Real-Time Video Anonymization in Smart City Intersections

Alex Angus*, Zhuoxu Duan**, Gil Zussman^ and Zoran Kostic^
*Qualcomm Technologies, Inc., **Rensselaer Polytechnic Institute, ^Columbia University

Session on Security and Privacy
19th IEEE International Conference on Mobile Ad-Hoc and Smart Systems (MASS 2022).
October 20 - 22, 2022. Denver, Colorado.

COSMOS

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Presentation Outline

Introduction

Privacy Concerns in Smart City Intersections

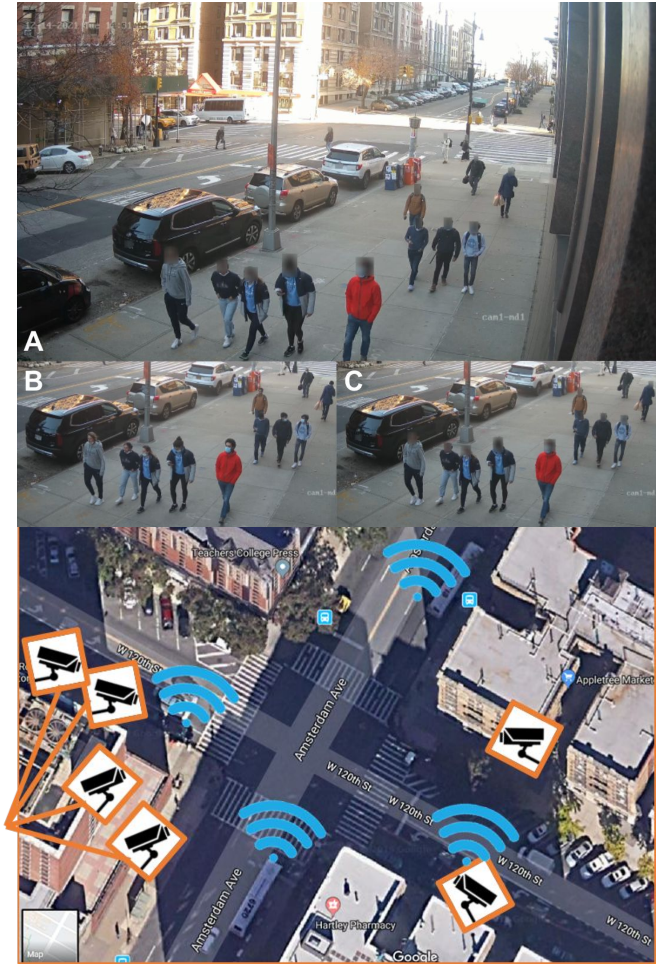
Methodology

Evaluation Results

Conclusion

COSMOS Research Testbed

- Cloud Enhanced Open Software Defined Mobile Wireless Testbed for City-Scale Deployment
- Pilot site at 120th St. and Amsterdam Ave in New York City
- Experimentation testbed for advanced wireless research and applications
- Sensing and high speed communication
- Edge computing clusters with scalable CPU and GPU resources
    - T4 and A100 Nvidia GPUs



COSMOS

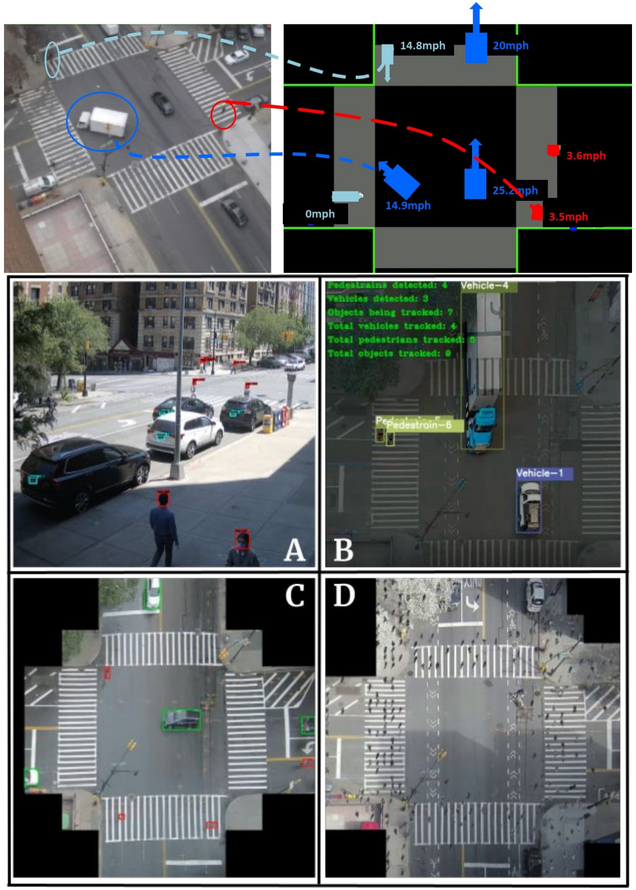
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Real Time Video Feeds in Smart City Intersections

Why do we need real time video feeds?

Real-time use cases:

- Traffic analytics
- Communication and feedback with cloud-connected vehicles
- Social distancing analysis in pandemics
- Radar screen



COSMOS

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

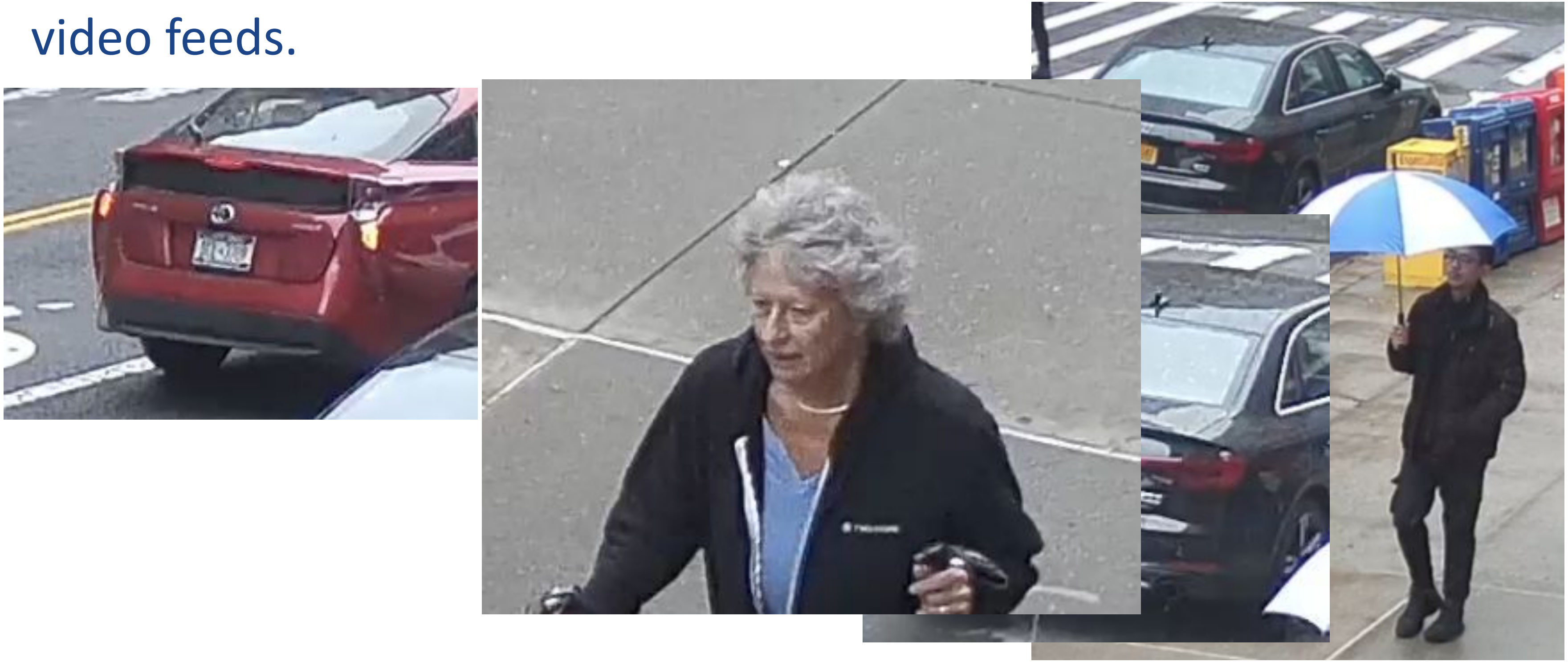
Personal privacy is inherently compromised when using ground floor video feeds.

