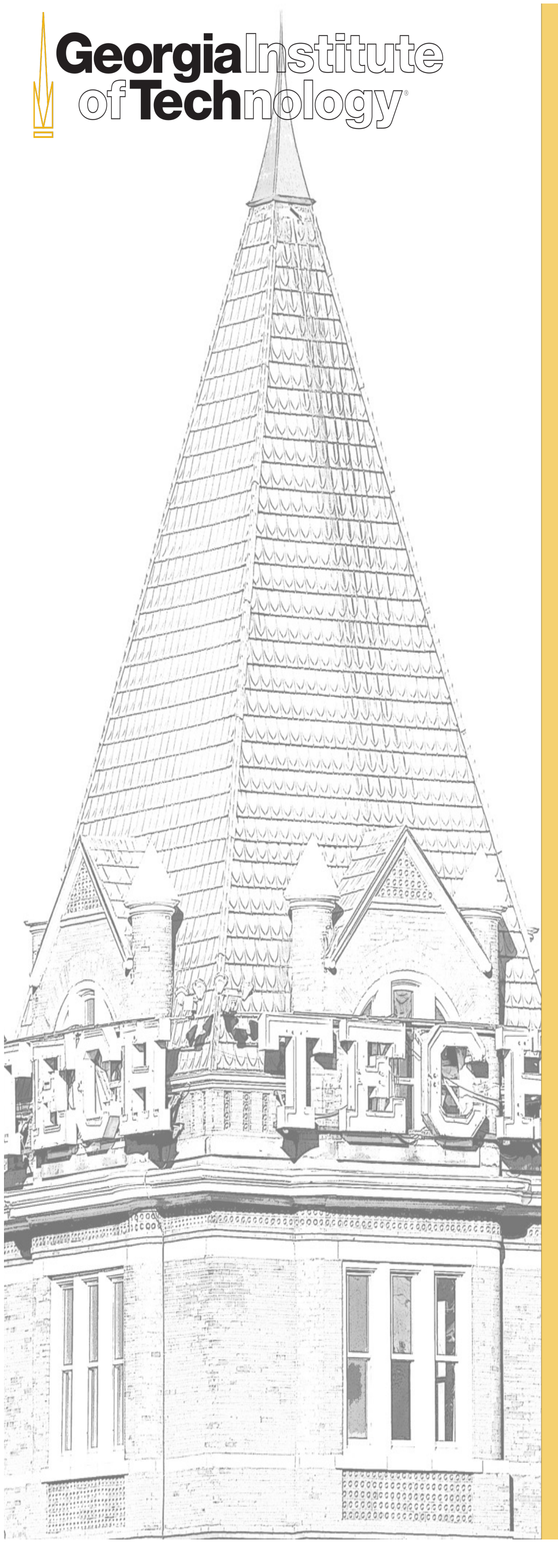


Edge Video Analytics

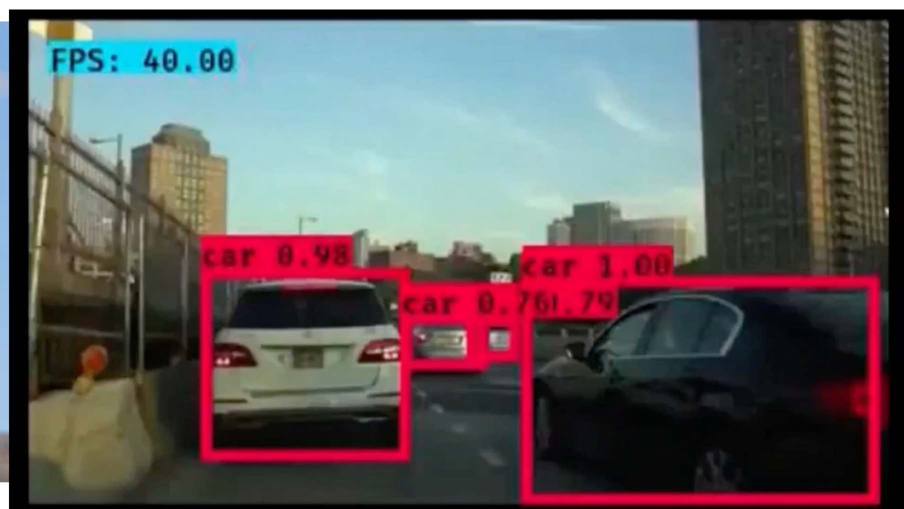
Prof. Dr. Ling Liu
School of Computer Science
Georgia Institute of Technology
www.cc.gatech.edu/~lingliu/

