Bandwidth-Efficient Inferencing at the Edge -- An Experimental Approach to Analyze the Effect of VSR on Compressed Video

Yu-Chien Tsai 蔡予謙
Network and Systems Laboratory
Graduate Institute of Electrical Engineering
National Taiwan University
Jun 11th, 2024

Outline

- Introduction
- Related Work
- Methodology
- Results

Outline

- **Introduction**
- Related Work
- Methodology
- Results

Motivation

Edge-assisted inferencing solutions

- Strain on limited upload bandwidth
  - Reducing quality saves bandwidth but impacts performance.
- The role of super-resolution
  - Prior works highlights the impact of image/video quality on object detection
  - Super-resolution can enhance the user experience in live-streaming.
  - Can Super-resolution also enhance object detection accuracy?

Outline

- Introduction
- **Related Work**
- Methodology
- Results

Super-resolution (SR)

- Convert low-resolution images into high-resolution images.
- Number of input images
  - Single Image Super Resolution (SISR)
  - Video Super Resolution (VSR)
    - Leverages both intra-frame and inter-frame information.

Video Codec

- H.264/AVC[4][5]
  - Although H.264 has been around for many years and newer codecs like HEVC and AV1 continue to emerge, H.264 still dominate and compatible with most of the devices nowadays.
  - Similar to previous standards, H.264 divide input images input macroblocks. These macroblocks are further coded in Intra or Inter mode.

Video Codec (Compression)

- Quantization parameter (QP)
  - QP defines how much information to give up in a given block of pixels.
  - Higher QP, lower quality.

Video Codec (Compression)

- Two modes to utilize QP
  - Constant quantization parameter (CQP)
    - Same QP for each frame.
    - Higher CQP, lower quality.
  - Constant rate factor (CRF)
    - QP is adjusted for each frame to provide maximum compression efficiency.
    - Higher CRF, lower quality.

Object detection

- Yolo[3]
  - There are many CNN-based object detectors today, e.g. R-CNN, YOLO, SSD, RetinaNet, EfficientNet.
  - YOLO is a single-stage detector that view detection problems as regression problems. YOLO is extremely fast and good in learning generalizable representation of objects.
- Pre-trained Yolov5x6 model
  - strongest/largest yolov5 model

Related Work[1][2]

How video/image quality affect object detection

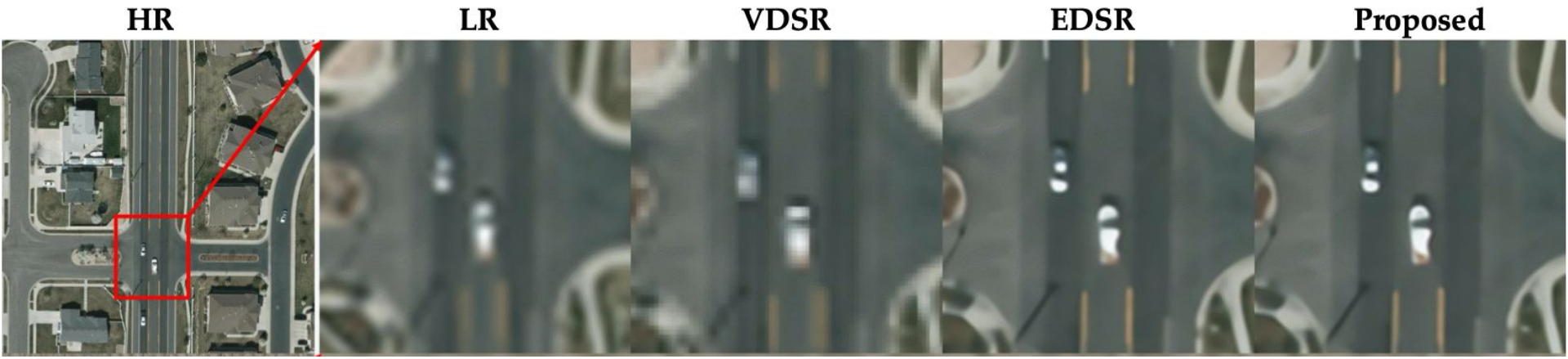
- Without any video/image enhancement
- Resolution is not the focus

