NNU · 南京师范大学
NANJING NORMAL UNIVERSITY

2025 icassp HYDERABAD, INDIA
Celebrating Signal Processing

Online Optimization of Offloading Video Analytics Tasks to Multiple Edges for Accuracy Maximization

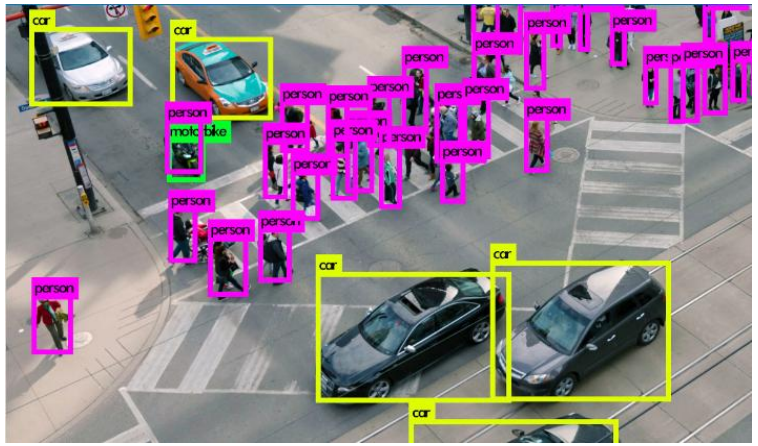
Yu Liang (Nanjing Normal University)
Sheng Zhang (Nanjing University)
Jie Wu (Temple University)

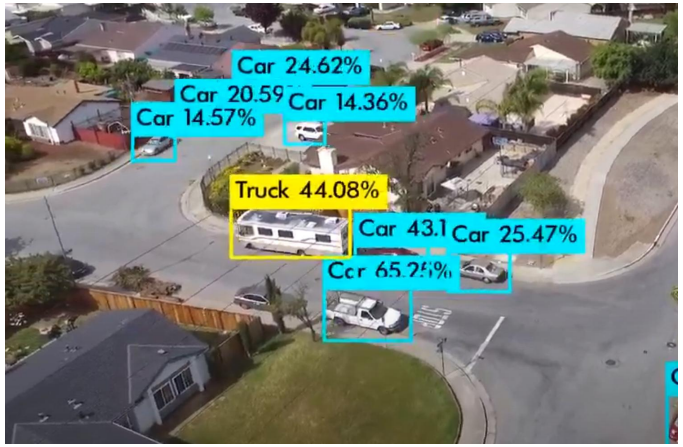
Outline

- Background
- System Model
  - Key Idea
  - Problem Formulation
- Solution
  - Online Algorithm: Design and Analysis
- Experiments

Real-time Video Analytics : Taking Object Detection as Example

- Target Scenarios





Real-time Video Analytics

Offloading Real-time Video Analytics Tasks

- Due to limited local resources of terminal devices, they often offload video analytics tasks to edge servers.
- Offloading introduces additional delays; however, in many typical application scenarios, there is an urgent requirement for high-accuracy, low-latency analytics results.

Design efficient offloading algorithms faces several challenges

Challenge #1: Computation-intensive analytics

Frame-by-frame video analytics is a computationally intensive task that consumes significant computing resources and can lead to notable delays

	Inference Time (on 2080Ti)	Inference Time (on Jetson)
Faster-RCNN	83ms	-
MobileNet	-	52ms
RetinaNet	-	54ms

Challenge #2: Limited resources on terminal devices

Computing resources on devices are typically limited, unable to support high-precision neural networks

- Terminals can only perform lightweight computational inference, failing to provide high-precision analytics results.
- Moreover, offloading video analytics tasks to edge servers results in longer delays. Thus, how to maximize video analytics accuracy while ensuring real-time performance is another pressing issue.

Challenge #3: Online and time-sensitive analytics

In real-time video stream analytics applications, video frames usually arrive online and need to be processed immediately upon arrival

- This requires preemptive offloading decisions for video frames, i.e., determining whether and which server to offload the incoming video frame in advance.
- Therefore, making preemptive offloading decisions based on existing video analytics results for upcoming video frames presents another research challenge.

