Overview of Visual Signal Compression towards Machine Vision

Shurun Wang[1], Yan Ye[1], Shiqi Wang[2]

[1] Alibaba Damo Academy, Beijing, China
[2] City University of Hong Kong, Hong Kong, China

Feb. 2024

Outline

- Introduction
- Framework
- Performance
- Conclusion and Envision

Introduction

Introduction

Human vision consumption

Live broadcast

Video conference

Video surveillance

...

AI algorithm

AI hardware

Intelligent transportation

Digital retina

Smart city

...

Machine vision consumption

➢ **Machine vision is replacing human vision**

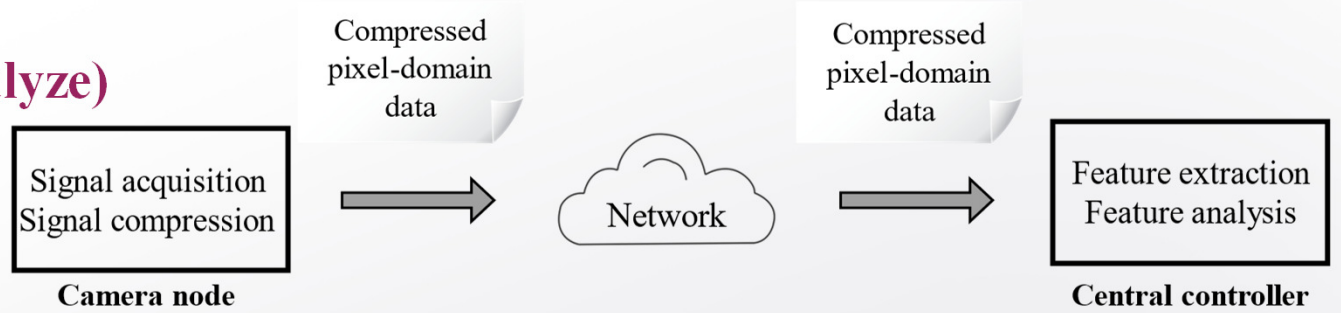
Video Coding for Machines

- **High visual data volume:** In recent years, IoVT (Internet of Video Things) has been deployed almost everywhere, producing prohibitively high visual data volume. It is impossible for human analysis.
- **Many machine vision based applications:** With the development of AI, there are numerous machine vision based applications, such as traffic, AI industry, autopilot, etc. Machine vision is replacing human vision.
- **Perceptually optimized video codecs:** Existing video codecs mainly target human perception. Not aligned with machine vision applications.

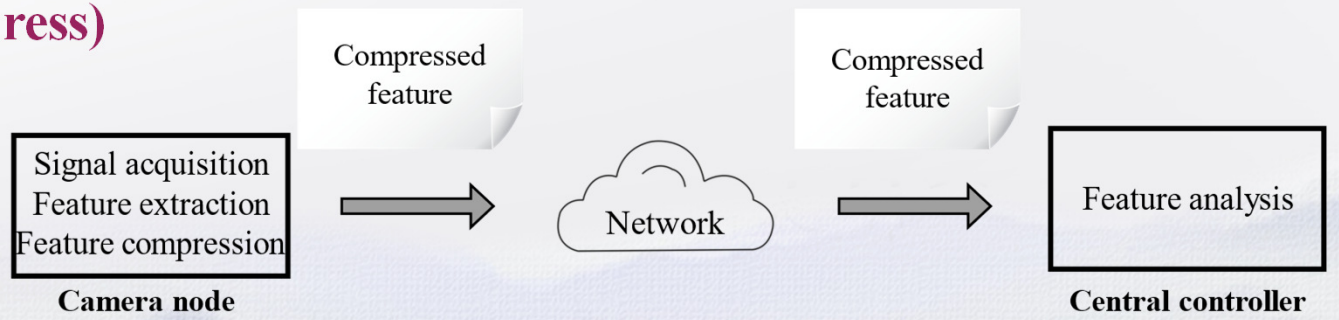
Video coding for machines (VCM) is important