	Bitrate 2Mb/s				Bitrate 1.5 Mb/s				Bitrate 1Mb/s			
	CRF-29	CRF-35	CRF-41	CRF-47	CRF-29	CRF-35	CRF-41	CRF-47	CRF-29	CRF-35	CRF-41	CRF-47
Faster R-CNN	31.5%	38.2%	54.0%	78.2%	33.3%	38.3%	54.2%	78.2%	38.7%	41.3%	54.2%	78.4%
SSD512	**16.8%**	25.5%	42.2%	69.3%	19.7%	25.4%	42.4%	69.5%	23.7%	27.4%	43.0%	70.2%
YOLOv3	17.9%	**22.6%**	**33.9%**	**55.4%**	**19.5%**	**23.0%**	**34.0%**	**55.4%**	**23.0%**	**24.6%**	**33.9%**	**55.6%**
RetinaNet	21.8%	29.1%	49.0%	77.7%	24.2%	29.7%	49.1%	77.8%	29.3%	32.8%	48.9%	78.1%

Related Work[7][8]

Task-driven super resolution

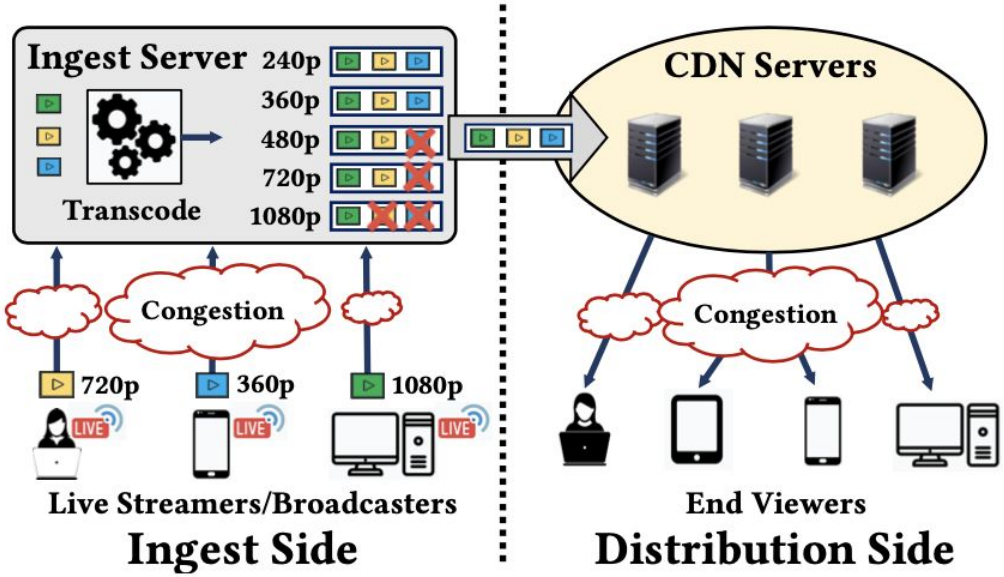
- Train SR models for specific tasks
- More specific targets like satellite images or surveillance cameras



Related Work[9][10][11]

SR-enhanced live streaming

- Consider only resolution but no quality degradation
- Consider only quality gain without any downstream task like object detection performance



Thesis Statement

- General super-resolution on low-quality videos can be helpful for object detection accuracy.
- But still ineffective in saving bandwidth and time.

Contribution

- Present tradeoffs between video average bitrate and object detection performance across different video qualities and resolutions.
- Present the (in)effectiveness of applying general super-resolution methods on lower-quality videos.