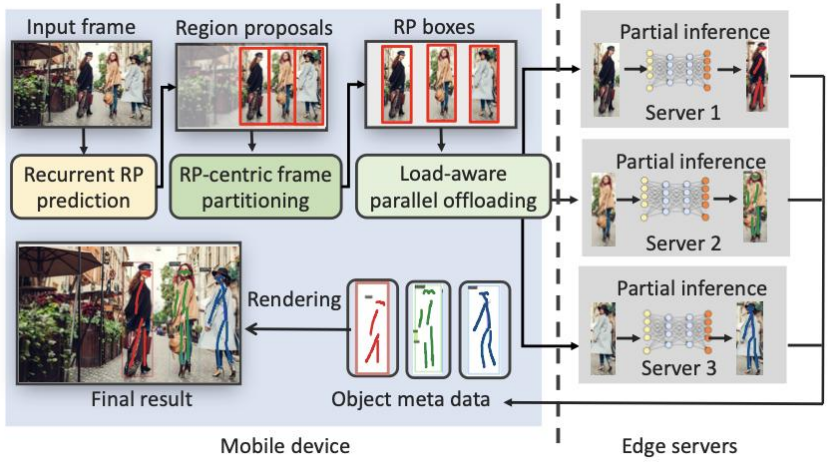
Prior Studies

edge-assisted video analytics task offloading

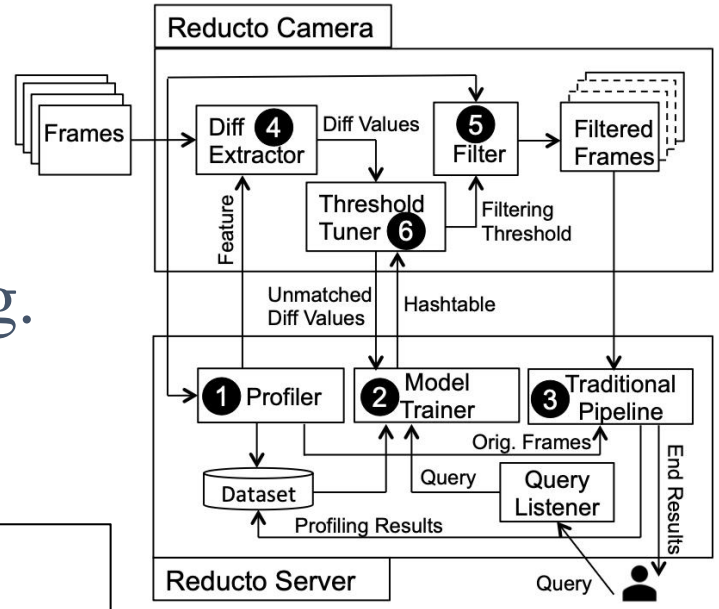
or

frame-filtering object detection

e.g.



