nsdi '25
22nd USENIX Symposium on Networked Systems Design and Implementation

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

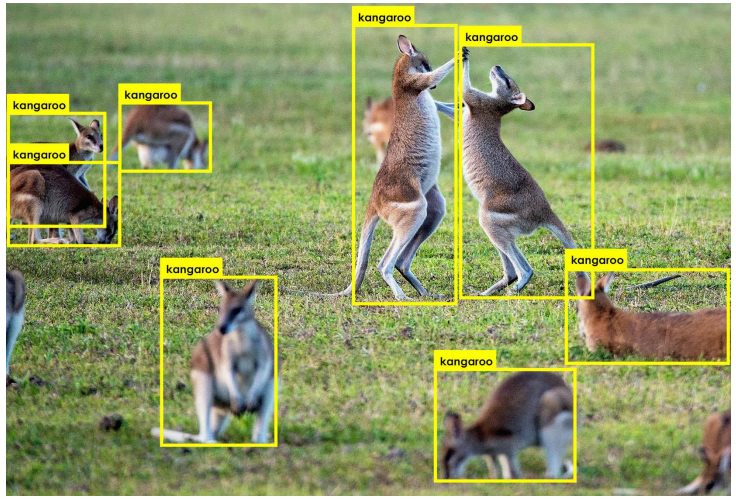
Region-based Content Enhancement for Efficient Video Analytics at the Edge

Weijun Wang*, **Liang Mi***, Shaowei Cen, Haipeng Dai, Yuanchun Li, Xiaoming Fu, Yunxin Liu

南京大学 NANJING UNIVERSITY
清華大學 Tsinghua University
UNIVERSITY OF GOTTINGEN GERMANY

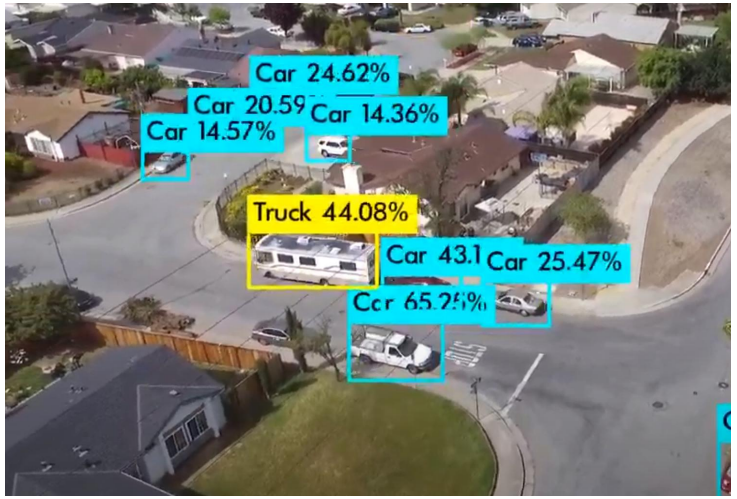
Video analytics is pervasive

Video analytics has been widely used in daily life



Wild-life camera
Learn about the habit of animals

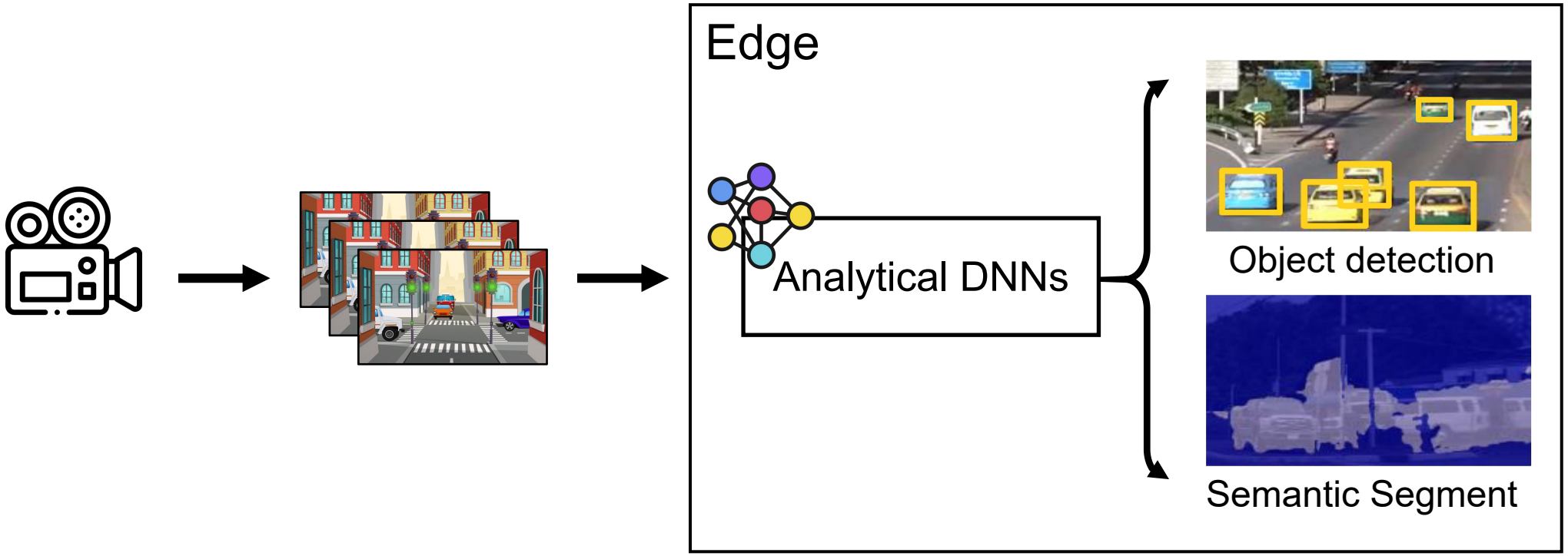
Traffic camera
Monitor the traffic condition



Drone camera
Estimate the number of cars

Video analytics pipeline

Accuracy↑, E2E latency↓ / Throughput ↑, Resource↓



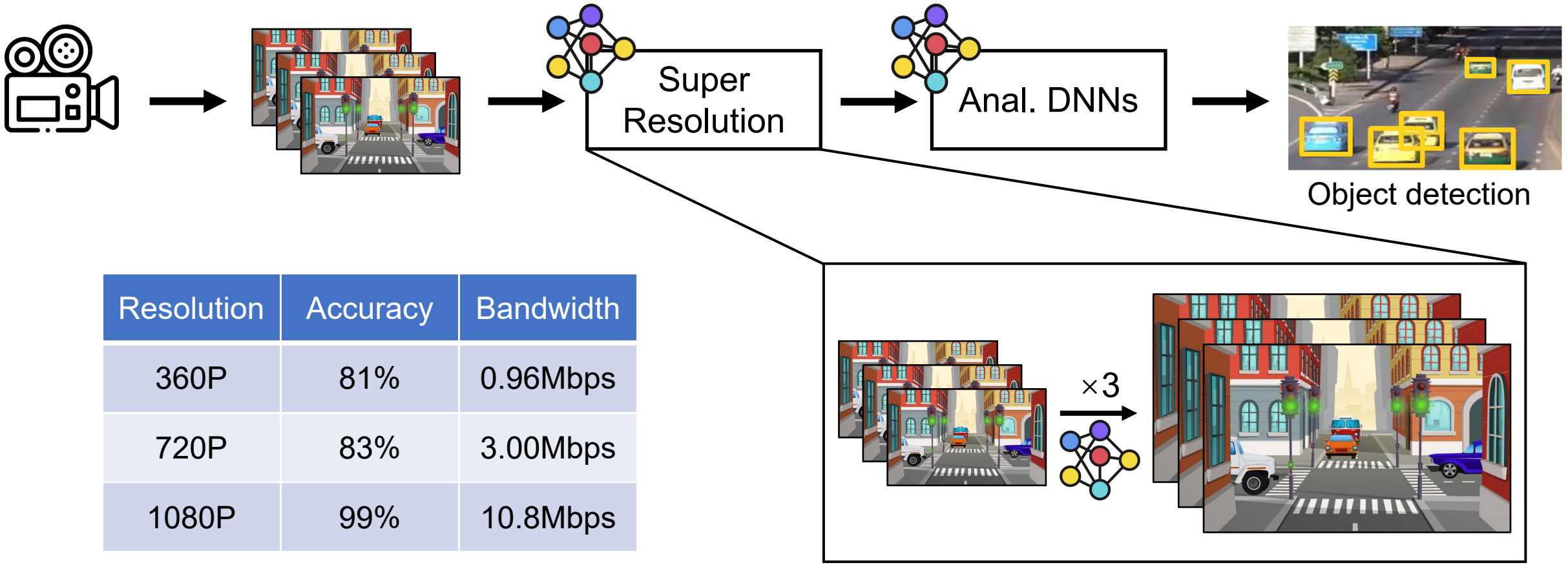
Video analytics is resource intensive

Networking Resource Intensive!

Resolution	Accuracy	Bandwidth
360P	81%	0.96Mbps
720P	83%	3.00Mbps
1080P	99%	10.8Mbps

Benefits of Super Resolution

Enhancing image content before analytics.

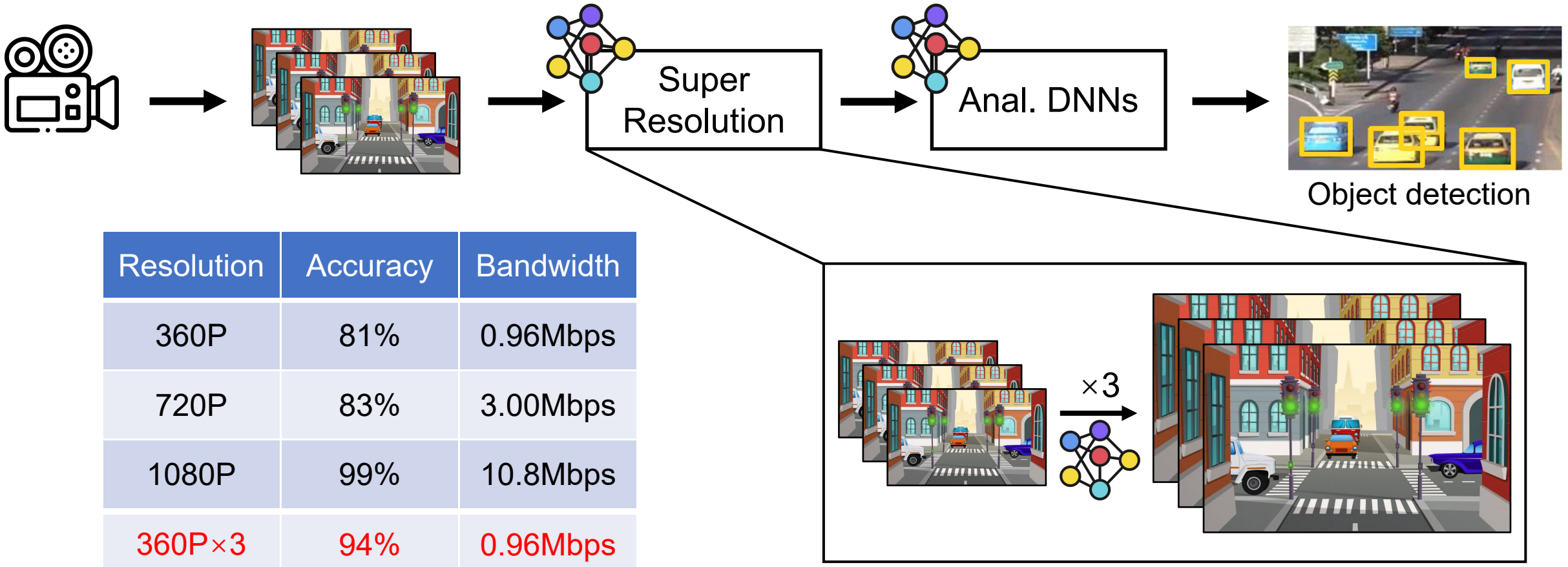


Resolution	Accuracy	Bandwidth
360P	81%	0.96Mbps
720P	83%	3.00Mbps
1080P	99%	10.8Mbps

Neural Adaptive Content-aware Internet Video Delivery, OSDI'18
Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning, SIGCOMM'20

Benefits of Super Resolution

Enhancing image content before analytics.

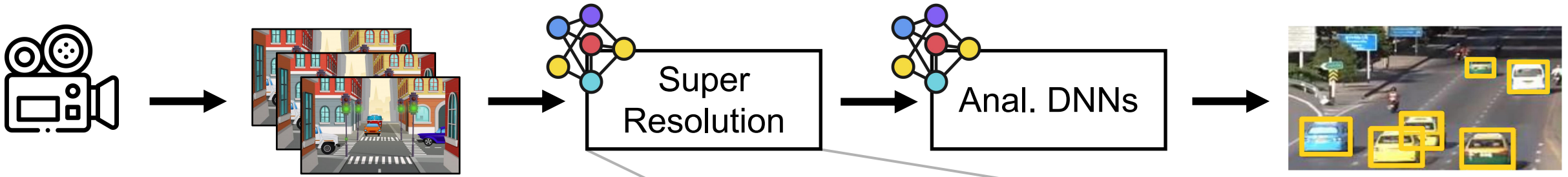


Resolution	Accuracy	Bandwidth
360P	81%	0.96Mbps
720P	83%	3.00Mbps
1080P	99%	10.8Mbps
360P×3	94%	0.96Mbps

Neural Adaptive Content-aware Internet Video Delivery, OSDI'18
Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning, SIGCOMM'20

Benefits of Super Resolution

Enhancing image content before analytics.