CTA and ATC

- **CTA (Compress-then-analyze)**



- **ATC (Analyze-then-compress)**

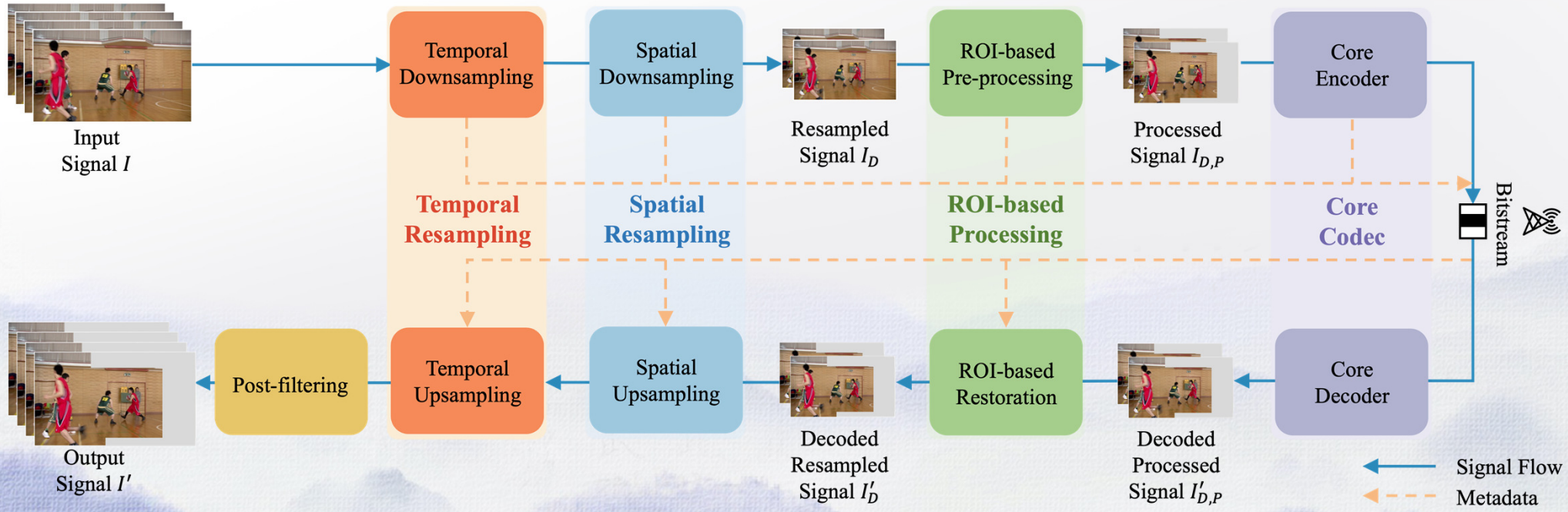


- **CTA**
  - Support visual reconstruction
  - Multiple machine tasks and human viewing

Visual signal compression towards machines (VSCM) is pragmatic

Framework

Overall Framework



Core codec

➢The core component for compression

- Bridge the gap between compact representation and the perceptible visual signal

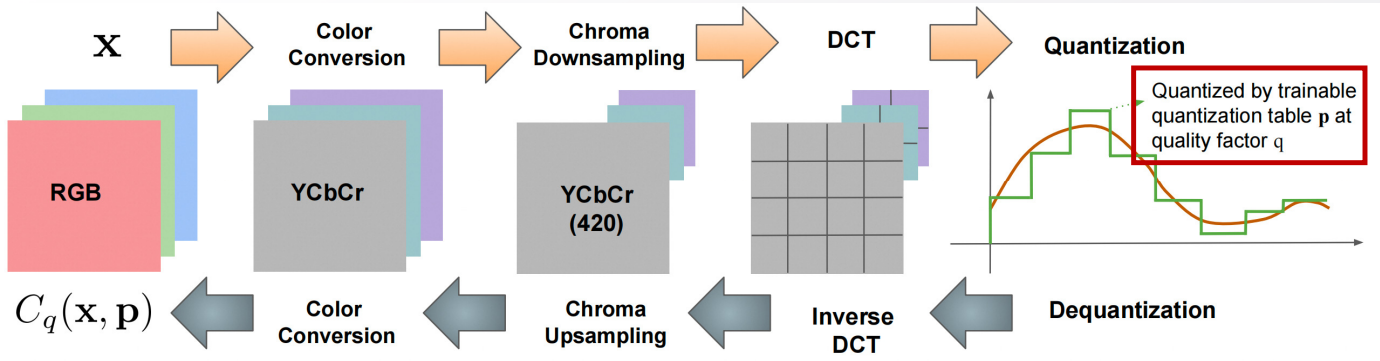
➢Including:

- Traditional codec
- Deep learning codec
- Combination of traditional codec and deep learning codec

Traditional codec

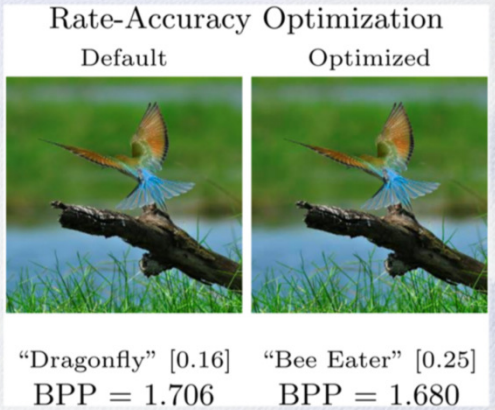
➢Image compression for VSCM

- Optimize the quantization model [1]



Optimize the quantization model with neural network

$$L = R + \lambda_1 D_{rec}\left(x, C_q(x,p)\right) + \lambda_2 D_{cls}(C_q(x,p))$$

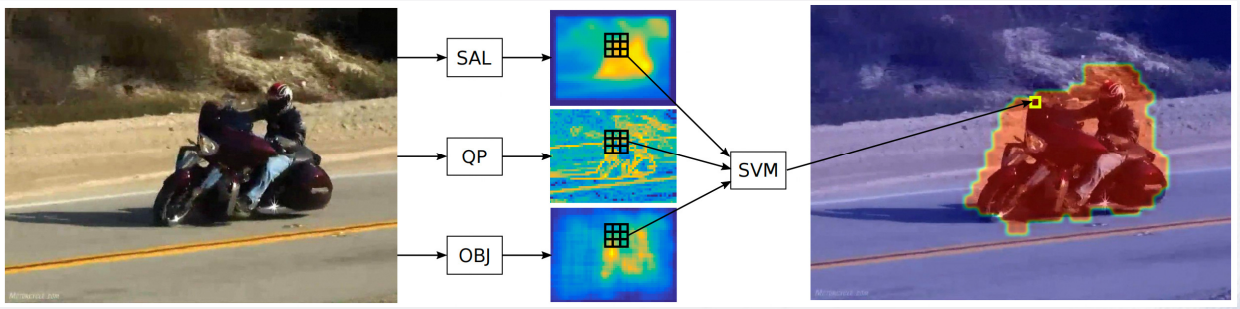


[1]Luo X, Talebi H, Yang F, et al. The rate-distortion-accuracy tradeoff: Jpeg case study[J]. arxiv preprint arxiv:2008.00605, 2020.