Outline

- Introduction
- Related Work
- **Methodology**
- Results

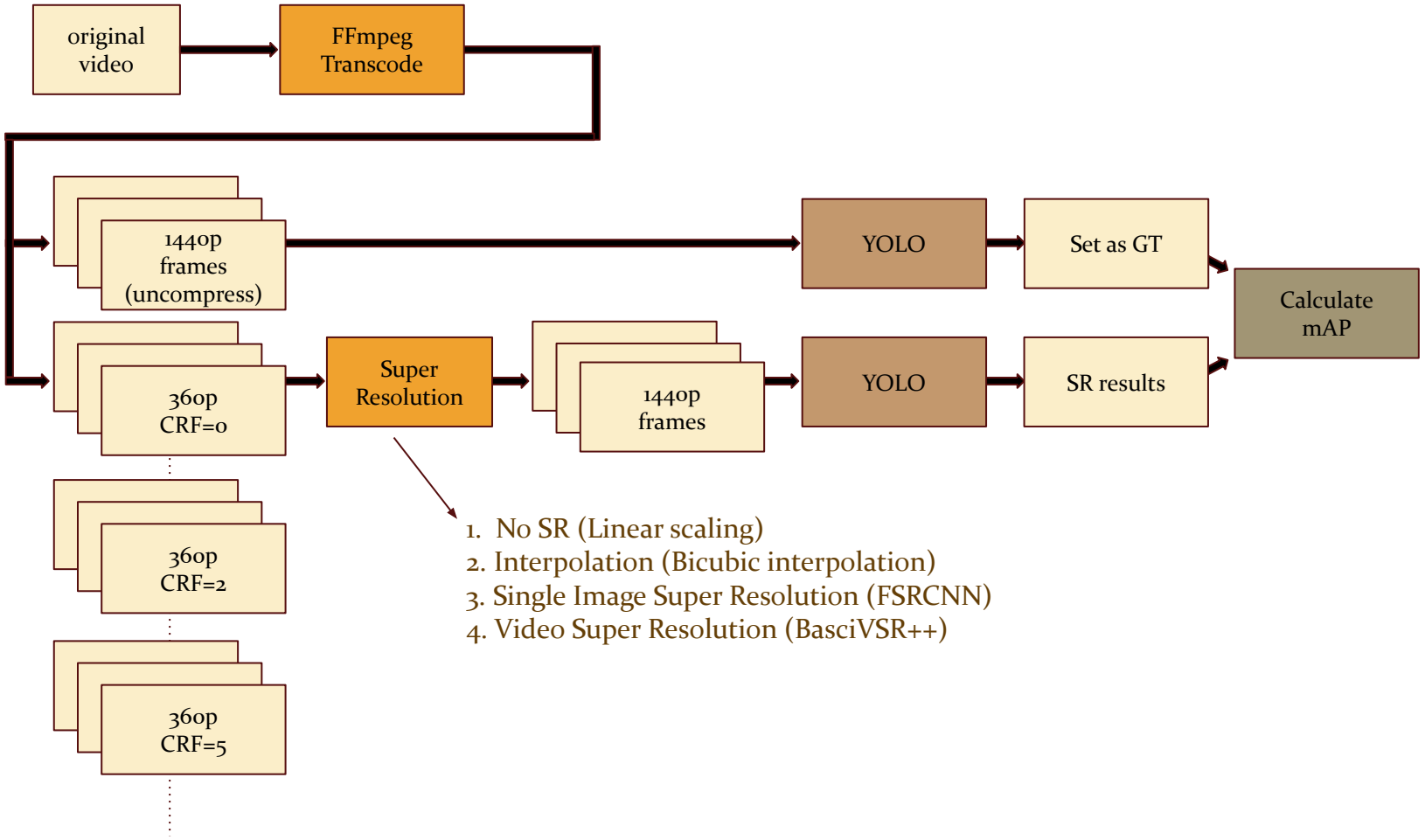
Video Data

- Inter4K:
  - 60fps
  - 5 second
  - high resolution (4K)
  - videos taken by mobile devices

Data Pre-processing

- Transcode with FFmpeg
  - Resolution pairs (4x)
    - 360p & 1440p
    - 270p & 1080p
  - Compression
    - CRF/CQP
      - 0/2/5/7/10/15/20/25/30/35/40
  - Preset
    - medium
  - Codec
    - H.264

