Enabling Edge-Cloud Video Analytics for Robotic Applications

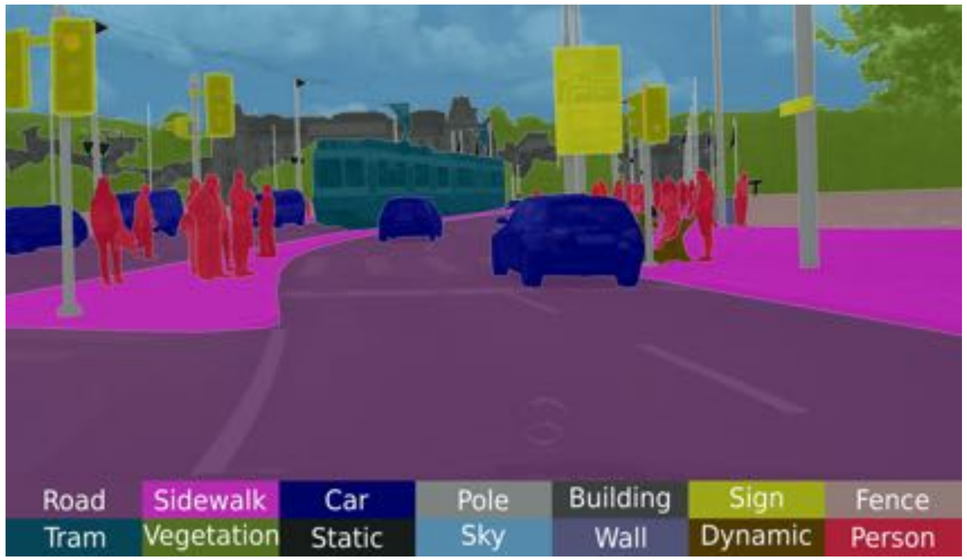
Yiding Wang, Weiyan Wang, Duowen Liu,
Xin Jin (PKU), Junchen Jiang (UChicago), Kai Chen

THE UNIVERSITY OF
CHICAGO

Edge-cloud video 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

Video analytics for robotics are demanding

- Autonomous robotics applications heavily rely on video analytics:

1. They require **scene understanding** capabilities for planning and control.

2. They use **high-frame-rate** and **high-resolution** video data.

3. They require **powerful DL models** and **computation resources** for high inference **accuracy** and **fast reaction**.

Video source: Autopilot AI | Tesla

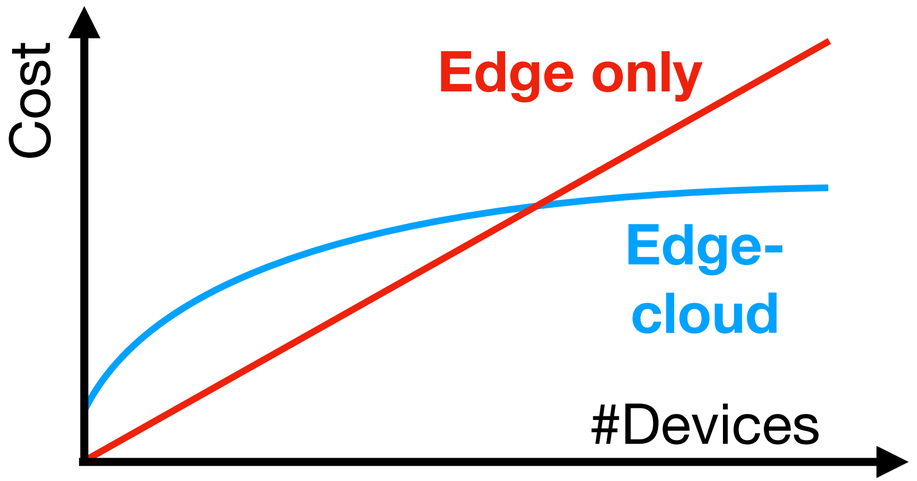
Example: delivery robots

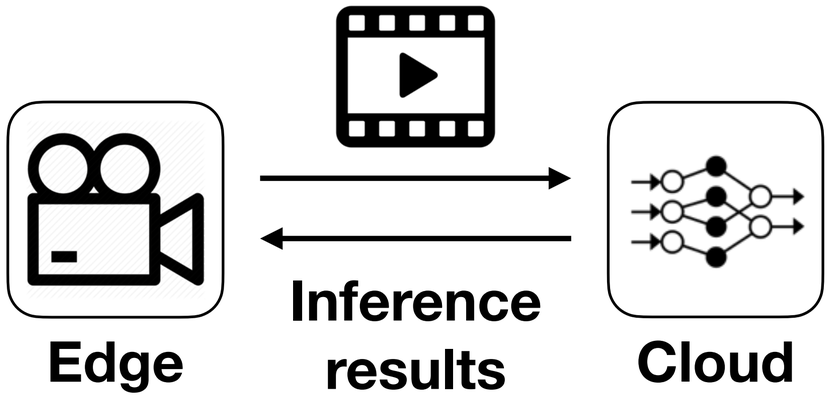
- Small-sized electric vehicles.
- A **large scale deployment** with **low cost**.
- But during the COVID-19 lockdown when they are needed the most, “the technology is **not ready at scale** to deploy”, said by Nuro president.