COSMOS

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

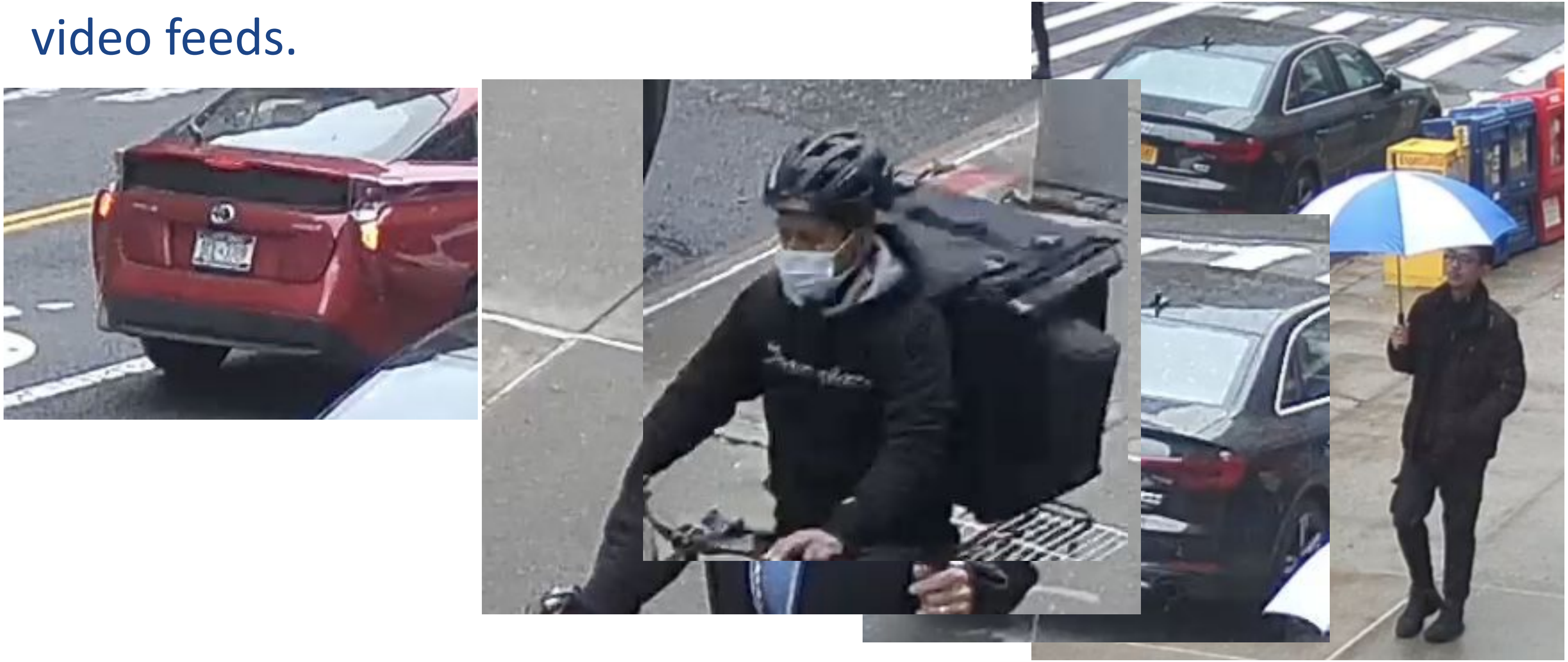
Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.



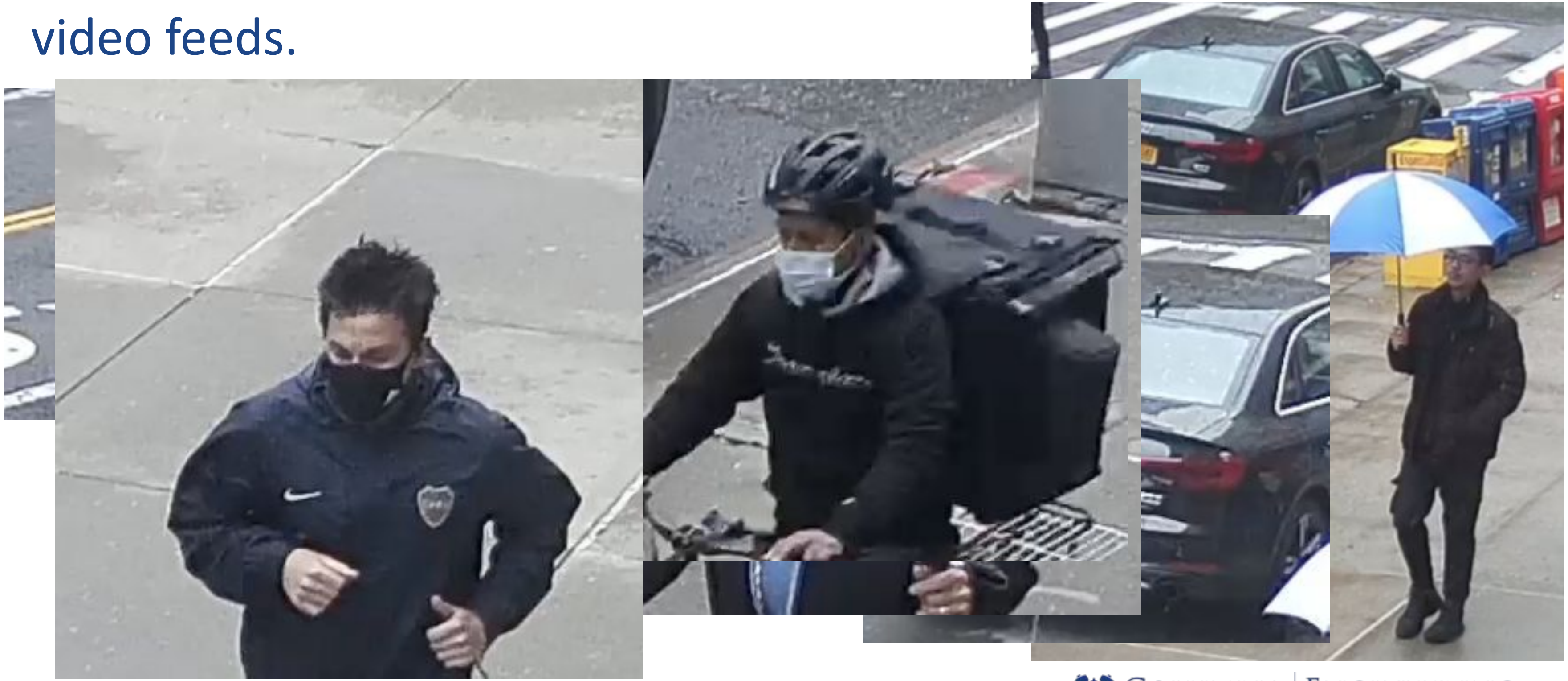
Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.



Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.



Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

Personal privacy is inherently compromised when using ground floor video feeds.

Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

Privacy Concerns in Smart City Intersections

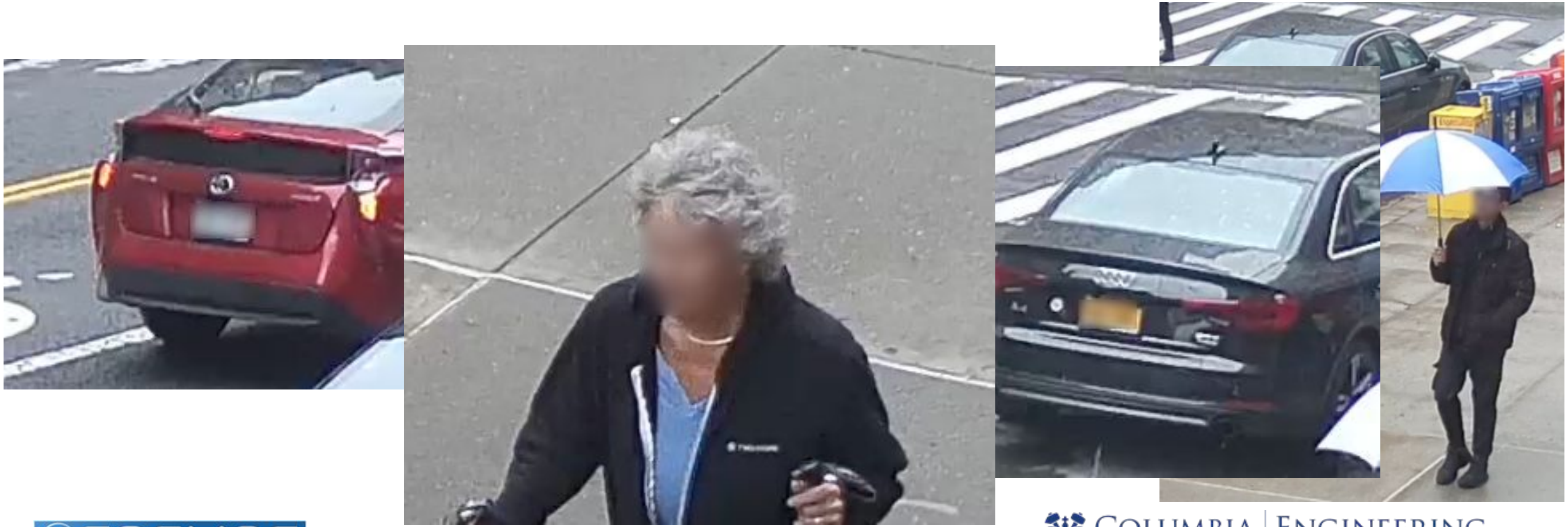
Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