Structure Overview - case 1

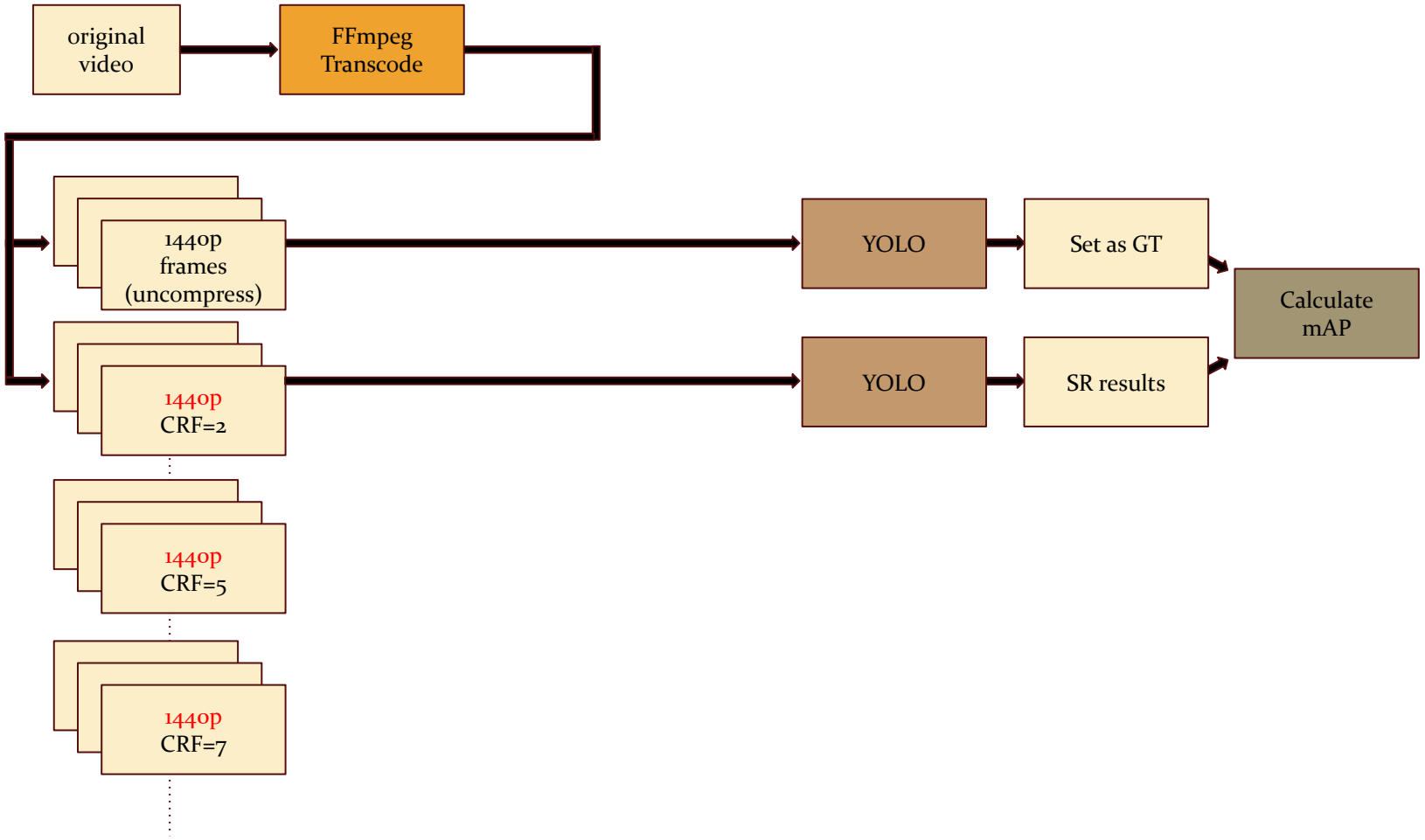


Video Enhancement

- No Super Resolution
  - scale (multiply) the bounding box coordinates
- Bicubic Interpolation (BI)
  - OpenCV resize function
- Single Image Super Resolution (SISR)
  - FSRCNN model
  - pre-trained on T91 dataset
- Video Super Resolution (VSR)
  - BasicVSR++ model
  - pre-trained on REDS4 dataset

Methods	Parameters	Runtime (ms)
Bicubic		0.05
FSRCNN	12809	5
BasicVSR++	7.7M	180

Structure Overview - case 2



Metrics

- mAP (mean average precision)
  - AP (average precision)
  - PRC (Precision-Recall Curve)
    - Confusion matrix
      - Recall = TP / (TP + FN)
      - Precision = TP / (TP + FP)
    - IOU (Intersection Over Union) > 50%
      - determine whether a boundbox is a correct prediction.