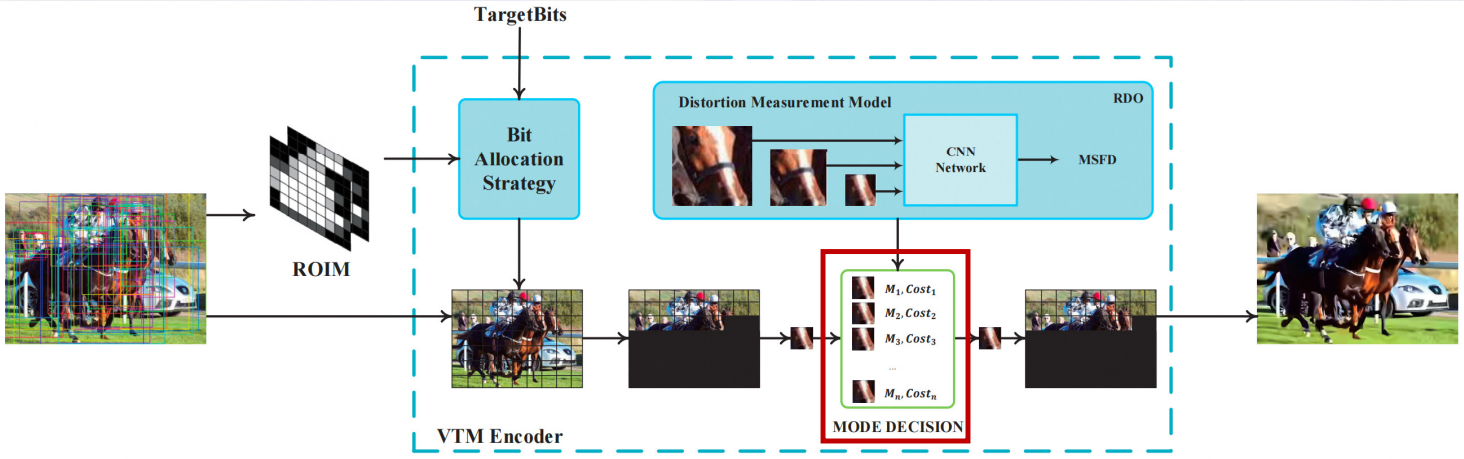
Traditional codec

➢Video compression for VSCM

- Optimize the quantization model [2]
  - CTU in object region?
    - Yes: No action
    - No: QP=51

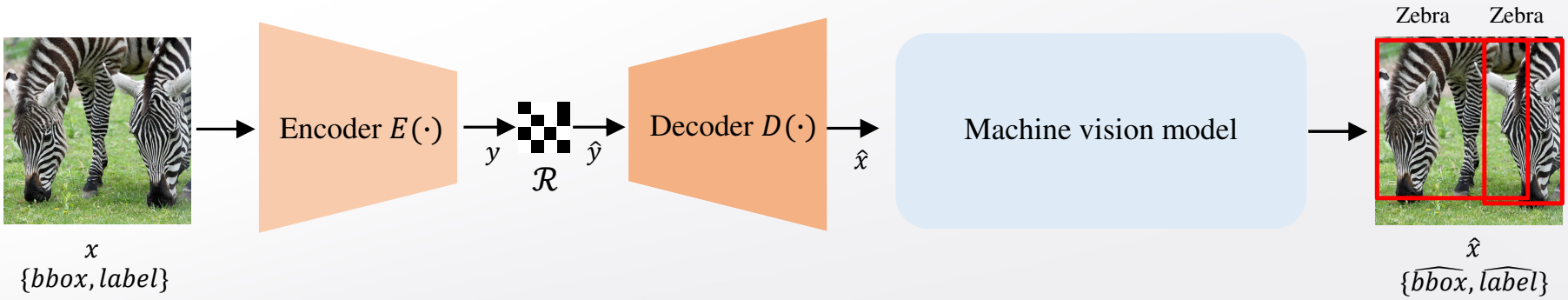


- Optimize the RDO (rate-distortion optimization) [3]



[2]Galteri L, Bertini M, Seidenari L, et al. Video compression for object detection algorithms[C]//2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 3007-3012.
[3]Huang Z, Jia C, Wang S, et al. Visual analysis motivated rate-distortion model for image coding[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

Deep learning codec



➢Loss function

- $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{feature}} + \lambda_2 \mathcal{L}_{\text{mse}} + \mathcal{R}$ [4]
- $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{task}} + \lambda_2 \mathcal{L}_{\text{mse}} + \mathcal{R}$ [5]

