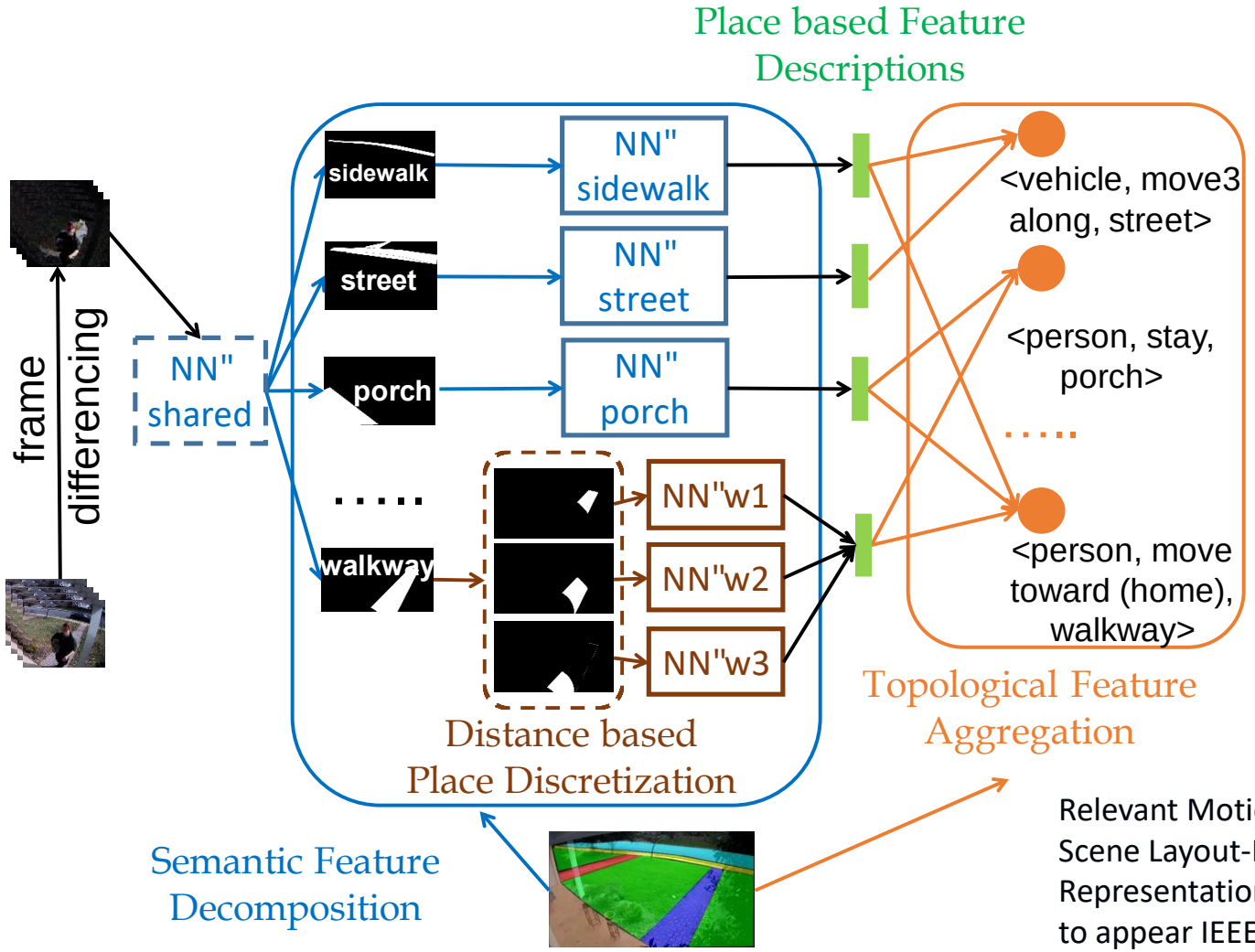
<person, move away (home), driveway>

INCORPORATE SCENE LAYOUT INTO A NETWORK



How can we learn general representations from limited training scenes, then generalize to new scenes?