Acknowledgement: This is a joint work with Ka Ho Chow,
Yanzhao Wu, Margaret Loper, Emre Gursoy., Stacey Truex, Wenqi Wei
Partial funding support: NSF, CISCO, IBM

Video Analytics and Object Detection

- Video Cameras are everywhere
  - every cellphone, every vehicle, every human
  - every building, every street, every highway …
- Object Detection: a core perception for video analytics

http://www.firsttoyreviews.com/drones-taking-the-future/



https://kjzz.org/content/1318066/phoenix-red-light-and-speed-cameras-end-dec-31

Video Analytics: Device-Edge-Cloud Continuum

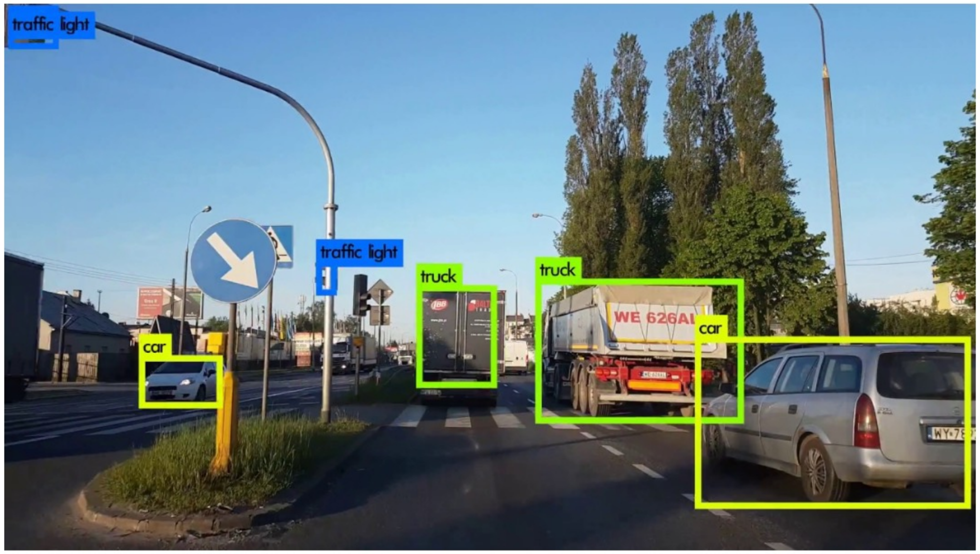
- **Video analytics is typically done in the Cloud**

(1) Overwhelming Demands for Bandwidth

- Shipping all the videos to the Cloud is NOT scalable
    - Netflix: ~ 3GB/hr of HD video → 6.8 Mbps per stream (recom: 25 Mbps)
    - London is estimated to have > 500,000 surveillance cameras

(2) Privacy concerns

- **Video Analytics on the Edge**
    - Distributed Learning & Inference on the Edge



Challenges of Edge Video Analytics

Unlike Cloud,

- Edge is resource limiting & little elasticity
- Edge is more exposed and more vulnerable to
  - **Systemic disruptions**
    - contention induced delay, performance/accuracy degradation
    - poor input data induced inference errors (e.g., poor lighting, foggy weather, convoluted objects, network jitter, …)
  - **Adversarial disruptions** (inference / training)
    - Security violation
    - Privacy violation

Systemic Disruption in Edge System (1)

Georgia Tech

- Edge Client may be sensitive to contention/load surge at edge server and WiFi bandwidth saturation.
  - **Degradation Effects:**
    - Server content induced random dropping of device-edge offloading operation
    - Bandwidth saturation induced blocking of device to edge offloading operations
  - **Solution Approach: Data Reduction Techniques**
    - utility-preserving importance sampling
    - utility-preserving region-of-interest based pruning