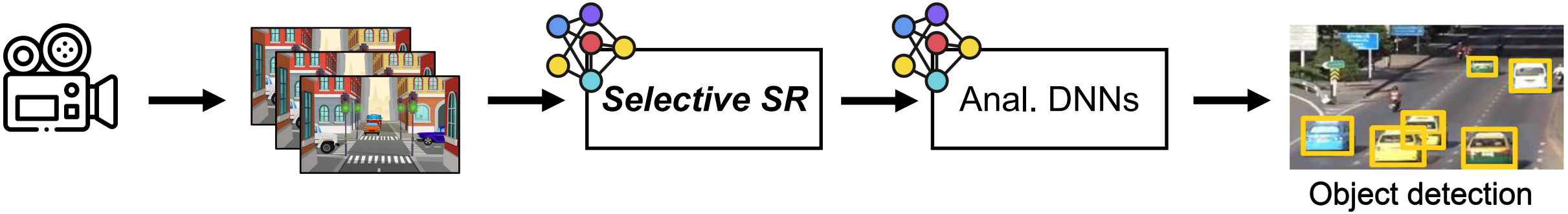
Resolution	Accuracy	Bandwidth
360P	81%	0.96Mbps
720P	83%	3.00Mbps
1080P	99%	10.8Mbps
360P×3	94%	0.96Mbps

SR introduces additional computing cost.
SR DNN is 7×FLOPS and 4×time cost than analytical DNN, e.g., YOLO.

Neural Adaptive Content-aware Internet Video Delivery, OSDI'18
Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning, SIGCOMM'20

Save computing resources with *selective SR*

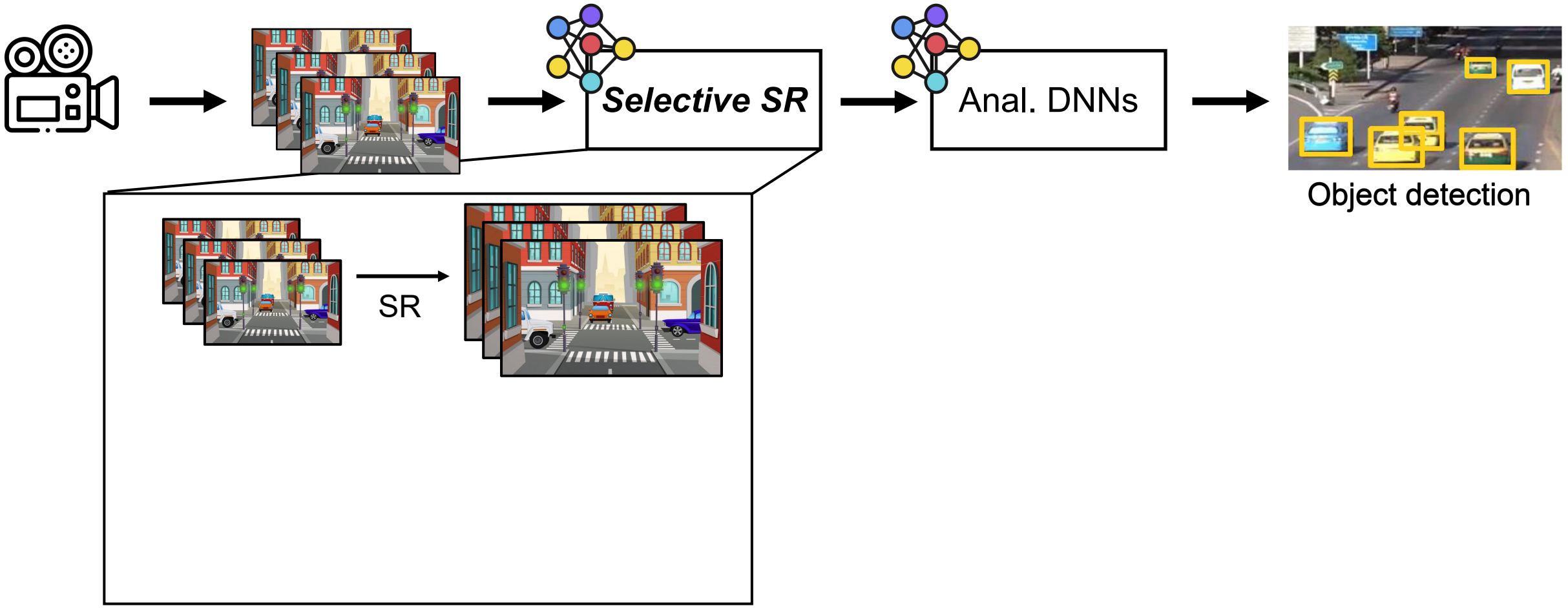
Enhance a few of frames and reuse them on continuous ones



NEMO: Enabling Neural-enhanced Video Streaming on Commodity Mobile Devices, MobiCom'20
NeuroScaler: Neural Video Enhancement at Scale, SIGCOMM'22

Save computing resources with *selective SR*

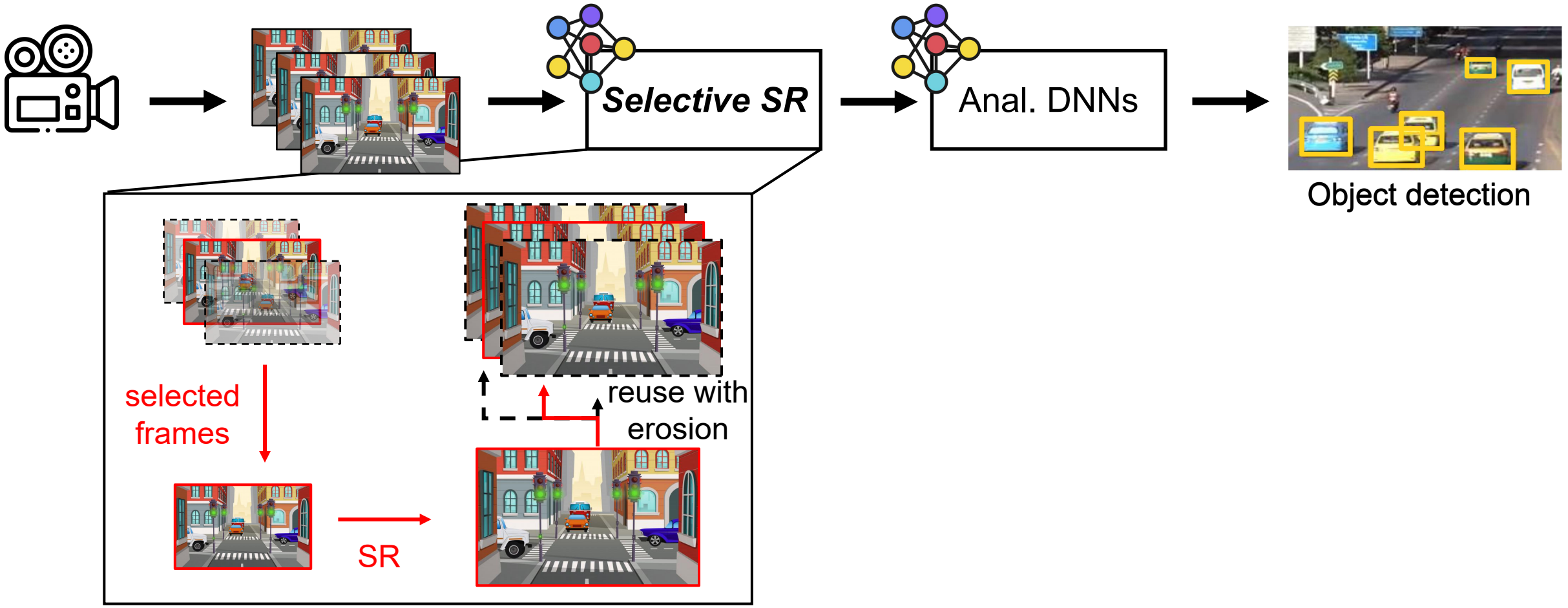
Enhance a few of frames and reuse them on continuous ones



NEMO: Enabling Neural-enhanced Video Streaming on Commodity Mobile Devices, MobiCom'20
NeuroScaler: Neural Video Enhancement at Scale, SIGCOMM'22

Save computing resources with *selective SR*

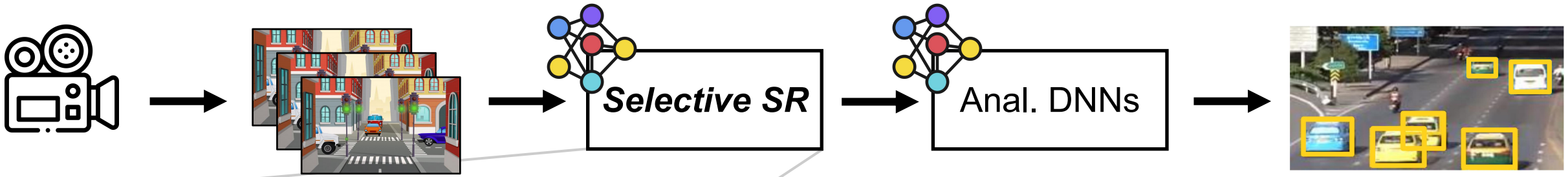
Enhance a few of frames and reuse them on continuous ones

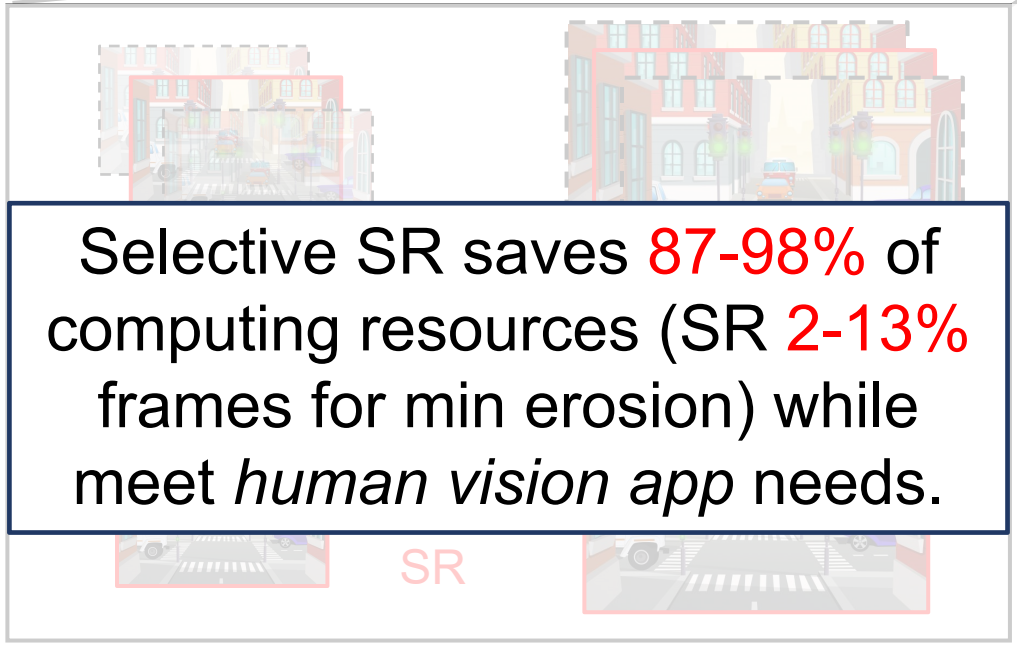


NEMO: Enabling Neural-enhanced Video Streaming on Commodity Mobile Devices, MobiCom'20
NeuroScaler: Neural Video Enhancement at Scale, SIGCOMM'22

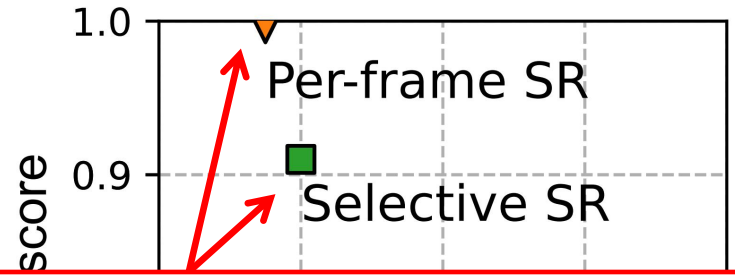
Save computing resources with *selective SR*

Enhance a few of frames and reuse them on continuous ones





Selective SR saves 87-98% of computing resources (SR 2-13% frames for min erosion) while meet *human vision app* needs.