➢Network structure

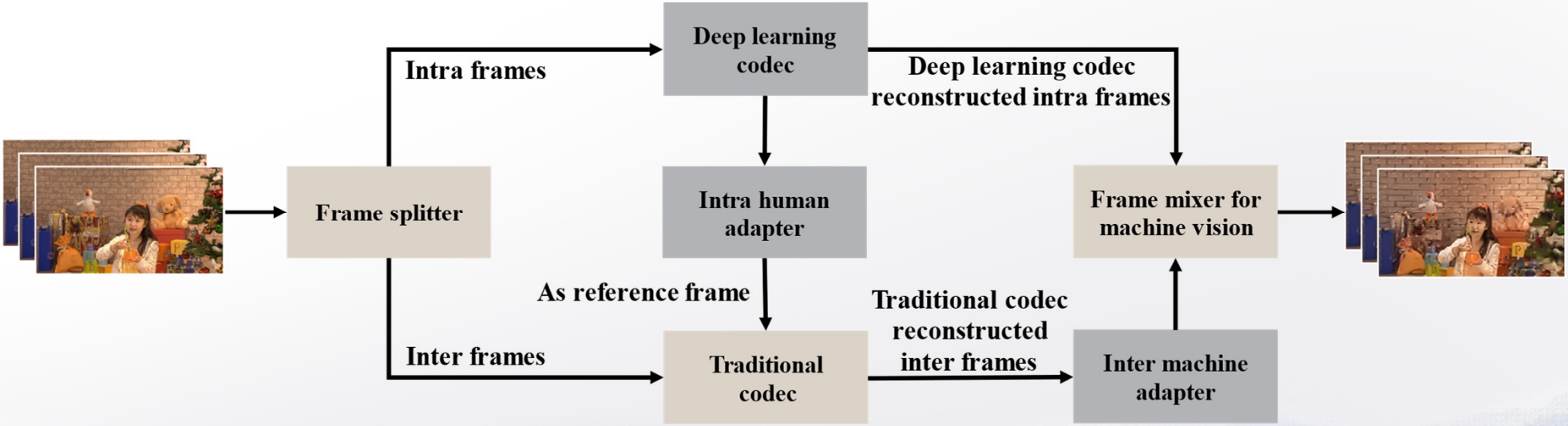
- Latent space masking network (LSMnet) [6]
- Variable bitrate structure [7]

➢Optimization range

- Joint optimize codec and machine vision model [7]

[4] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, and E. Rahtu. Image coding for machines: an end-to-end learned approach. In *ICASSP 2021-2021 IEEE Inter national Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1590–1594. IEEE, 2021.
[5] S. Wang, Z. Wang, S. Wang, and Y. Ye. End-to-end compression towards machine vision: Network architecture design and optimization. IEEE Open Journal of Circuits and Systems, 2:675–685, 2021.
[6] K. Fischer, F. Brand, and A. Kaup. Boosting neural image compression for machines using latent space masking. arXiv preprint arXiv:2112.08168, 2021.
[7] S. Wang, Z. Wang, S. Wang, and Y. Ye. Deep image compression towards machine vision: A unified optimization framework. IEEE Transactions on Circuits and Systems for Video Technology, 2022.

Combination of traditional codec and deep learning codec



Original
mAP 70.18

VVC RA QP=52
Bitrate 43.23, mAP 19.22

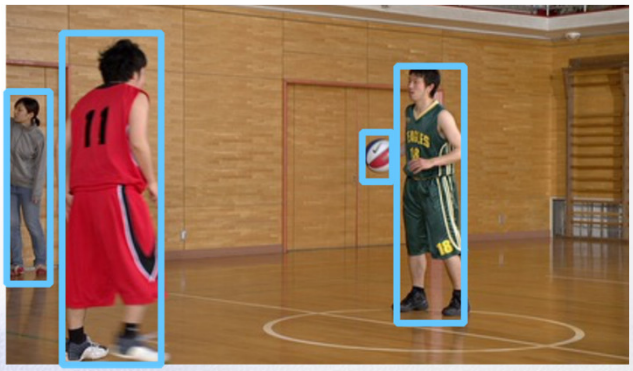
Proposed RA QP=52
Bitrate 42.19, mAP 22.33

[8] H. Zhang, N. Le, J. Ahonen, A. Hallapuro, A. Aminlou, L. Xu, R. Youvalari Ghaznavi, F. Cricri, H. Tavakoli, E. Aksu, and M. Hannuksela. [VCM] response to the cfp of the VCM by nokia. MPEG doc. m60753 and ISO/IEC JTC 1/SC 29/WG 2, 2022.

ROI processing

➢ ROI with Machine vision model

- Transfer with metadata



Original frame and objects

Background remove [9]

Foreground rescale [10]

Background blur [11]