LAYOUT-INDUCED VIDEO REPRESENTATION (LIVR)



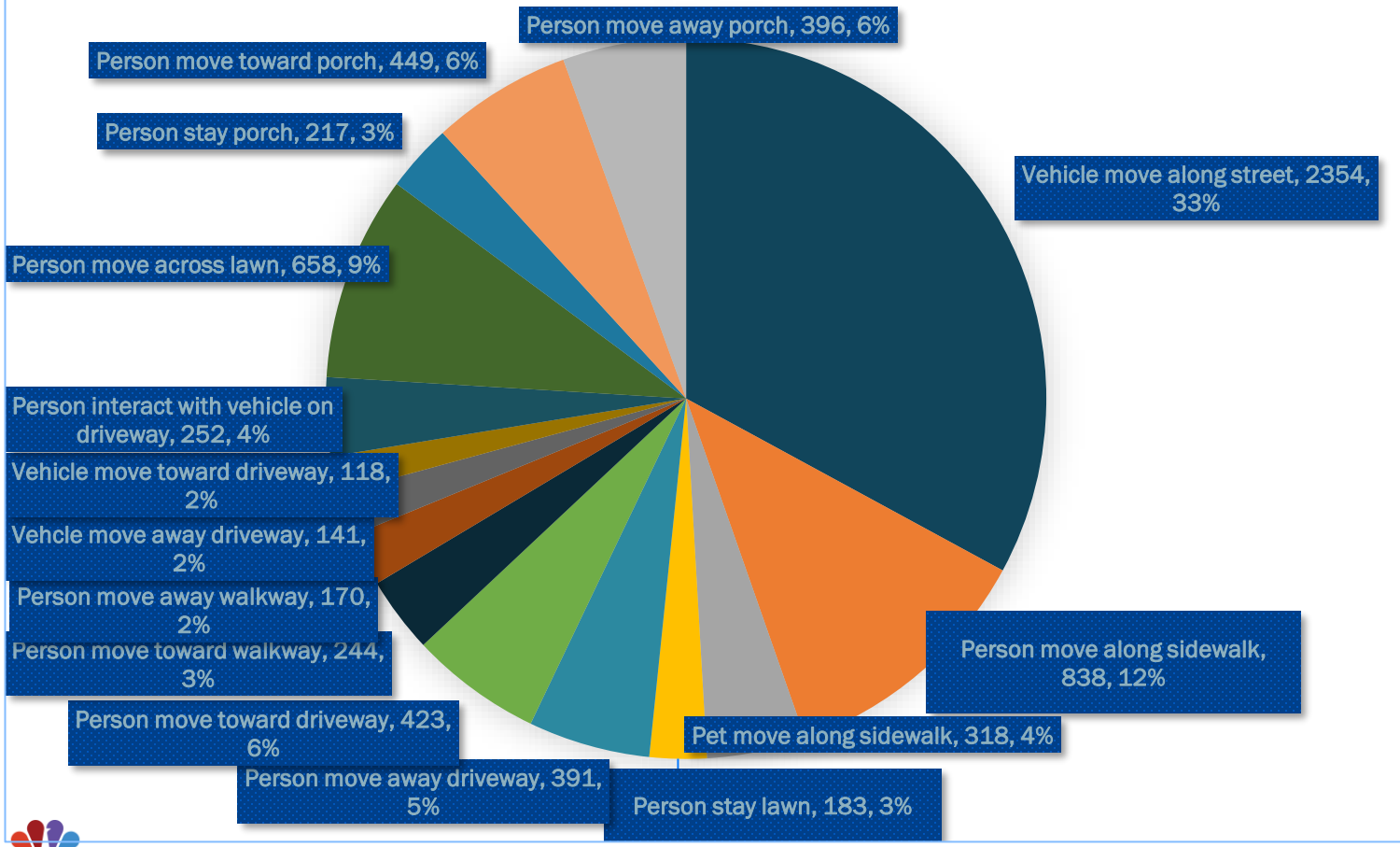
Relevant Motion Event Detection with Scene Layout-Induced Video Representation (LIVR), to appear IEEE ICCV'2019

COMCAST OBJECT-IN-PLACE ACTION DATASET

2020 embedded VISION summit

Events

- 26 outdoor cameras, 5 observed, 21 unseen
- 7152 videos, 57% observed, 43% unseen