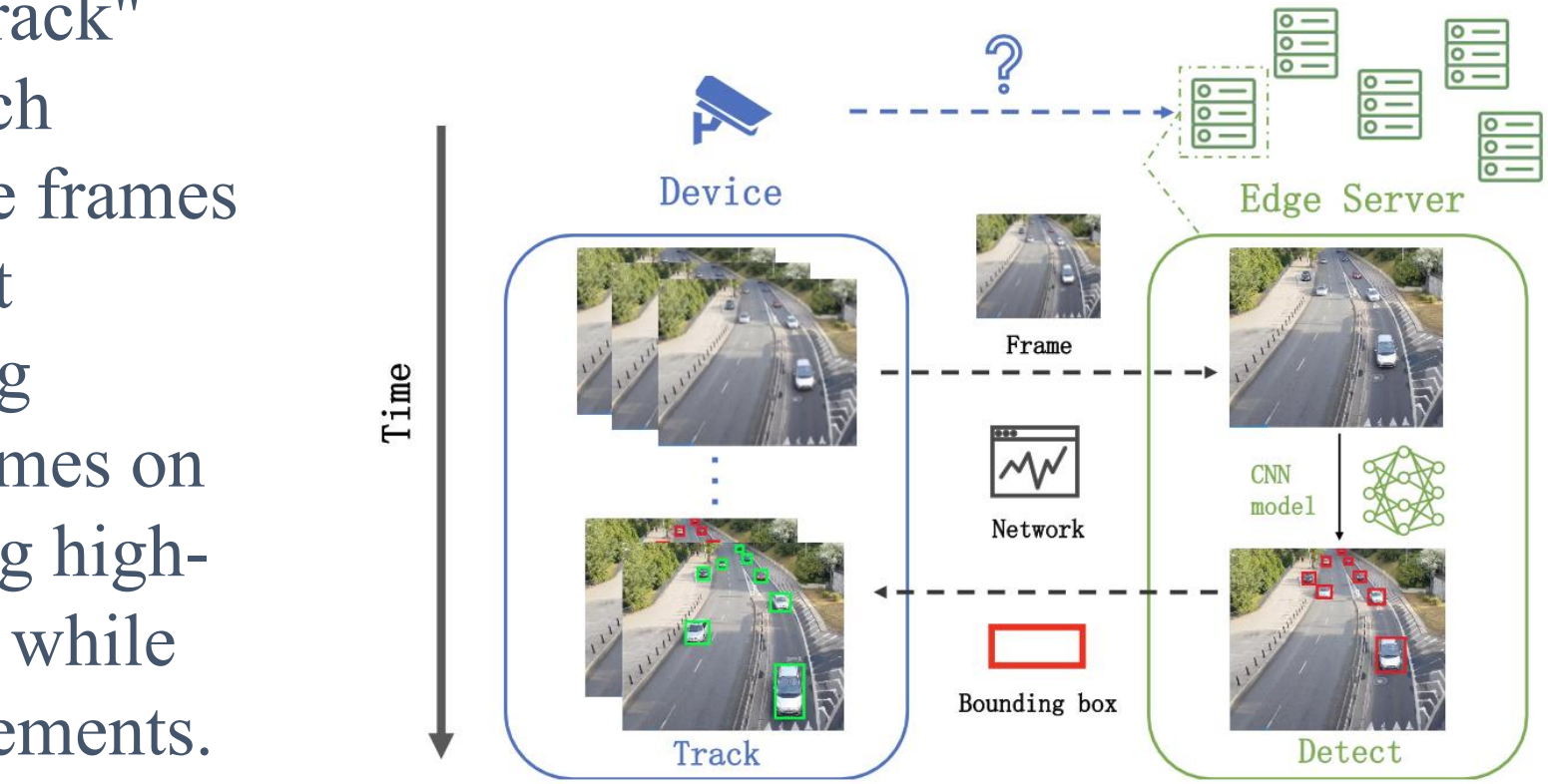
e.g.



none of them fully solved the aforementioned three challenges

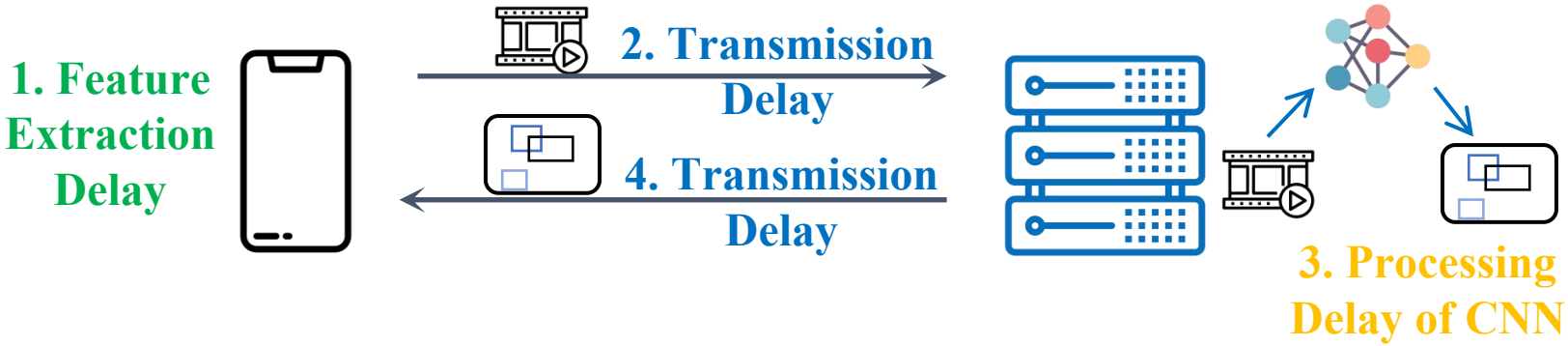
Our Solution: Overview

We adopt the "detect + track" processing strategy, which involves offloading some frames to edge servers for object detection, and performing tracking for the other frames on the device, thus achieving high-accuracy video analytics while meeting real-time requirements.