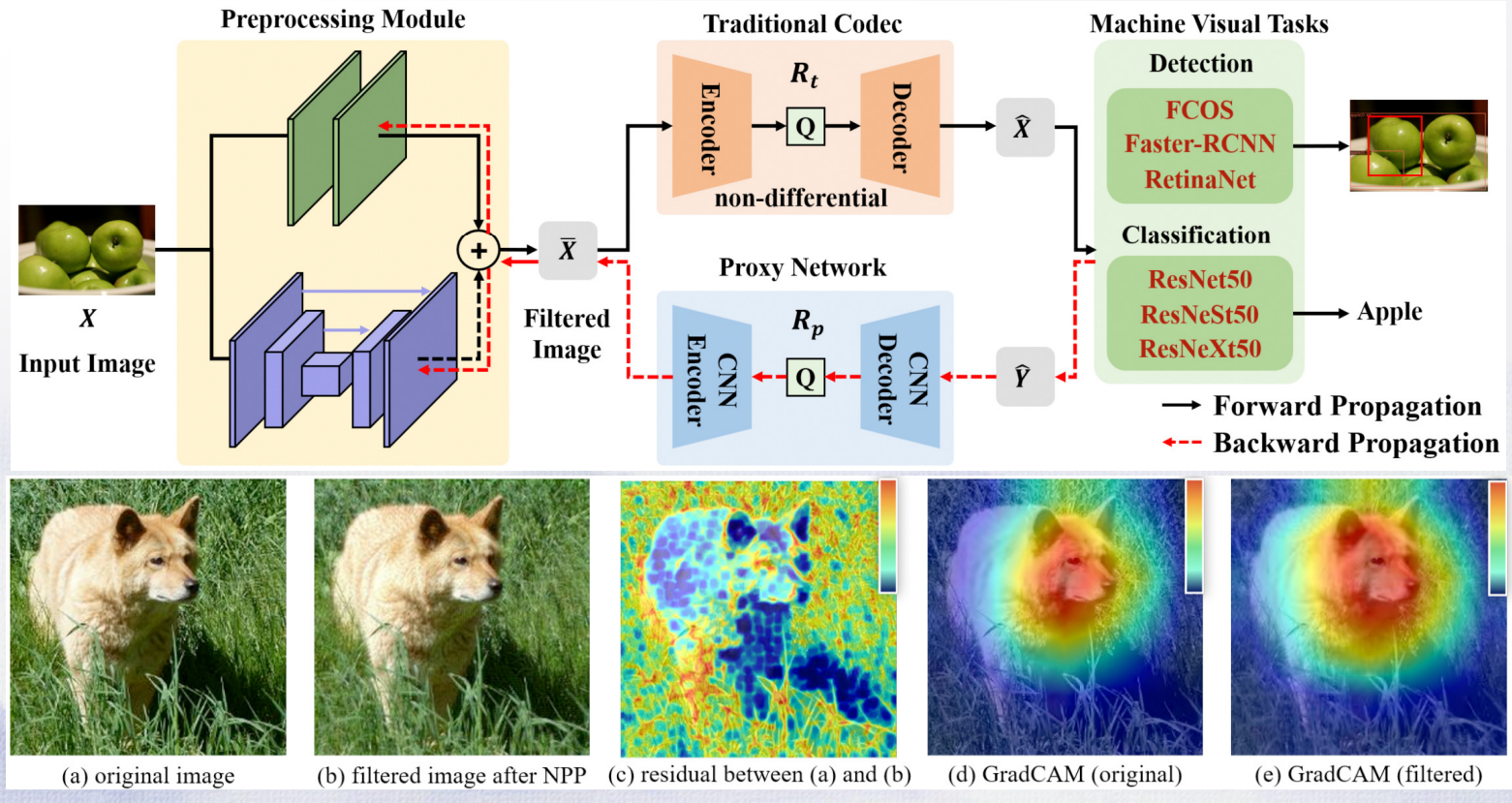
Repack [12]

[9] S. Wang, B. Li, Z. Wang, S. Wang, and Y. Ye. [VCM] video coding for machines cfp response from alibaba and city university of hong kong. *MPEG doc. m60737 and ISO/IEC JTC 1/SC 29/WG 2*, 2022.
[10] H. Kalva, V. Adzic, B. Furht, A. Krause, M. Eimon, and A. Perera. [VCM] response to VCM cfp from the florida atlantic university and op solutions, llc. MPEG doc. m60743 and ISO/IEC JTC 1/SC 29/WG 2, 2022.
[11] B. Li, S. Wang, S. Wang, and Y. Ye. Ahg15: Feature based encoder-only algorithms for the video coding for machines. JVET AC0086 and Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 2023.
[12] Y. Lee, S. Kim, K. Yoon, H. Lim, S. Kwak, H. Choo, and J. Seo. [VCM track 2] response to VCM cfp: Video coding with machine-attention. MPEG doc. M60378 and ISO/IEC JTC 1/SC 29/WG 2, 2022.