Systemic Disruption in Edge System (2)

- Edge Client (e.g., end-devices) may not be capable of running a full precision model for video analytics.
  - **Multiple Reasons:**
    - Limited resources (compute/storage)
    - Privacy concerns (sensing data are proprietary)
  - **Solution Approach: Model Reduction Techniques**
    - Model Reduction through gradient compression or NN pruning
      - to produce model of reduced sizes and complexity while maintaining good accuracy on-par to the high-fidelity model used in a centralized cloud setting
    - Distributed multi-fidelity collaborative DNN approach to learning and inference

Systemic Disruption: Model Compression

- **Low rank filter-based model compression**
  - All the gradients are sorted and only remove x% at low rank and the rest is zeroed. Only gradients larger than a threshold are to be transmitted in full precision. The rest is zeroed. x% is set as the control knob.

- **Model Reduction by gradients compression**

no gradient compression

Global model trained with higher accuracy using local model compression than no compression

benign acc	0	1%	10%	20%	30%	40%	50%	70%	80%	90%
LFW	0.695	0.697	0.705	0.701	0.71	0.709	0.713	0.711	0.683	0.676
CIFAR100	0.67	0.673	0.679	0.685	0.687	0.695	0.689	0.694	0.676	0.668
CIFAR10	0.863	0.864	0.867	0.872	0.868	0.865	0.868	0.861	0.864	0.859
MNIST	0.9568	0.9567	0.9577	0.957	0.9571	0.9575	0.9572	0.9576	0.9573	0.9556

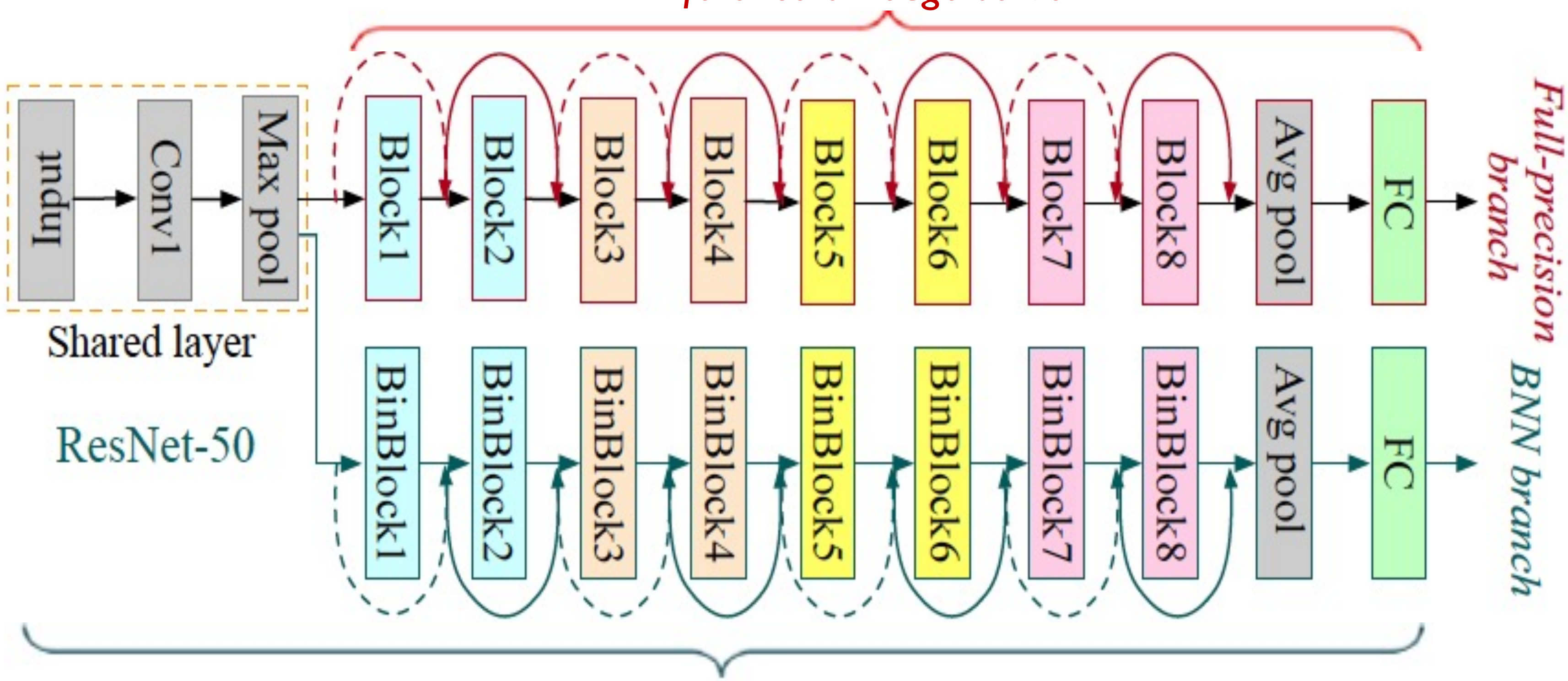
Benign accuracy of four datasets with varying compression rates