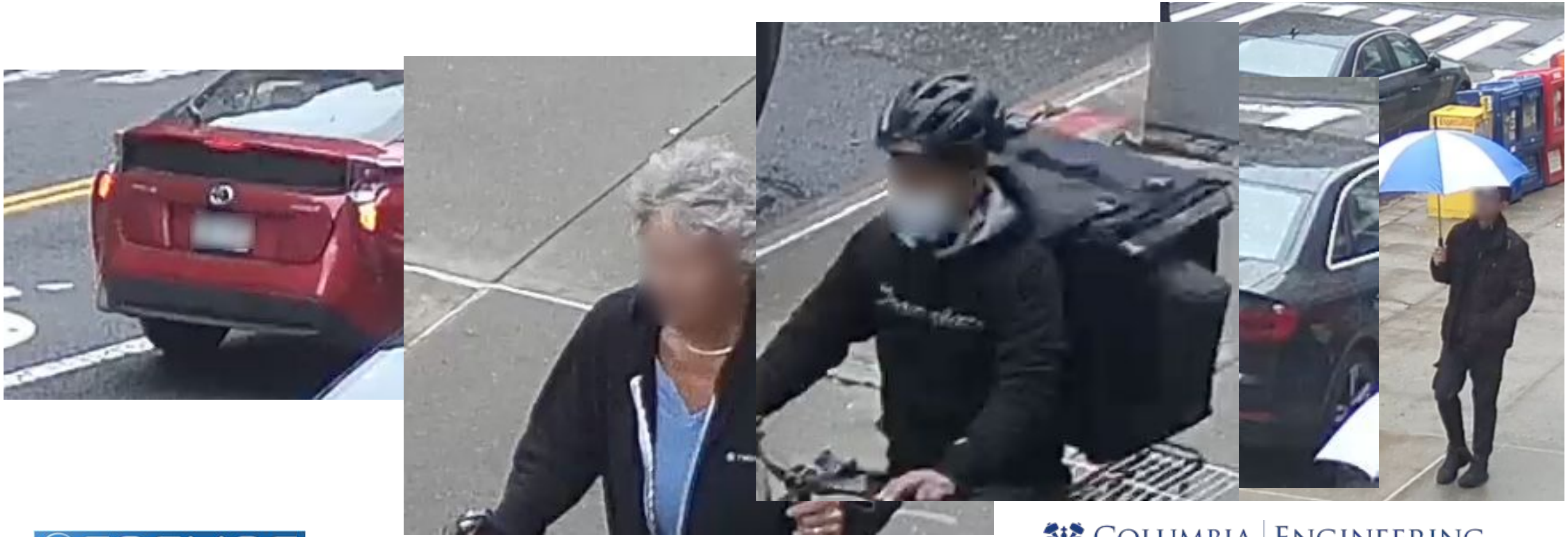
Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**



Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

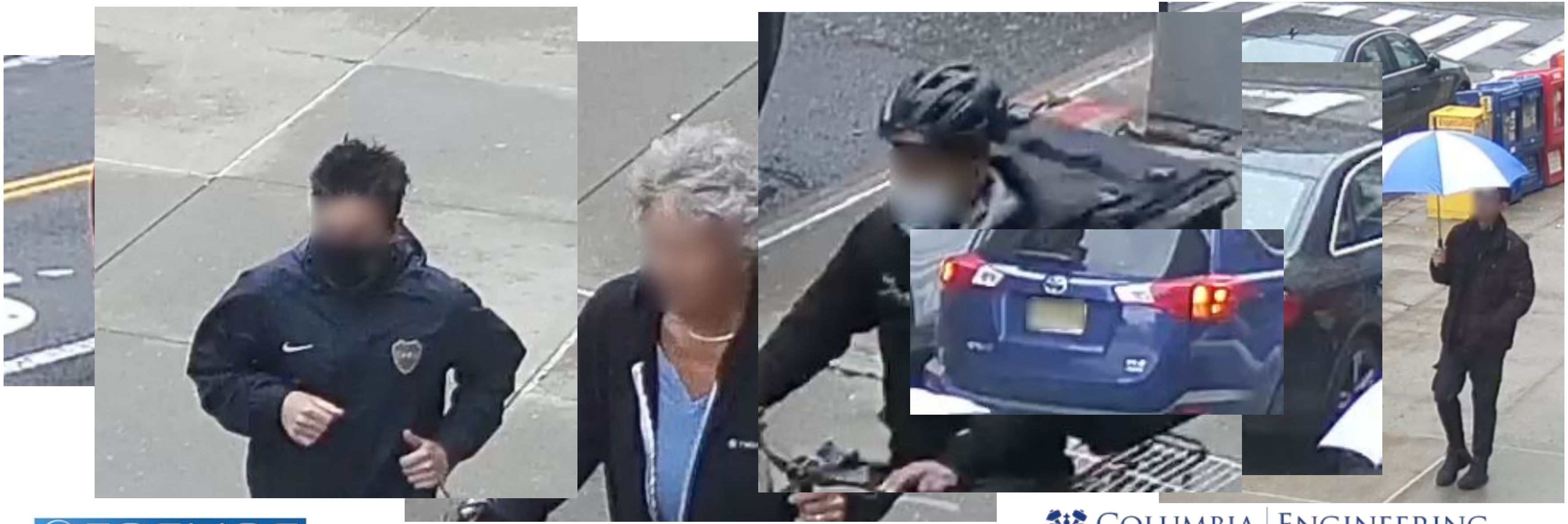


Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

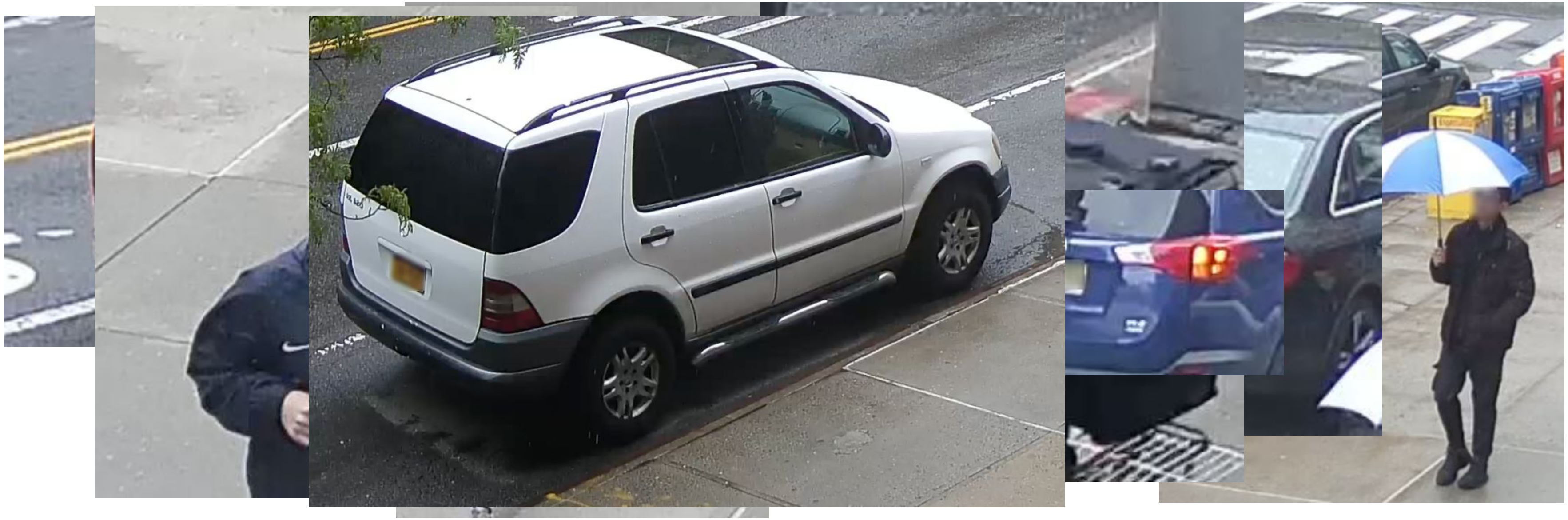


Privacy Concerns in Smart City Intersections

Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**

Privacy Concerns in Smart City Intersections

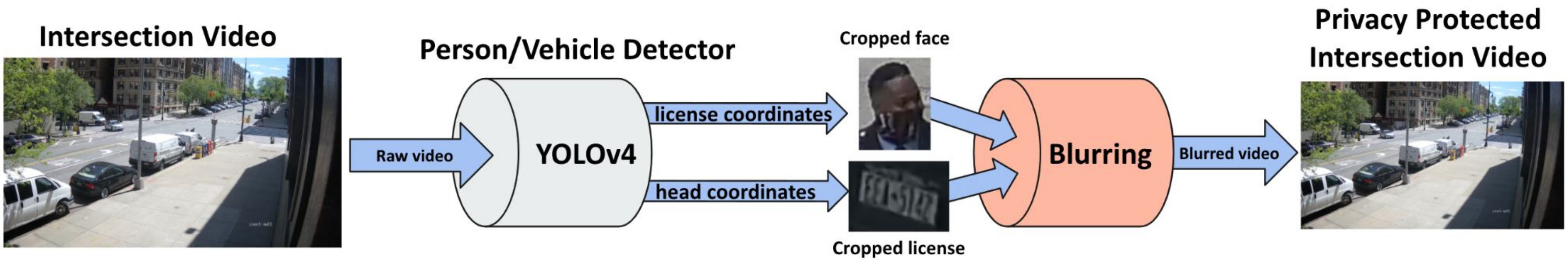
Goal: Build a pipeline for **anonymization** of faces and license plates in **intersection videos.**