Source: Delivery Robots Aren't Ready—When They Could Be Needed Most | WIRED

Edge-cloud can help scale up video analytics

- Unlike auto. vehicles, delivery robots are **budget and battery constrained.**
- Run the same demanding **scene-understanding DNN tasks**.
- A popular way: **offload** the heavy DL inference tasks to the **cloud**.
- Trade-off: **accuracy/data quality** and **latency/bandwidth**





Popular edge-cloud optimizations are limited

- To manage the trade-off between accuracy and bandwidth consumption:

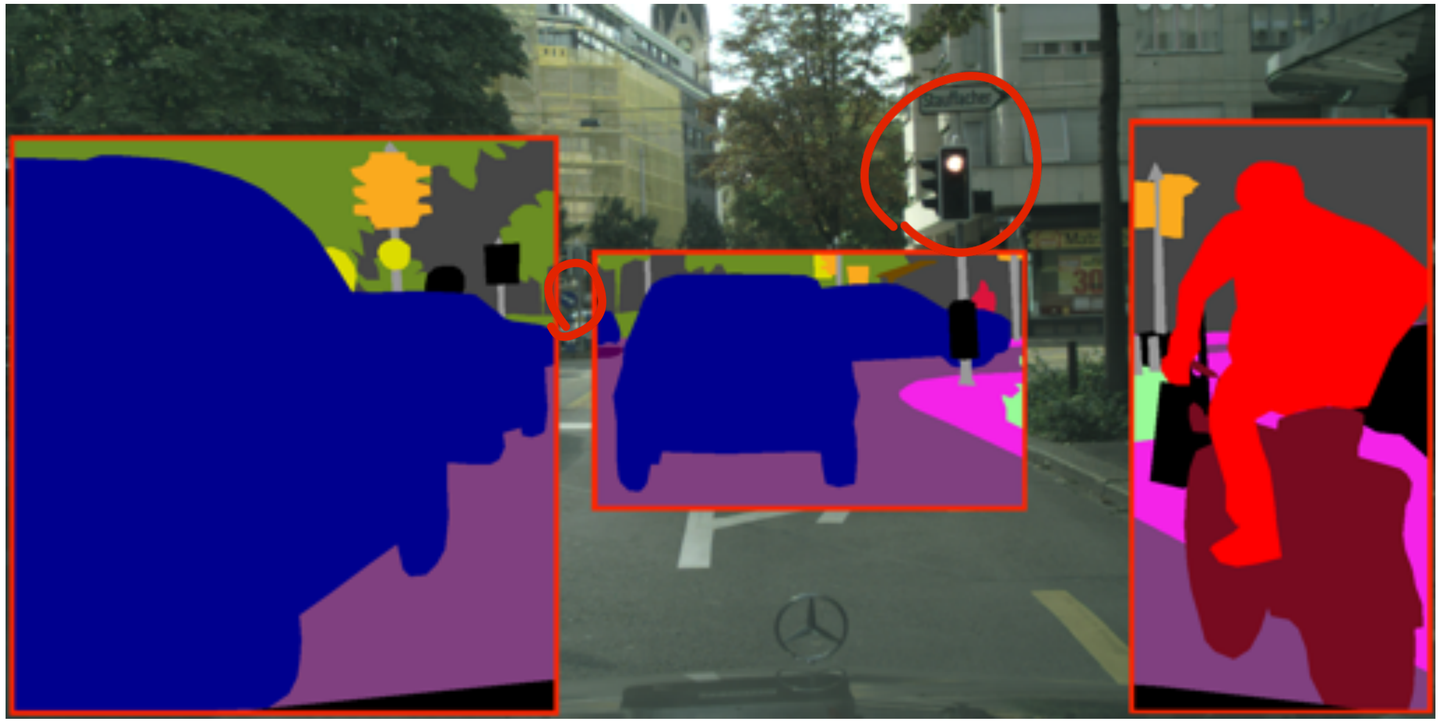
1. Frame filtering: only sending the interesting frames to the cloud
2. Cropping: only sending the interesting regions to the cloud
3. Harmless degradation: tuning the task-specific video knobs to balance

- These techniques may not fit robotics applications because of the different requirements and they are not effective to improve ***tail accuracy***.

Filter and cropping may miss details

- Frame filtering is more suitable for **stationary** rather than **moving cams**. 88%-94% of frames contain important objects.

- Frame cropping may miss important details.
  For robotics applications, region-of-interest (ROI) is the full frame.