[8] ESORCS 2020
Code :https://github.com/git-disl/ESORICS20-CPL

Systemic Disruption: Multi-Fidelity Adaptation

- Model Reduction through Multi-Fidelity Adaptation
  - Use the independent light weight BNN branch to focus on the simple tasks
  - Uses the full precision backbone to correct the error of the BNN branch through dynamic adaptation.

Inference on edge server



Inference on edge client (end device)

[13] Infocom2020

Systemic Disruption in Edge System (3)

- Object detection at edge may result in low throughput and high latency due to the mismatch between incoming video streaming rate (FPS) and detection processing rate (FPS)

Throughput Problems in Edge Video Analytics

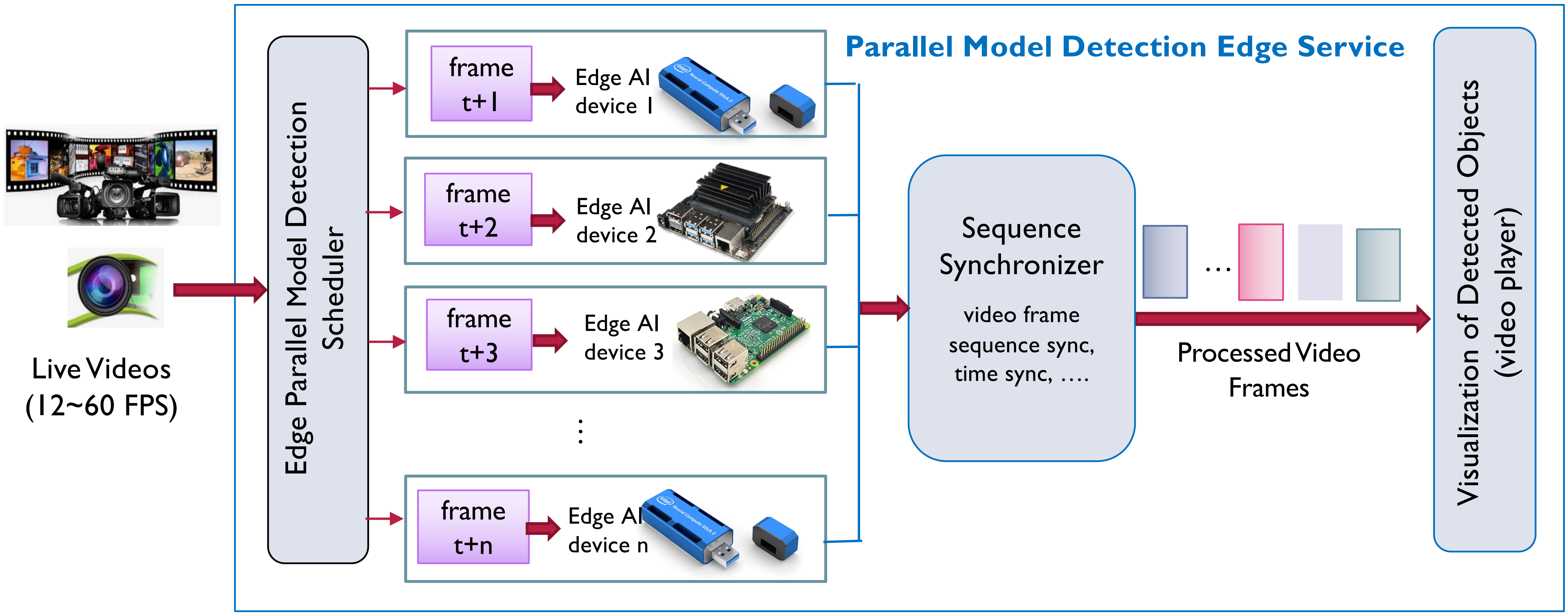
https://drive.google.com/file/d/13nOsA-9RMeYdeAG5nmTvuwzPESVTwmNa/view?usp=sharing