COMCAST

COMCAST OBJECT-IN-PLACE ACTION DATASET

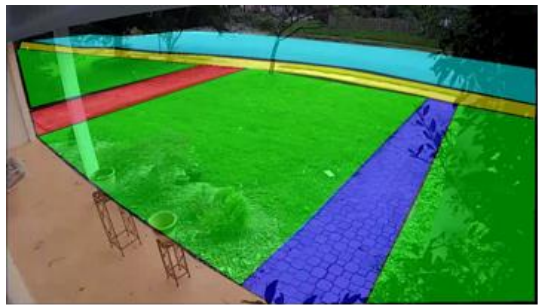
pet_move along_sidewalk
person_move along_sidewalk

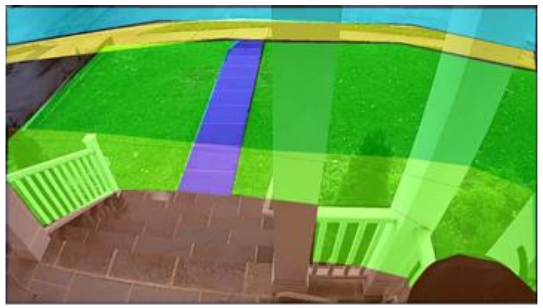
person_move toward (home)
_walkway

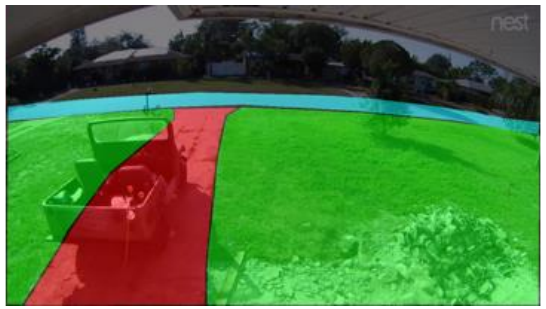
person_stay_lawn

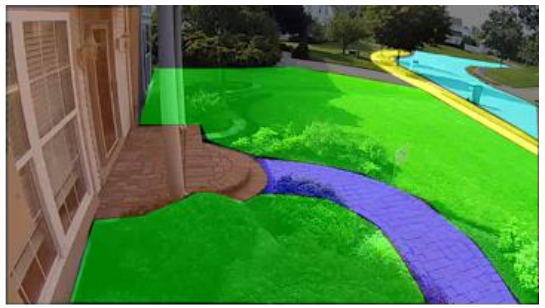