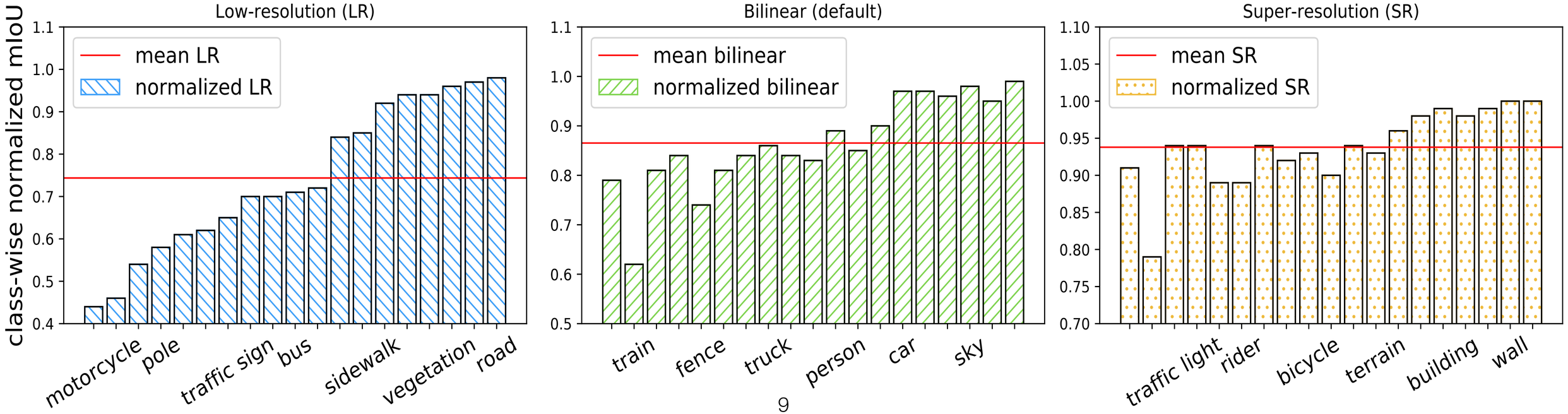
Frame cropping

Tail accuracy

- More importantly, managing accuracy and bandwidth is not enough.
- The tail performance means the worst performance, e.g., 90% or 95% tails
- Accuracy (**m**IoU, **m**AP) does not reflect the worst performance.
- 3-way trade-off: bandwidth v.s. accuracy v.s. **tail accuracy**
- **Class-wise** tail accuracy and **frame-wise** tail accuracy

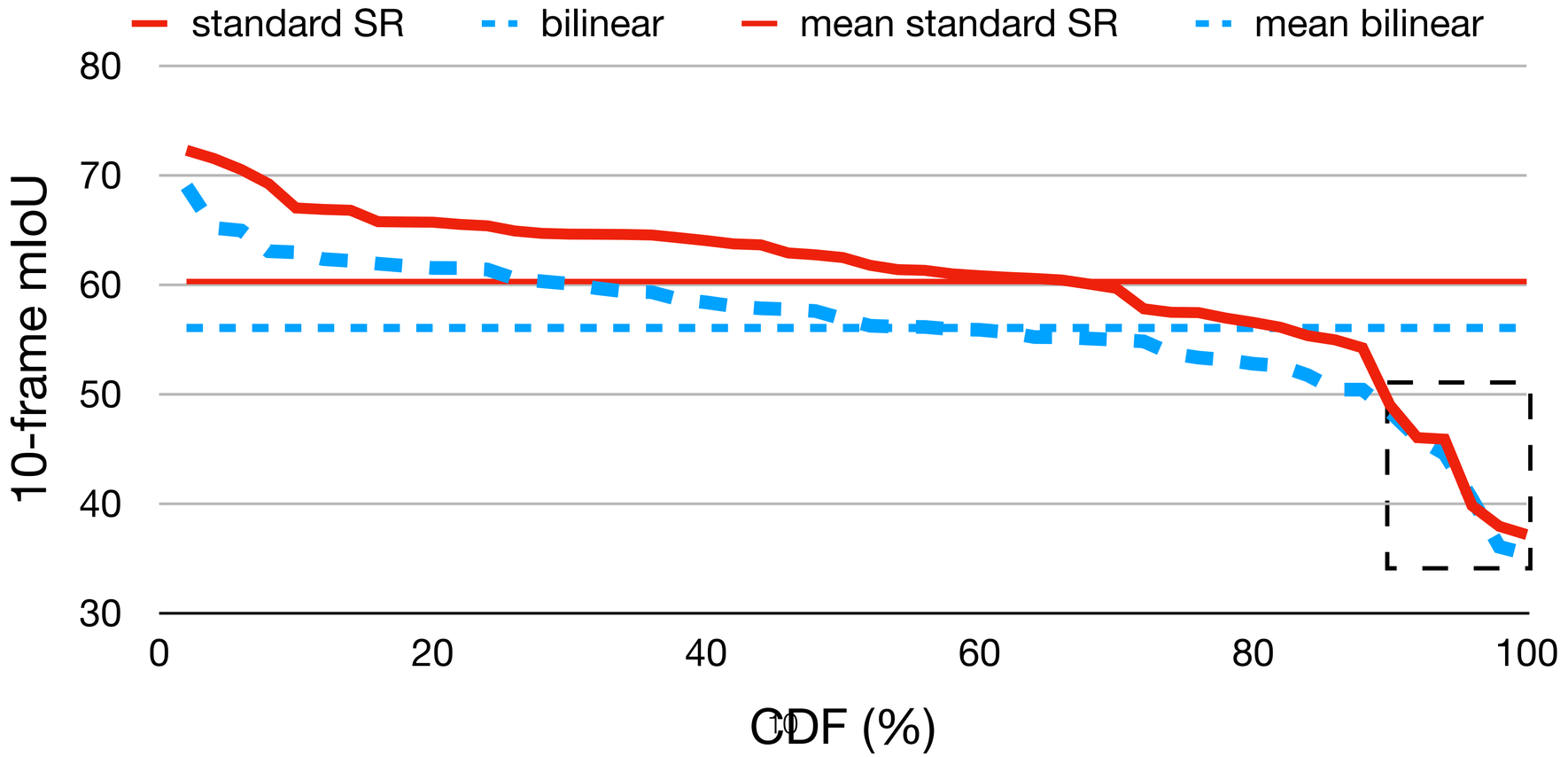
Class-wise tail accuracy

- Quality reconstruction techniques can improve the mean accuracy, but does little help to the hard-to-identify classes.
- “Tail classes” include motorcycles, bicycles, riders, traffic lights/signs, and buses.



Frame-wise tail accuracy

- Unbalanced performance on degraded data over a temporal series of frames
- This is caused by the hard-to-analyze, complex, and quality-sensitive frames.