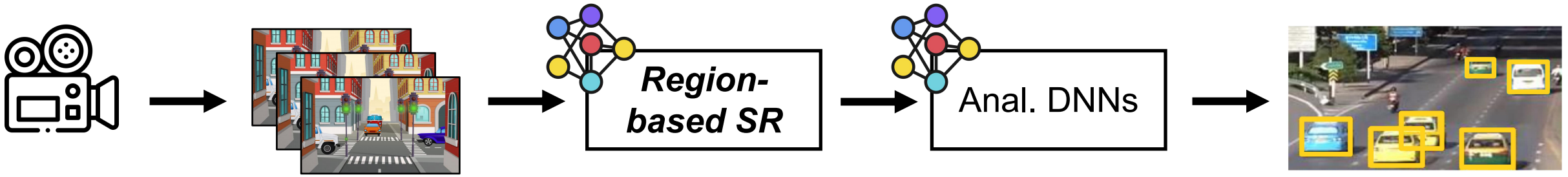


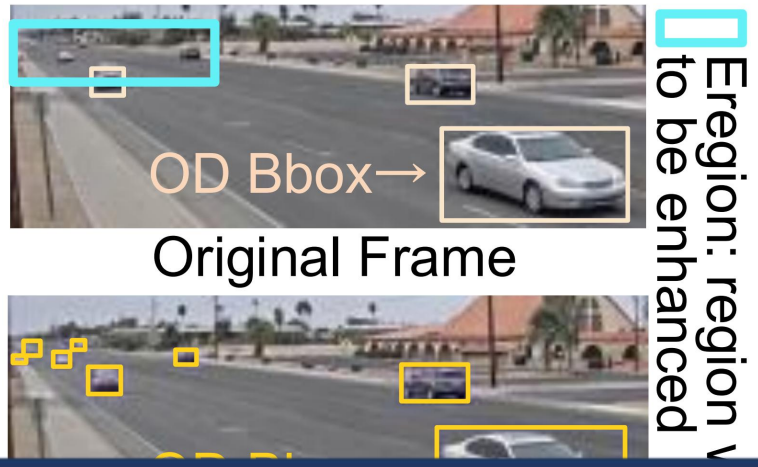
Setting accuracy target >90%, needs SR 41-60% frames! *As analytical tasks are more sensitive on image erosion.*

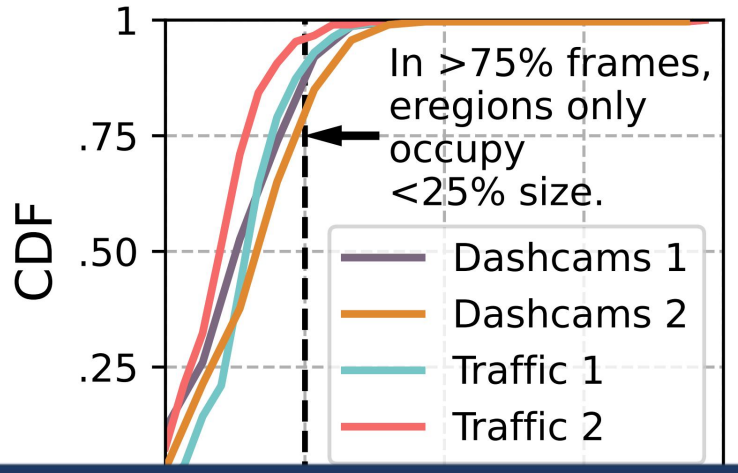
NEMO: Enabling Neural-enhanced Video Streaming on Commodity Mobile Devices, MobiCom'20
NeuroScaler: Neural Video Enhancement at Scale, SIGCOMM'22

Our idea: save computing resources with *Region-based SR*

Enhance *a small portion of content* in each frame







Question: How to execute region-based SR efficiently at the edge?

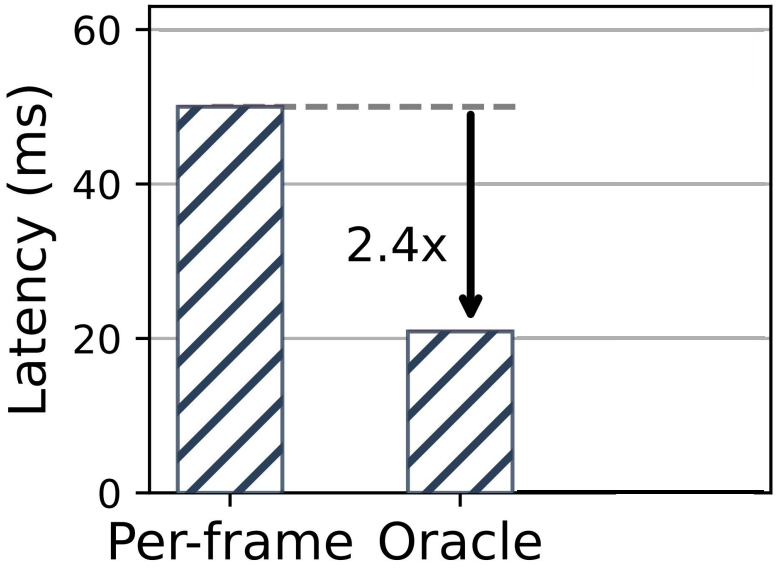
RegenHance

Goal: achieve *high-throughput* video analytics at the edge via *region-based content enhancement*

Challenges:

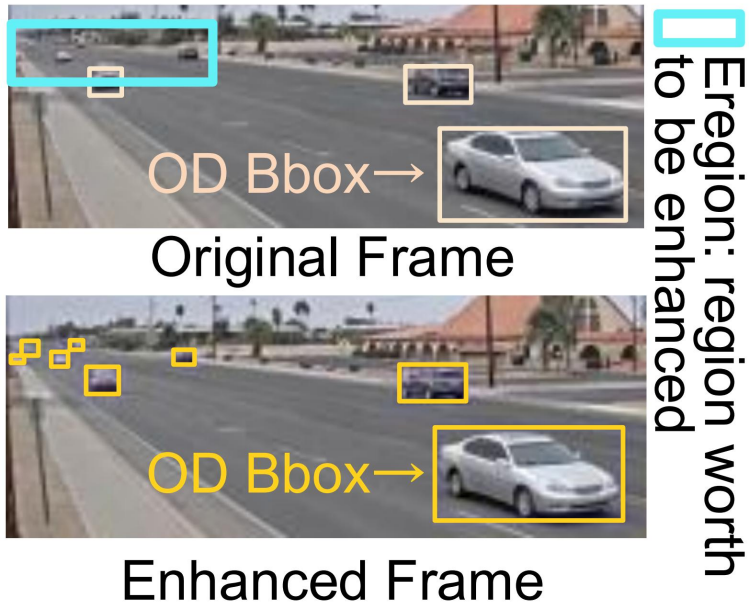
- How to fast and accurately identify eregions on original frames?
- How to enhance eregions across streams in high throughput?
- How to allocate resources among components on edge?

Challenge 1: How to efficiently identify eregions?



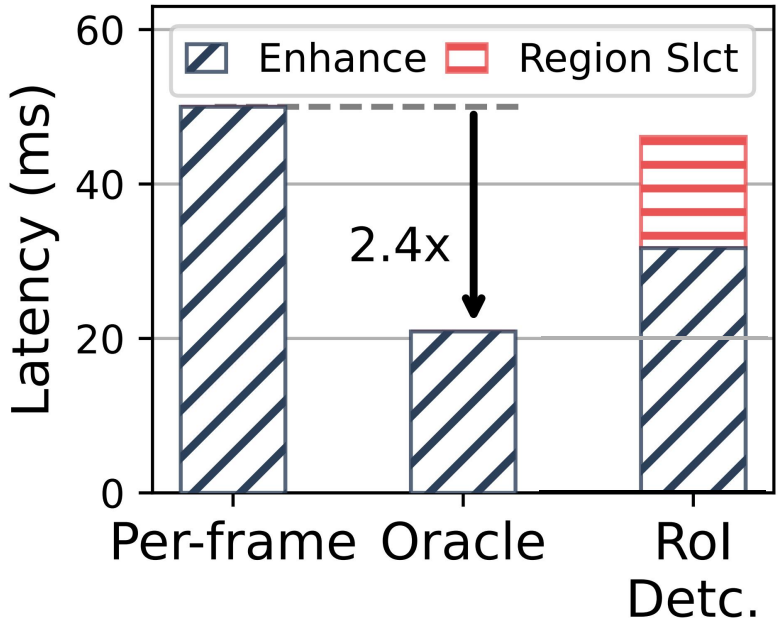
- Only enhancing Eregions contributes 2.4x time saving.

Challenge 1: How to efficiently identify eregions?



- Only enhancing Eregions contributes 2.4x time saving.
- A chicken-egg problem: eregion, $I(F_{SR}) \oplus I(F_{LR})$, is calculated with the frame already SRed.

Challenge 1: How to efficiently identify eregions?

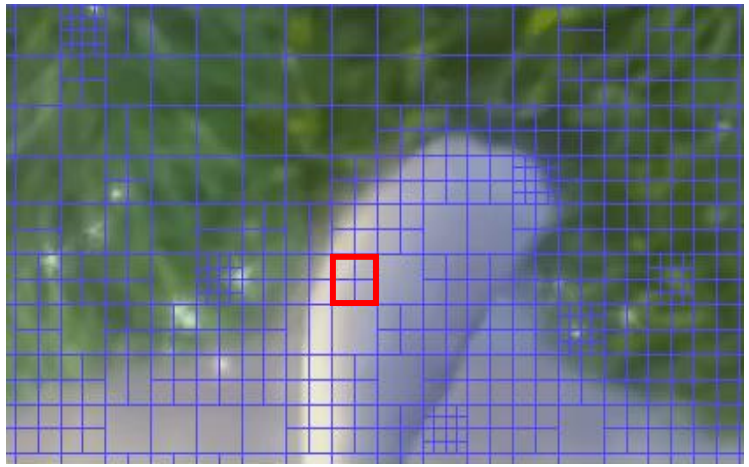


- Only enhancing Eregions contributes 2.4x time saving.
- A chicken-egg problem: eregion, $I(F_{SR}) \oplus I(F_{LR})$, is calculated with the frame already SRed.
- *RoI detection* may identify eregions, but too costly.

Server-Driven Video Streaming for Deep Learning Inference. SIGCOMM'20
Edge Assisted Real-time Object Detection for Mobile Augmented Reality, MobiCom'19

Technique 1: MB-based Region Importance Prediction

Observation #1: Previous RoI methods are costly as its pixel-grained detection.



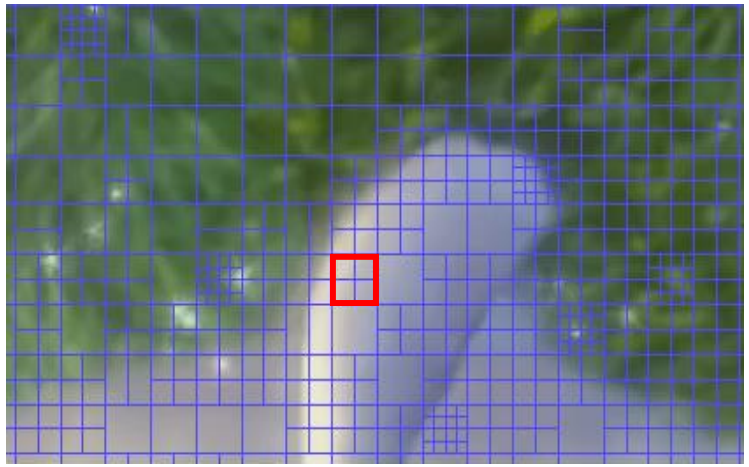
Macroblock (16x16) -grained.

Identify important regions on MB-grained can be 256x lighter than RoI detection.

Technique 1: MB-based Region Importance Prediction

Identify eregions in **MB-grained**

Design #1: Setting Macroblock as the basic unit is both efficient and accurate.



Macroblock (16x16) -grained.

Identify important regions on MB-grained can be 256x lighter than RoI detection.

Technique 1: MB-based Region Importance Prediction

Identify eregions in **MB-grained**

Design #1: Setting Macroblock as the basic unit is both efficient and accurate.

Observation #2: Identifing eregion is an semantic segmentation problem.

Identifing eregion is equavilant to assign each MB one label, i.e., importance score.

Technique 1: MB-based Region Importance Prediction

Identify eregions in **MB-grained** with **semantic segmentation model**

Design #1: Setting Macroblock as the basic unit is both efficient and accurate.

Design #2: Leverage ultra-light SS models to predict region importance on MB level.

Identifing eregion is equavilant to assign each MB one label, i.e., importance score.

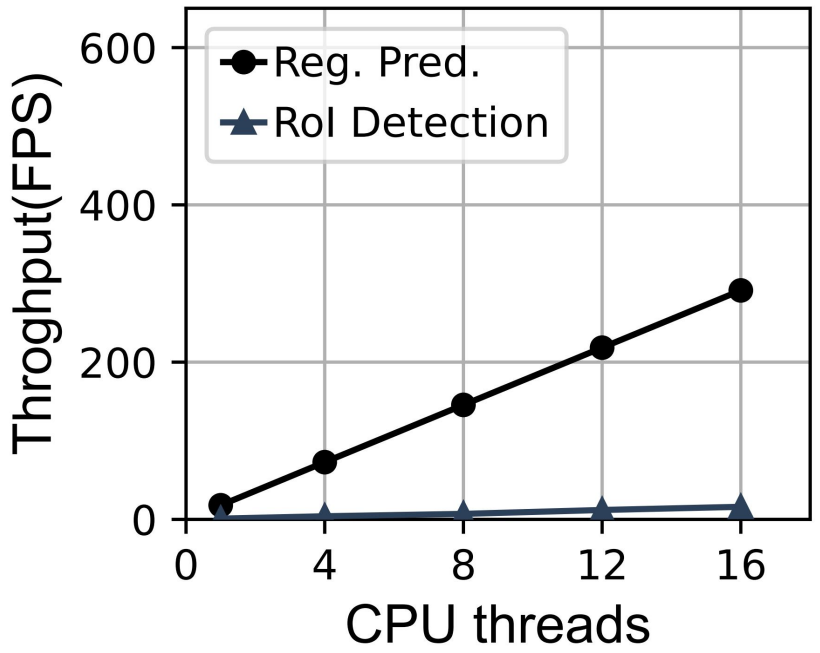
Technique 1: MB-based Region Importance Prediction

Identify eregions in **MB-grained** with **semantic segmentation model**

Design #1: Setting Macroblock as the basic unit is both efficient and accurate.