ROI processing

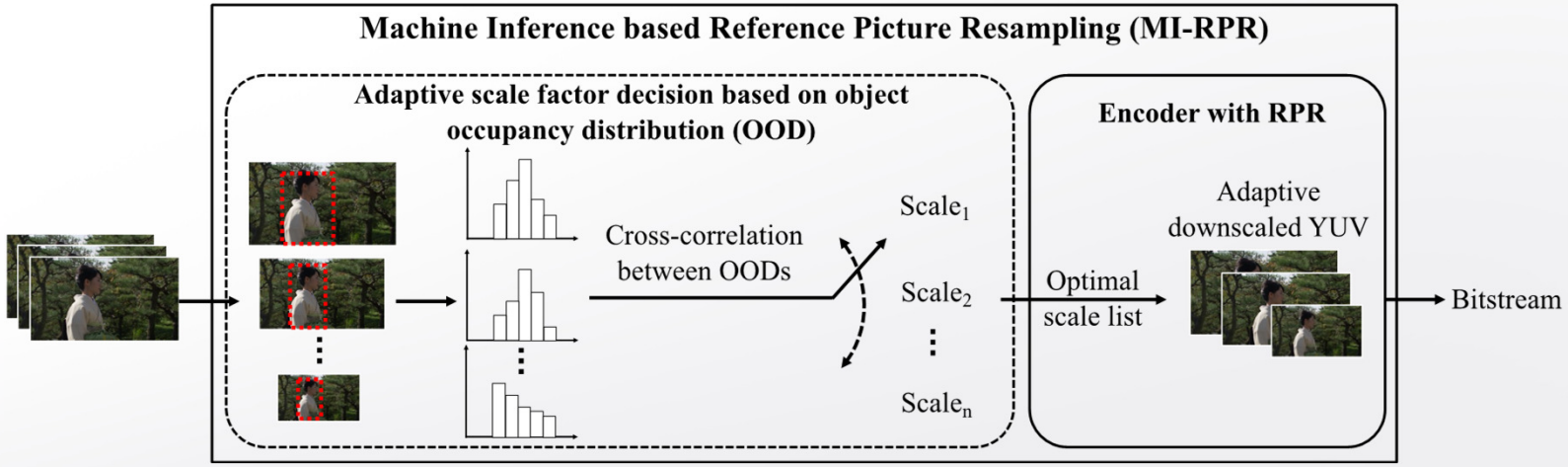
➢ ROI without Machine vision model

- Transfer without metadata



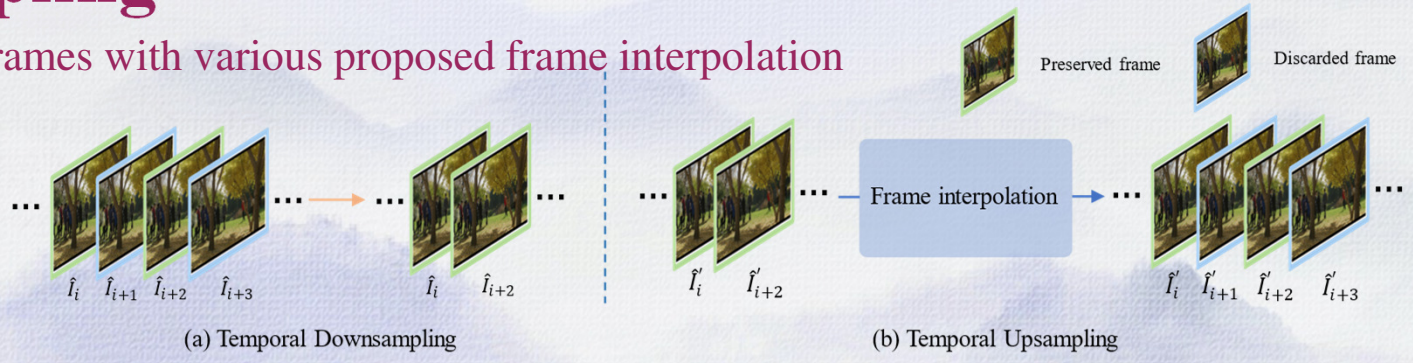
[13] Lu G, Ge X, Zhong T, et al. Preprocessing enhanced image compression for machine vision[J]. arXiv preprint arXiv:2206.05650, 2022

Spatial resampling



Temporal resampling

- Recover the discarded frames with various proposed frame interpolation models [15,16]