Our work Runespoor

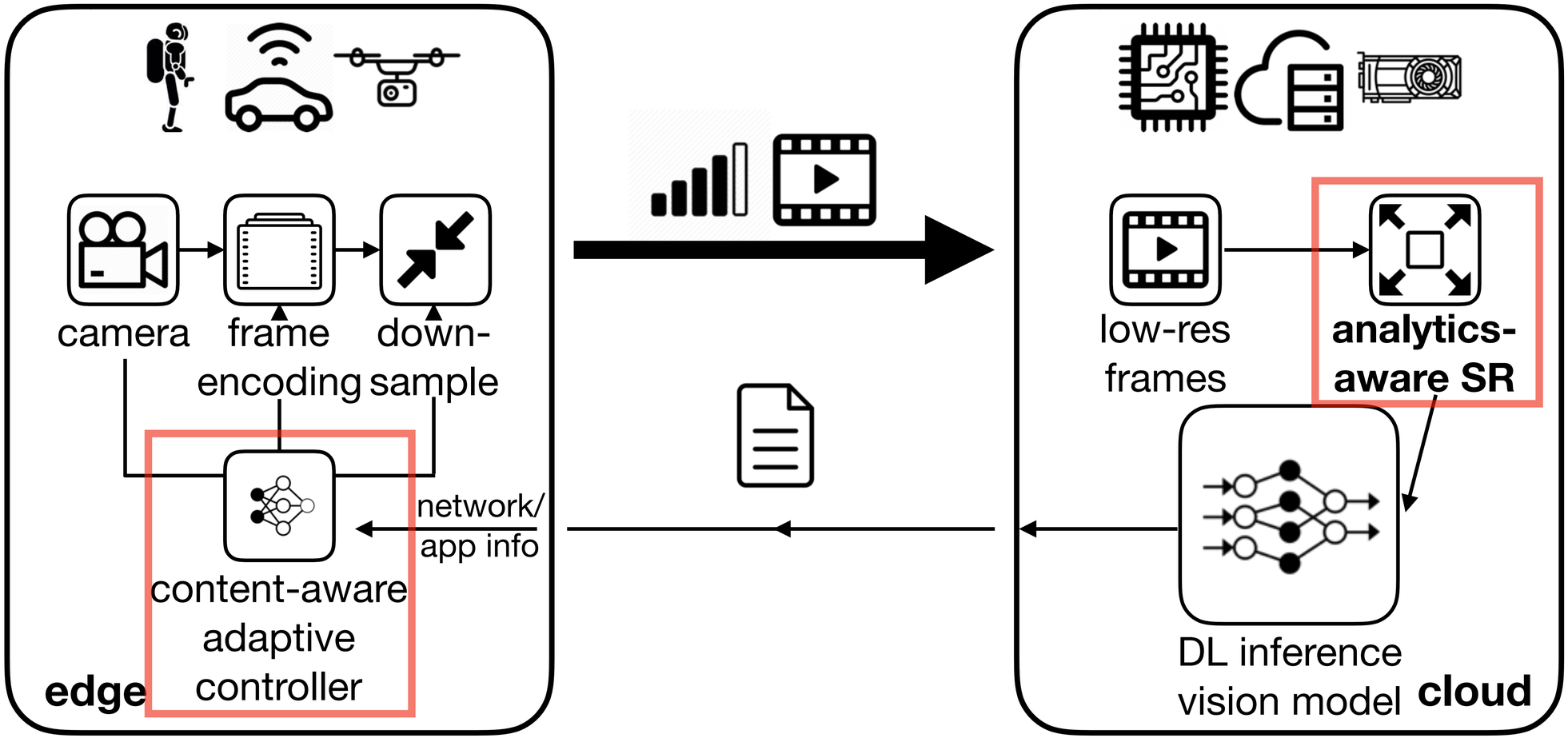
- Goal: Achieving **high accuracy** and **high tail accuracy** under **limited bandwidth** for advanced video analytics tasks.

1. We expose and define the important **tail accuracy problem** in edge-cloud video analytics.

2. We propose **Analytics-aware Super-Resolution** (ASR) that fixes tail accuracy by focusing on detailed information reconstruction.

