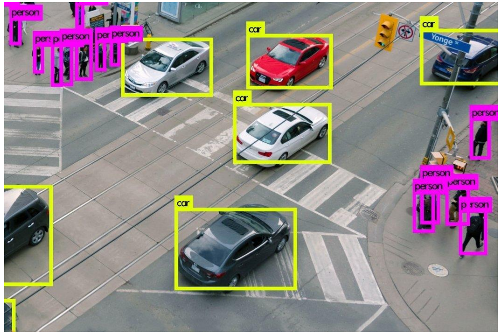
Understanding the Potential of Server-driven Edge Video Analytics

Qizheng Zhang, Kuntai Du, Neil Agarwal, Ravi Netravali, Junchen Jiang

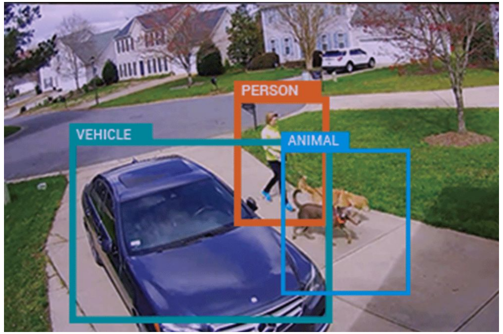
THE UNIVERSITY OF CHICAGO

PRINCETON UNIVERSITY

Edge video analytics are everywhere



Smart cities
Traffic status monitoring

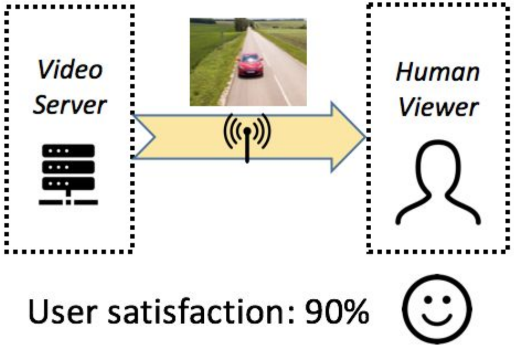


Smart homes
Security surveillance

Industrial settings
Production management

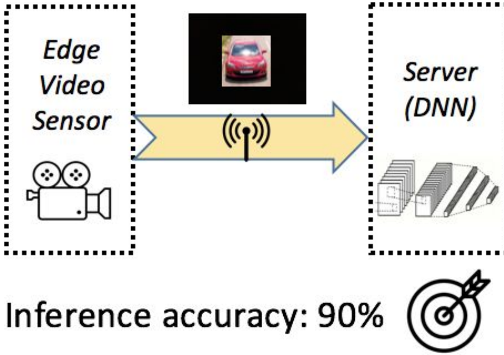
Goal: **Highly accurate** video analytics systems **with less network resource usage**

Serving computer-vision DNNs poses new requirements



Previous: video streaming for **human users**

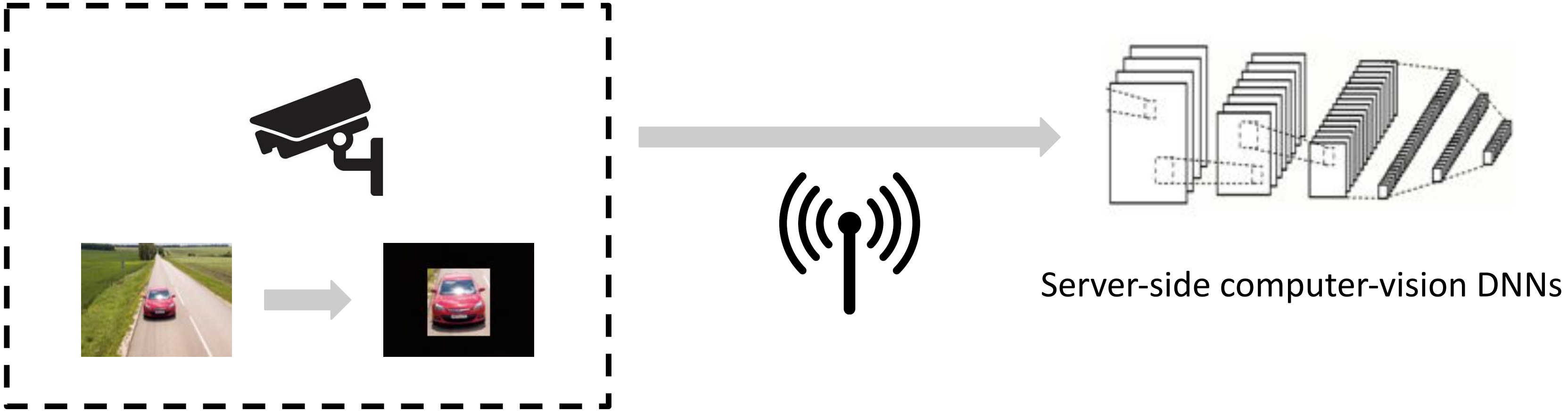
High user satisfaction under bandwidth constraints