System Model

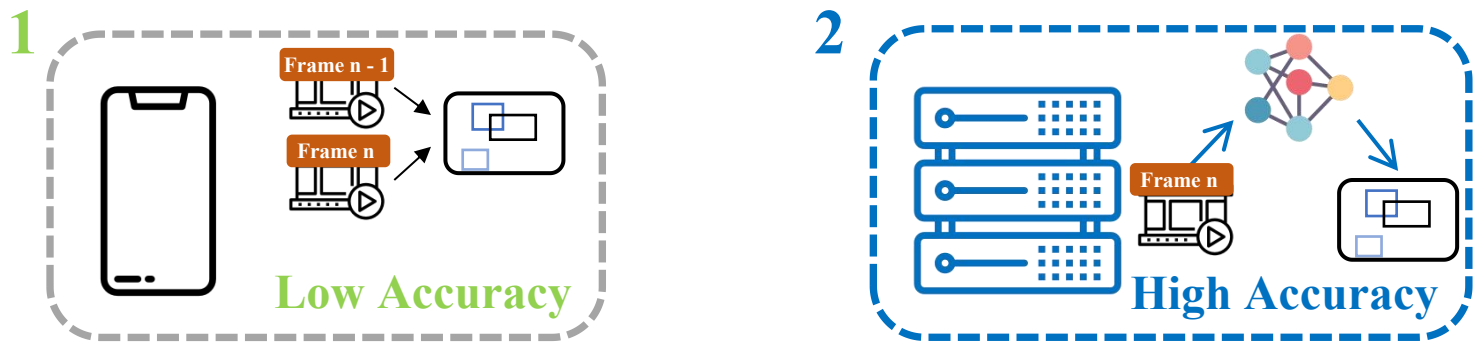
- **Key Idea**
  - **Find offloading decisions to**
    - ✓ **Ensure that the overall delay is within the time constraint.**



    - ✓ **Minimize the overall accuracy loss in the long term.**

System Model

- **Key Idea**
  - Find offloading decisions to
    - ✓ Ensure that the processing time is within the time constraint.
    - ✓ **Minimize the overall accuracy loss in the long term.**



Decide offloading decisions to maximize accuracy under a given latency budget

System Model

Problem Formulation

accuracy of edge-side CNN inference

accuracy of on-device object tracking

$$\min \sum_{f,t,s} \left(1 - \sum_{b^* \in \mathcal{B}_f^G} \max_{b \in \mathcal{B}_f^I} \{ x_{f,s} \cdot IoU_{b,b^*}^{s} + (1 - x_{f,s}) \cdot IoU_{b,b^*}^{D} \}\right)$$

minimize the long-term accuracy loss of the detection of objects

$$s.t. \quad C_1 : \sum_{t \in \mathcal{T}} \sum_{f,s} x_{f,s} \geq \sum_{t \in \mathcal{T}} \sum_{f} \tau,$$

minimum detection frequency constraint

$$C_2 : \sum_{f,s} \{ x_{f,s} L_s^S + (1 - x_{f,s}) L^D \} \leq L_{max}, \forall t \in \mathcal{T},$$

time on edge: $x_{f,s} L_s^S$; time on device: $(1 - x_{f,s}) L^D$

real time constraint

$$C_3 : \sum_{s} x_{f,s} \leq 1, \forall f \in \mathcal{F}_t,$$

only one edge can be selected per online decision

$$var. \quad C_4 : x_{f,s} \in \{0, 1\}, t \in \mathcal{T}, f \in \mathcal{F}_t, s \in \mathcal{S}.$$