		Predicted Label: Positive	Predicted Label: Negative
Actual Label	Positive	TRUE POSITIVE TP	FALSE NEGATIVE FN
	Negative	FALSE POSITIVE FP	TRUE NEGATIVE TN

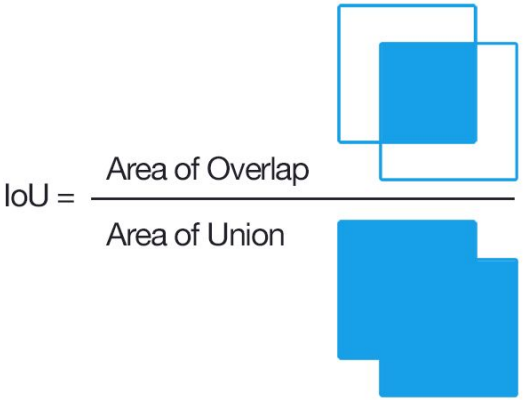


figure reference: https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/ https://www.researchgate.net/publication/329526806_A_novel_heuristic_method_for_misclassification_cost_tuning_in_imbalanced_data

Metrics

- PSNR (peak signal-to-noise ratio)

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right)$$

  - MAX: 255

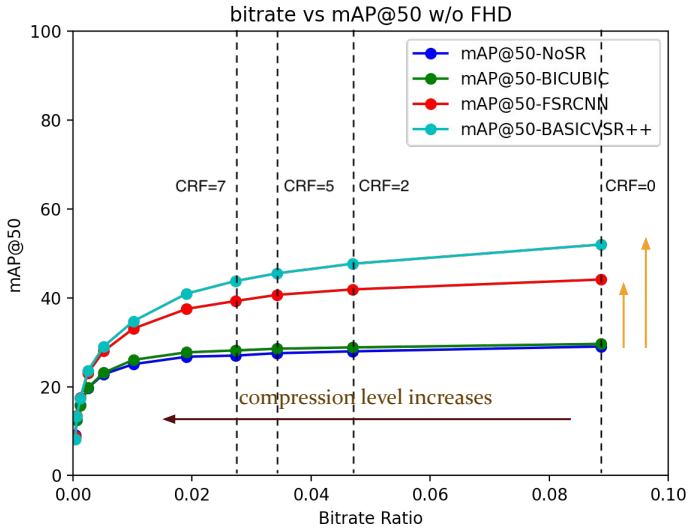
Outline

- Introduction
- Related Work
- Methodology
- **Results**

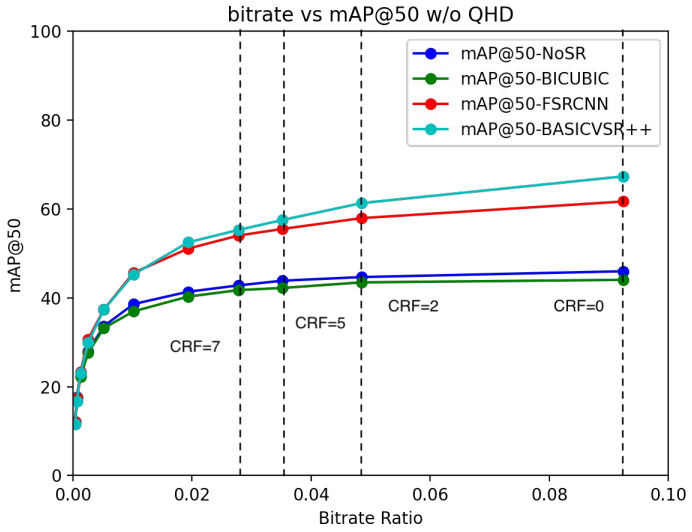
Compressed LR + SR (case 1)