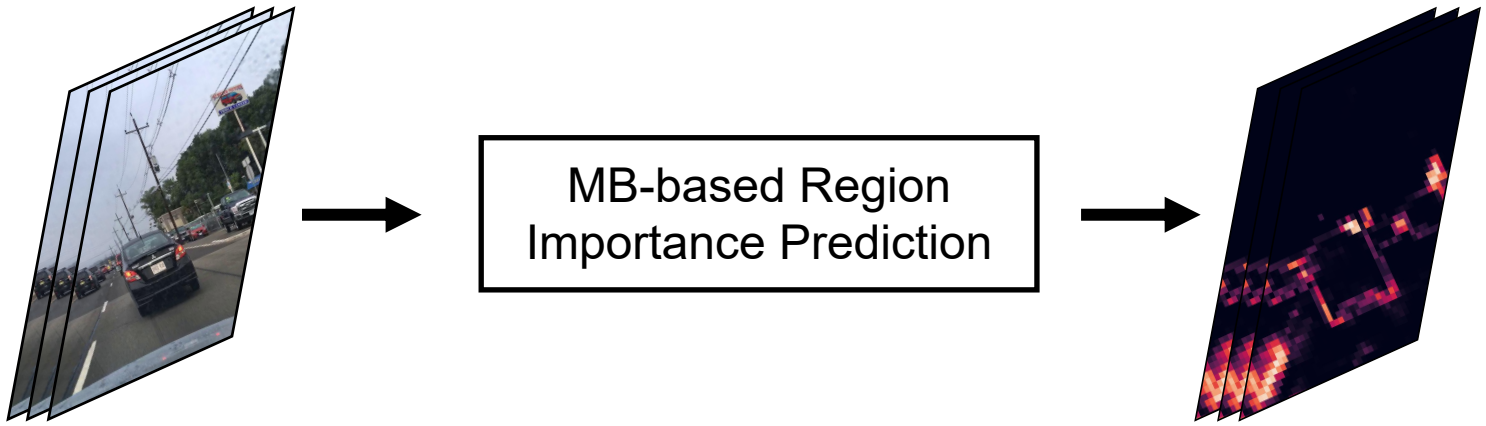
Design #2: Leverage ultra-light SS models to predict region importance on MB level.

30fps per CPU thread,
60× of RoI detection.



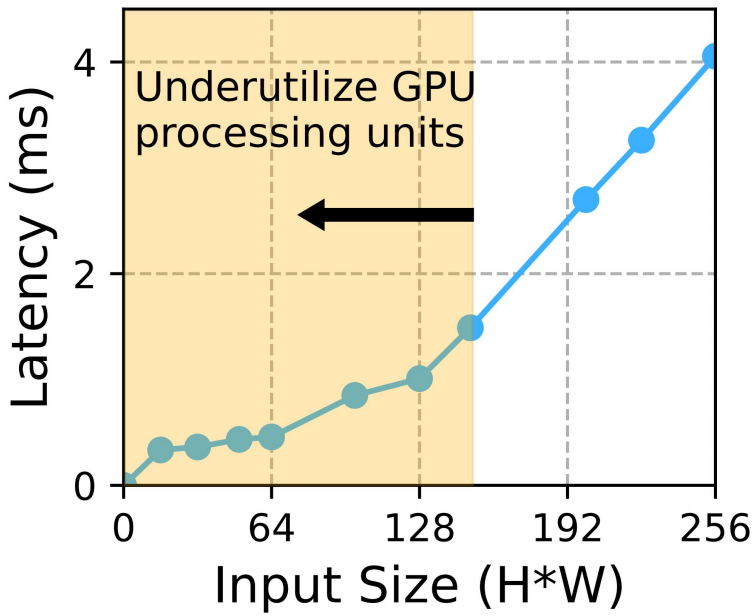
Challenge 2: How to enhance eregions in high throughput?

- Eregions are irregular and sparsely distributed in each frame.



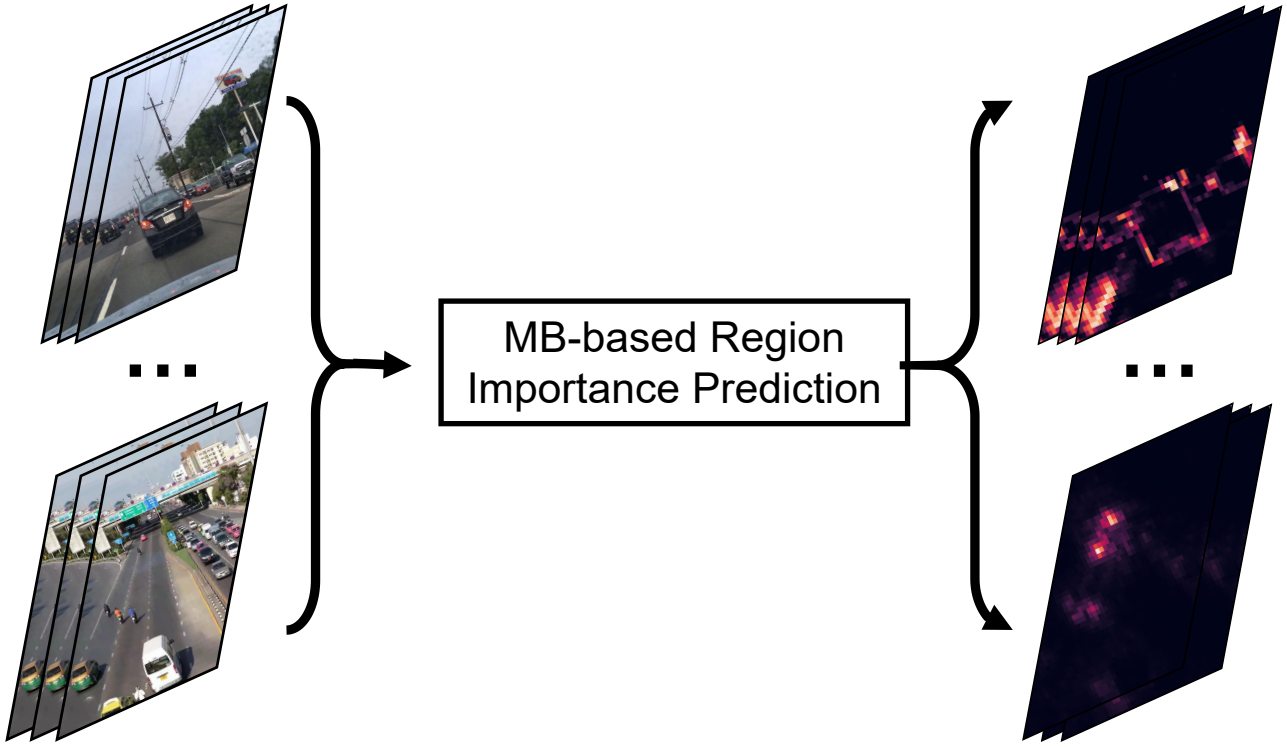
Challenge 2: How to enhance eregions in high throughput?

- Eregions are irregular and sparsely distributed in each frame.
- Efficient batching enhancement needs the *same shape* and *enough size*.



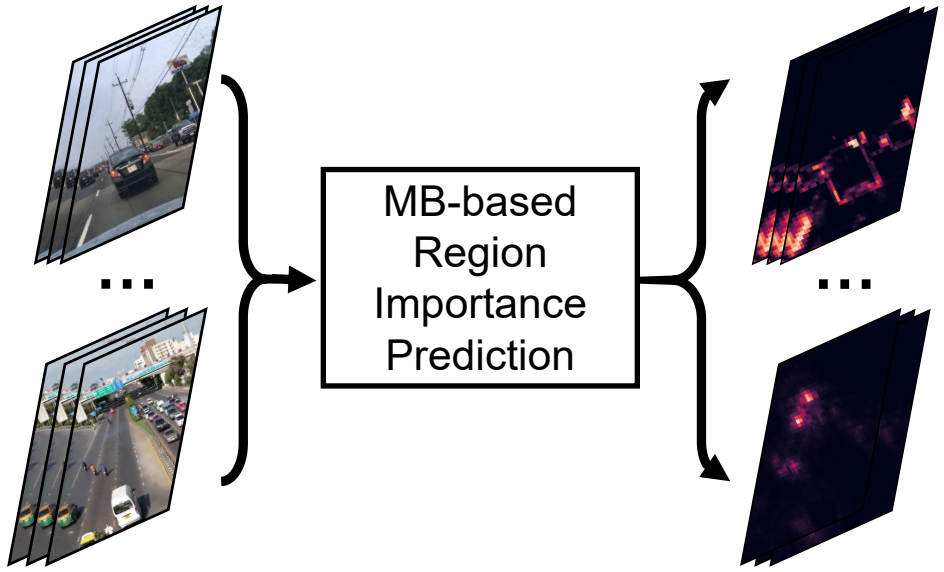
Challenge 2: How to enhance eregions in high throughput?

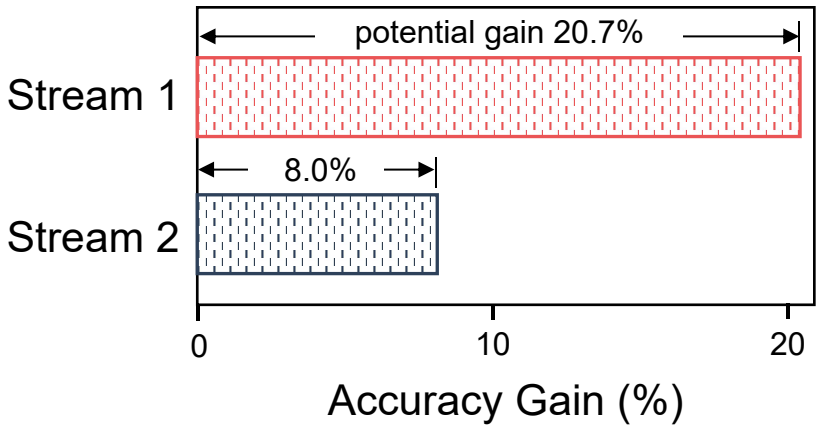
- Eregions are irregular and sparsely distributed in each frame.
- Efficient batching enhancement needs the *same shape* and *enough size*.



Challenge 2: How to enhance eregions in high throughput?

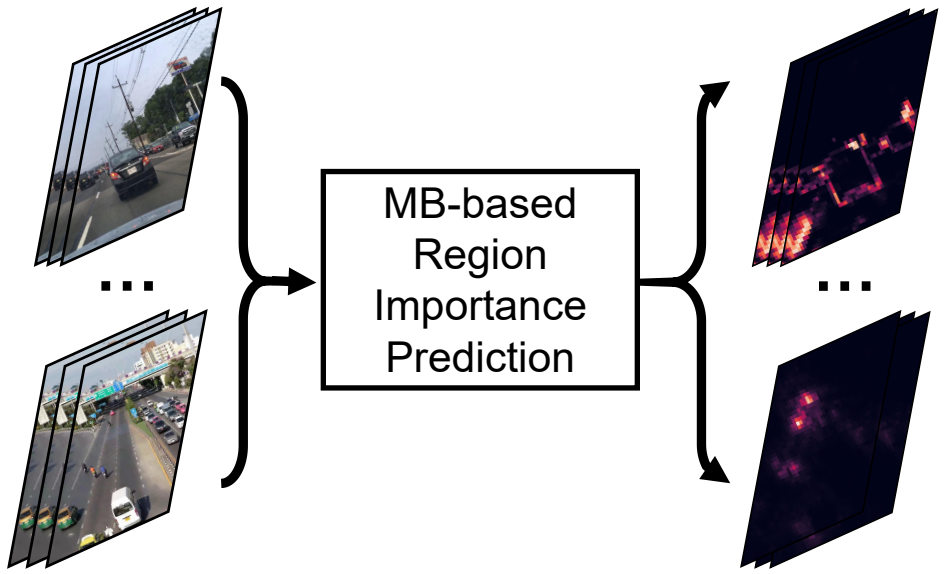
- Eregions are irregular and sparsely distributed in each frame.
- Efficient batching enhancement needs the *same shape* and *enough size*.
- Regions to be enhanced must be considered holistic across streams.