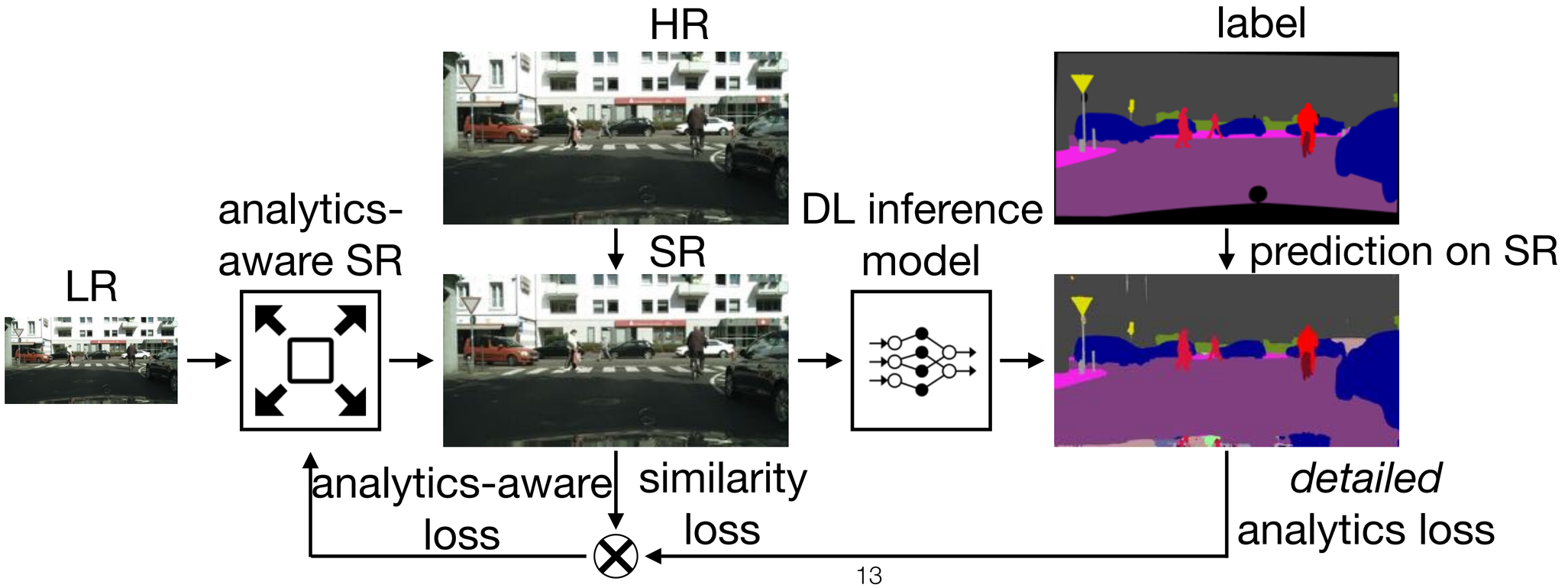
3. We use **Content-aware Adaptive Controller** (CAC) that adapts to fast-changing scenes with DL outputs in an end-to-end system.

Runespoor overview



Analytics-aware Super-Resolution (ASR)

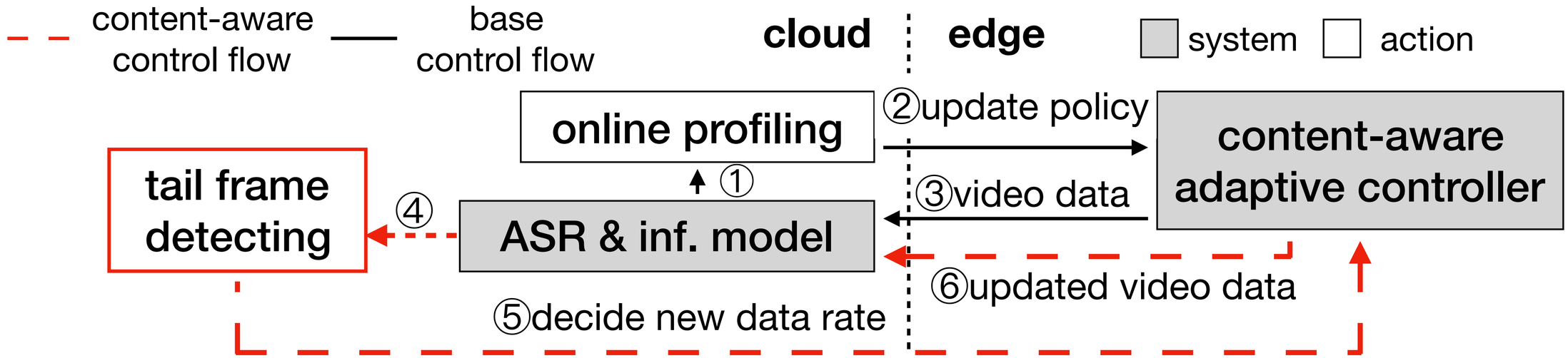
- Training a SR model uses a pair of high- and low-resolution images.
- Analytics-aware training uses the DL inference model and labeled data.
- Quality video for machine analytics, instead of human viewers.



Content-aware Adaptive Controller (CAC)

- CAC monitors the content and determines hard-to-analyze frames on the cloud, and adjusts the data rate on the edge.

- Without ground truth, we profile the relation between the ratio of small regions and inference accuracy use it to detect “tail frames”.