$x_{f,s} = 1$: offloaded to edge

$x_{f,s} = 0$: processed locally

Our goal is to minimize accuracy loss under a given latency budget

Solution Analysis

- **Challenge #1**
  - **Since the ground truth cannot be obtained before the analytics is completed, the existing detection results need to be utilized to replace the ground truth.**

$$\min \sum_{f,t,s}(1-\sum_{b^* \in \mathcal{B}_f^G} \max_{b \in \mathcal{B}_f^I}\{x_{f,s} \cdot IoU_{b,b^*}^{s} + (1-x_{f,s}) \cdot IoU_{b,b^*}^{D}\})$$

$$s.t. \quad C_1: \sum_{t \in \mathcal{T}} \sum_{f,s} x_{f,s} \geq \sum_{t \in \mathcal{T}} \sum_{f} \tau,$$

$$C_2: \sum_{f,s} \{x_{f,s} L_s^S + (1-x_{f,s}) L^D\} \leq L_{max}, \forall t \in \mathcal{T},$$

$$C_3: \sum_{s} x_{f,s} \leq 1, \forall f \in \mathcal{F}_t,$$

$$var. \quad C_4: x_{f,s} \in \{0,1\}, t \in \mathcal{T}, f \in \mathcal{F}_t, s \in \mathcal{S}.$$

$$\min \sum_{f,t,s}(1-\sum_{b^* \in \mathcal{B}_f^G} \max_{b \in \mathcal{B}_f^I}\{x_{f,s} \widetilde{IoU}_{b,b^*}^{s} + (1-x_{f,s}) \widetilde{IoU}_{b,b^*}^{D}\})$$

$$s.t. \quad C_1, C_2, C_3, C_4.$$

Solution

- **Challenge #2**
  - The **long-term detection frequency bound** is hard to analyze.
    - ✓ Use queue-based approach to decouple the original problem into subproblems per interval.

$$Q_{t+1} = \max\{-g_t(\boldsymbol{x}_{t+1}), Q_t + g_t(\boldsymbol{x}_{t+1})\}$$

$$\begin{aligned} \min_{\boldsymbol{x}\in\widetilde{\mathcal{X}}} \quad & \{f_t(\boldsymbol{x}) + (Q_t + g_{t-1}(\boldsymbol{x}_t))g_t(\boldsymbol{x}) + \alpha\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2\} \\ s.t. \quad & \boldsymbol{h}_t(\boldsymbol{x}) \leq 0. \end{aligned}$$

  - **Integral variables** exist.
    - ✓ Relax the domain from integral to rational.

$$\boldsymbol{h}_t(\tilde{\boldsymbol{x}}_t) \leq 0, \tilde{\boldsymbol{x}}_t \in \widetilde{\mathcal{X}}$$

Solution (Sketch)

$$\min \sum_{f,t,s} (1 - \sum_{b^* \in \mathcal{B}_f^G} \max_{b \in \mathcal{B}_f^I} \{x_{f,s} \widetilde{IoU}^{s}_{b,b^*} + (1 - x_{f,s}) \widetilde{IoU}^{D}_{b,b^*}\})$$

$$s.t. \quad C_1, C_2, C_3, C_4.$$

Algorithm 1 Queue-based OnLine Optimization Alg. (OLA)

Input: τ, L_{max}

Initialize a proper step size α, $Q_1 = 0$, $\boldsymbol{g}_0(\cdot) = 0$;
Initialize $\hat{\boldsymbol{x}}_1$ by offloading video frames to fixed edge server, $\dot{\boldsymbol{x}}_1 = \ddot{\boldsymbol{x}}_1 = \hat{\boldsymbol{x}}_1$;
for $t = 1$ to T **do**
  Deploy the provisioning $\hat{\boldsymbol{x}}_t$;
  Solve subproblem $\mathbb{P}^0_{t+1}$ to obtain $\dot{\boldsymbol{x}}_{t+1}$;
  Solve subproblem $\mathbb{P}_{t+1}$ to obtain $\ddot{\boldsymbol{x}}_{t+1}$;
  Update the virtual queue:
  $Q_{t+1} = \max\{-g_t(\dot{\boldsymbol{x}}_{t+1}), Q_t + g_t(\dot{\boldsymbol{x}}_{t+1})\}$;
  if $(\dot{\boldsymbol{x}}_{t+1} - \ddot{\boldsymbol{x}}_{t+1})^{\mathsf{T}}(\dot{\boldsymbol{x}}_t - \ddot{\boldsymbol{x}}_{t+1}) \geq 0$ **then** $\tilde{\boldsymbol{x}}_{t+1} = \ddot{\boldsymbol{x}}_{t+1}$;
  else $\tilde{\boldsymbol{x}}_{t+1} = \dot{\boldsymbol{x}}_{t+1}$;
  Obtain $\hat{\boldsymbol{x}}_{t+1}$ by rounding $\tilde{\boldsymbol{x}}_{t+1}$ randomly;
end for