vehicle_move away (home)
_driveway

Example Actions



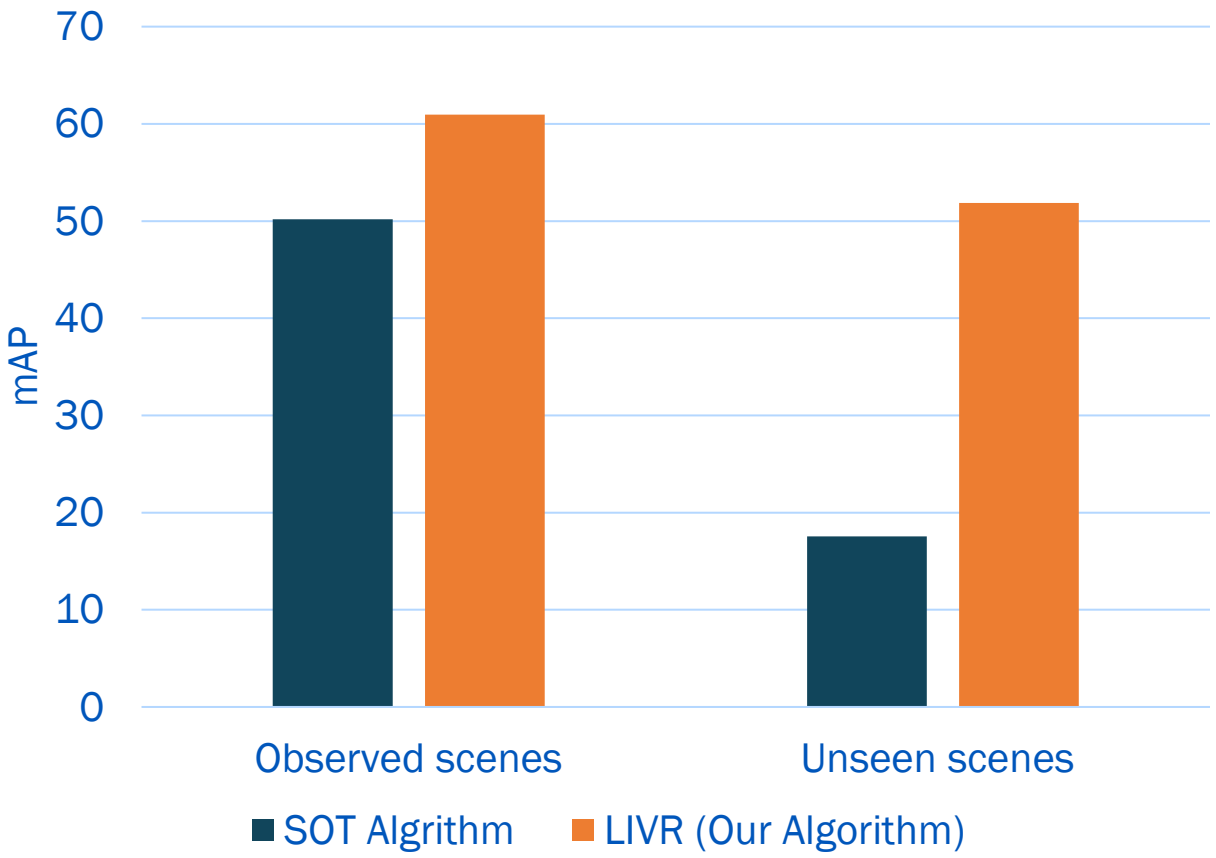






Segmentation Maps

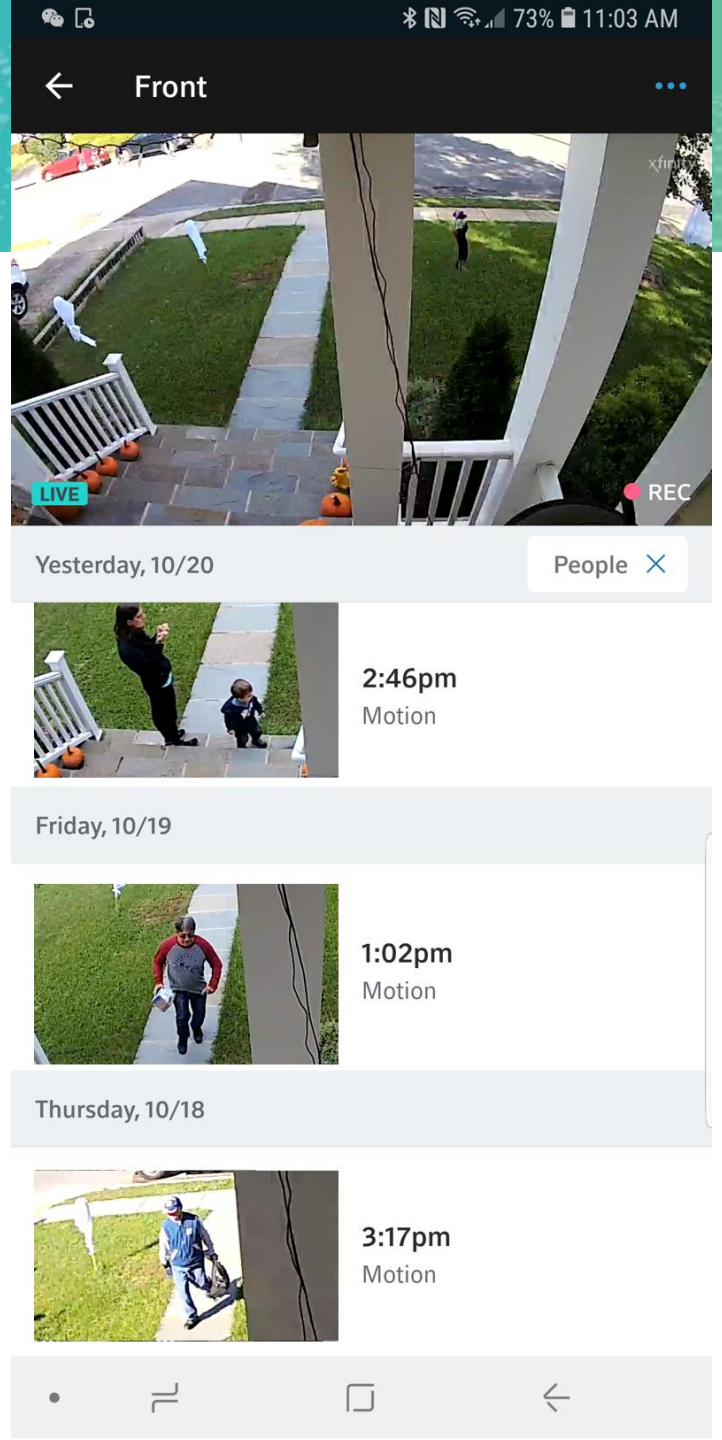
PERFORMANCE EVALUATION



WHAT WE HAVE SO FAR

- Simultaneous understanding of (Object, Place, Action) is important to for smart home video monitoring