Systemic Disruption in Edge System (3)

- Object detection at edge may result in low throughput and high latency due to the mismatch between incoming video streaming rate (FPS) and detection processing rate (FPS)
  - ✓ Solution Approach: Parallel Detection Processing

    Leveraging AI hardware or fast network like 5G, 6G

Fast Edge Video Analytics by Exploiting Multi-model Detection Parallelism



Single Edge node attached with multiple AI-hardware devices, each runs one detection model

1. Round Robin
2. FCFS

[1] IEEE CogMI 2021

Performance of a single NCS (mAP)

Model		YOLOv3
mAP (%, No dropping)	ADL-Rundle-6	62.5
	ETH-Sunnyday	86.9
mAP (%, Dropping)	ADL-Rundle-6	42.7
	ETH-Sunnyday	66.1

Impact on mAP

Original Video (ETH-Sunnyday, 14 FPS)

With no frame dropping (slow, 2.6 FPS)

With frame dropping (low precision, 14 FPS)

Experimental Results (round robin scheduler)

ADL-Rundle-6

Input Video FPS (λ): 30

#Frames: 525

Single NCS2: μ=2.3 FPS

Offline mAP (%): SSD300: **54.4**, YOLOv3: **62.5**

$$n = \lceil \lambda/\mu \rceil \sim \lceil 30/2.3 \rceil \geq 13$$

Table 2: Experiments with Multiple NCS2 Sticks (ADL-Rundle-6)

Processing		Offline	Online						
Model	#NCS2	1	1	2	3	4	5	6	7
SSD300	Detection FPS	2.3	2.3	4.6	6.9	9.1	11.5	13.7	16.0
	mAP (%)	54.4	46.7	**56.2**	**55.8**	**55.4**	**55.7**	**55.7**	**54.7**
YOLOv3	Detection FPS	2.5	2.5	5.1	7.5	10.0	12.5	14.8	17.3
	mAP (%)	62.5	42.7	56.7	61.2	**62.7**	**62.7**	**62.7**	**62.7**

No-frame dropping

Random dropping

Experiment setup: 7 Intel NCS2 sticks, installed on an edge node
with an Intel i7-10700K CPU, 24GB main memory and Ubuntu 20.04.

Experiments: Multiple detection models on heterogeneous AI hardware devices

- Object Detection Hardware
  - Fast edge node:
    - CPU: Intel i7-10700K (8 cores, desktop)
    - CPU Memory: 24 GB
  - Slow edge node:
    - CPU: AMD A6-9225 (2 cores, laptop)
    - CPU Memory: 12 GB
  - 7 Intel NCS2 sticks
- Test Videos:
  - ETH-Sunnyday
  - https://motchallenge.net/vis/ETH-Sunnyday/
  - Video FPS: 14
  - #Frames: 354
- Evaluation Metrics
  - Detection FPS

Experiments with 8 detection models in parallel (Round Robin Schedule v.s. FCFS Scheduler)

Georgia Tech

Detection FPS	#NCS2	0	1	2	3	4	5	6	7
Round-Robin	NCS2	-	2.5	5.1	7.5	10.0	12.4	14.8	17.3
	Fast CPU + NCS2	13.5	5.1	7.6	10.1	12.7	15.0	17.6	**20.1**
	Slow CPU + NCS2	0.4	0.9	1.3	1.8	2.2	2.6	3.1	3.4
FCFS	NCS2	-	2.5	5.1	7.5	9.9	12.5	15.0	17.3
	Fast CPU + NCS2	13.5	16.0	17.1	19.4	22.0	24.3	26.7	**29.0**
	Slow CPU + NCS2	0.4	3.0	5.5	7.8	10.3	12.7	14.9	17.9