Theoretical Analysis

Lemma 1. *With the definition of dynamic regret and constraint violation in integral and real domains, we have:*

$$Reg_T \leq \widetilde{Reg}_T, \quad Vio_T = \widetilde{Vio}_T. \tag{11}$$

Lemma 2. *The solution of the objective function f_t over the real domain is bounded:*

$$\sum_{t=1}^{T} f_t(\tilde{\boldsymbol{x}}_t) \leq \sum_{t=1}^{T} f_t(\dot{\boldsymbol{x}}_t) + 2\max_t\{Q_t\} \sum_{t=1}^{T} \theta_t. \tag{12}$$

Theorem 1. *With previous assumptions and lemmas, the integral dynamic regret is upper-bounded as:*

$$Reg_T \leq \widetilde{Reg}_T = o(T), \tag{13}$$

if $V_{\boldsymbol{\lambda}}$, V_f, V_g, $V_{\boldsymbol{x}}$ are subliner with respect to T.

Theorem 2. *With previous assumptions and lemmas, the integral constraint violation is upper-bounded as:*

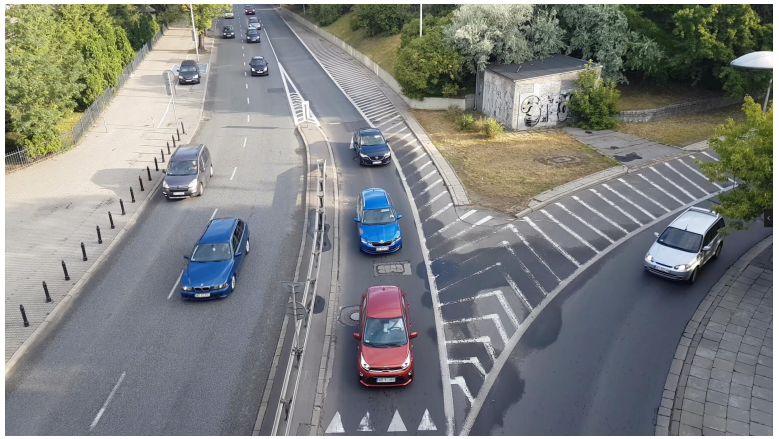
$$Vio_T = \widetilde{Vio}_T = o(T), \tag{14}$$

if $\widetilde{V}_g$, $V_{\boldsymbol{\lambda}}$ are sublinear; V_f, V_g, $V_{\boldsymbol{x}}$ are subquadratic with T.

Via rigorous proof, both dynamic regret regarding detection accuracy and the real-time requirement are ensured.

Evaluation Settings

- **Hardware**
  - Jet AGX Orin as edge server, Jetson AGX Xavier as terminal device
- **Software**
  - Yolov5s, v5m, v5l deployed on edge server for object detection
  - OpenCV for object tracking
  - CVXPY for convex optimization
- **Data**
  - Real-world monitoring videos [14] and bandwidths [15]
- **Baselines**
  - Frame-by-frame offloading
  - Fixed-Frame offloading