This work: video analytics with **DNNs**

High inference accuracy under bandwidth constraints

Typical design #1: Camera-side heuristics



Camera-side heuristics leverage **local compute resources** to decide how videos should be encoded by a sender.

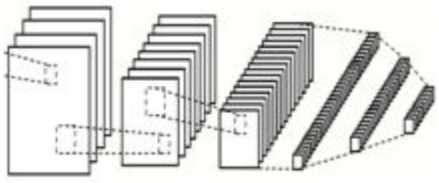
Typical design #2: Server-driven

Server-driven systems utilize **server-side feedback** to guide how videos should be encoded by a sender.

Why not using camera-side heuristics?

Limitation of edge cameras

Incapable of running expensive DNN inference.



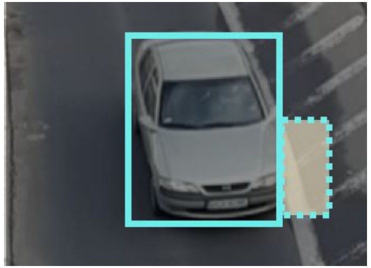
Benefit of being driven by server-side DNNs

Sufficient memory and computation power to support DNN inference.

Server-side DNNs are allowed to **directly** determine what to encode in high quality.

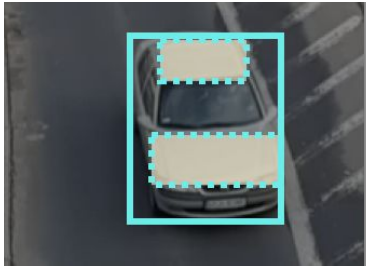
Problems with current server-driven systems

Input video frame　　Region proposal　　Server-side feedback



Some pixels **outside** region proposals are **influential** to accuracy.

Some pixels **within** region proposals are **not influential** to accuracy.



Current systems rely exclusively on region proposals for extracting server-side feedback, which is **sub-optimal**.

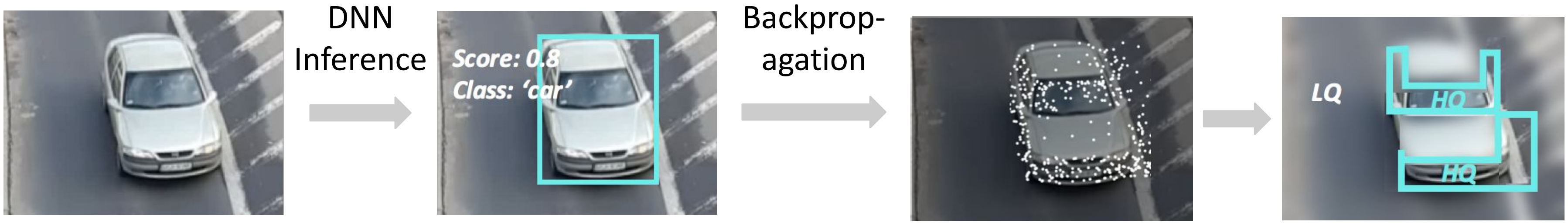
Why is region-proposal-based feedback sub-optimal?

Region proposals are derived from **intermediate feature map results** from DNNs.

We would like a way to extract feedback **directly** from **final** inference results.