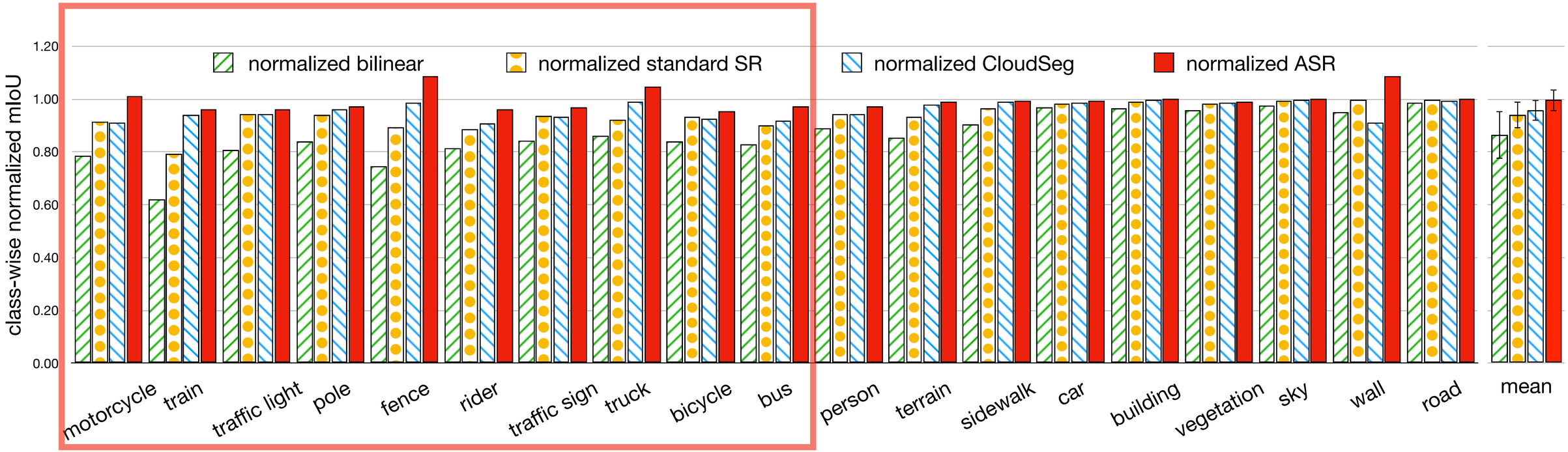
Evaluation

- Serving the latest CV models on high-quality datasets (~2K resolution, 17 fps)
- Cityscapes for semantic segmentation and VisDrone for object detection
- Downsampled by 2×-8× to save bandwidth with standard codecs.

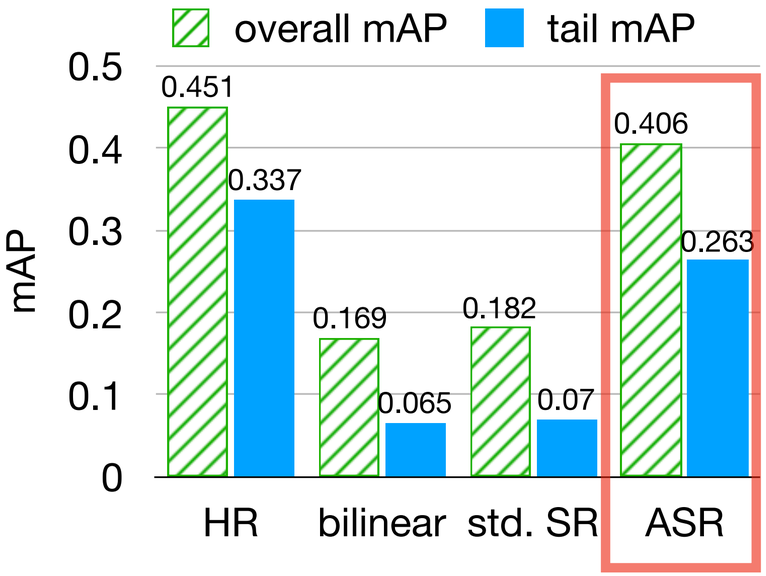
Higher accuracy for tail classes

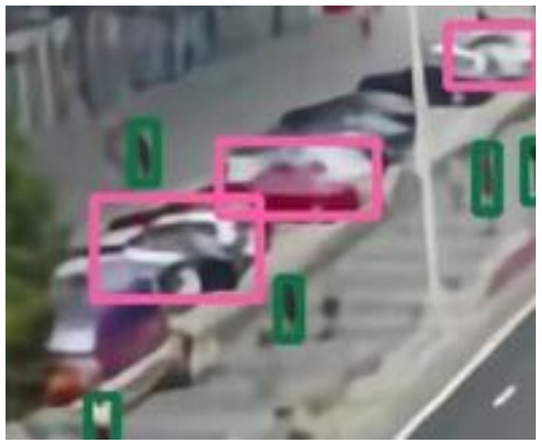
- For 4× downsampled data, ASR can largely improve the inference accuracy of hard-to-label classes, compared to other recent image reconstruction works.

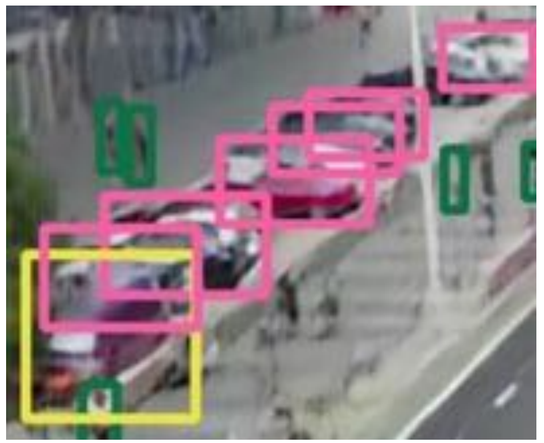
- For 50%-100% classes (the worst 9/19), the improvement is > 20%.



- For object detection, we can correctly detect more details (people, cars).