- No SR: baseline
- Bicubic: ~0%
- FSRCNN: 7%~15%
- BasicVSR++: 7%~20%

1080p



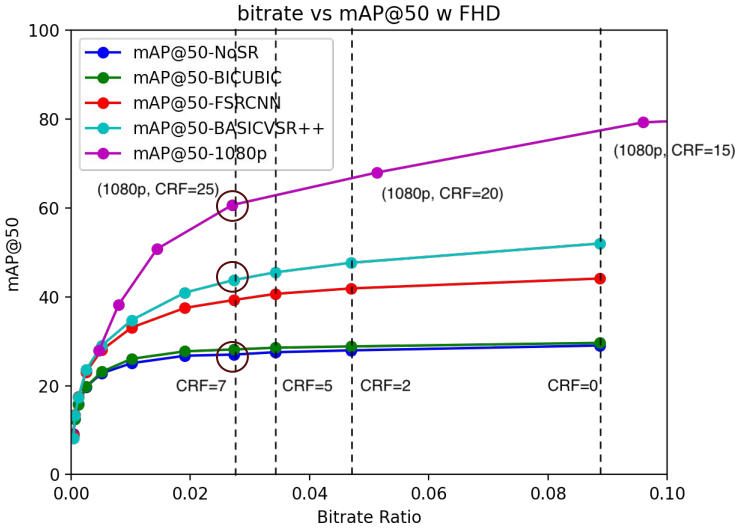
1440p



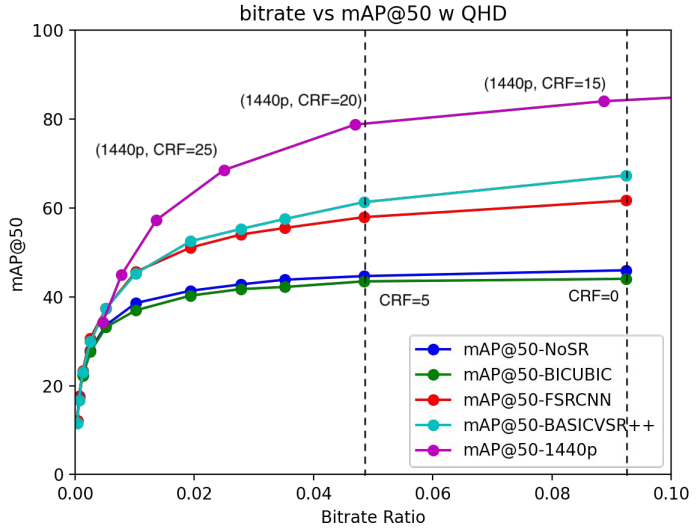
Compressed HR (case 1+2)

- mAP of compressed high resolution video
  - Bitrates: (1080p, CRF=25) close to (270p, CRF=7)
  - mAP: (1080p, CRF=25) better than (270p, CRF=7)
  - Reduce the resolution to save bandwidth might not be the greatest idea!

1080p



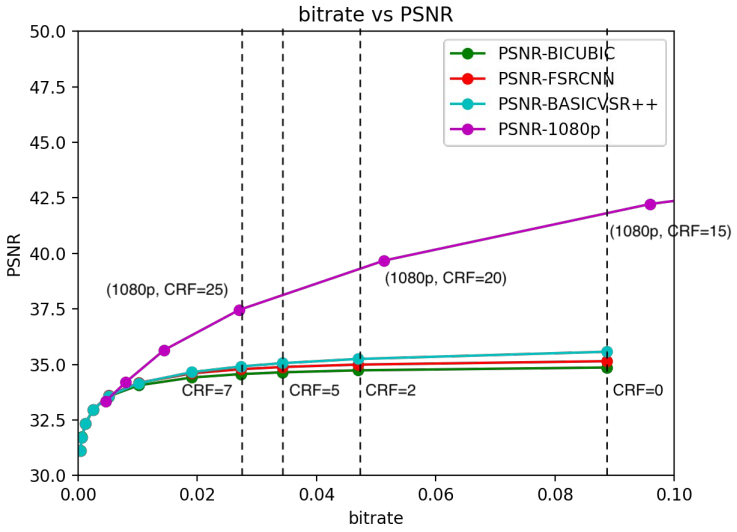
1440p



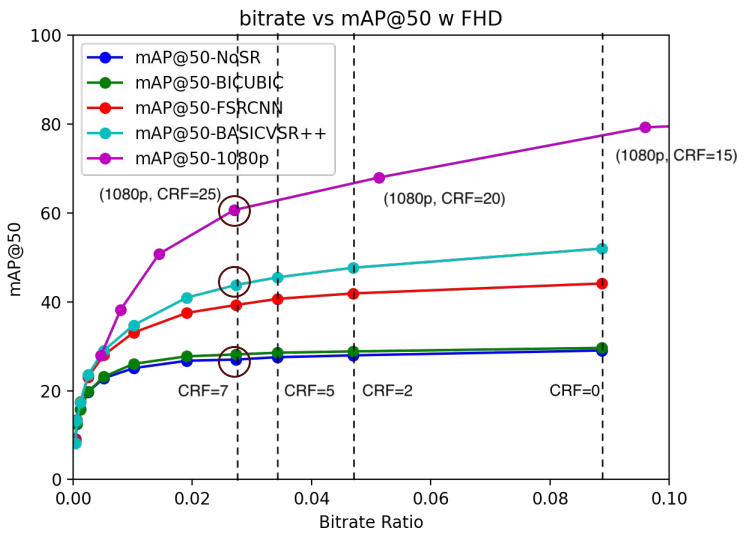
PSNR

- SISR and VSR have little difference with Bicubic
  - SR can be are helpful for object detection

1080p PSNR



1080p mAP@50



Conclusions

- Directly scales videos with interpolation have no benefit for mAP and PSNR.
- SISR and VSR methods attain maximum 15~20% of mAP improvement.
- Adjusting CRF (CQP) alone could be more effective than adjusting both CRF (CQP) and resolution, even with the help of general SR methods, in terms of both time and bandwidth.