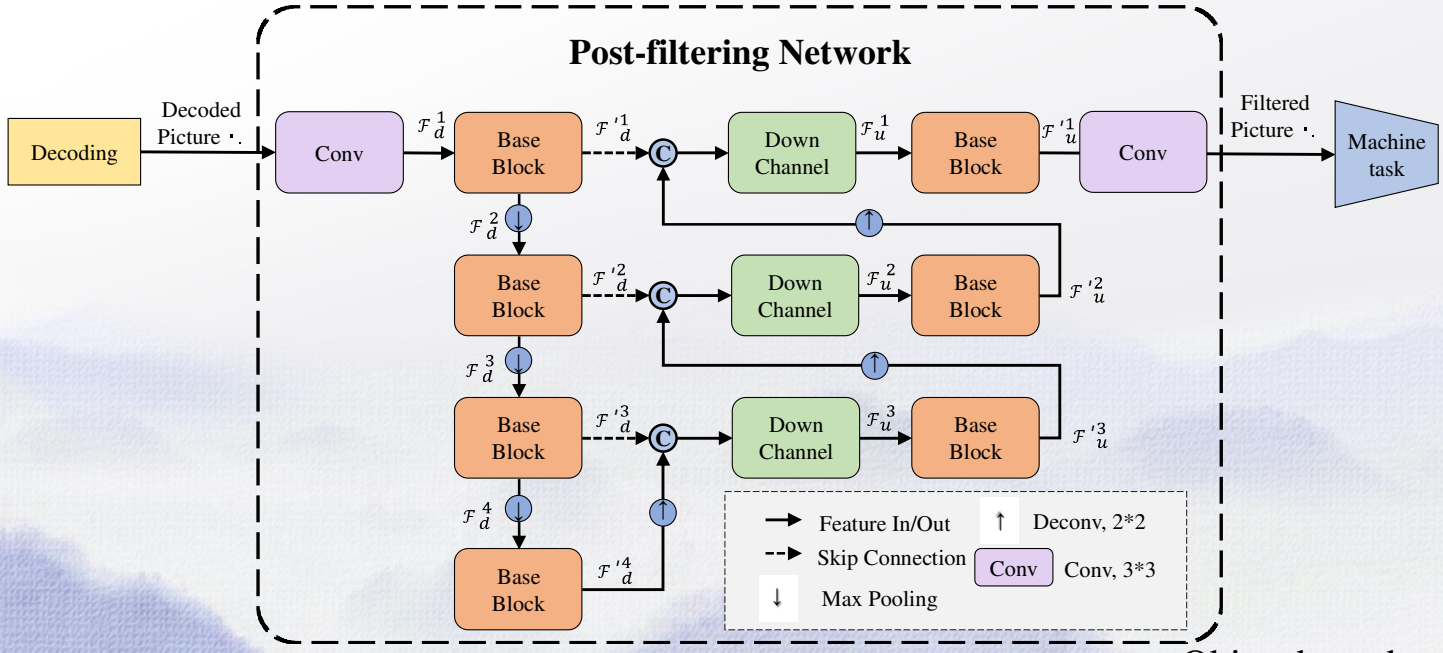


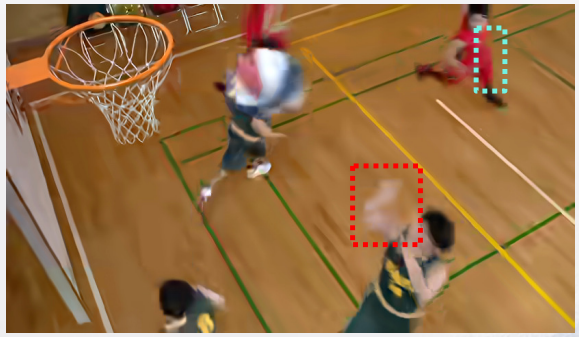
(a) Temporal Downsampling (b) Temporal Upsampling

[14]A. Kim, E. An, K. Seo, S. Jung, W. Cheong, J. Lee, and H. Choo. [VCM] adaptive spatial resampling based on mi rpr for vcm. *MPEG doc. m64124 and ISO/IEC JTC 1/SC 29/WG 4*, 2023.
[15] Z. Liu, Y. Zhang, R. Chernyak, H. Zhu, X. Xu, J. Jia, S. Liu, M. Park, and K. Choi. [VCM] response to VCM call for proposals - an evc based solution. MPEG doc. m60779 and ISO/IEC JTC 1/SC 29/WG 2, 2022.
[16] H. Wang, J. Xue, Y. Zhang, W. Ji, and J. Liu. [VCM] response to cfp on video coding for machines from china telecom. MPEG doc. m60775 and ISO/IEC JTC 1/SC 29/WG 2, 2022.

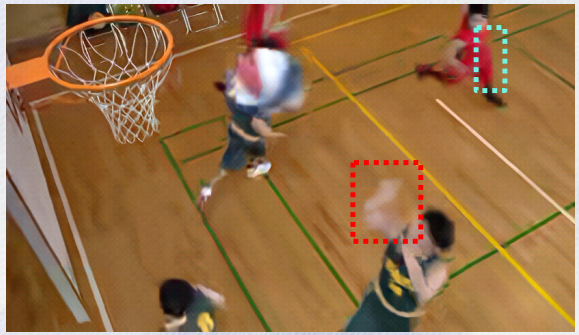
Post filtering

- Machine-oriented post-filtering network to enhance the decoded video before feeding it to the machine task networks [17]