- Detection Model: YOLOv3
- Edge node + 7 NCS2 attached via USB 3.0:
  - RR: balanced workloads, but the slowest device will be the bottleneck
  - FCFS: better performance thanks to workloads-aware adaptation

[1] IEEE CogMI 2021

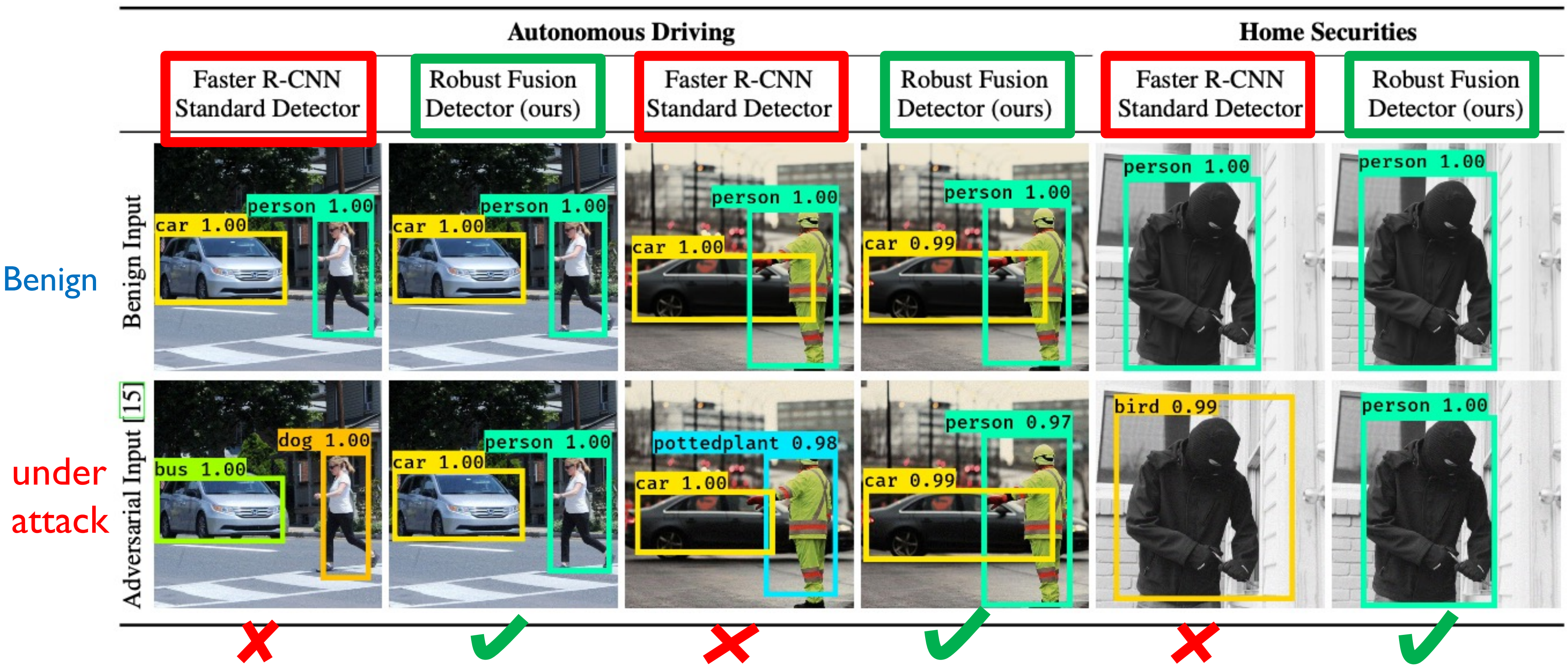
Challenges of Edge Video Analytics

- Systemic disruptions
  - Contention induced delay, performance/accuracy degradation
  - Low-value data offloading induced inference errors (e.g., poor lighting, foggy weather, convoluted objects, network jitter, …)
  - Mismatch between incoming stream rate and the detection processing throughput (#frames per second – FPS)
- Adversarial disruptions (inference phase + training phase)
  - Security violation [2-6]
  - Privacy violation [7-11]