Implications

- Mobile-AR or Live Streaming
  - Bandwidth
    - Consider both resolution and CRF (or other rate control mode) when encoding videos.
    - Optimize SR without directly fitting specific datasets.

Q&A

References

[1] Borel-Donohue, Christoph, and S. Susan Young. "Image quality and super resolution effects on object recognition using deep neural networks."

[2] Aqqa, Miloud, Pranav Mantini, and Shishir K. Shah. "Understanding How Video Quality Affects Object Detection Algorithms."

[3] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection."

[4] Ostermann, Jörn, et al. "Video coding with H. 264/AVC: tools, performance, and complexity."

[5] Chen, Jian-Wen, Chao-Yang Kao, and Youn-Long Lin. "Introduction to H. 264 advanced video coding."

[6] H. S. Malvar, A. Hallapuro, M. Karczewicz and L. Kerofsky, "Low-complexity transform and quantization in H.264/AVC."

References

[7] Haris, Muhammad, Greg Shakhnarovich, and Norimichi Ukita. "Task-driven super resolution: Object detection in low-resolution images."

[8] Wang, Yi, et al. "Remote sensing image super-resolution and object detection: Benchmark and state of the art."

[9] Chen, Ying, et al. "Higher quality live streaming under lower uplink bandwidth: an approach of super-resolution based video coding."

[10] Wang, Zelong, et al. "Revisiting super-resolution for internet video streaming."

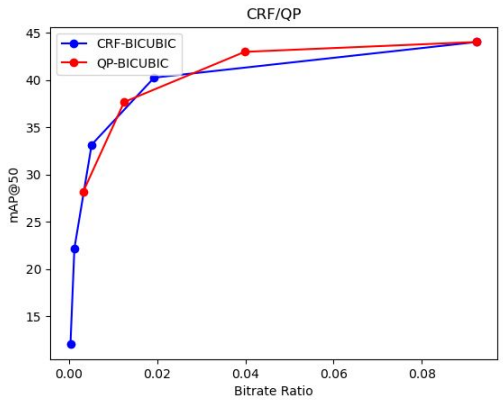
[11] Kim, Jaehong, et al. "Neural-enhanced live streaming: Improving live video ingest via online learning."