VVC RA QP=47



Filtered picture

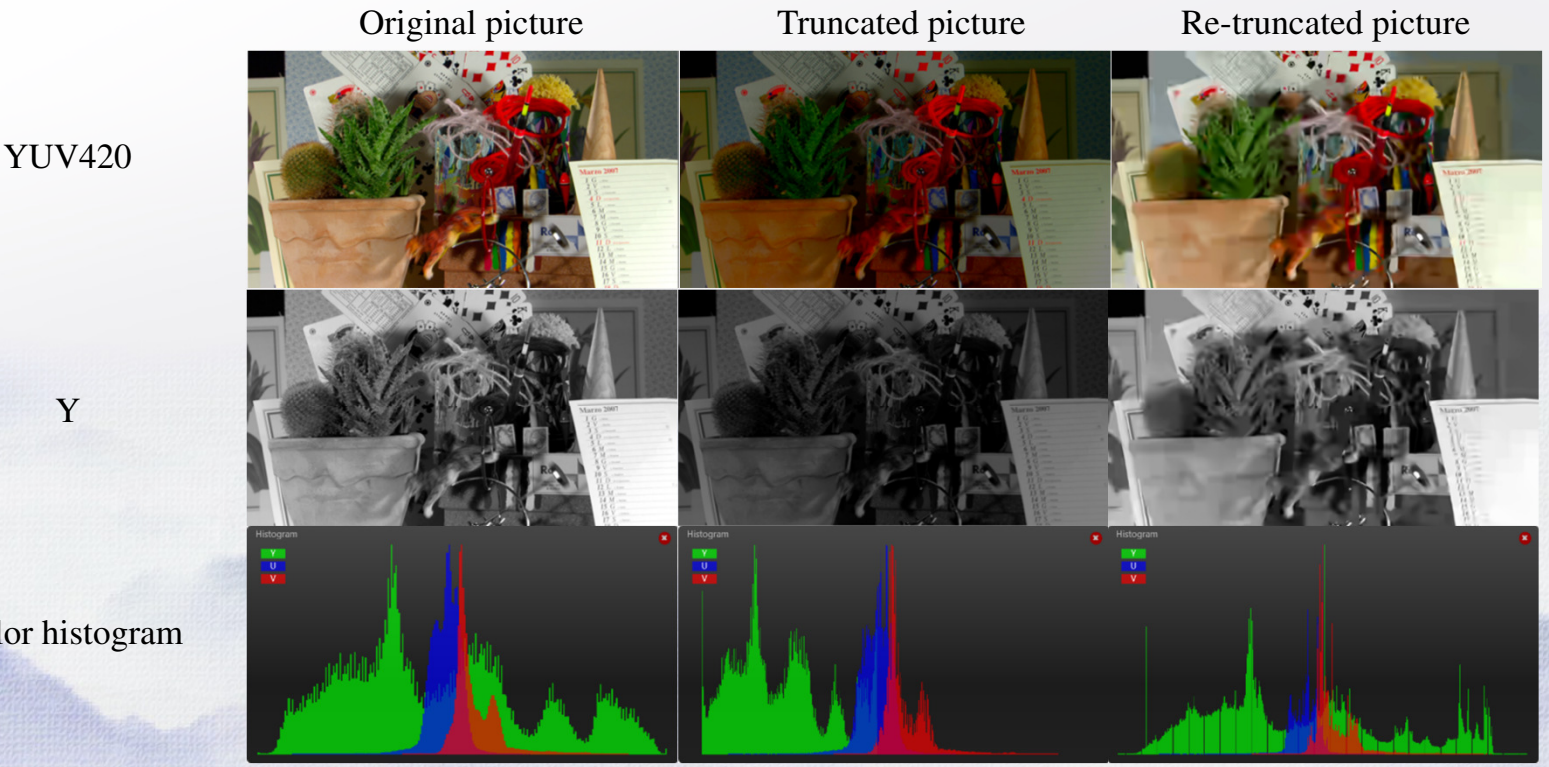
Object boundary more sharp

[17] B. Li, S. Wang, Y. Ye and S. Wang. Ahg8: A post-processing algorithm for machine consumption. JVET AG0212 and Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 2024.

Others

➢Bitdepth truncation [18]

- Before encoding, **truncate** the bitdepth of **luma** component with 1 bit
- After decoding, **recover** the bitdepth of **luma** component with 1 bit



[18] D. Ding, X. Zhao, Z. Liu, S. Liu. [VCM] Bitwise Efficiency: Truncating Bit Depth for Machine Video Coding. MPEG doc. m65525 and ISO/IEC JTC 1/SC 29/WG 2, 2023.

Performance

Performance evaluation

➢Test datasets and evaluation tasks

Dataset Name	Test sequence number	Definition	Machine Task	Machine Task Metric
SFU-HW	13	2560x1440, 1920x1080, 832x480, 416x240	Object Detection	mAP
TVD	7	1920x1080	Object Tracking	MOTA

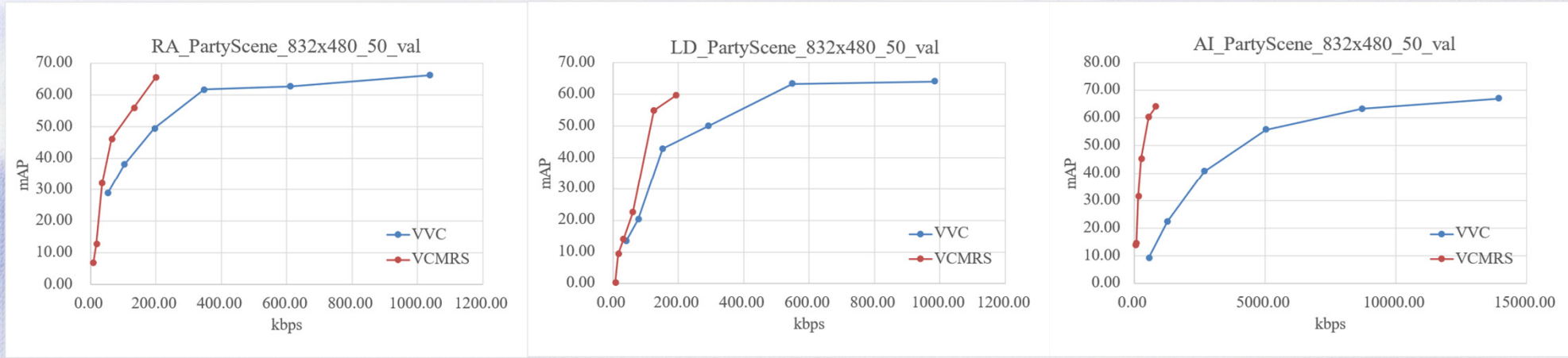
➢Anchor and test

- Anchor: VVC (VTM-22.2)
- Test: VCMRS-0.7 (https://mpeg.expert/software/MPEG/Video/VCM/VCM-RS)
- Configs: RA, LD, AI
- Performance metric: BD-rate, Pareto BD-rate and BD-quality (mAP/MOTA)

Performance

➢Obvious performance improvements compared with VVC

Task (Dataset)	Configs	BD-rate	Pareto BD-rate	BD-quality
Object detection (SFU-HW)	RA	/	-49.65%	8.90
	LD	/	-49.50%	11.90
	AI	-84.83%	-84.86%	/
Object tracking (TVD)	RA	/	-76.43%	/
	LD	/	-60.21%	11.23
	AI	-91.83%	-91.83%	/



Conclusion and Envision

Conclusion and Envisions

- Remarkable performance improvements for VSCM have been achieved.
- The visual signal compression for multiple machine tasks.
- The compact representation for machine vision features.
- The joint compression for both machine vision and human perception.

達摩院 ALIBABA DAMO ACADEMY CityU 香港城市大學 City University of Hong Kong

Thanks