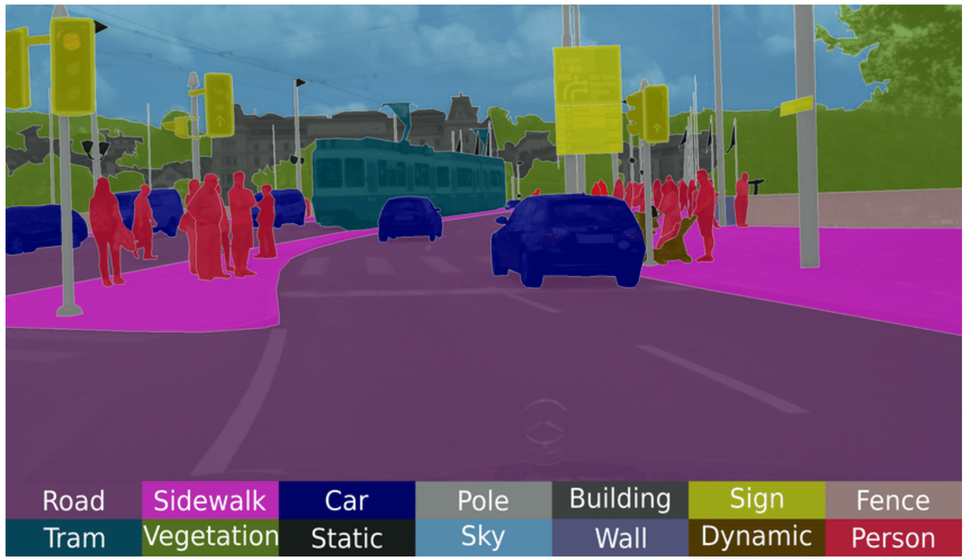
Bridging the Edge-Cloud Barrier for Real-time Advanced Vision Analytics

Yiding Wang, Weiyan Wang, Junxue Zhang,
Junchen Jiang (UChicago), Kai Chen

(Edge-to-cloud) vision analytics are ubiquitous

- Large scale deployment of cameras: traffic monitoring, event detection
- Vehicles/robots with cameras: autonomous driving vehicles/robotics/drones

Object detection



Semantic segmentation

Advanced applications are demanding

- Example: segmentation and object detection tasks for autonomous driving
  - Real-time-level interaction requires **low latency**
  - High inference accuracy requires high fidelity **data** and **computing resource**
- Currently advanced applications run heavy vision inference tasks on the **edge**.

"Real-time video analytics: the killer app for edge computing"
–Ganesh Ananthanarayanan etc.

- But it makes sense to consider a cloud-based solution.

Potential benefits of the cloud

- Reducing the requirements for edge devices, thus making the large-scale deployment cheap.