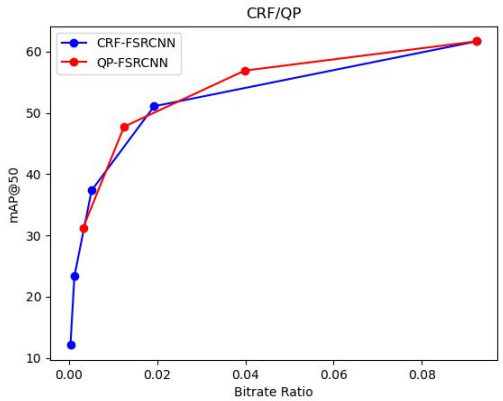
Thank you for listening

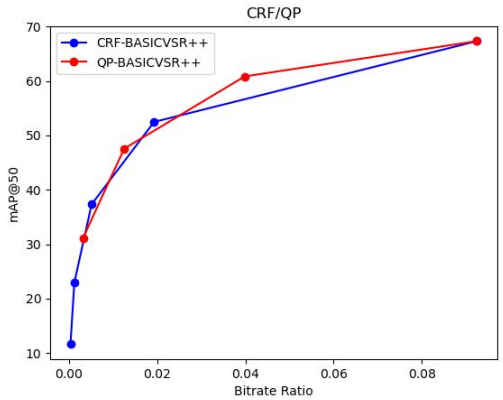
Backup slides

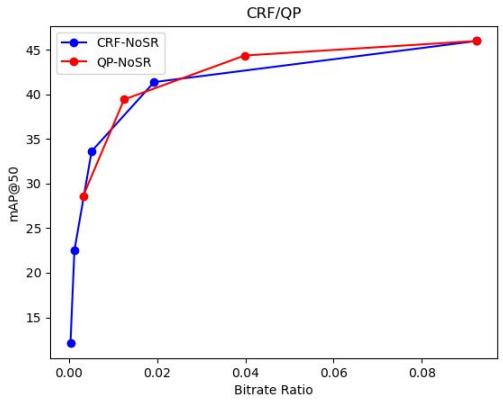
Results

- CRF vs CQP









Metrics

- mAP (mean average precision)
  - mAP@50: IOU > 0.5

GT	BB	IoU	Confidence	Is TP
GT1	BB1	0.7	0.9	Yes
GT2	BB2	0.9	0.8	Yes
GT2	BB7	0.6	0.7	No
GT3	BB3	0.6	0.6	Yes
GT4	BB4	0.4	0.6	No

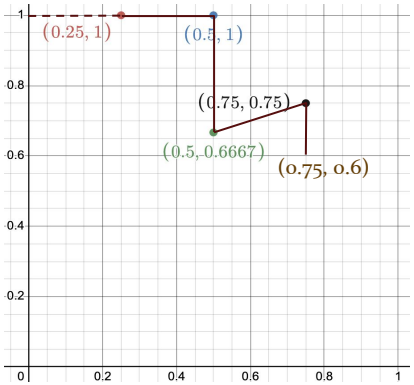
Precision = TP / (TP+FP) = 1/1
Recall = TP / total GT = 1/4

Precision = TP / (TP+FP) = 2/2
Recall = TP / total GT = 2/4

Precision = TP / (TP+FP) = 2/3
Recall = TP / total GT = 2/4

Precision = TP / (TP+FP) = 3/4
Recall = TP / total GT = 3/4

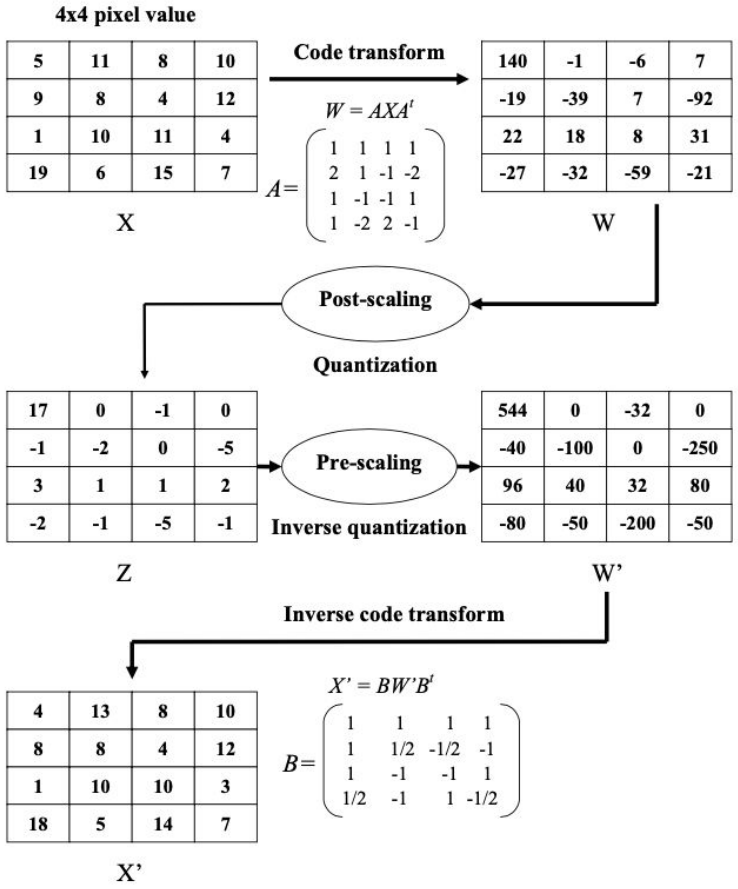
Precision = TP / (TP+FP) = 3/5
Recall = TP / total GT = 3/4



Precision: Y-axis
Recall: X-axis

Quantization parameter

- H.264, H.265,…
  - I/B/P-type frames
  - residual encoding using DCT or Integer transform



encode[6]:

$$X_q(i,j) = \text{sign}\{X(i,j)\}\frac{|X(i,j)| + f(Q_s)}{Q_s}$$

decode[6]:

$$X_r(i,j) = Q_s X_q(i,j).$$