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
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
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
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

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 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, sunny
- 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, sunny
- 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

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, sunny
- 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

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, sunny
- 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

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

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

![Chart showing recall at visible object thresholds for faces and license plates.
Example Objects at and below the visible thresholds.
Detection Recall at and above the visible thresholds.
Object area distributions.]

- **Faces**
  - 608 x 608: 0.837
  - 960 x 960: 0.989
  - 1440 x 1440: 0.997

- **License Plates**
  - 608 x 608: 0.939
  - 960 x 960: 0.999
  - 1440 x 1440: 0.999
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?

Recall = \frac{tp}{tp + fn}
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?

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**

<table>
<thead>
<tr>
<th>Model Resolution</th>
<th>Face Recall</th>
<th>License Plate Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>960x960</td>
<td>98.24%</td>
<td>98.61%</td>
</tr>
<tr>
<td>1440x1440</td>
<td>98.61%</td>
<td>98.62%</td>
</tr>
</tbody>
</table>

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 Anonymization Pipeline

Serialize

Fine Tuned Darknet YOLOv4

TensorRT GPU Inference Optimization

Serialized Engine

Deserializing Inference Engine

Input Video

Preprocessed Batch

Deserialize

Deserialize

Anonymized Video

Columbia Engineering

The Fu Foundation School of Engineering and Applied Science
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
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

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

\[25.94 + 563.34 + 8.23 = 597.5 \text{ ms} = 1.674 \text{ FPS}\]

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 →

\[9.7 + 16.33 + 0.51 = 26.54 \text{ ms} = 37.68 \text{ FPS}\]

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

\[10.13 + 12.39 + 0.49 = 23.01 \text{ ms} = 43.46 \text{ FPS}\]
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