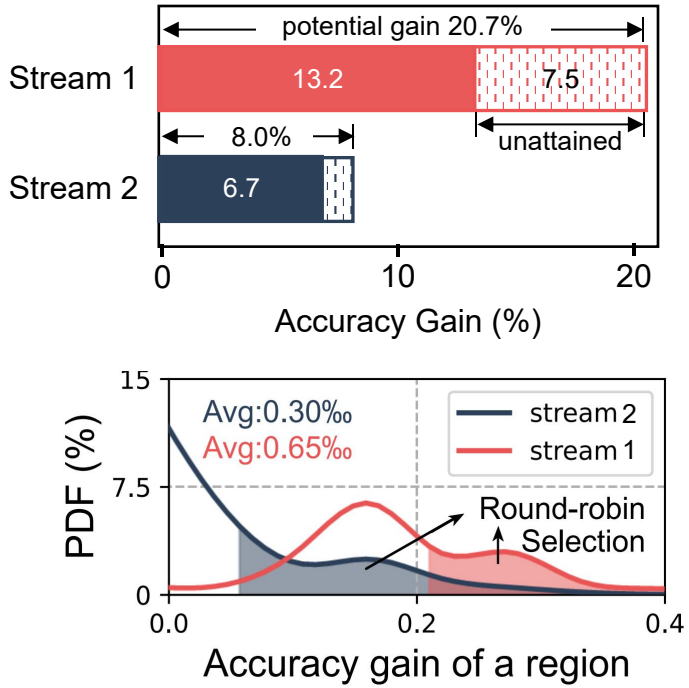




Challenge 2: How to enhance eregions in high throughput?

- Eregions are irregular and sparsely distributed in each frame.
- Efficient batching enhancement needs the *same shape* and *enough size*.
- Regions to be enhanced must be considered holistic across streams.

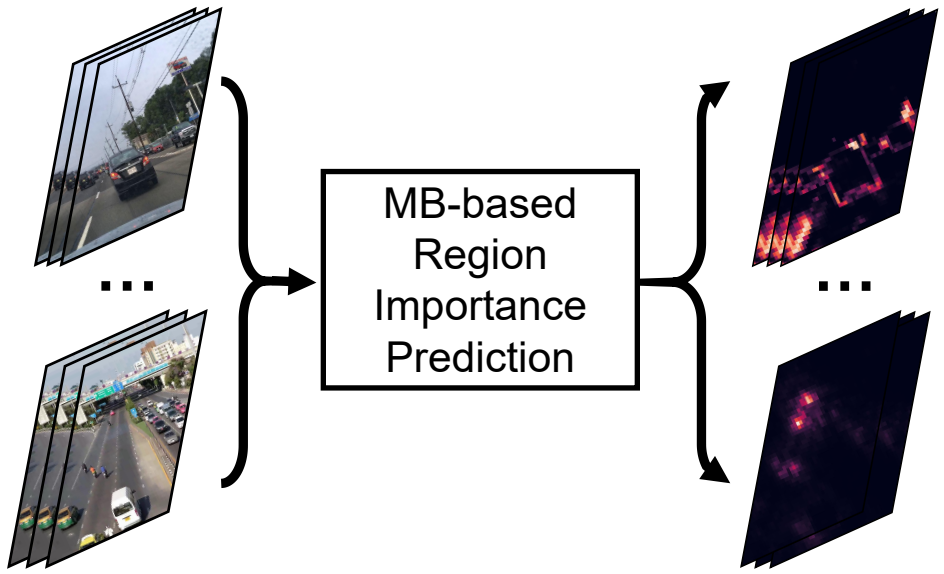


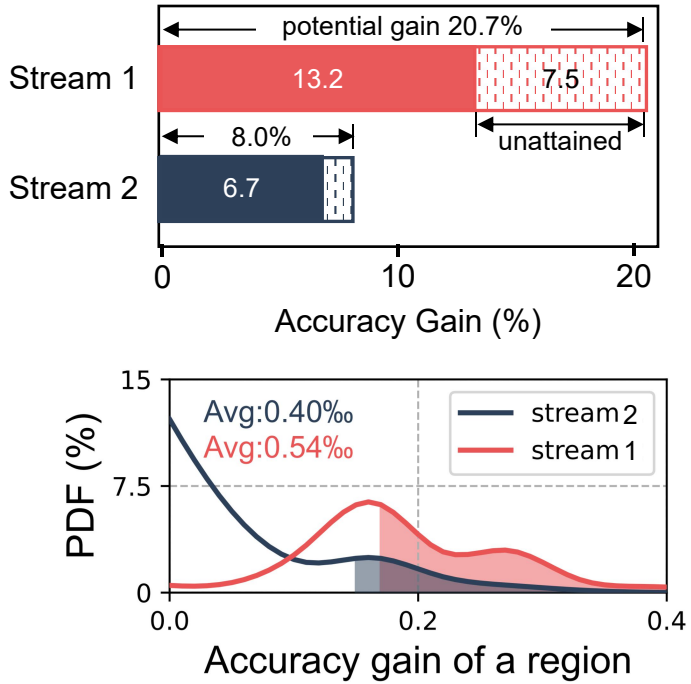


Individual region enhancement across streams leads to 19.9 overall acc gain.

Challenge 2: How to enhance eregions in high throughput?

- Eregions are irregular and sparsely distributed in each frame.
- Efficient batching enhancement needs the *same shape* and *enough size*.
- Regions to be enhanced must be considered holistic across streams.

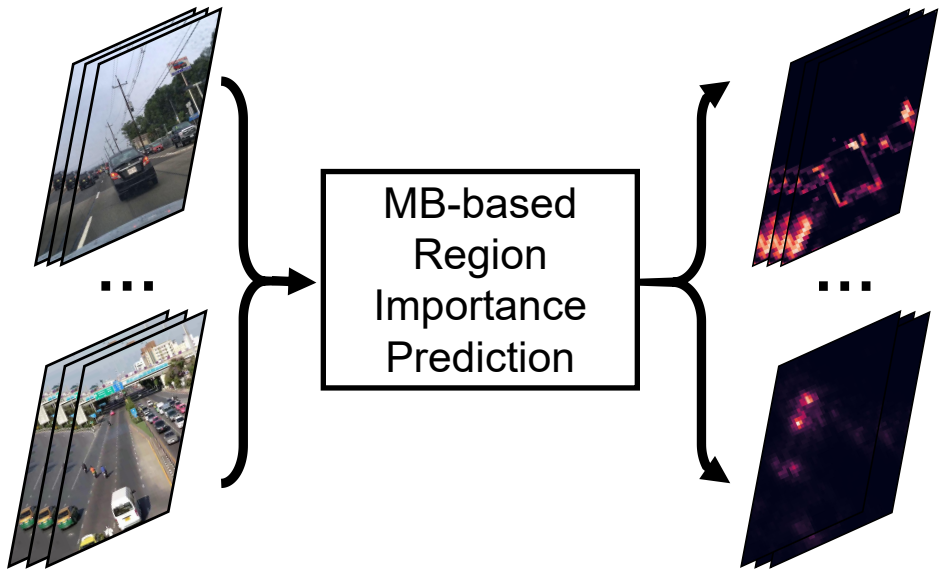


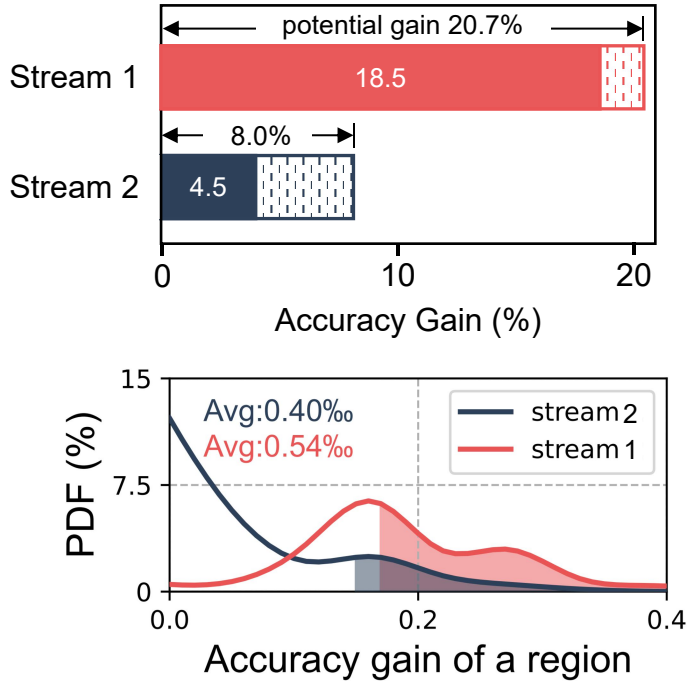


Individual region enhancement across streams leads to 19.9 overall acc gain.

Challenge 2: How to enhance eregions in high throughput?

- Eregions are irregular and sparsely distributed in each frame.
- Efficient batching enhancement needs the *same shape* and *enough size*.
- Regions to be enhanced must be considered holistic across streams.



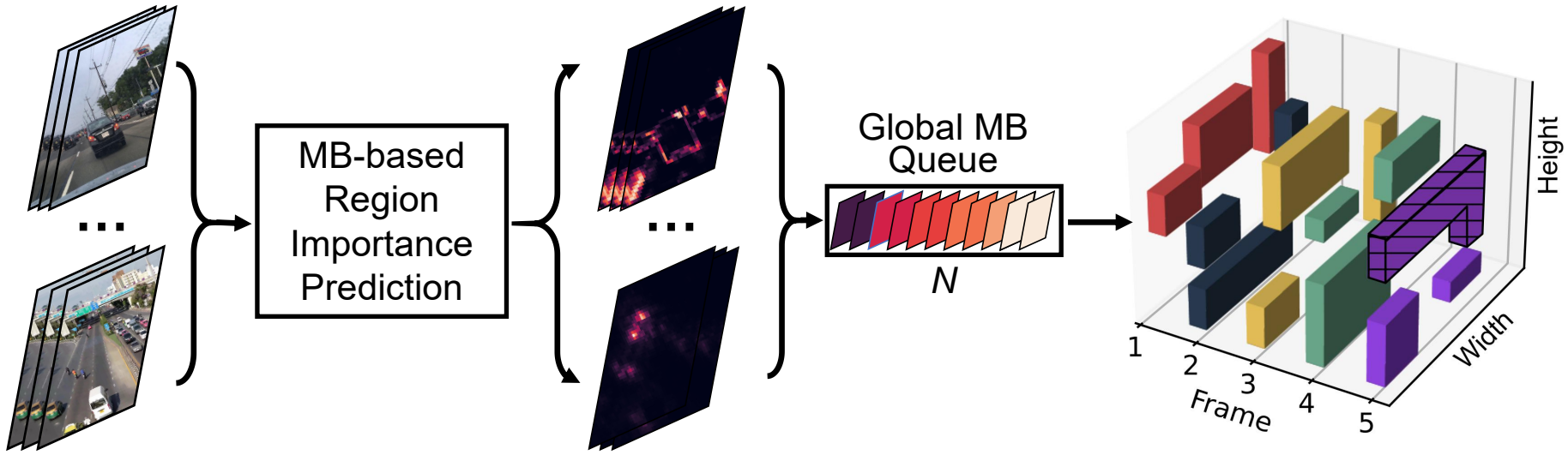


Holistic region enhancement across streams achieves 23.0 overall acc gain.

Technique 2: Region-aware Enhancement

Select and **enhance** most benefitial eregions

Design #1: Aggregates and sorts MBs across all video streams in order of importance.

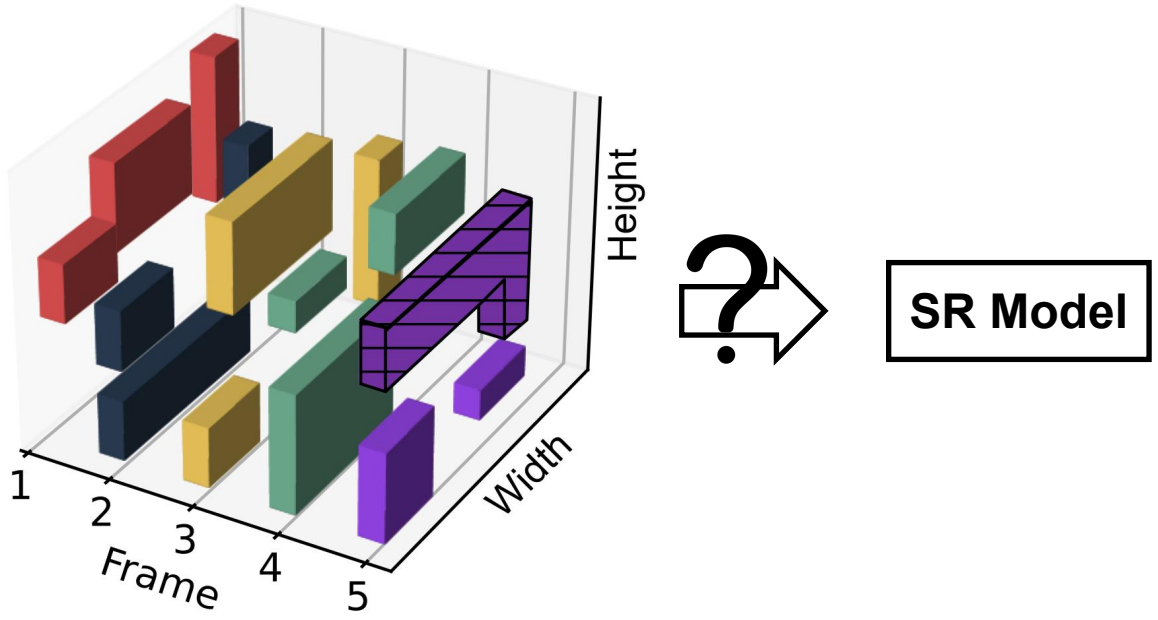


Technique 2: Region-aware Enhancement

Select and **enhance** most benefitial eregions

Design #1: Aggregates and sorts MBs across all video streams in order of importance.