Privacy Protection in Smart City Intersections

Deep learning based anonymization pipeline

- custom dataset collection
- supervised training of customized YOLOv4 models in Darknet framework
- inference optimization with TensorRT to achieve real time performance



COSMOS

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

COSMOS Ground Floor Intersection Dataset

COSMOS pilot site:

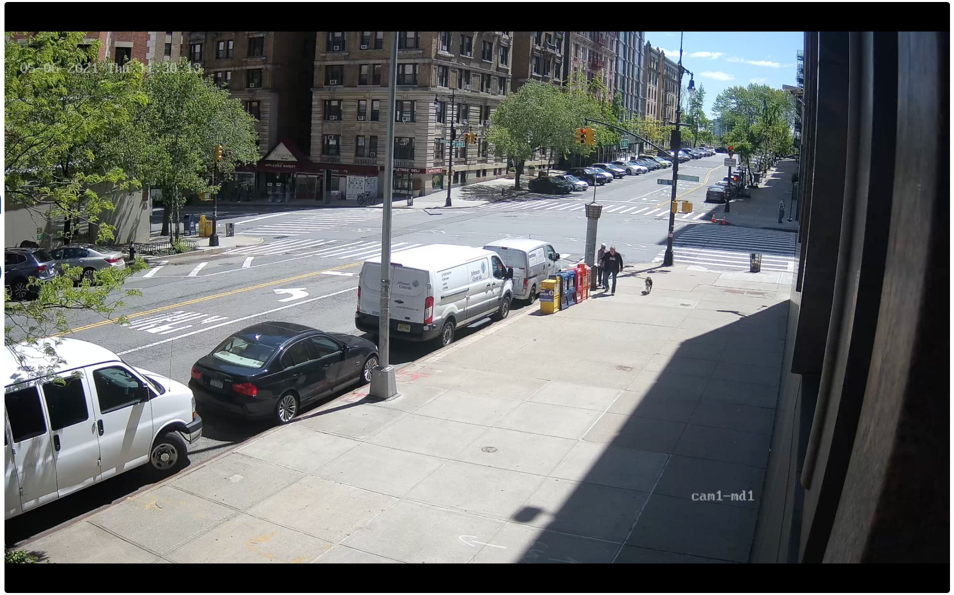
- 16 videos, 180 seconds each
- 30 FPS, 3840 x 1920 pixels
- Weather conditions
  - daytime, nighttime, cloudy, sunny, rainy
- Every 6th frame is annotated ⟶ over 14,000 ground truth frames
  - 70,186 faces
  - 124,614 licenses
- Median object areas ⟶ small
  - faces: 198 pixels
  - licenses: 83 pixels

COSMOS Ground Floor Intersection Dataset

COSMOS pilot site:

- 16 videos, 180 seconds each
- 30 FPS, 3840 x 1920 pixels
- Weather conditions
  - daytime, nighttime, cloudy, sunn
- Every 6th frame is annotated ⟶ o
  - 70,186 faces
  - 124,614 licenses
- Median object areas ⟶ small
  - faces: 198 pixels
  - licenses: 83 pixels

Daytime sunny

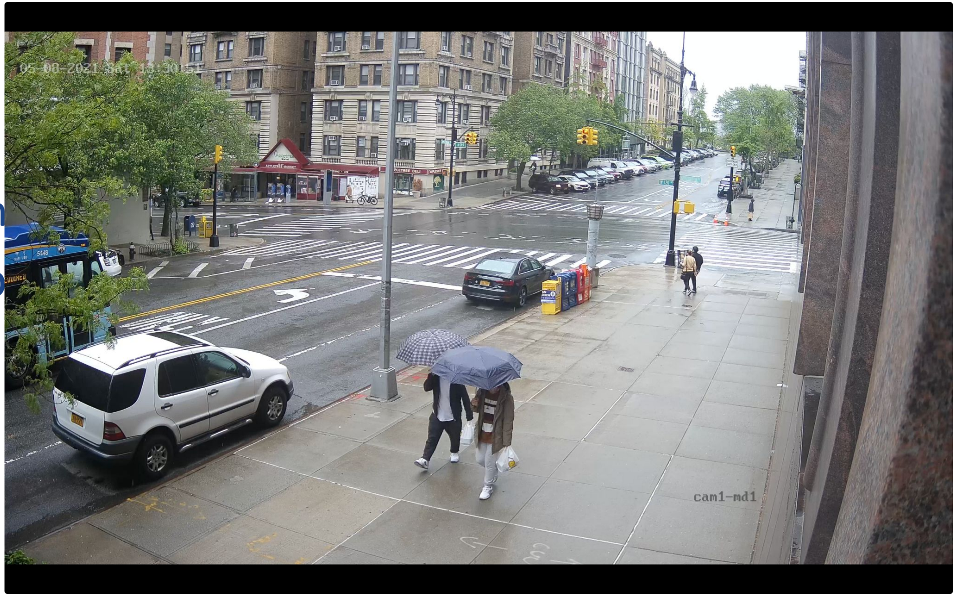


COSMOS Ground Floor Intersection Dataset

COSMOS pilot site:

- 16 videos, 180 seconds each
- 30 FPS, 3840 x 1920 pixels
- Weather conditions
  - daytime, nighttime, cloudy, sunn
- Every 6th frame is annotated → o
  - 70,186 faces
  - 124,614 licenses
- Median object areas → small
  - faces: 198 pixels
  - licenses: 83 pixels

Daytime rainy

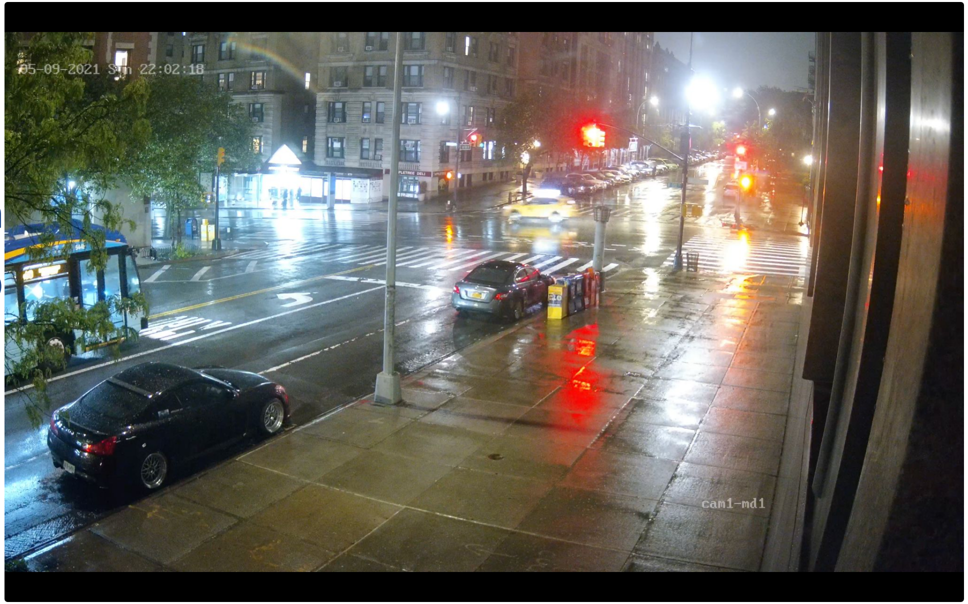


COSMOS Ground Floor Intersection Dataset

COSMOS pilot site:

- 16 videos, 180 seconds each
- 30 FPS, 3840 x 1920 pixels
- Weather conditions
  - daytime, nighttime, cloudy, sunn
- Every 6th frame is annotated ⟶ o
  - 70,186 faces
  - 124,614 licenses
- Median object areas ⟶ small
  - faces: 198 pixels
  - licenses: 83 pixels