- High-end autonomous driving vehicles vs. delivery robots/vehicles

Challenges

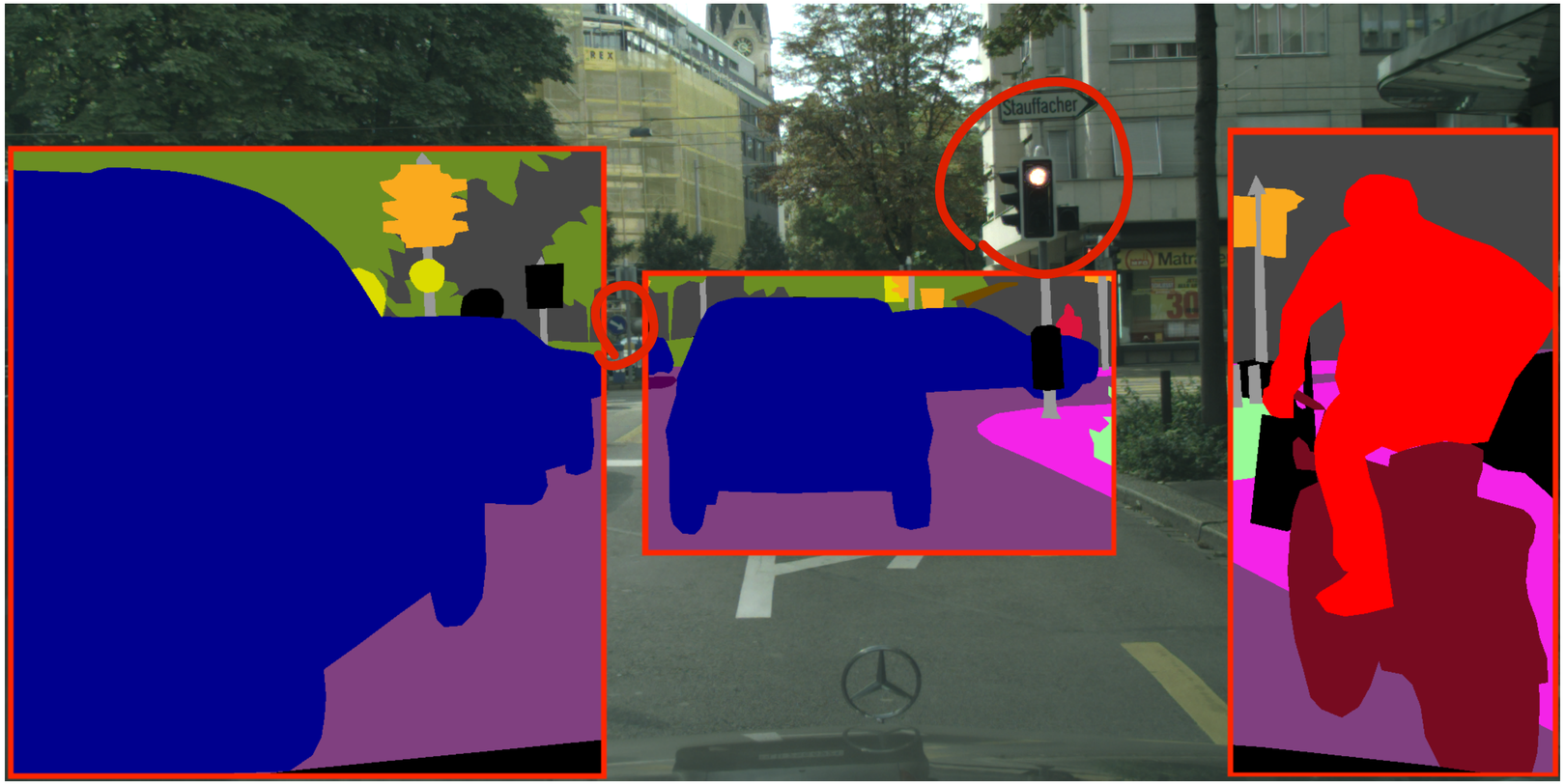
The edge-to-cloud real-time advanced vision applications face **strict bandwidth-accuracy trade-off**:

1. **Accuracy**: demanding applications → high accuracy → high quality data

2. **Bandwidth**: high quality camera feeds → high network bandwidth usage

Idea #1: cropping

Sending only cropped areas of region-of-interest (ROI). (Reinventing Video Streaming for Distributed Vision Analytics, Pakha et al., HotCloud 2018).



Limitation: For advanced applications, ROI is is the full frame. → Cannot crop.