Observation: enhancement latency is pixel-value-agnostic but shape-aware.

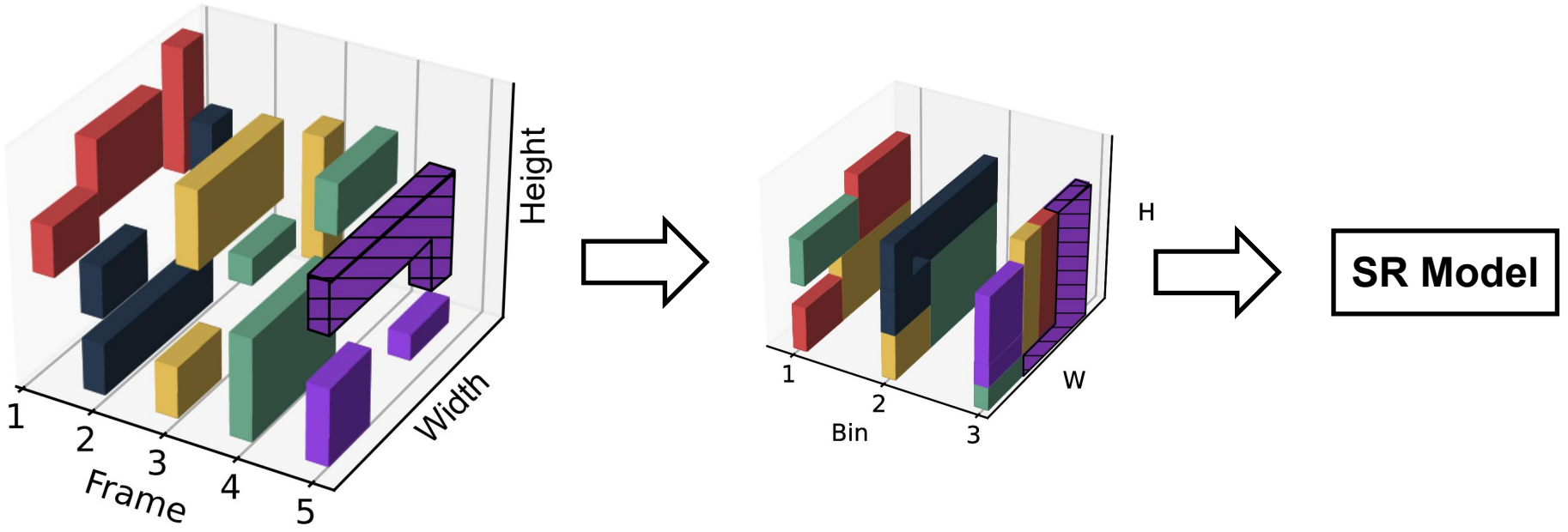


Technique 2: Region-aware Enhancement

Select and **enhance** most benefitial eregions

Design #1: Aggregates and sorts MBs across all video streams in order of importance.

Design #2: Stitch eregions into smaller shape before enhancement.



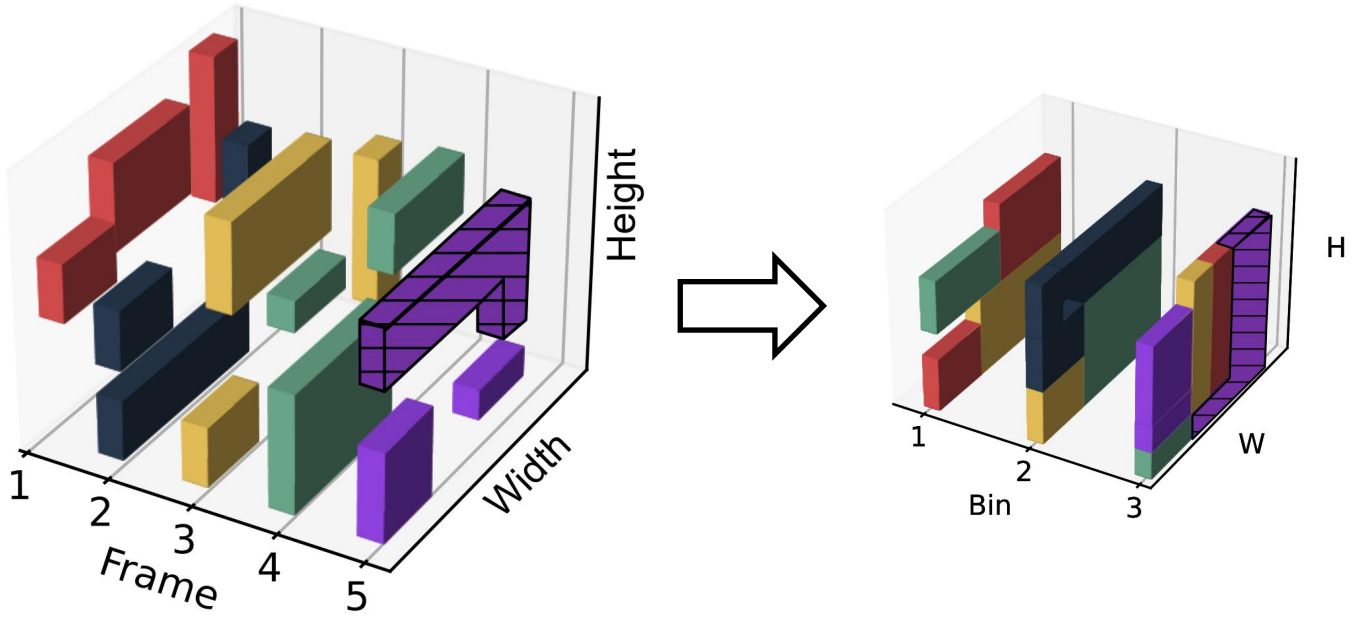
Technique 2: Region-aware Enhancement

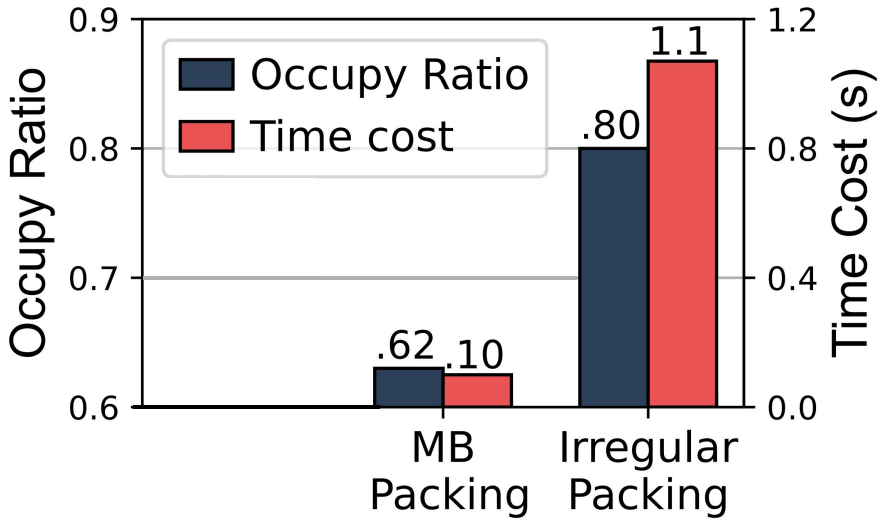
Select and **enhance** most benefitial eregions

Design #1: Aggregates and sorts MBs across all video streams in order of importance.

Design #2: Stitch eregions into smaller shape before enhancement.

Strawmen: *MB packing* stitches individual MBs with extended pixels, *Irregular packing* stitches regions consisted of connective MBs





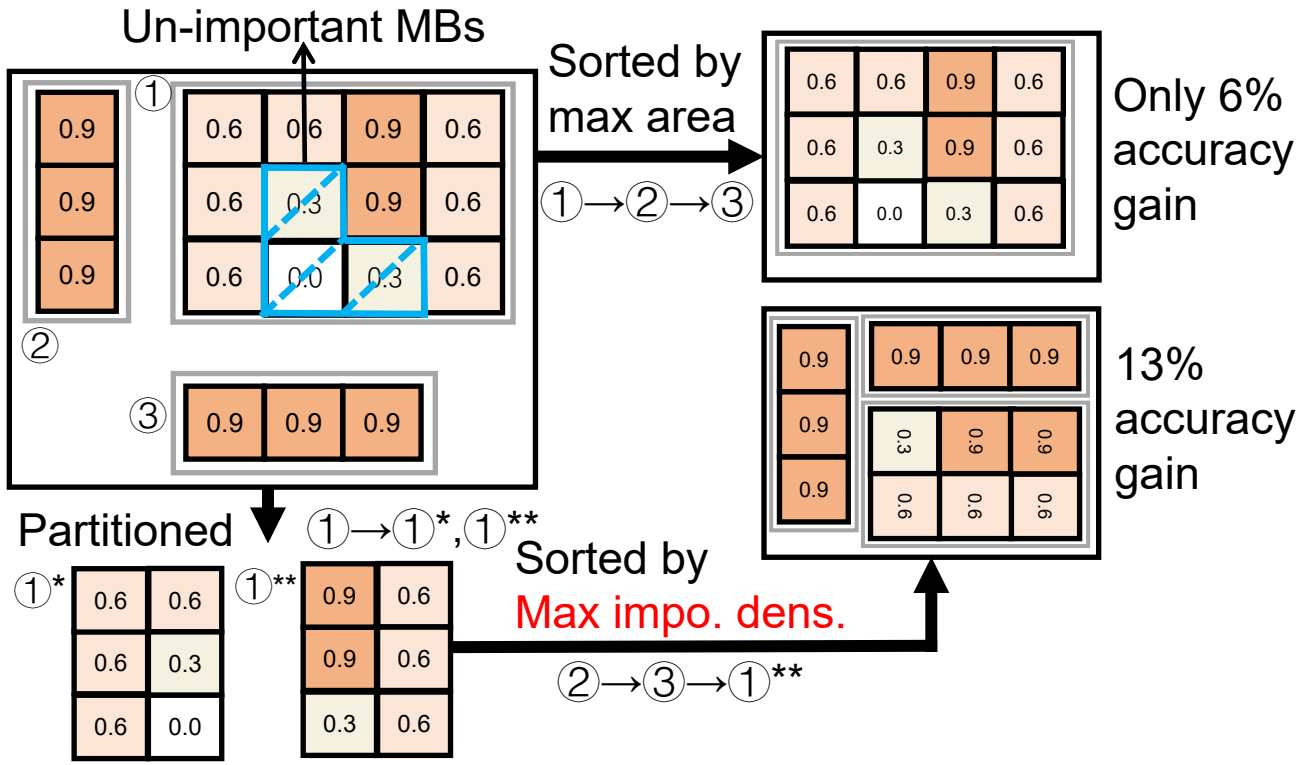
Technique 2: Region-aware Enhancement

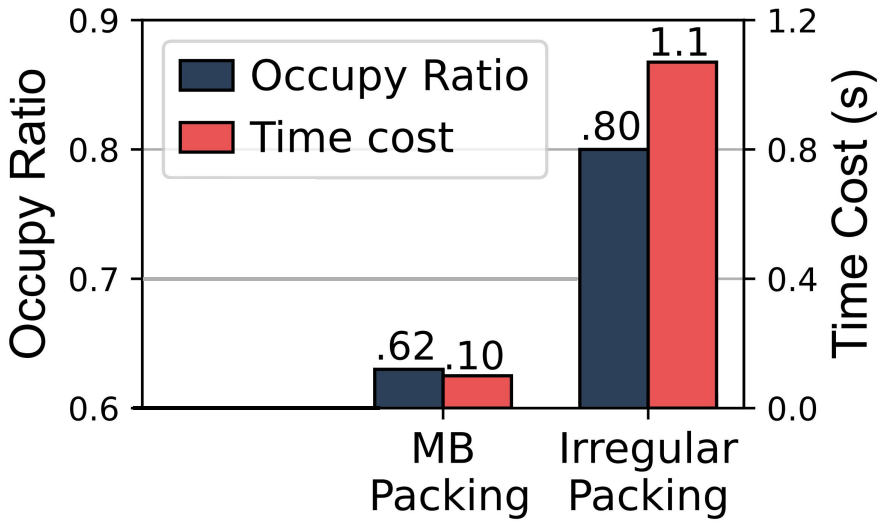
Select and **enhance** most benefitial eregions with **region-aware bin packing**

Design #1: Aggregates and sorts MBs across all video streams in order of importance.

Design #2: Stitch eregions into smaller shape before enhancement.

"Tetris" bin packing: Packs eregions (*N* MBs) in order of importance density into *B H*W* bins.