Adversarial Robustness of Object Detection

Object Detection on three images with the standard detector Faster RCNN

Object Detection on the same images **with the robust fusion detector**



[5] SIGKDD 2021
[6] CVPR 2021

Reference

1. Yanzhao Wu, Ling Liu, and Ramana Kompella. Parallel Detection for Efficient Video Analytics at the Edge. Proceedings of the IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, 2021. http://arxiv.org/abs/2107.12563
2. Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data Poisoning Attacks Against Federated Learning Systems. **Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS 2020)**, Guildford, United Kingdom, September 14-18, 2020.
3. Ka Ho Chow, Ling Liu, Emre Gursoy, Stacey Truex, Wenqi Wei and Yanzhao Wu. Understanding Object Detection Through An Adversarial Lens. **Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS 2020)**, Guildford, United Kingdom, September 14-18, 2020.
4. Ka Ho Chow, Ling Liu, Margaret Loper, James Bae, Emre Gursoy, Stacey Truex, Wenqi Wei and Yanzhao Wu. "Adversarial Objectness Gradient Attacks on Real-time Object Detectioni **Intelligent Systems and Applications (TPS 2020)**, Dec. 1-3, 2020, Virtual Conference.
5. Ka Ho Chow and Ling Liu. ``Robust Object Detection Fusion Against Deception", Proceedings of **ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2021)**
6. Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka Ho Chow, Wenqi Wei. ``Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics", Proceedings of **IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2021)**, June 19-25, 2021.
7. Wenqi Wei, Ling Liu, Margaret Loper, Emre Gursoy, Stacey Truex, Wenqi Wei and Yanzhao Wu. A Framework for Evaluating Client Privacy Leakages in Federated Learning. **Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS 2020)**
8. Wenqi Wei, Ling Liu, Gong Su and Arun Iyengar. ``Gradient Leakage Resilient Federated Learning", Proceedings of **IEEE 2021 International Conference on Distributed Computing Systems (ICDCS 2021)**, Washington DC, Washington DC, United States, July 7-July 10, 2021.
9. Stacey Truex, Ling Liu, Emre Gursoy, Wenqi Wei, Lei Yu. Effect of Differential Privacy and Data Skewedness on Membership Inference Vulnerability. Proceedings of the **IEEE 2019 International Conference on Trust, Privacy and Security in Intelligent Systems and Applications** (TPS 2019). Dec. 12-14, Los Angeles, California USA.
10. Lei Yu, Ling Liu, Calton Pu, Emre Gursoy, Stacey Truex. ``Differentially Private Model Publishing for Deep Learning", **Proceedings of the 2019 IEEE Symposium on Security and Privacy (IEEE S&P 2019)**. May 20-22, 2019, The Hyatt Regency, San Francisco, CA.
11. Stacey Truex, Ling Liu, Emre Gursoy, Wenqi Wei, Lei Yu. ``Towards Demystifying Membership Inference Attacks", **IEEE Transaction on Services Computing,** (accepted in Feb. 2019). https://ieeexplore.ieee.org/document/8634878. Also available at https://arxiv.org/abs/1807.09173.
12. Yakun Huang, Xiuquan Qiao, Jian Tang, Pei Ren, Ling Liu, Calton Pu, Junliang Chen. DeepAdapter: A Collaborative Deep Learning Framework for the Mobile Web Using Context-Aware Network Pruning, **Proceedings of IEEE International Conference on Computer Communications (Infocom) 2020**
13. Yakun Huang, Xiuquan Qiao, Hongru Zhao, Jian Tang, Ling Liu. `` Towards Video Streaming Analysis and Sharing for Multi-Device Interaction with Lightweight DNNs", **Proceedings of IEEE International Conference on Computer Communications (Infocom 2021).**

Thank You

Contact: Prof. Dr. Ling Liu
lingliu AT cc.gatech.edu