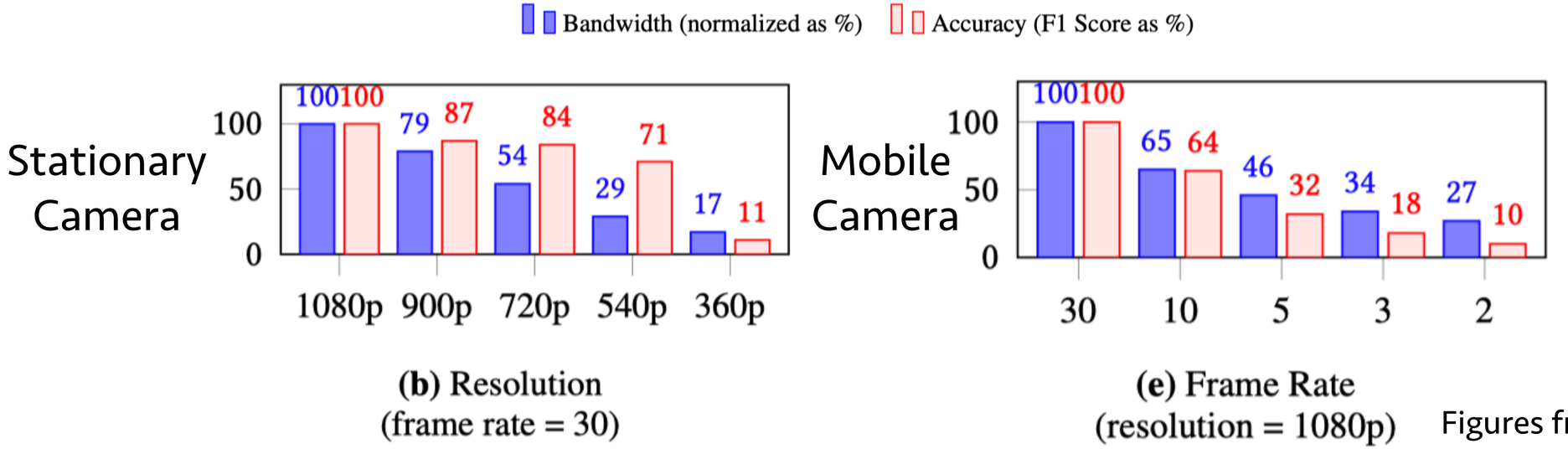
Idea #2: frame filtering

Filtering the relevant frames and streaming them to the cloud. (Scaling Video Analytics On Constrained Edge Nodes, Canel et al., SysML 2019)

Limitation: Works well for always-on **stationary traffic camera** feeds, but not for a **moving vehicle/robot**: relevant objects are always in the scene.

Idea #3: harmless degradation

Using a task-specific degradation config. (AWStream: Adaptive Wide-Area Streaming Analytics, Zhang et al., SIGCOMM 2018)



Figures from AWStream

Mobile cameras: value **frame rate**. Stationary cameras: value **resolution**

Limitation: **Advanced** applications require both high frame rate and resolution. 4× downsampling→13% loss on mIoU, 20% on small yet critical object classes

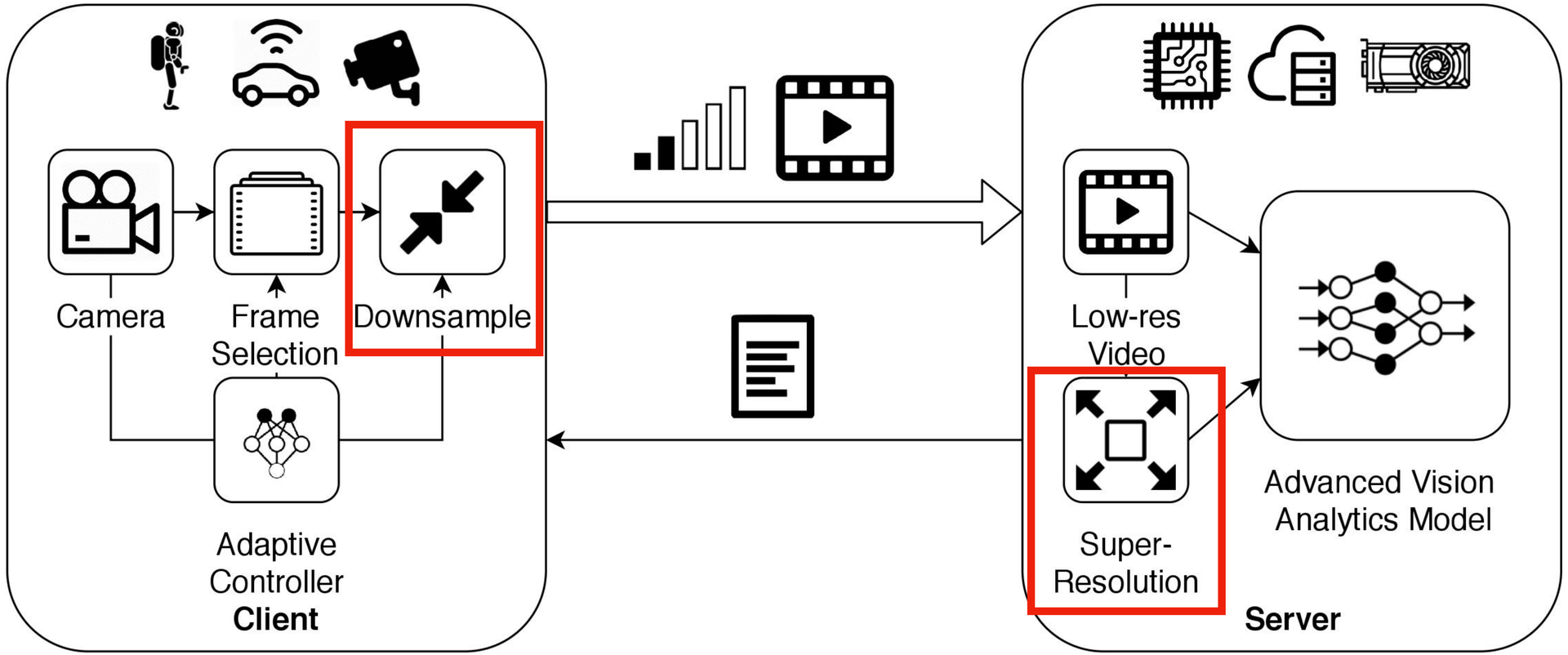
Our work

- CloudSeg, an edge-to-cloud framework for **advanced vision analytics** that achieves both low latency and high inference accuracy with **analytics-aware super-resolution**.