Our approach: Saliency-based feedback through backpropagation

Saliency: Gradient of confidence scores sum with regard to input pixel values



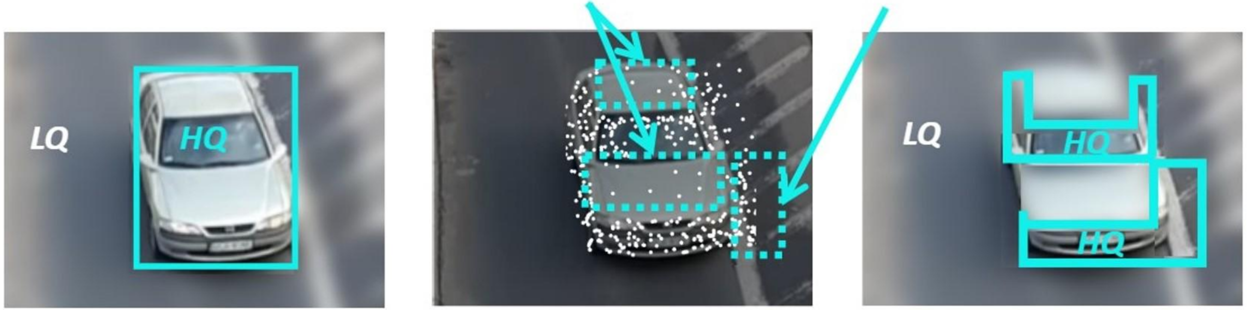
Input video frame Inference results Saliency values Server-side feedback

Saliency-based feedback can be extracted **directly** from inference results through backpropagation.

Advantage of using saliency

*Low saliency: Could have been encoded in **low quality***

*High saliency: Should have been encoded in **high quality***



(a) Assign HQ to region proposals (Confidence drop: 0.3)

(b) Pixels of high saliency returned by server-side DNN

(c) Assign HQ to high-saliency macroblocks (Confidence drop: 0.2)

- ✓ Saliency can capture how much **changing each pixel value** can influence accuracy.
- ✓ Saliency-based feedback enables us to encode videos **at a finer-grained level.**

Practical system design

Challenge: Obtaining saliency values with **uncompressed** video frames is too expensive

For practical system design, we consider **two key parameters**:

- ➢ Video quality for feedback extraction
- ➢ Frame rate of feedback extraction

Finding the “sweet spot” in design trade-offs

Extracting saliency…

- ➢ From uncompressed frames
- ➢ On every frame

high bandwidth usage

correct saliency

Extracting saliency…

- ➢ From greatly compressed frames
- ➢ On every 30 (or more) frames

low bandwidth usage

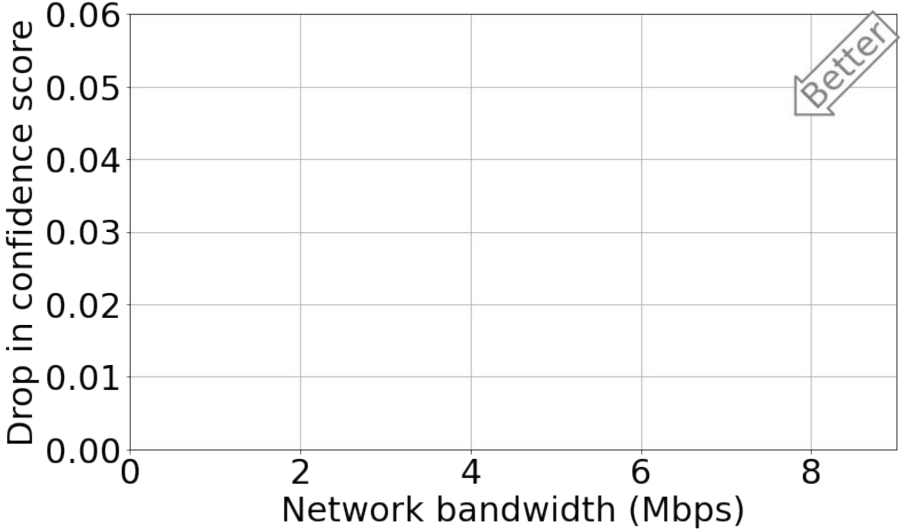
incorrect saliency

“Sweet spot”

