SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

VISUAL CODING FOR HUMANS AND MACHINES

Ivan V. Bajić

School of Engineering Science
Simon Fraser University
Burnaby, BC, Canada

Workshop on Image/Video/Audio Quality @ WACV 2024

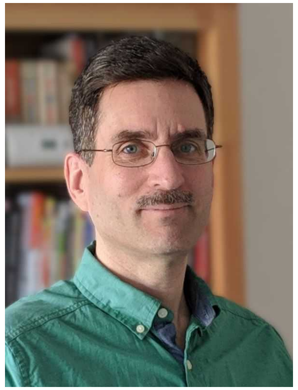
SPECIAL THANKS

People @ SFU Multimedia Lab (multimedia.fas.sfu.ca) whose work contributed to this presentation – thank you!


Anderson de Andrade


Bardia Azizian


Hyomin Choi


Robert A. Cohen


Yalda Foroutan


Alon Harell

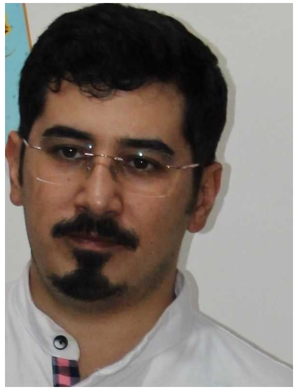

Hadi Hadizadeh


Elahe Hosseini


Suemin Lee


Saeed Ranjbar Alvar


Mateen Ulhaq


Rashid ZamanshoarHeris

OVERVIEW

Introduction

- Why coding for machines?

Part 1 – Coding for machines

- Rate-distortion results
- Examples

Part 2 – Coding for humans and machines

- Image coding
- Video coding

Part 3 – Standardization

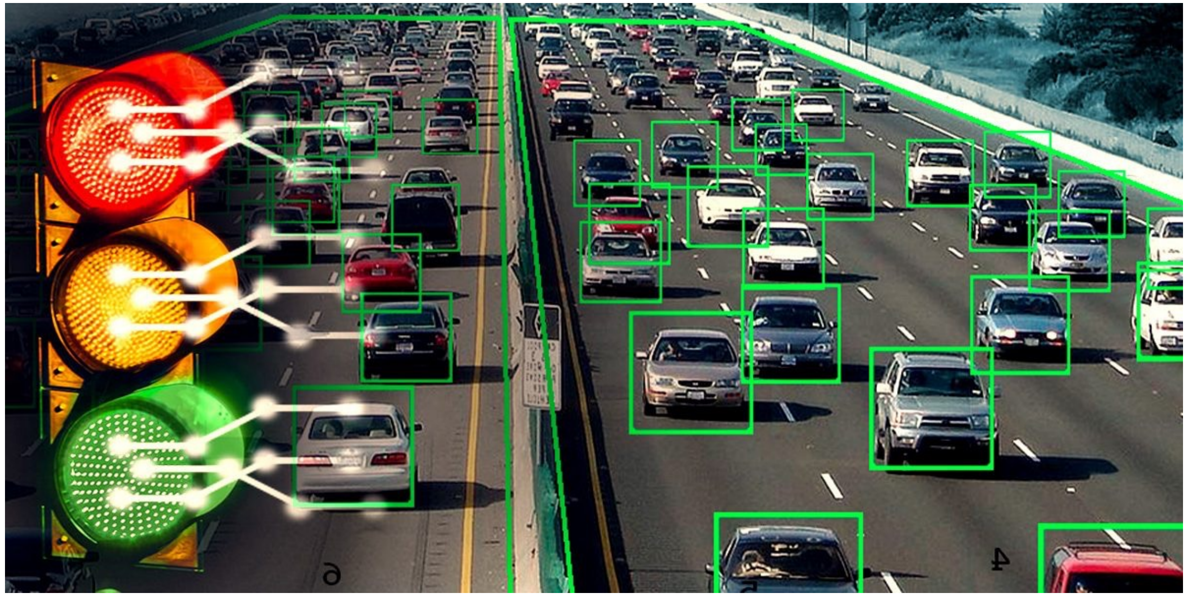
- CDVS and CDVA
- JPEG AI
- MPEG-VCM (Video Coding for Machines)

Introduction

EMERGING APPLICATIONS

Automatic traffic monitoring & management

- Cameras (and other sensors) along roads and intersections
- Counting vehicles, pedestrians, etc.
- Estimating their speed, traffic intensity, detecting violations and emergencies
- Help manage traffic
- Tasks:
  - Object detection
  - Object tracking
  - Human viewing (occasionally)