- CloudSeg saves **bandwidth** by recovering the high-resolution frames from the low-resolution stream via **super-resolution**.

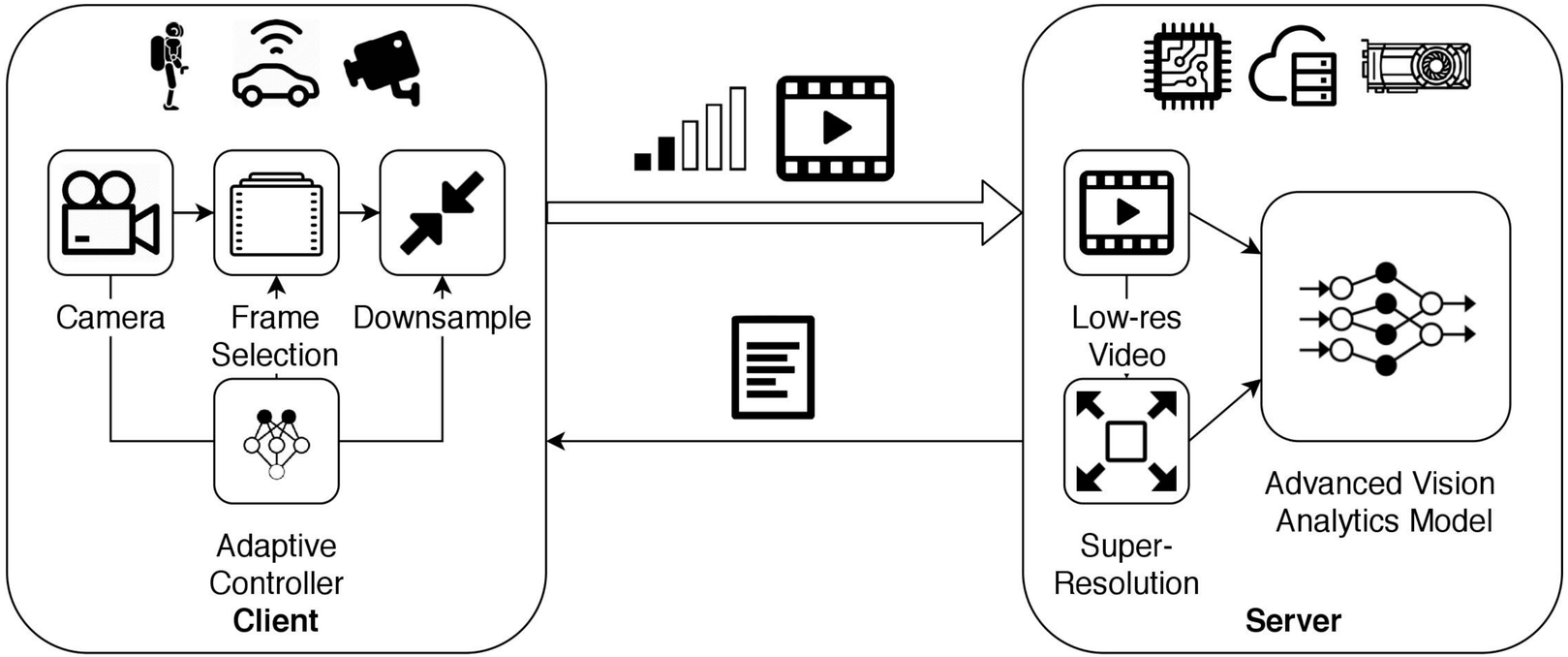
- CloudSeg keeps high **accuracy** via SR tailored for the actual analytics tasks.

Design of CloudSeg framework



Low extra latency (6.2ms) with an efficient SR model on the GPU server.