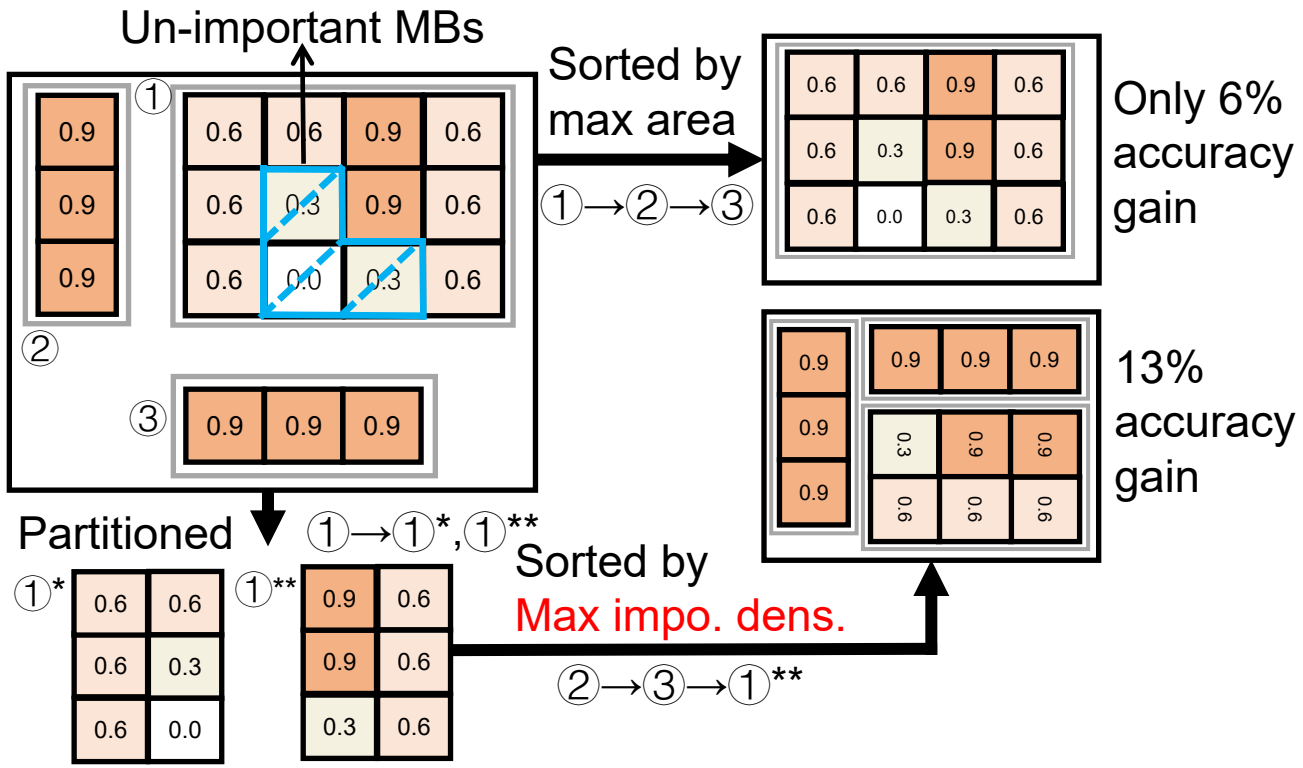
Technique 2: Region-aware Enhancement

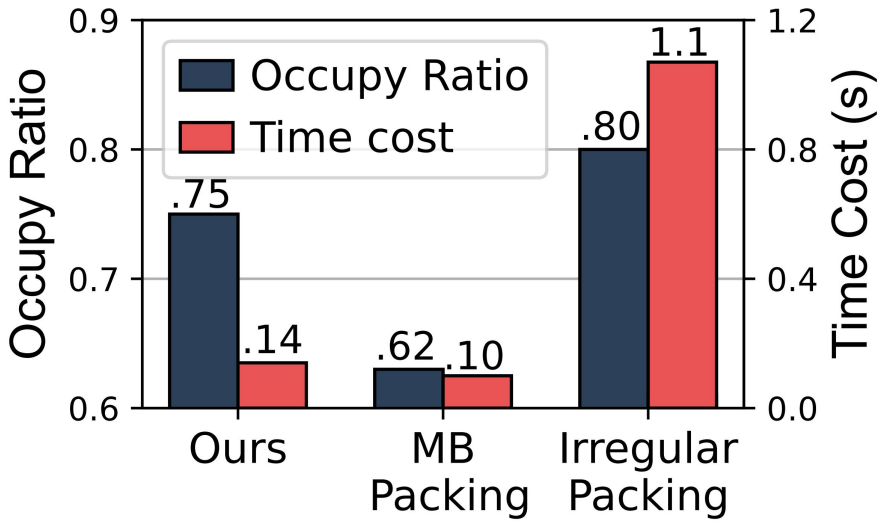
Select and **enhance** most benefitial eregions with **region-aware bin packing**

Design #1: Aggregates and sorts MBs across all video streams in order of importance.

Design #2: Stitch eregions into smaller shape before enhancement.

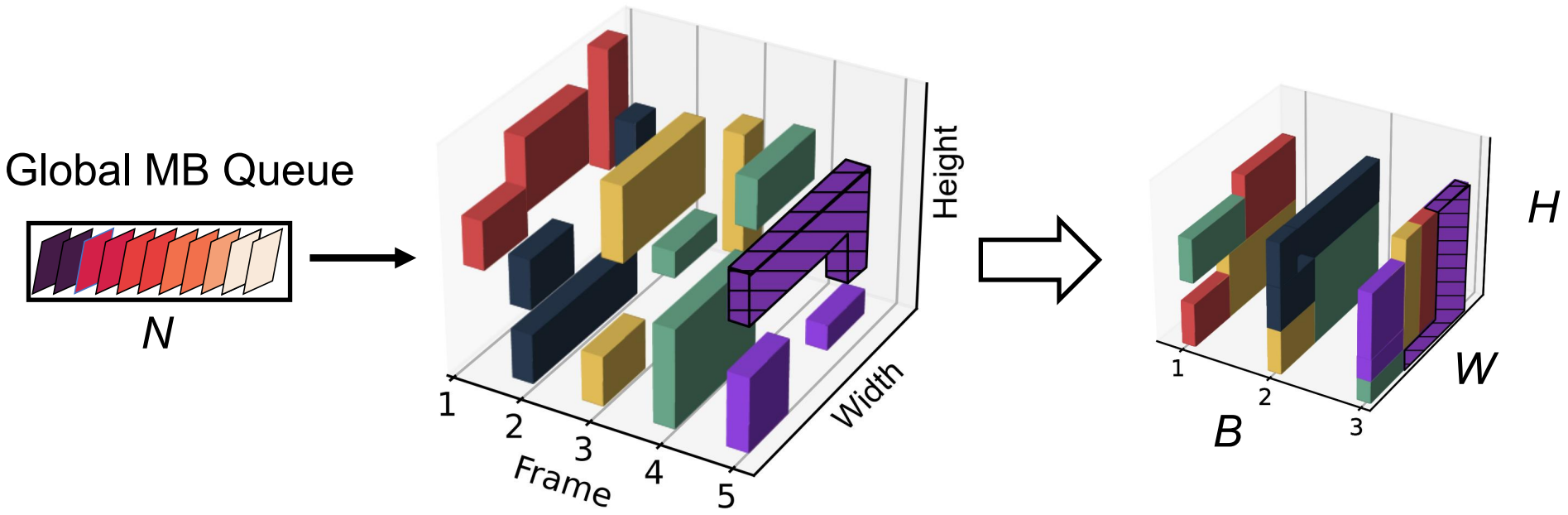
"Tetris" bin packing: Packs eregions (*N* MBs) in order of importance density into *B H*W* bins.





Challenge 3: How to allocate resources among comp.?

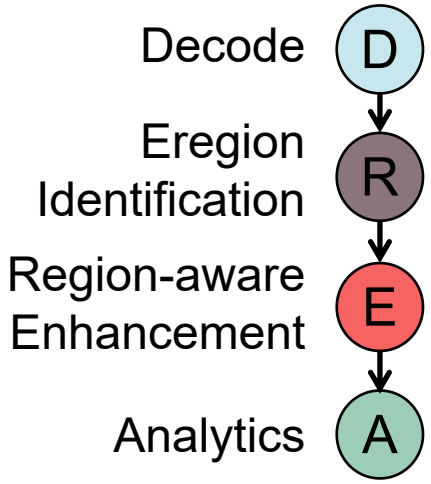
- Hyper-parameters need to be rational configged based on hardware.



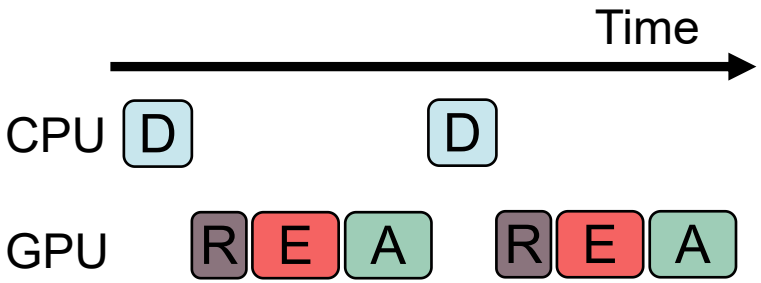
Challenge 3: How to allocate resources among comp.?

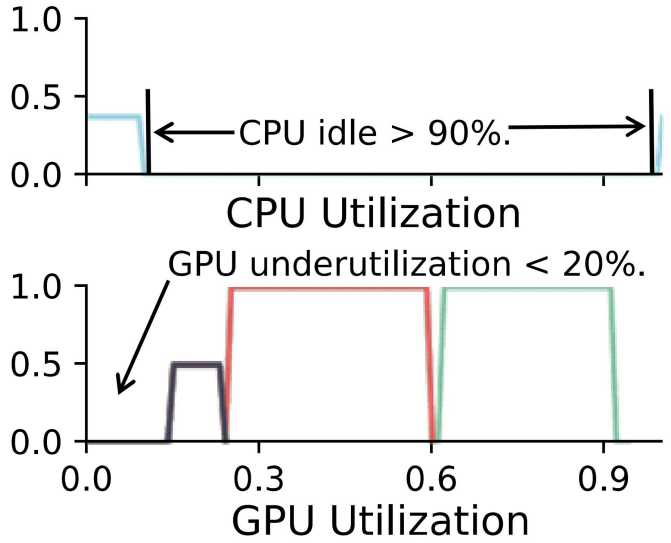
- Hyper-parameters need to be rational configged based on hardware.
- Computing resources must be carefully allocated among runtime components.

Data Flow Graph (DFG)



Execution Plan



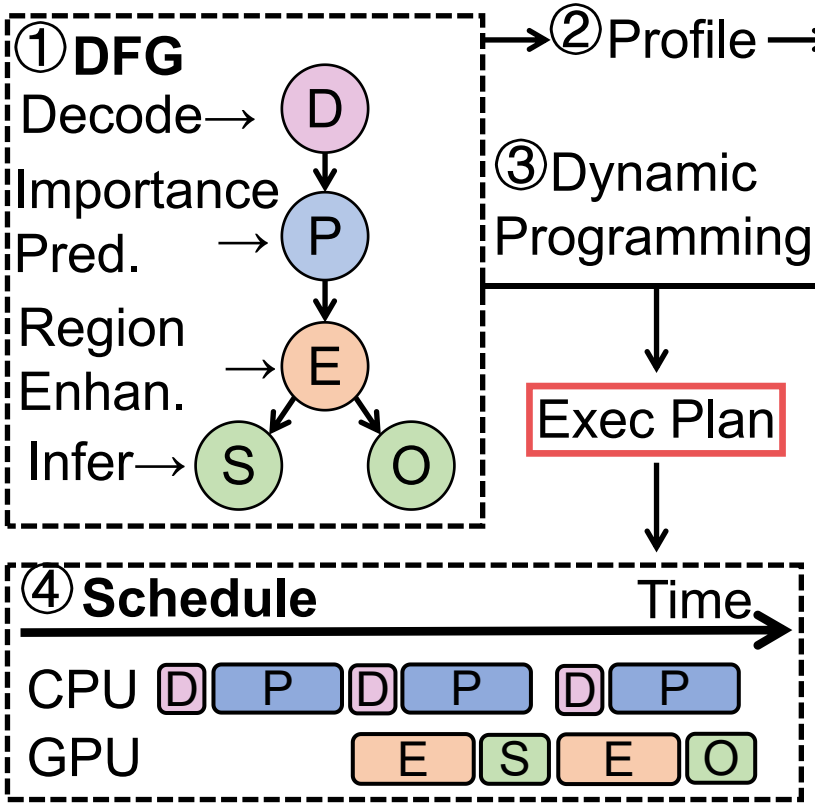


Region-agnostic scheduler leaves >90% CPU and >15% GPU idle time.

Technique 3: Profile-based Execution Planning

Draw up the execution plan **offline** with profile-based method.

Design: parse the DFG, profile throughput of each component on given edge with represent video, and generate plan.