Nighttime rainy

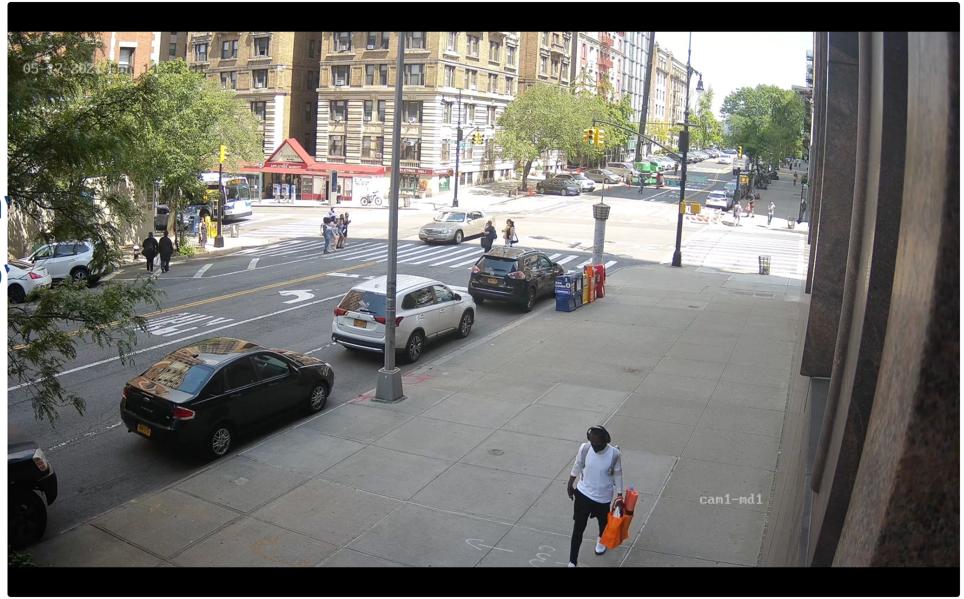


COSMOS Ground Floor Intersection Dataset

COSMOS pilot site:

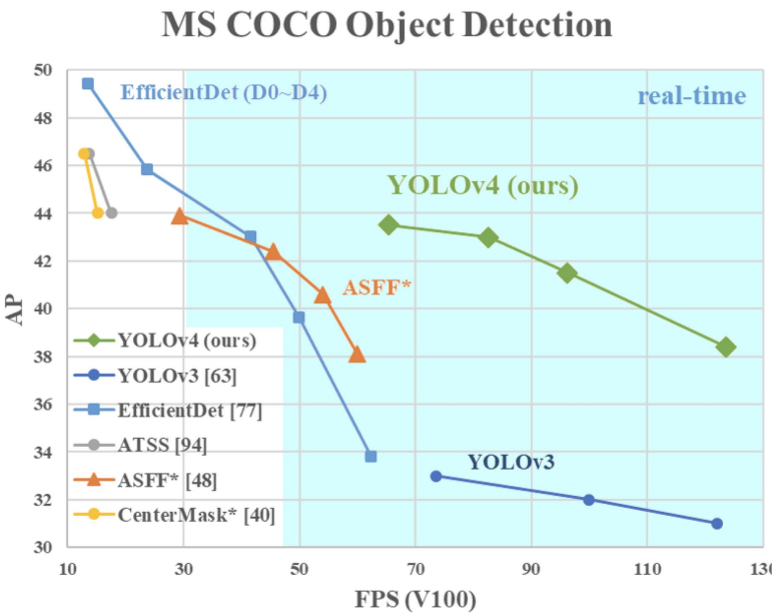
- 16 videos, 180 seconds each
- 30 FPS, 3840 x 1920 pixels
- Weather conditions
  - daytime, nighttime, cloudy, sunn
- Every 6th frame is annotated → o
  - 70,186 faces
  - 124,614 licenses
- Median object areas → small
  - faces: 198 pixels
  - licenses: 83 pixels

Daytime overcast



YOLOv4 Object Detection

- YOLOv4 is a single stage model that detects, localizes, and classifies relevant objects
- There is a trade off between inference speed and detection accuracy
- Small objects (faces and license plates) require large input resolution models
  - 608 x 608
  - 960 x 960
  - 1440 x 1440

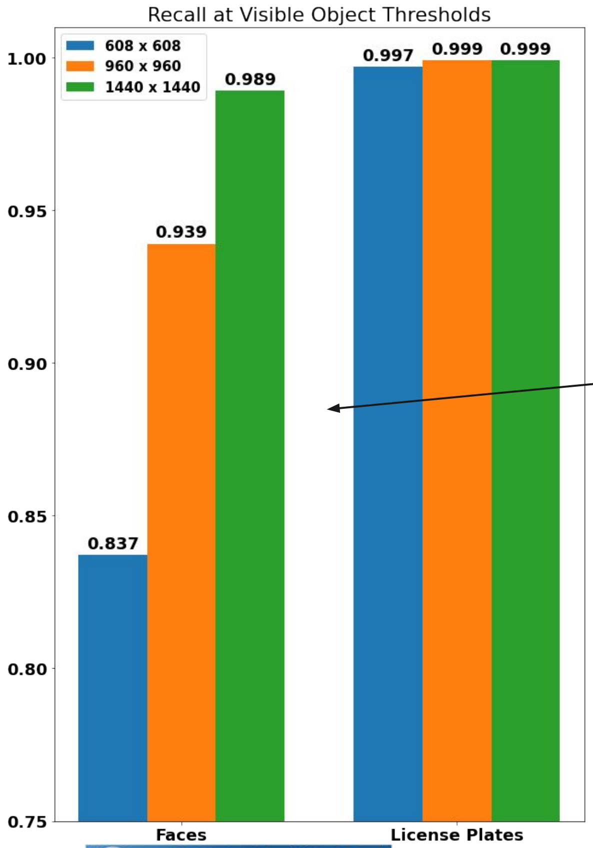


Bochkovskiy, A., Wang, C., & Liao, H.M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv, abs/2004.10934*.

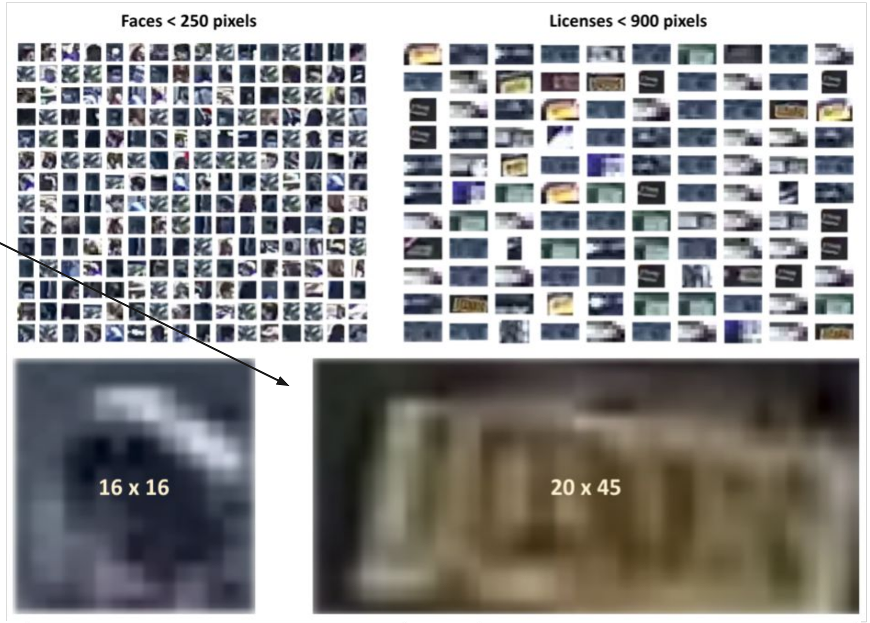
YOLOv4 Model Training and Validation

- Pre-trained on MS COCO object detection dataset
- 2 class detection → faces and license plates
- 10,000 iterations on custom ground floor intersection dataset
- Training completed using NVIDIA A100 and T4 GPUs hosted on Google Cloud Platform
- 2 out of 16 videos are left out of training for validation
- Weights yielding the highest validation mAP are chosen as the final weights
- CIoU loss function
- DropBlock regularization
- 64 frame batch size