- Learning general place based feature descriptions to improve generalization



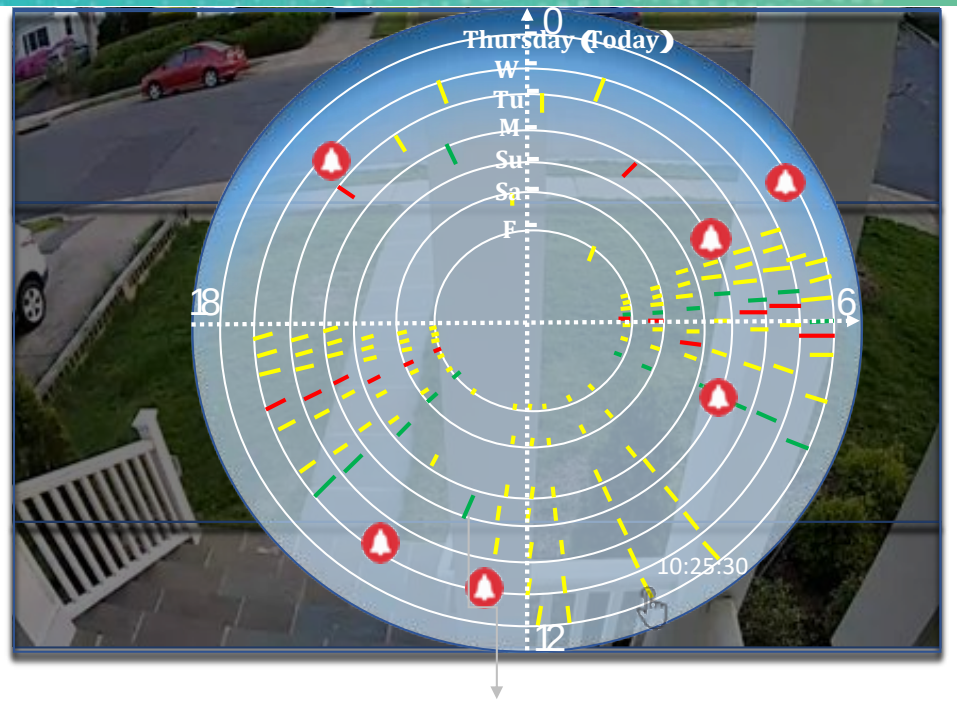
Temporal Relevancy

COMCAST

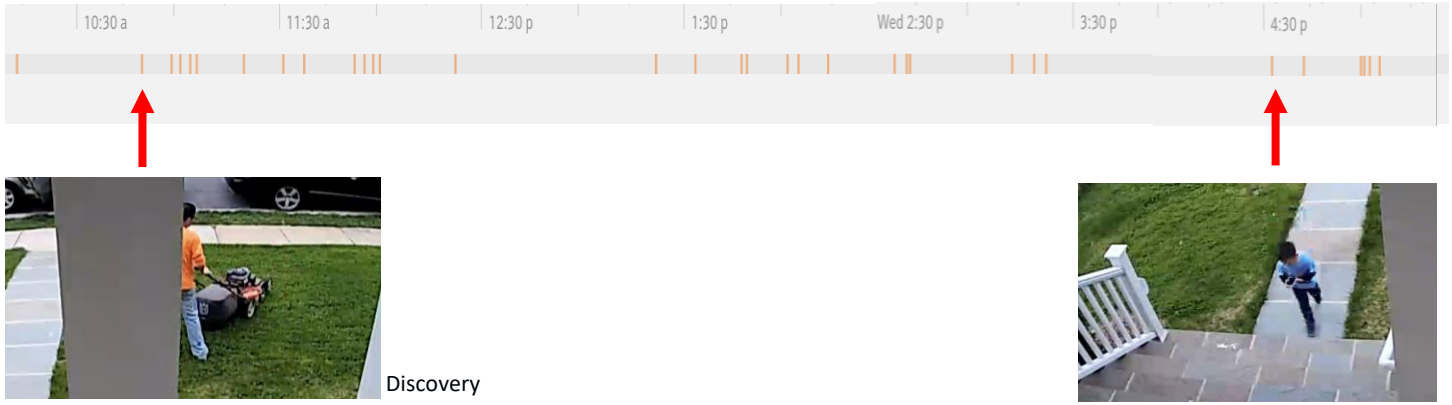
3. TEMPORAL RELEVANCY

- What: object/event relevancy
- Where: spatial relevancy
- **When: temporal relevancy**

Temporal regularity and anomaly:



Temporal sparsity:



Discovery

COMCAST

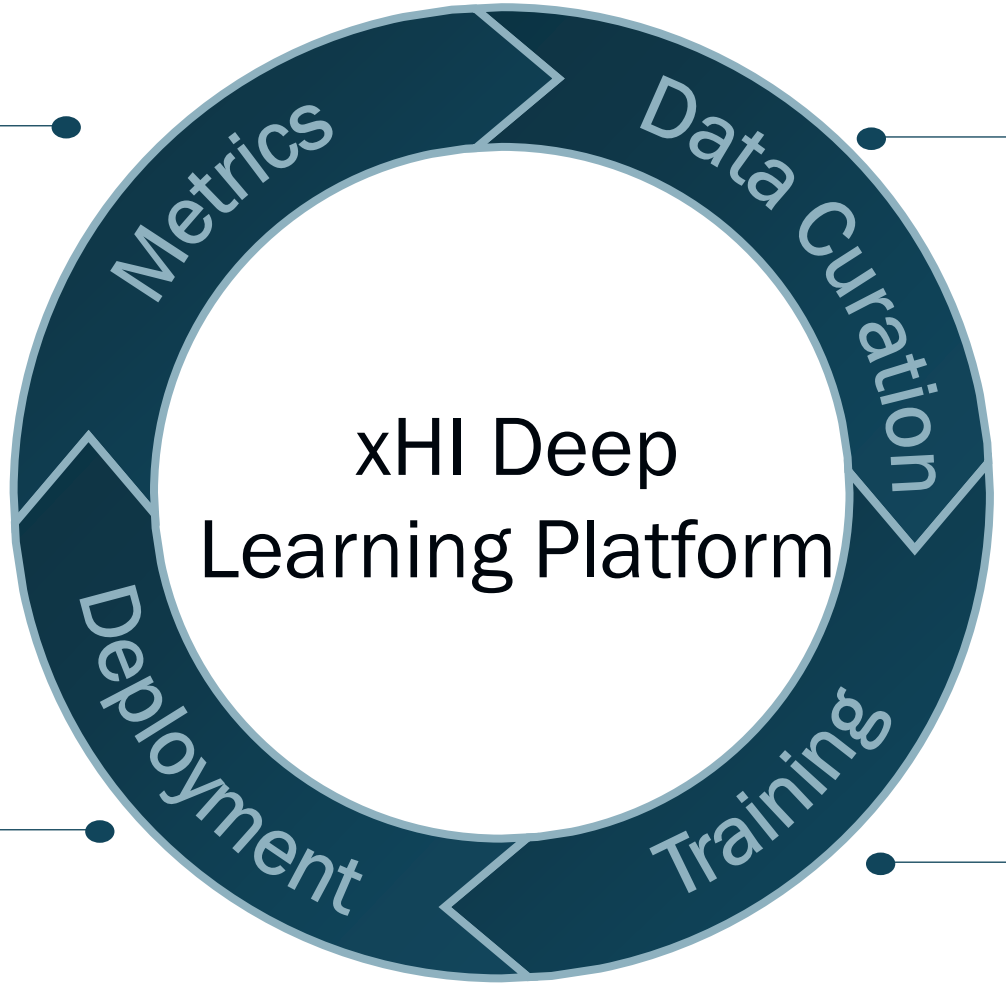
OBJECT/EVENT RELEVANCY
+ SPATIAL RELEVANCY
+ TEMPORAL RELEVANCY

= RELEVANT EVENTS
FOR NOTIFICATION

COMCAST

XHI MACHINE/DEEP LEARNING PLATFORM

xHI ML platform allows us to train, deploy and evaluate our machine learning models efficiently