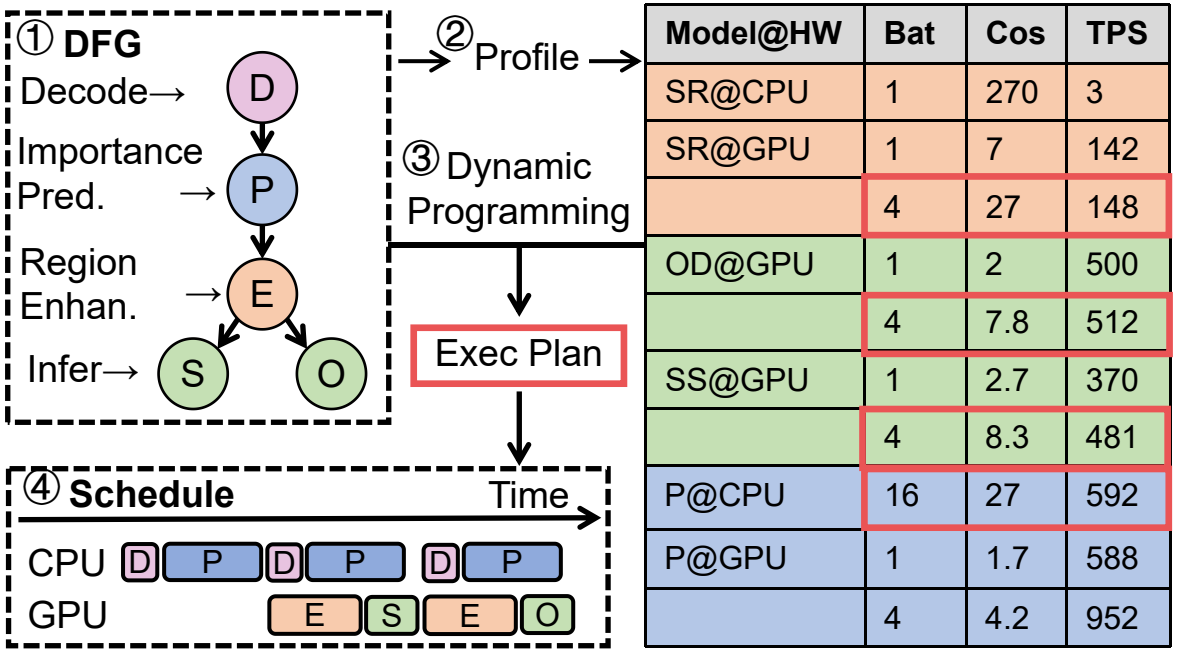


Model@HW	Bat	Cos	TPS
SR@CPU	1	270	3
SR@GPU	1	7	142
	4	27	148
OD@GPU	1	2	500
	4	7.8	512
SS@GPU	1	2.7	370
	4	8.3	481
P@CPU	16	27	592
P@GPU	1	1.7	588
	4	4.2	952

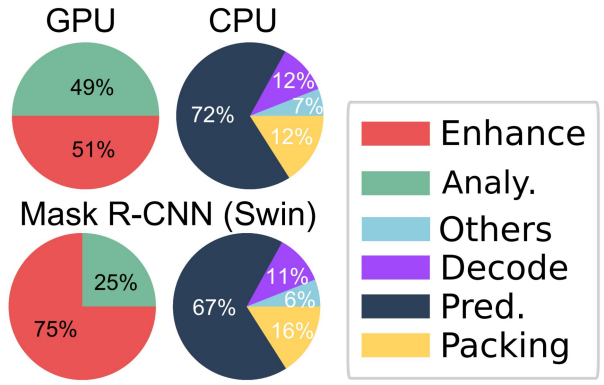
Technique 3: Profile-based Execution Planning

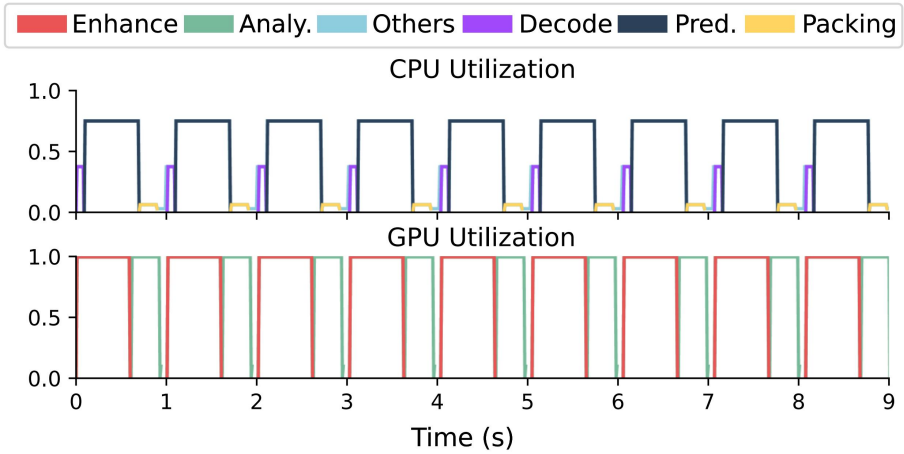
Draw up the execution plan **offline** with profile-based method.

Design: parse the DFG, profile throughput of each component on given edge with represent video, and generate plan.

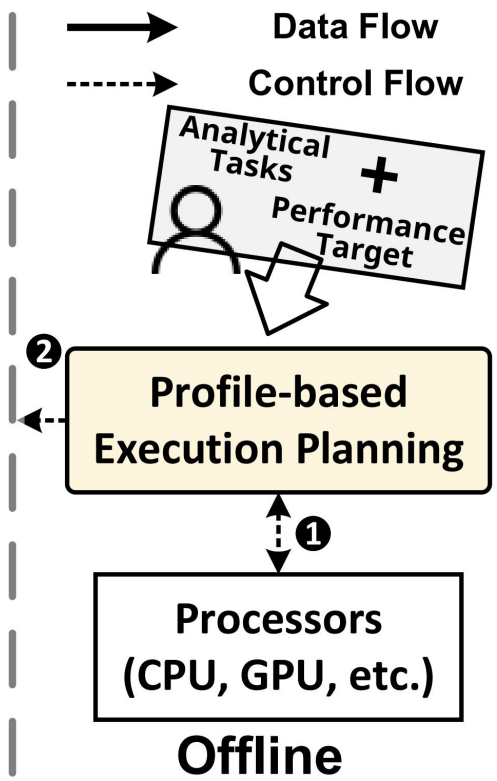


Model@HW	Bat	Cos	TPS
SR@CPU	1	270	3
SR@GPU	1	7	142
	4	27	148
OD@GPU	1	2	500
	4	7.8	512
SS@GPU	1	2.7	370
	4	8.3	481
P@CPU	16	27	592
P@GPU	1	1.7	588
	4	4.2	952





Put everything together



Evaluation Setup

Devices: cloud comparison (A100 + i9-12900K), edge devices (T4 + i7-8700, 4090/3090Ti + i9-13900K, Jetson AGX Orin 64GB)

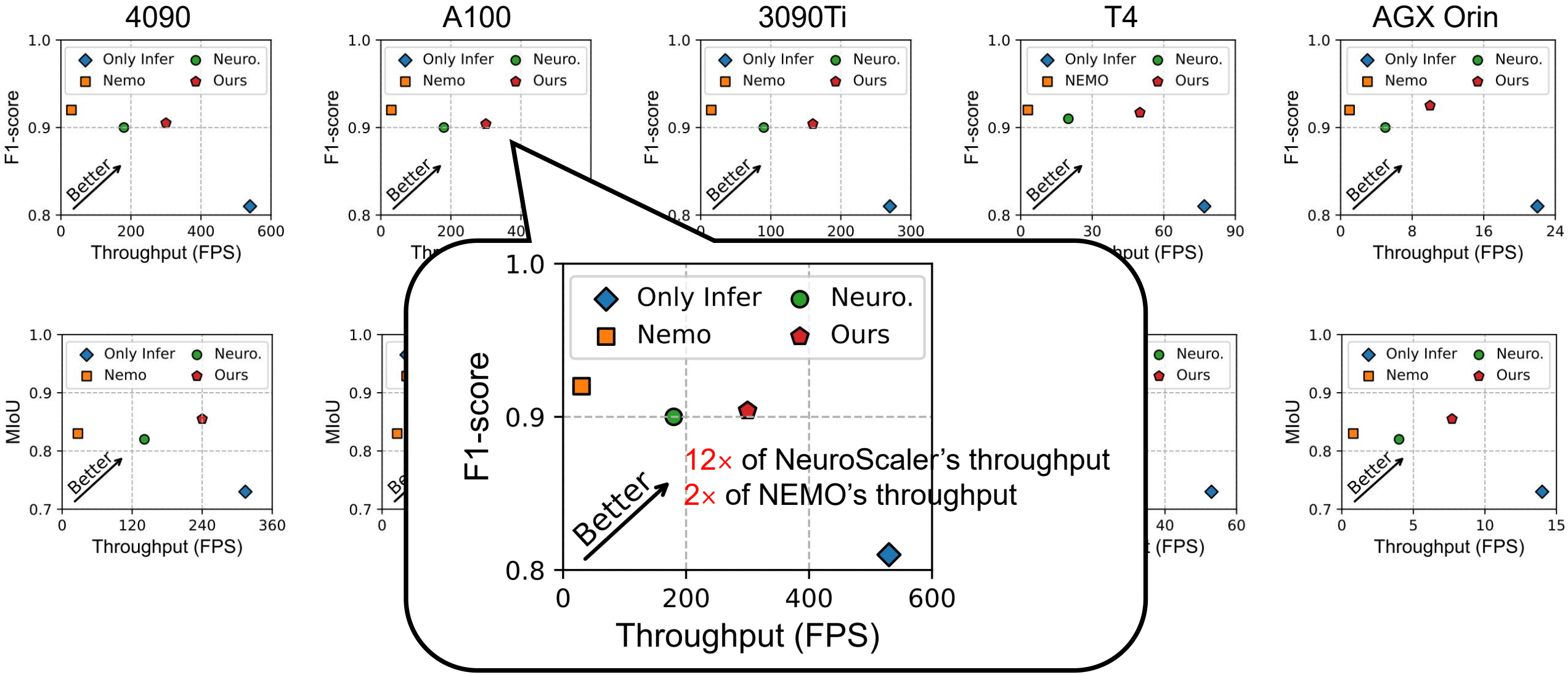
Downstream tasks and dataset:

Tasks	Metric	Models	Dataset
Object Detection	F1-score Stream #	Mask R-CNN (Swin) YOLO	YODA VideoClips
Semantic Segmentation	mIoU Stream #	FCN HarDNet	Cityscape BDD100K

Baselines: Only infer, Selective SR (NeuroScaler, NEMO)

NEMO: Enabling Neural-enhanced Video Streaming on Commodity Mobile Devices, MobiCom'20
NeuroScaler: Neural Video Enhancement at Scale, SIGCOMM'22

Performance on heterogeneous devices



Others

More details in our paper

- Implementation details
- Robustness of various video resolutions
- Contributions by individual components
- Ablation experiment: Accuracy-throughput-resource analysis
- Scalability
- Visualization
- .....

Region-based Content Enhancement for Efficient Video Analytics at the Edge

Weijun Wang[1]* Liang Mi[2*†] Shaowei Cen[2†] Haipeng Dai[2] Yuanchun Li[1] Xiaoming Fu[3] Yunxin Liu[1‡]
[1]*Institute for AI Industry Research (AIR), Tsinghua University*
[2]*State Key Laboratory for Novel Software Technology, Nanjing University*
[3]*University of Göttingen*

Abstract

Video analytics is widespread in various applications serving our society. Recent advances of content enhancement in video analytics offer significant benefits for the bandwidth saving and accuracy improvement. However, existing content-enhanced video analytics systems are excessively computationally expensive and provide extremely low throughput. In this paper, we present region-based content enhancement, that enhances only the important regions in videos, to improve analytical accuracy. Our system, RegenHance, enables high-accuracy and high-throughput video analytics at the edge by 1) a macroblock-based region importance predictor that identifies the important regions fast and precisely, 2) a region-aware enhancer that stitches sparsely distributed regions into dense tensors and enhances them efficiently, and 3) a profile-based execution planer that allocates appropriate resources for enhancement and analytics components. We prototype RegenHance on five heterogeneous edge devices. Experiments on two analytical tasks reveal that region-based enhancement improves the overall accuracy of 10-19% and achieves 2-3× throughput compared to the state-of-the-art frame-based enhancement methods.

Content enhancement offers a promising solution to tackle this issue. It delivers a remarkable accuracy improvement [68,70,80,82,84,96,100] and bandwidth saving [41,60,79,93–95,97] by leveraging neural enhancement models (*e.g.*, super-resolution [24, 64], generative adversarial network [30, 48], image restoration [63]) to enhance the informative details of video frames prior to feeding them into the final analytical models. Different from previous model optimization methods (*e.g.*, model merging [55, 71], model updating [60, 74], and model switching [59, 95]), data-optimization content enhancement does not require modifying user-provide models [23, 31, 73]. Besides, even if the city government updates current low-end cameras to the latest ones that can offer high-quality video, content enhancement can still improve the details of small objects or blurred content.

Unfortunately, naively employing content enhancement in practice is excessively computationally expensive. Such straightforward ways of enhancements not only cause high latency but also compete for computational resources with final analytical models. For example, applying enhancement on one tail-accuracy frame or executing generative adversarial networks on hard-recognized human faces causes hundreds to thousands of milliseconds of latency [84, 96]. The state-of-

Conclusion

- Content enhancement brings great benefits to video analytics
- Per-frame enhancement or temporal selective enhancement fall short, but **spatial region enhancement** is spot on
- **Our contribution:** RegenHance - region-based content-enhanced video analytics at heterogeneous edges
- Results: better accuracy-throughput tradeoff and lower E2E delay
- Code repository: https://github.com/mi150/RegenHance

Thanks
Q&A