Programmatic Accuracy Evaluation

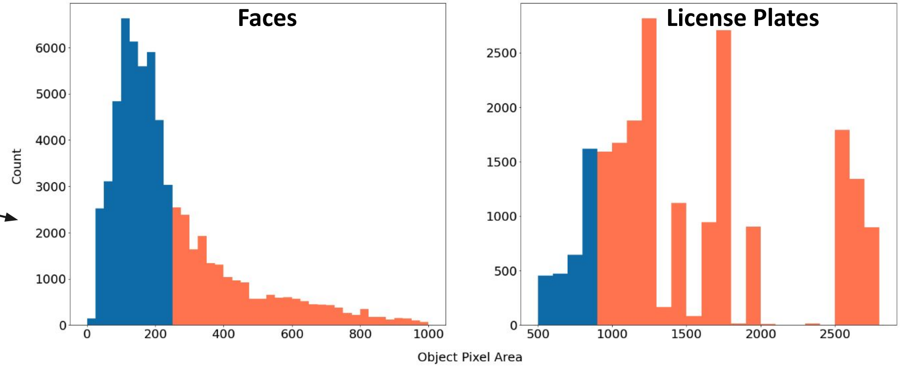


Example Objects at and below the visible thresholds

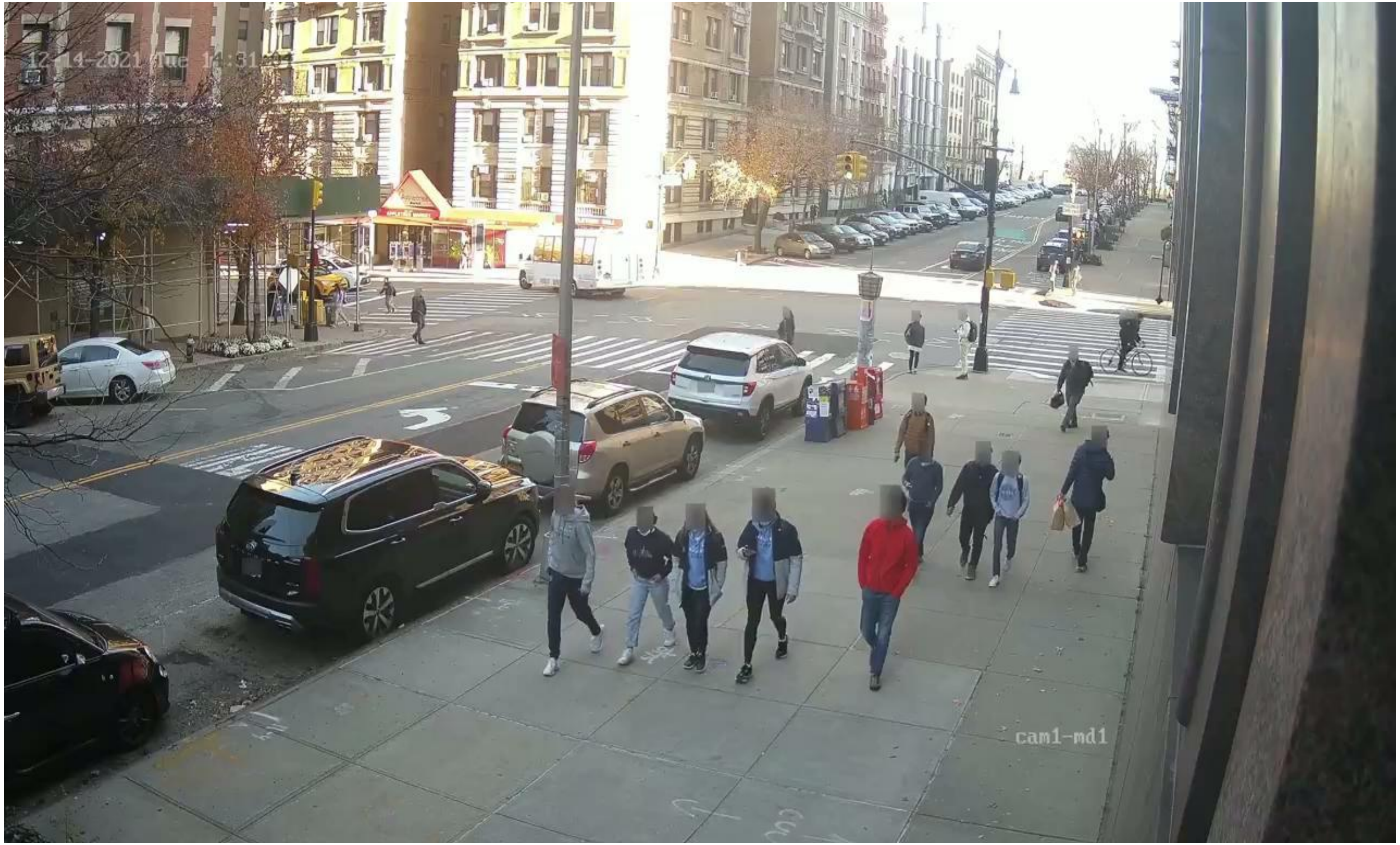


Detection Recall at and above the visible thresholds

Object area distributions



Example Output



Programmatic Accuracy Evaluation - Results (Visible Face Recall)

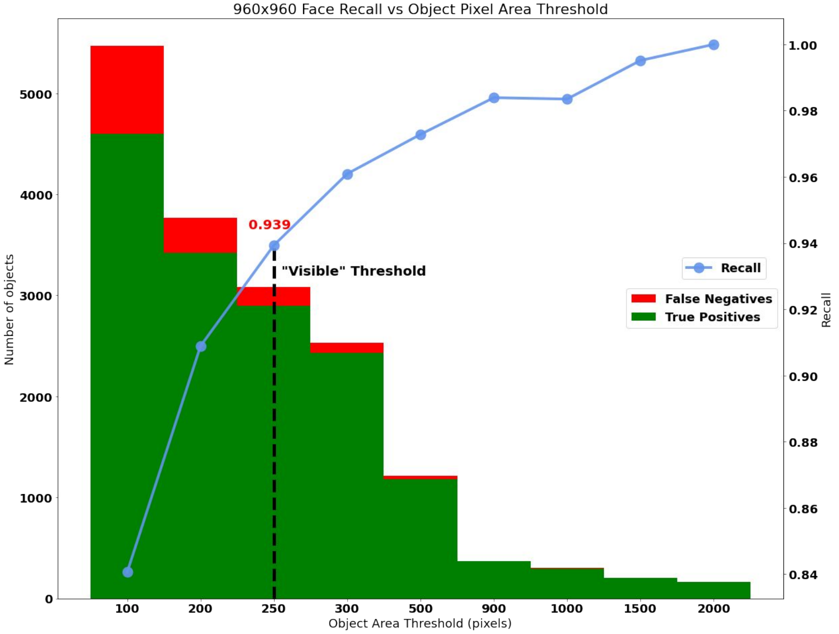
608 x 608 **"visible"** face recall: **83.72%**

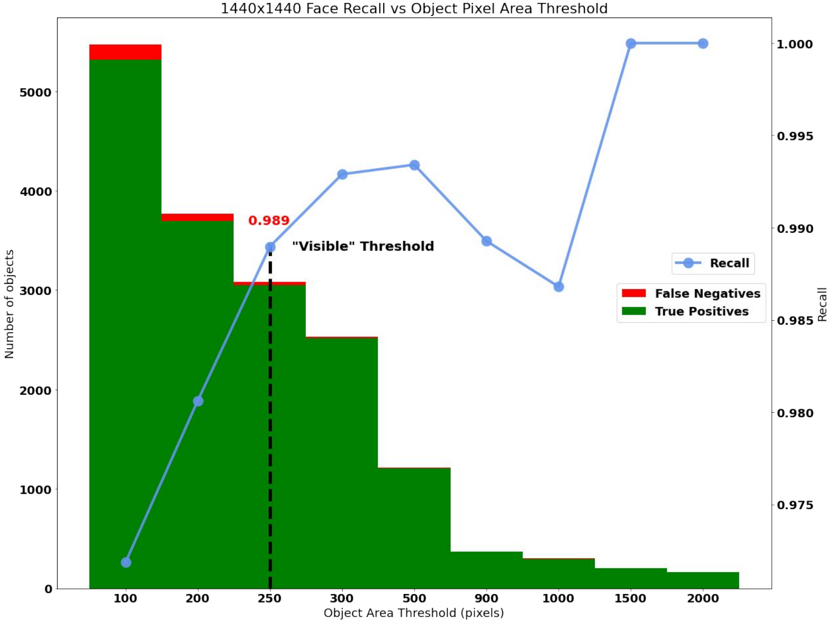
960 x 960 **"visible"** face recall: **93.93%**

1440 x 1440 **"visible"** face recall: **98.90%**

How many relevant objects are detected?

$$\text{Recall} = \frac{tp}{tp + fn}$$





COSMOS

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Programmatic Accuracy Evaluation - Results (Visible License Recall)

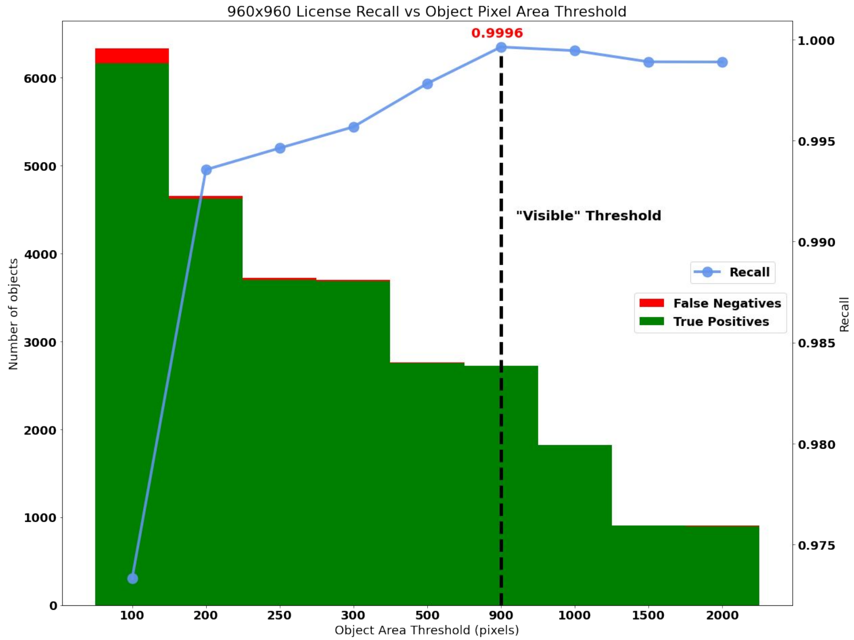
608 x 608 **"visible"** license recall: **99.71%**