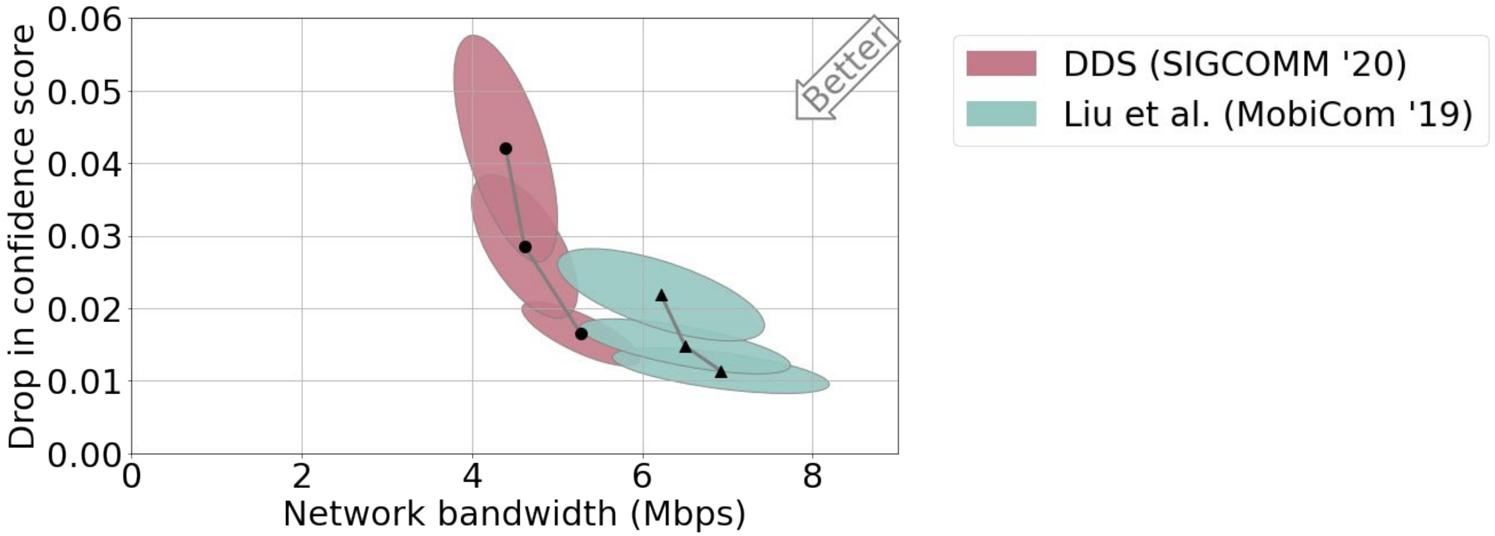
low bandwidth usage

sufficiently correct saliency

Evaluation of practical design



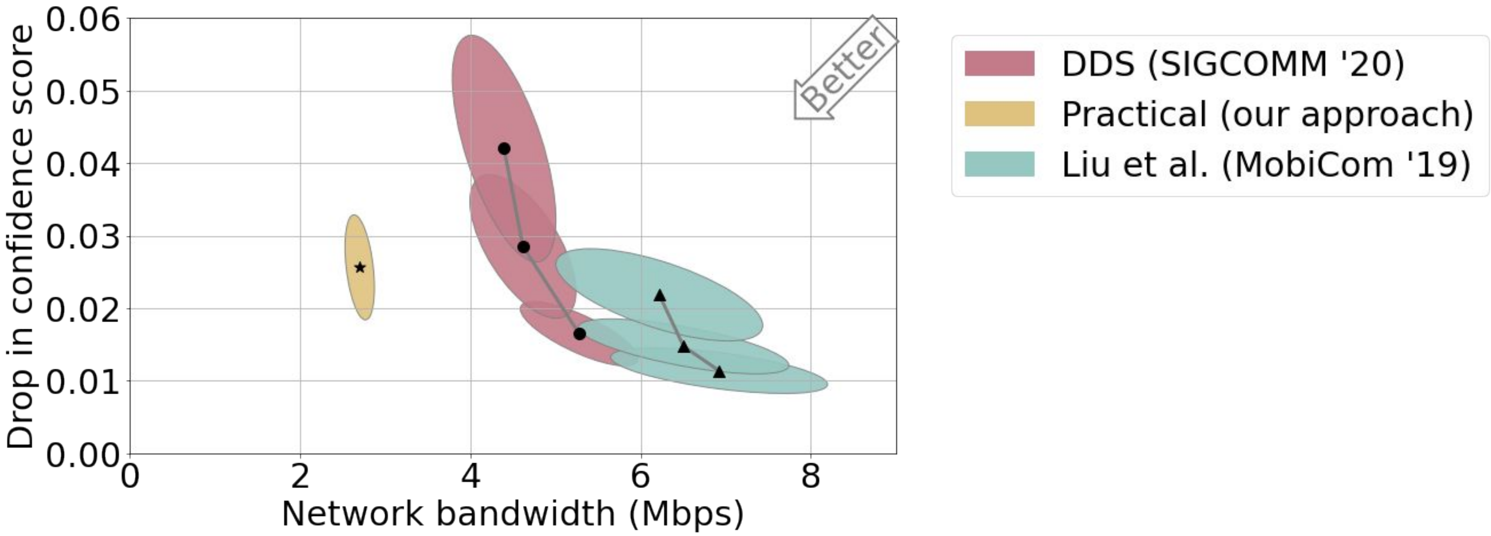
Evaluation of practical design



Inference accuracy degradation v.s. Network bandwidth plot on one of our video datasets

The aforementioned design trade-offs can be clearly observed in the two baselines we choose.