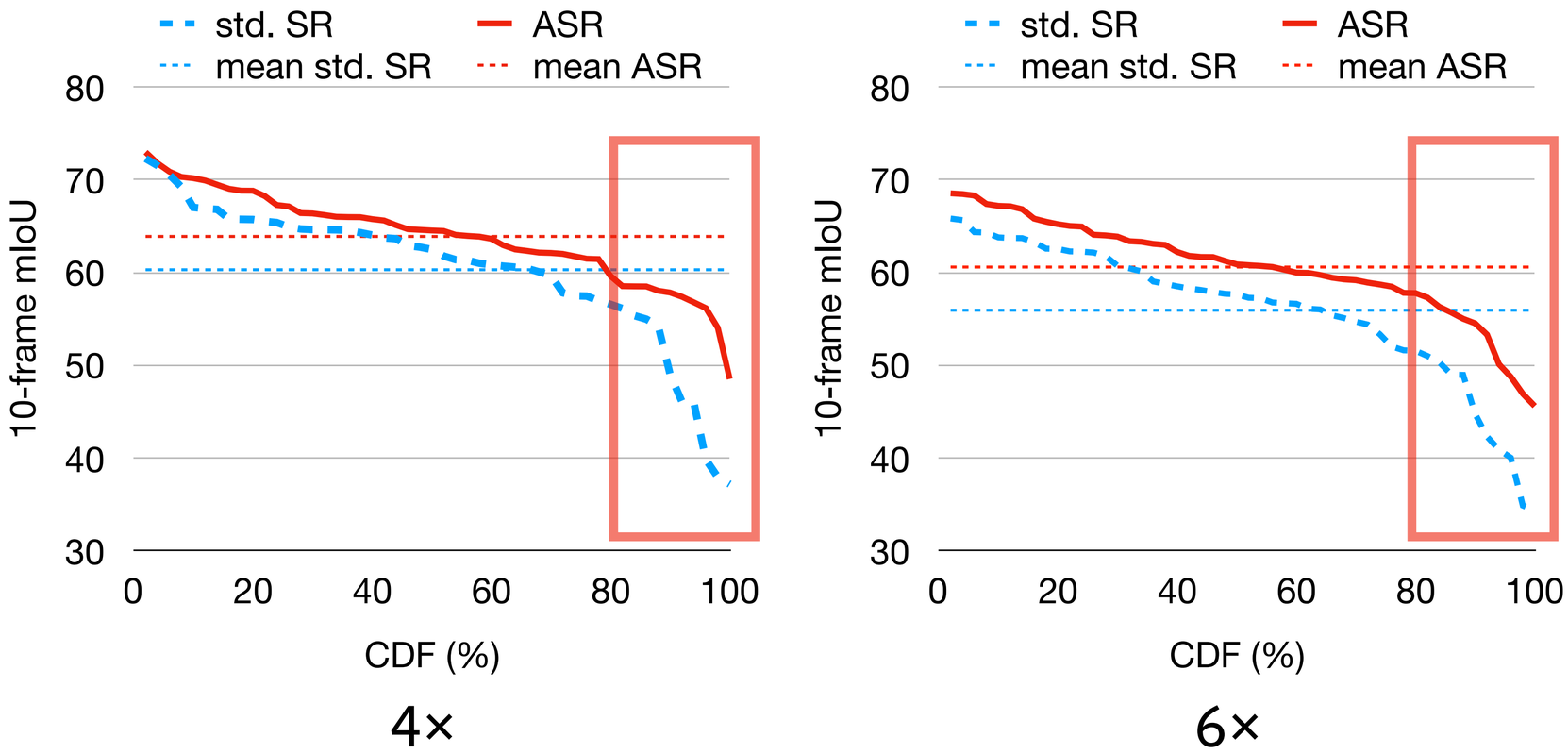






Improve the most difficult frames

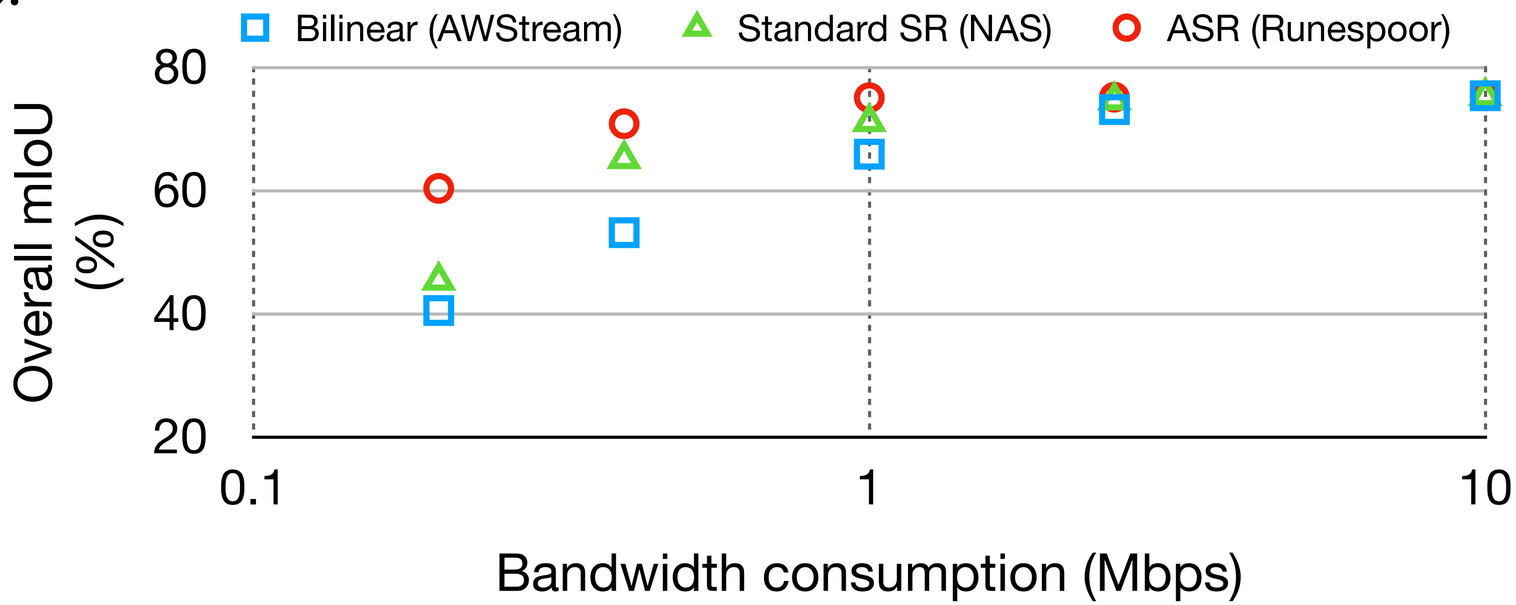
- For semantic segmentation, we improve the 90% and 99% accuracies by 18%-22% and 35%-54%, compared to the latest ML-based technique.