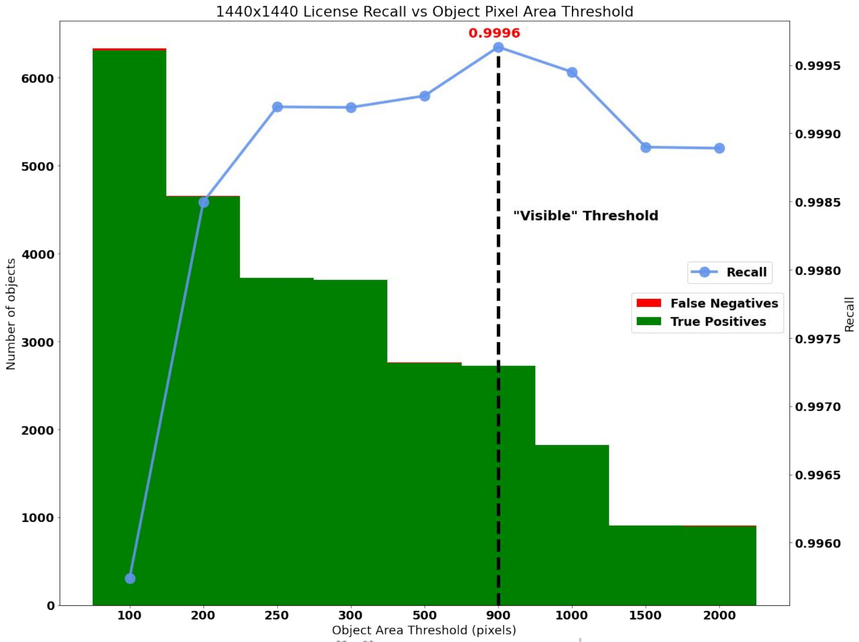
960 x 960 **"visible"** license recall**:** **99.96%**

1440 x 1440 **"visible"** license recall: **99.96%**

How many relevant objects are detected?

$$\text{Recall} = \frac{tp}{tp + fn}$$





Manual Pipeline Validation – Overview

- Ground truth labels are scarce and must be prioritized for training
- Anonymization accuracy is validated by visually inspecting output on new intersection videos
- Areas are defined where an exposed face or license is counted as a “miss”

Manual Pipeline Validation – Results

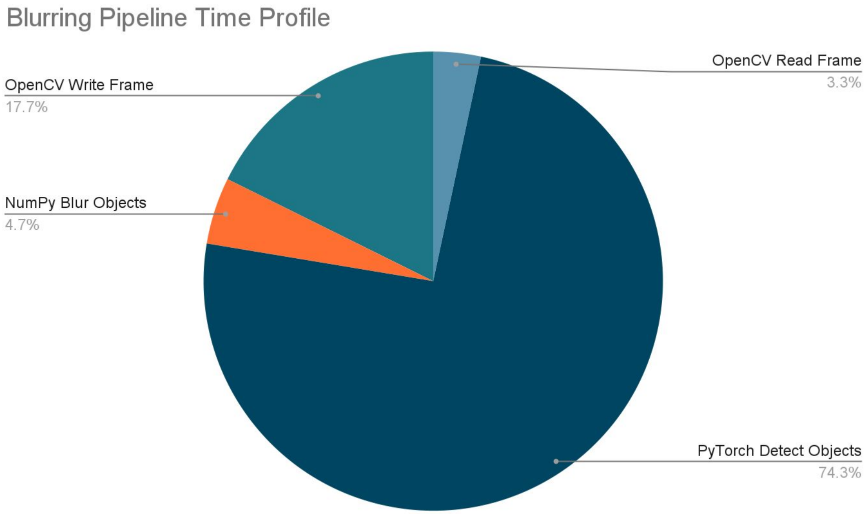
TABLE I: Manual Accuracy Evaluation Results

Model Resolution	Face Recall	License Plate Recall
960x960	98.24%	98.61%
1440x1440	98.61%	98.62%

Manual evaluation results confirm generalization to new intersection scenes.

Anonymization in Real Time

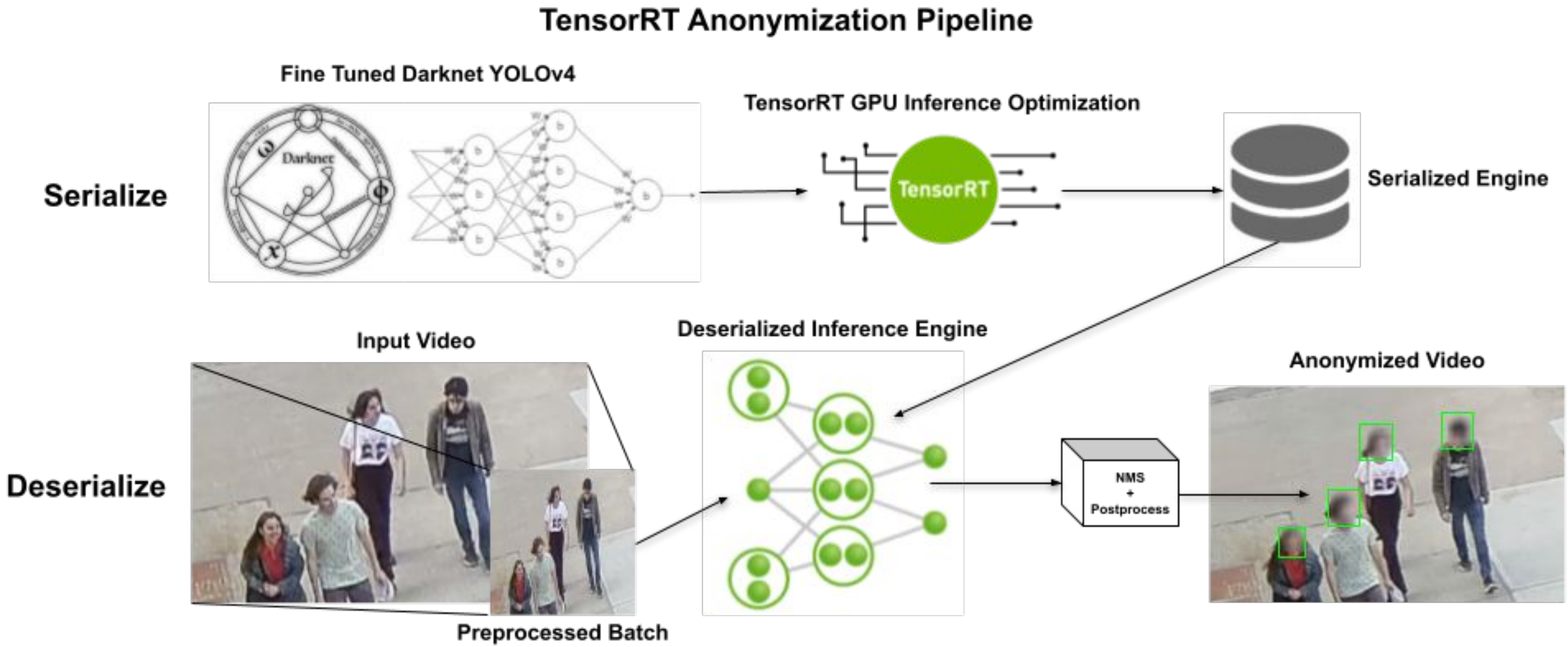
- To operate the pipeline in real-time, inference latency needs to be minimized → Computational complexity of forward pass is immense
- Real-time target is 33ms end-to-end latency. This includes:
  - frame read
  - preprocessing
  - inference
  - nms/postprocessing
  - anonymization
  - frame write



Operating at Real Time

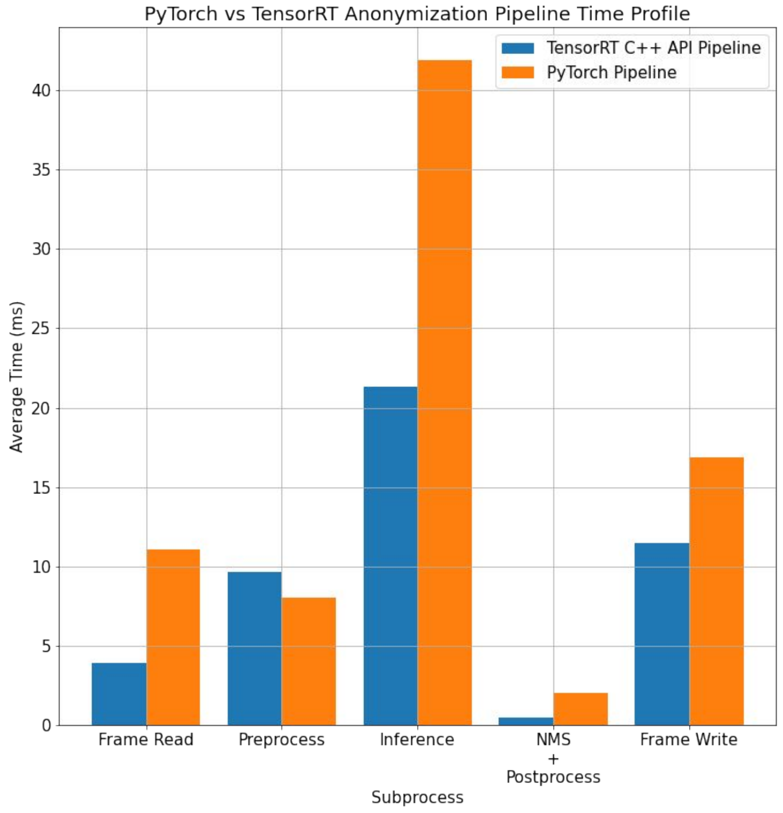
- TensorRT is an inference optimization framework for deep learning models on Nvidia GPUs
  - FP16 quantization
  - Layer and tensor concatenation
  - Tuned GPU kernel selection
  - Dynamic tensor memory