Bandwidth usage saving

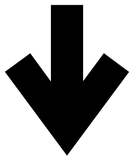


With CloudSeg, downsampling 2K frames to 512×256 can reduce 13.3× bandwidth usage with **2.6%** accuracy (mIoU) loss for semantic segmentation. To achieve the same accuracy, AWStream needs to stream 720p feed, thus CloudSeg can reduce bandwidth use by **6.8×** compared with AWStream.

Inference accuracy on critical details

- Some classes in Cityscapes dataset are very insensitive to input resolutions: sky, road, wall, building, etc.

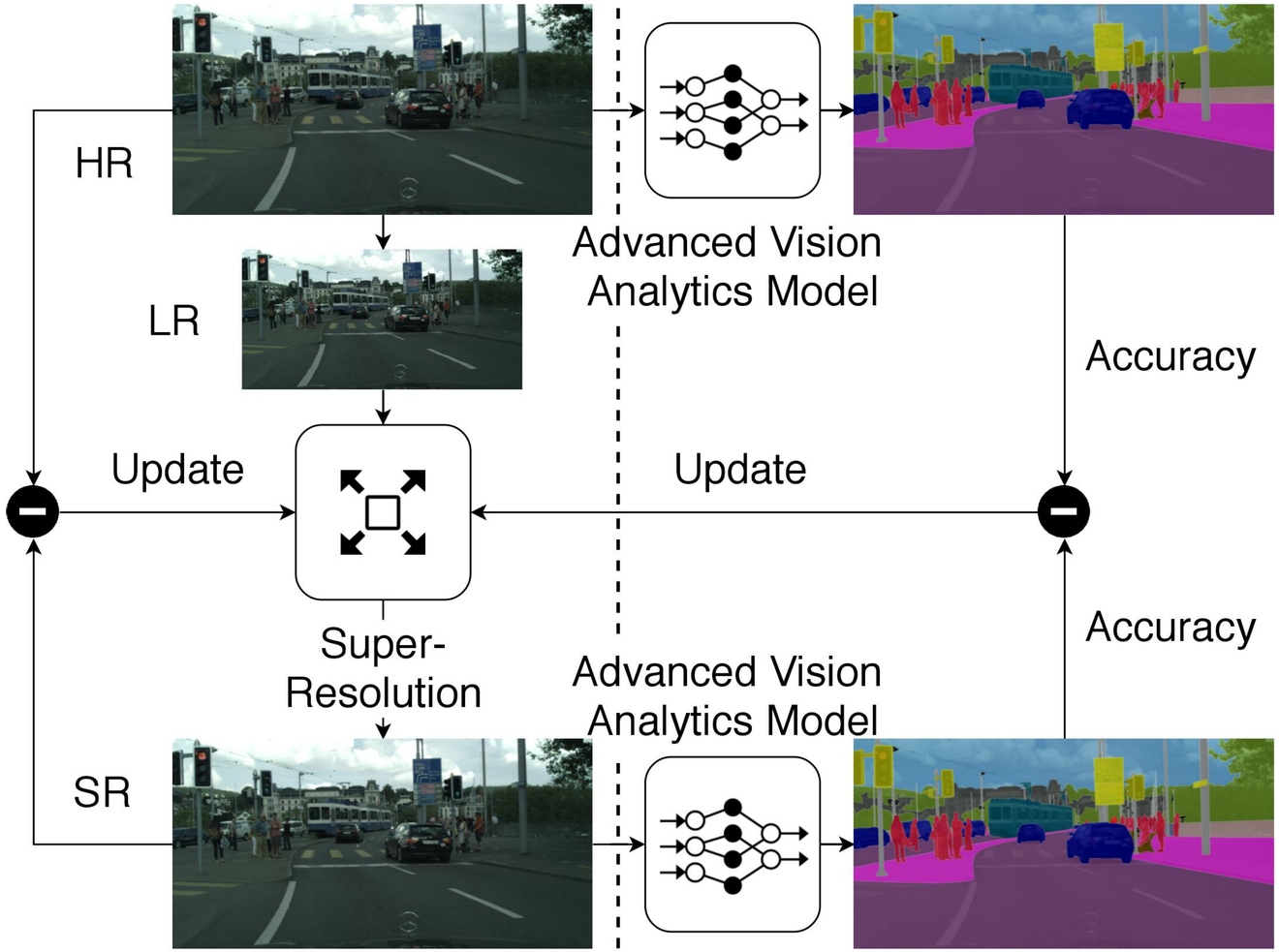
- Others e.g. rider, bicycle, motorcycle, traffic sign/light, person are sensitive to the input quality and also critical to the real-world applications.