Overall accuracy and bandwidth saving

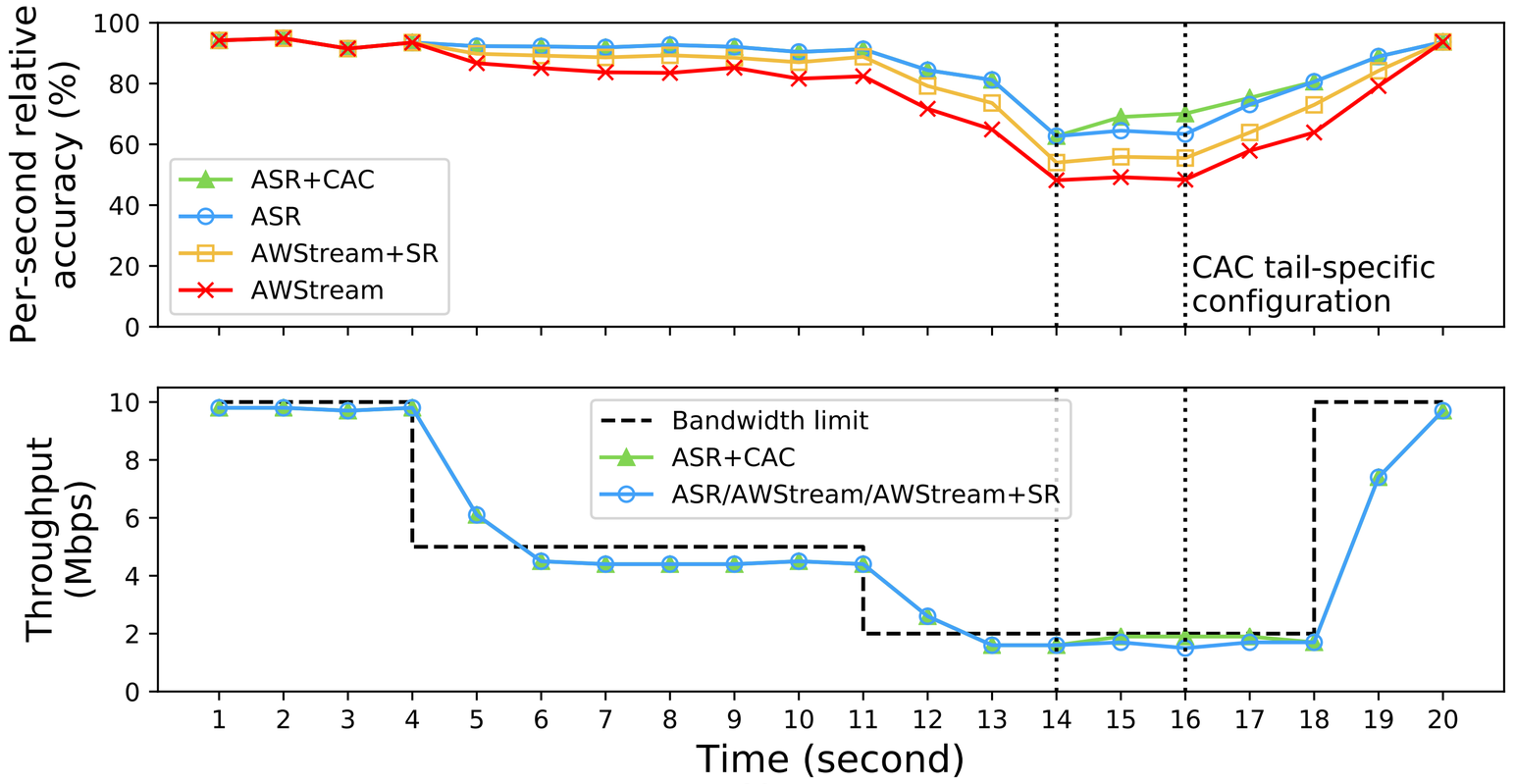
- For overall accuracy, we improve the overall accuracy by 1%-33% compared to the latest ML-based techniques.

- To reach the same high accuracy (70% mIoU), Runespoor saves 2.1× and 6.9× bandwidth consumption compared to the latest streaming & analytics systems.



End-to-end performance with CAC

- Content-aware adaptive controller (CAC) detects the hard frames and improves the frame-wise tail performance under same bandwidth constraints.



Summary

- To enable edge-cloud video analytics for robotics applications, tail accuracy is important in addition to overall accuracy and bandwidth consumption.
- We propose Runespoor, an edge-cloud system to reconstruct degraded video on the cloud and ensure the online performance of system.
- **Analytics-aware Super-Resolution** (ASR) improves the ML technique by focusing on detailed information reconstruction for analytics tasks.
- **Content-aware Adaptive Controller** (CAC) reuses DNN inference results to adapt to fast-changing scenes with fine-grained data rate control.