Inference Optimizations with TensorRT



TensorRT Optimized Pipeline vs. Non-Optimized

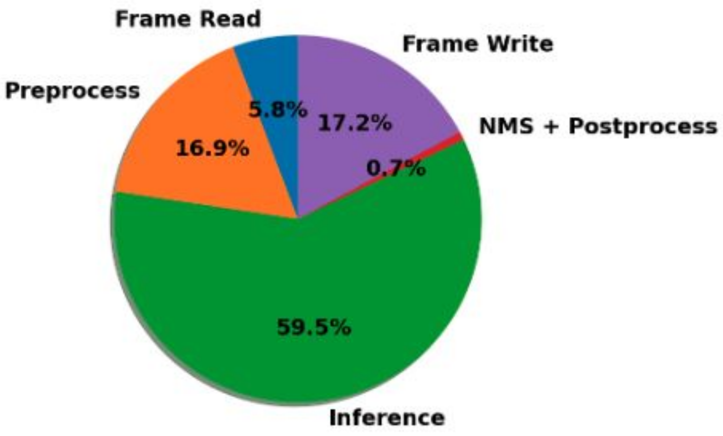
- Pipeline configurations
  - 960 x 960 model
  - batch size = 1
  - FP32 precision
  - 1 x A100 GPU
- **TensorRT C++ Pipeline reduces inference bottleneck**
- Frame read/write operations are also faster in C than in Python

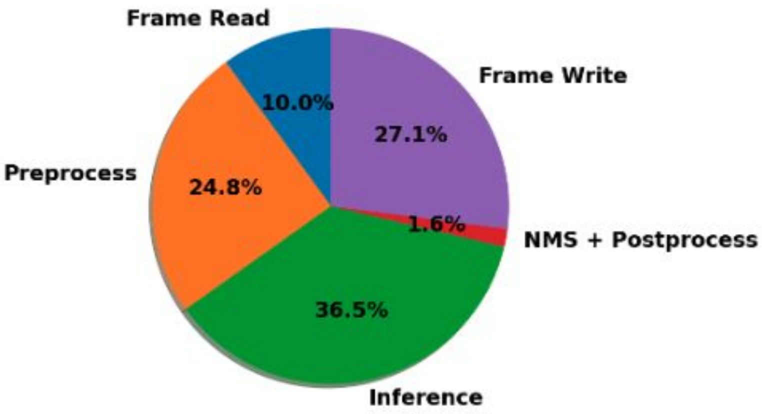


COSMOS

Columbia Engineering
The Fu Foundation School of Engineering and Applied Science

Anonymization Pipeline Timing Profiles





- 960 x 960 input resolution
- T4 GPU
- batch size = 1
- FP16
- 63.34 ms/frame

- 608 x 608 input resolution
- A100 GPU
- batch size = 8
- FP16
- 18.28 ms/frame

Latency Analysis Takeaways

TABLE II: Anonymization Pipeline Timing with Various Configurations

Model Input Resolution (pixels)	GPU	Precision	Batch Size	Full Pipeline	Frame Read	Preprocess	Inference	NMS + Postprocess	Frame Write
				TensorRT C++ Pipeline					
608x608	TegraX1	FP16	1	653.79	5.13	25.94	563.34	8.23	51.13
960x960	TegraX1	FP16	1	1491.49	10.79	64.74	1305.81	10.80	99.32
1440x1440	TegraX1	FP16	1	3285.98	23.74	144.34	2899.58	13.40	204.89
608x608	TeslaT4	FP16	1	29.06	1.74	4.40	17.94	0.27	4.70
960x960	TeslaT4	FP16	1	63.34	3.65	10.67	37.68	0.47	10.86
960x960	TeslaT4	FP16	4	63.71	3.91	10.73	38.32	0.45	10.28
960x960	TeslaT4	FP16	8	63.37	3.87	10.96	38.48	0.43	9.63
1440x1440	TeslaT4	FP16	1	139.35	7.64	23.43	84.97	0.76	22.55
1440x1440	TeslaT4	FP16	4	139.93	7.88	23.51	85.97	0.75	21.81
608x608	TeslaT4	FP32	1	44.75	1.59	4.34	33.99	0.24	4.58
960x960	TeslaT4	FP32	1	97.46	3.66	10.52	72.41	0.44	10.43
960x960	TeslaT4	FP32	4	99.34	3.89	11.05	73.51	0.45	10.43
1440x1440	TeslaT4	FP32	1	223.01	7.65	23.43	168.4	0.76	22.78
608x608	A100	FP16	1	21.82	1.90	4.41	9.83	0.33	5.34
960x960	**A100**	**FP16**	**1**	**42.44**	**4.05**	**9.7**	**16.33**	**0.51**	**11.83**
960x960	A100	FP16	4	38.82	4.00	9.75	13.16	0.51	11.39
960x960	A100	FP16	8	38.1	4.03	10.13	12.39	0.49	11.05
1440x1440	A100	FP16	1	83.19	8.2	21.22	28.26	0.83	24.67
1440x1440	A100	FP16	4	79.35	8.14	21.34	24.67	0.80	24.39
608x608	A100	FP32	1	23.64	1.74	4.28	12.17	0.28	5.15
960x960	A100	FP32	1	46.88	3.9	9.66	21.34	0.50	11.47
960x960	A100	FP32	4	42.47	3.88	9.92	17.09	0.49	11.08
1440x1440	A100	FP32	1	91.62	8.07	21.06	37.22	0.81	24.45
				PyTorch Python Pipeline					
608x608	TeslaT4	FP32	1	78.43	3.63	4.91	61.01	0.01	7.65
960x960	TeslaT4	FP32	1	173.31	9.52	10.93	134.77	0.01	16.08
608x608	A100	FP32	1	63.05	3.02	3.89	46.86	0.01	8.01
960x960	A100	FP32	1	79.94	11.07	8.06	41.89	0.01	16.89
960x960	A100	FP32	2	63.73	7	8.1	30.22	0.02	16.62
1440x1440	A100	FP32	1	130.89	14.12	20.11	58.86	0.02	34.41

All values are average execution time per frame measured in milliseconds. Timing operations incur negligible overhead (≈ $10\mu s$).

1. Jetson Nano (TegraX1) can't operate the pipeline anywhere close to real-time. Even the 608x608 model operates at:

25.94 + 563.34 + 8.23 = 597.5 ms = 1.674 FPS

2. Several configurations (GPU/FP precision/batch size) operate under 33.3 ms time constraint, excluding frame read/write. For example: 960x960, A100, FP16, BS=1 →

9.7 + 16.33 + 0.51 = 26.54 ms = 37.68 FPS

3. Average latencies improve if we can tolerate batch inference. For example: 960x960, A100, FP16, BS=8 →

10.13 + 12.39 + 0.49 = 23.01 ms = 43.46 FPS

Assessment of Risk of Violating Privacy - Edge Cases

Licenses passing poles

People passing license

Assessment of Risk of Violating Privacy - Edge Cases

Faces superimposed on other objects

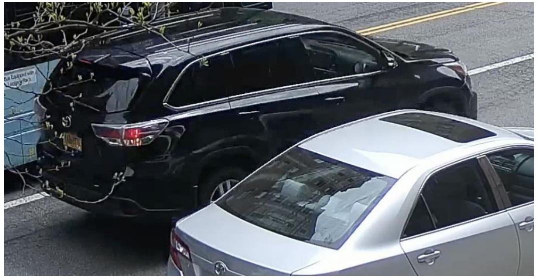






Assessment of Risk of Violating Privacy - Edge Cases

Buses and branches



Person holding object

Babies

Conclusion

- **The blurring pipeline anonymizes up to 99% of faces and license plates**
  - Edge cases can be reduced with more (and better) **training data** and **data augmentation**
- **The blurring pipeline operates in real time**
  - TensorRT **inference optimizations**, datacenter **GPUs**, and **reduced precision** calculations drastically **increase throughput**

Future work could explore **unsupervised detection** and **model reduction** for edge devices

Questions