- Observation: There is a mismatch between goals of SR and analytics tasks.

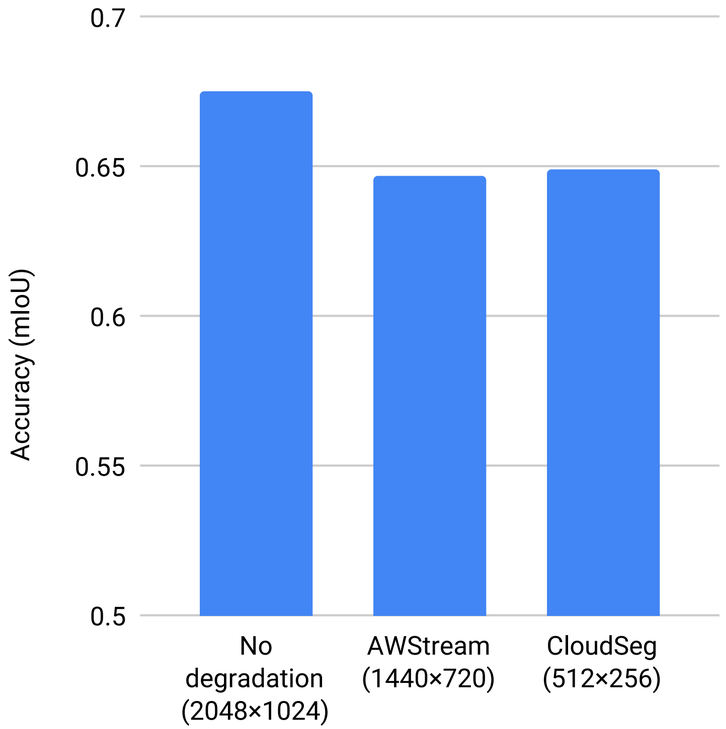
Analytics-aware super-resolution

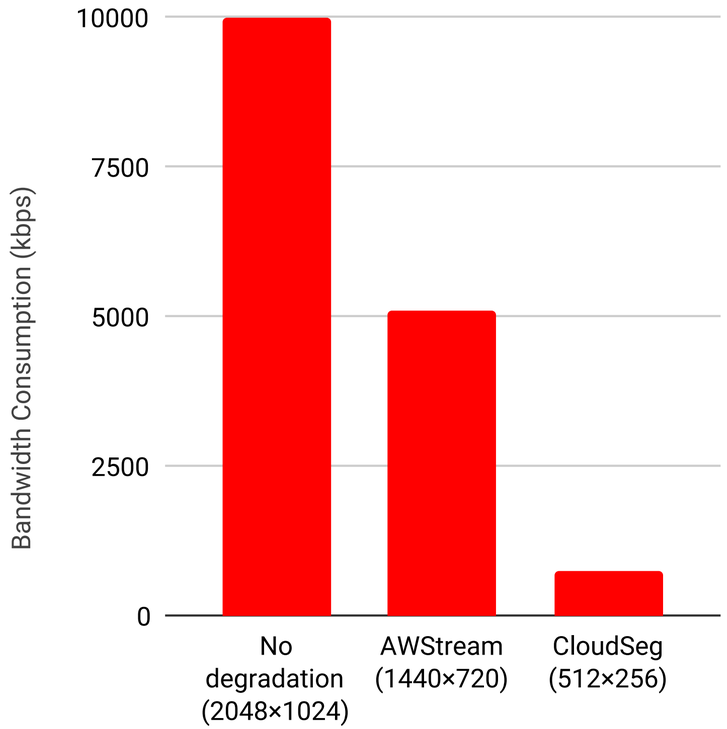
- Target of SR training: reduce the backend model inference accuracy loss
- Especially improve the accuracy on small objects, compared with vanilla SR
- 2.6% accuracy loss compared with HR frames



Promising results

- **6.8×** bandwidth saving compared with AWStream, at same inference accuracy
  2.6% accuracy (mIoU) loss compared with original 2K frames (13.3× larger)





Summary

- Enabling edge-to-cloud real-time advanced vision analytics is meaningful. The key technical challenge is the **strict bandwidth-accuracy trade-off**.

- The design of CloudSeg is a first step to tackle the trade-off with **analytics-aware super-resolution**.

- Promising results: 6.8× bandwidth saving compared with directly downsampling, with negligible drop in accuracy compared with original video.