xHI test platform

RDK AWS

xHI dashboard

CPU/GPU/ARM

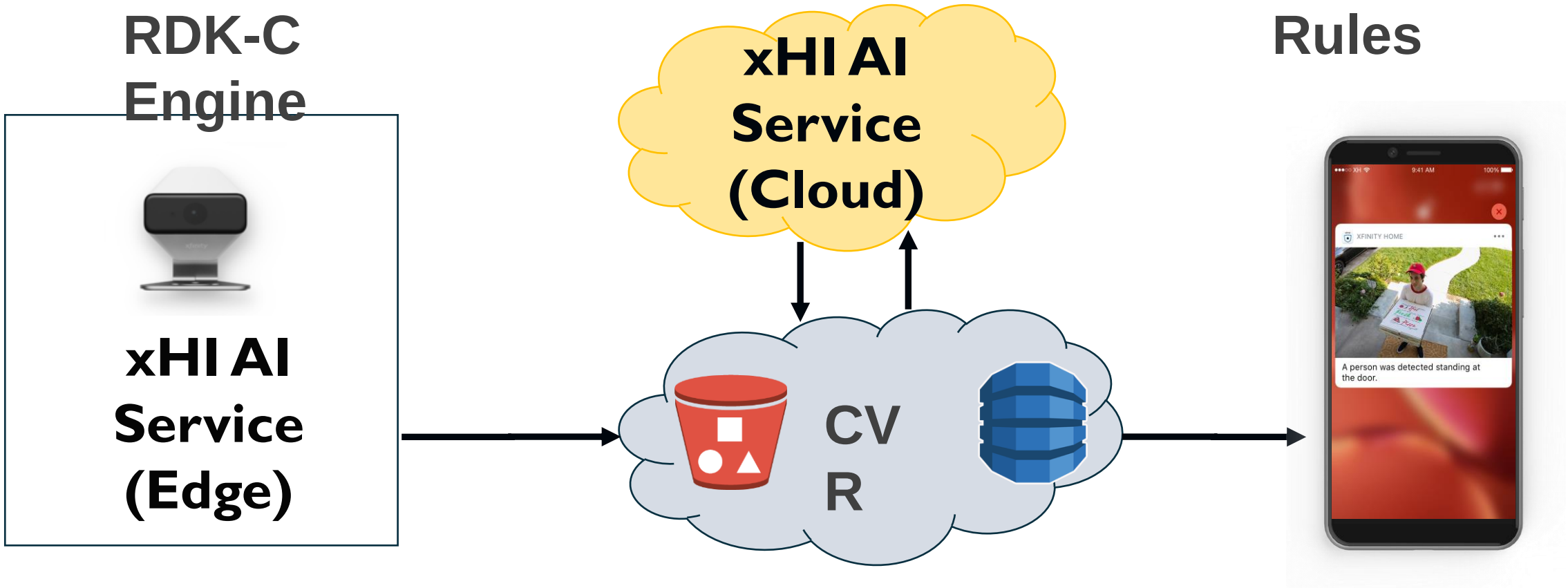
GPU

jupyter

COMCAST

XFINITY HOME INTELLIGENCE (XHI)

xHI is the home-grown AI service using machine learning and computer vision to generate the magic features behind xFinity Home products



Conclusions

- Incorporating object, spatial and temporal context for activity recognition, when data are limited

- Critical to deliver relevant events to smart home customers

COMCAST

Resources

Comcast AI Pages:

https://jobs.comcast.com/ml-ai-team-page

Papers & Data:

1. RemoteNet: **https://arxiv.org/pdf/1801.02031.pdf**
2. LIVR: **https://arxiv.org/pdf/1804.01429.pdf**
3. VideoSSL: **https://arxiv.org/pdf/2003.00197.pdf**

COMCAST

Questions?