Experiment Result

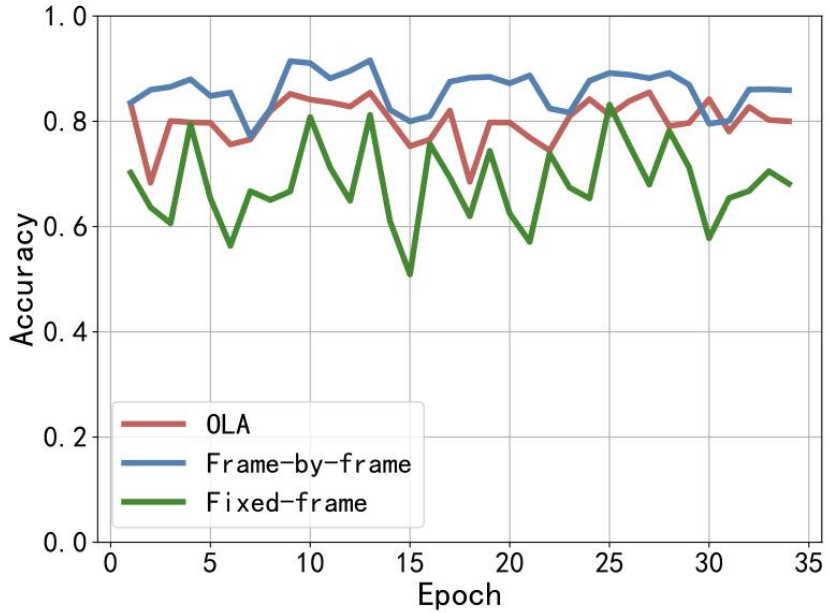


Fig. 2. Video analytics accuracy per epoch

OLA achieves an average accuracy of 79.9%, representing a 17.3% improvement over Fixed-Frame (68.12%)

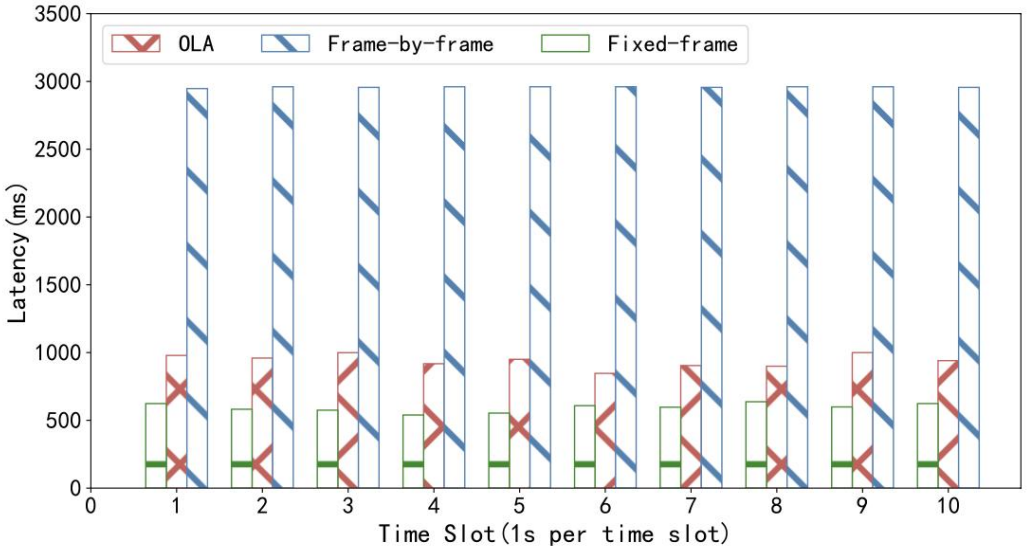


Fig. 3. Overall latency

Although Frame-by-Frame achieves the highest accuracy, its overall latency reaches 2,997ms, which is 3.17 times that of OLA, with only a 6% increase in accuracy. Since the Fixed-Frame offloads fewer video frames, it has the lowest latency but also the lowest accuracy.

Conclusions

- This work adopts a "detect + track" approach, offloading video frames to edge servers for object detection at a relatively high frequency while continuously tracking detected objects on the terminal device using tracking technology between consecutive frames.
- We propose a queue-based online optimization algorithm, OLA.
- Experiments demonstrate that OLA can achieve high-accuracy video analytics results while meeting real-time requirements.
- Rigorous theoretical proof verifies that the dynamic regret bound and constraint violation bound grow sublinearly over time.

Online Optimization of Offloading Video Analytics Tasks to Multiple Edges for Accuracy Maximization

Yu Liang (Nanjing Normal University)
Sheng Zhang (Nanjing University)
Jie Wu (Temple University)