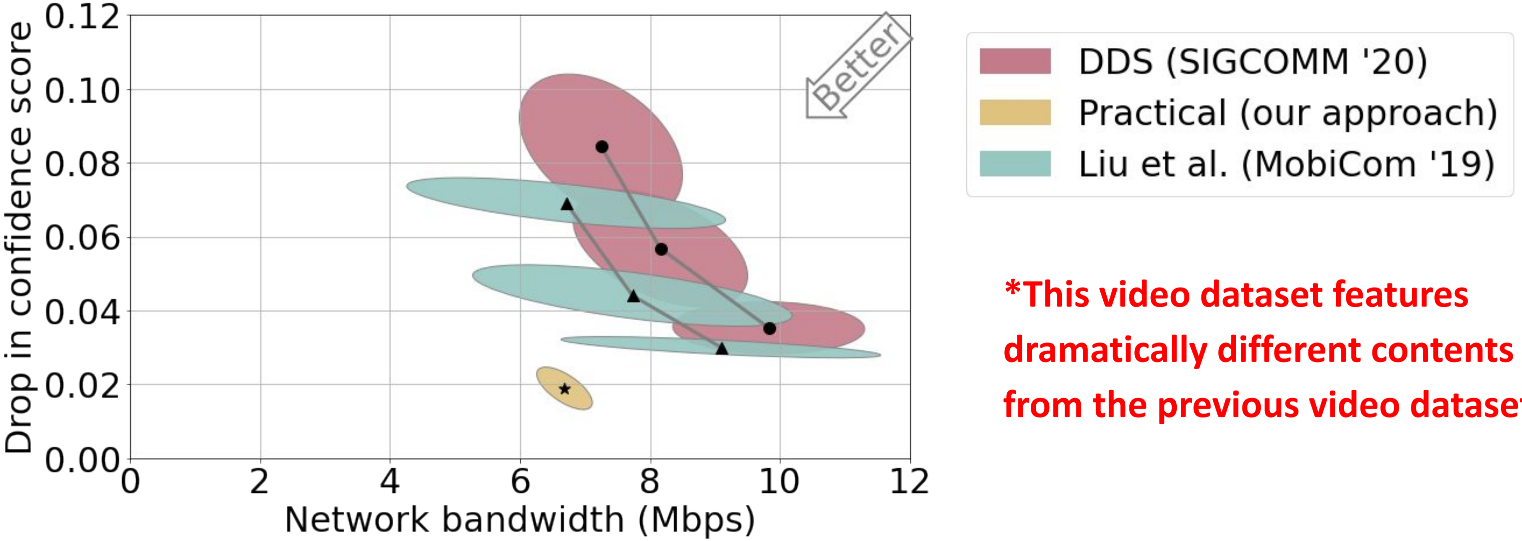
Evaluation of practical design



Inference accuracy degradation v.s. Network bandwidth plot on one of our video datasets

Our saliency-based approach saves **43-57%** bandwidth usage without sacrificing confidence scores.

Evaluation of practical design



***This video dataset features dramatically different contents from the previous video dataset.**

Inference accuracy degradation v.s. Network bandwidth plot on the other video dataset

Our saliency-based approach shows improvements on datasets with a variety of video contents.

Limitations

- Would our approach work on more vision tasks?
  - Yes, but we do not guarantee substantial performance gain.
- Would our approach incur significant extra system usage?
  - Saliency computation incurs 82% more GPU memory usage than forward inference.
- Could our approach work for temporal video encoding?
  - In this work, we only explore spatial video encoding. Temporal encoding would be the next step.

Conclusions

- Current server-driven edge video analytics systems rely exclusively on **region proposals** for extracting feedback, which is sub-optimal as region proposals are derived from **intermediate feature map results** from DNNs.
- We introduce **saliency-based feedback** to **directly** model each pixel's contribution to the inference accuracy from **final inference results**.
- We explore what frame quality and frequency at which saliency should be extracted, and our practical design shows decent performance gain on diverse video contents.