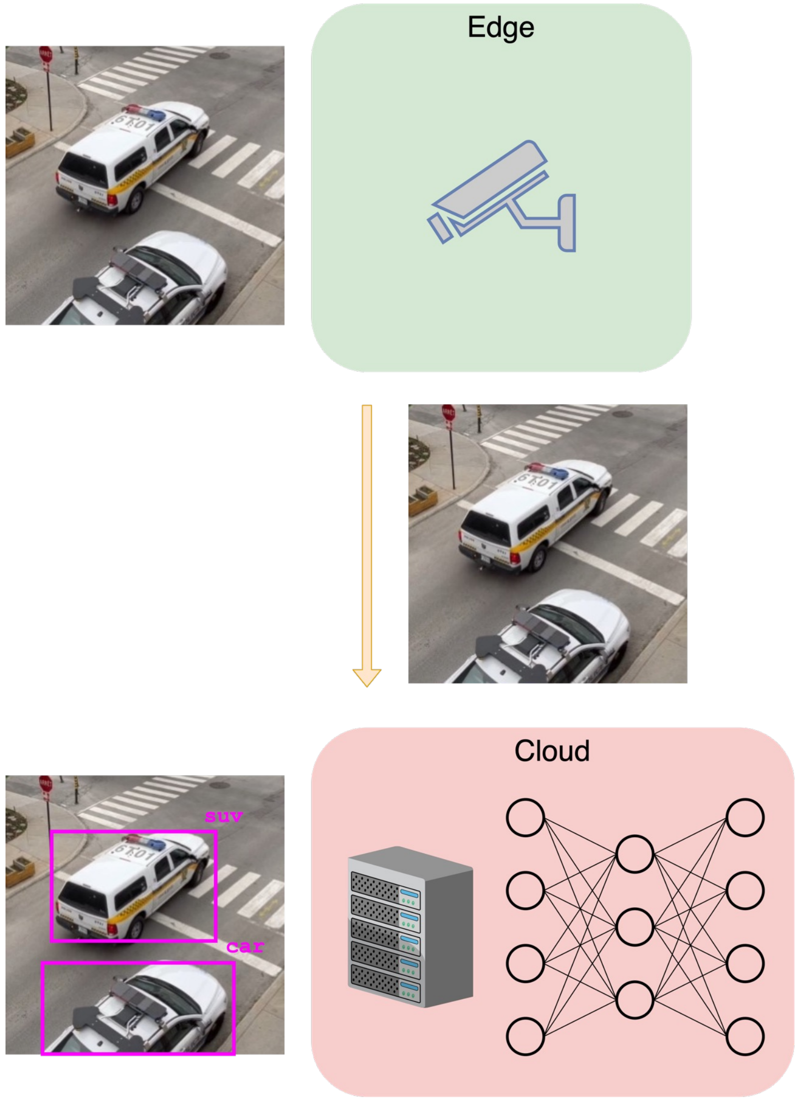
entrackr.com

The traditional approach

- Camera captures the image
- Encoded image sent to the cloud
- Analysis ("intelligence") performed in the cloud
- Result sent back to the edge (if needed) or to other systems in the cloud

Challenges:

- Concerns over privacy
- Does not take full advantage of capabilities of modern edge devices
- High bitrate

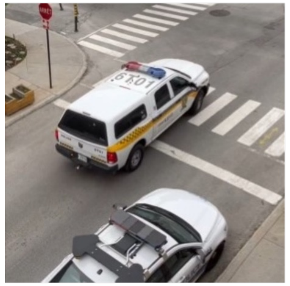


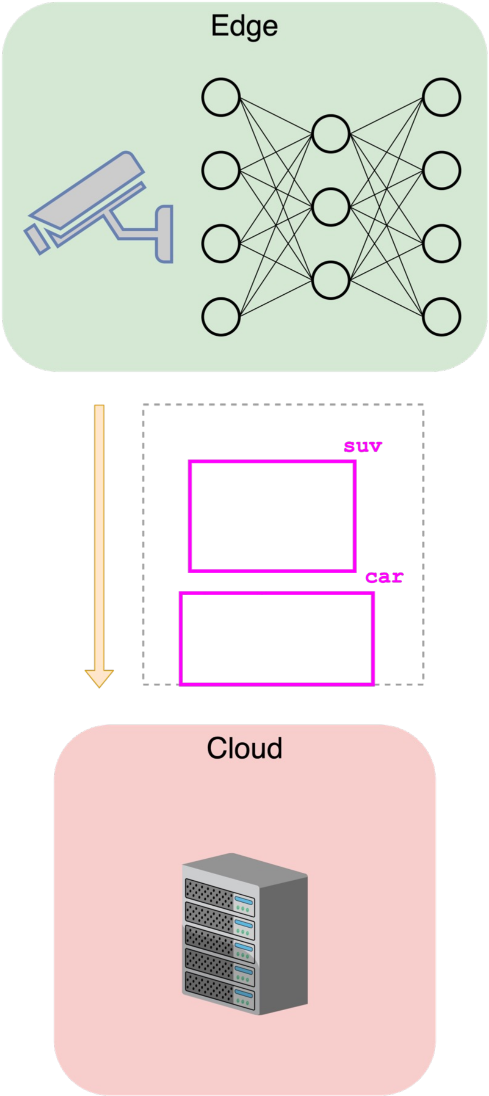
The new approach

- Analysis ("intelligence") performed at the edge
- Only the result sent to the cloud, if needed
- Makes the edge device "smart"
- Addresses many privacy concerns
- Lowest bitrate

Challenges:

- Can be energy-intensive (at the edge)
- Model complexity limited by the resources of the edge device
    - Cloud will always be able to host larger, more complex models
- What if a different type of analysis is needed?





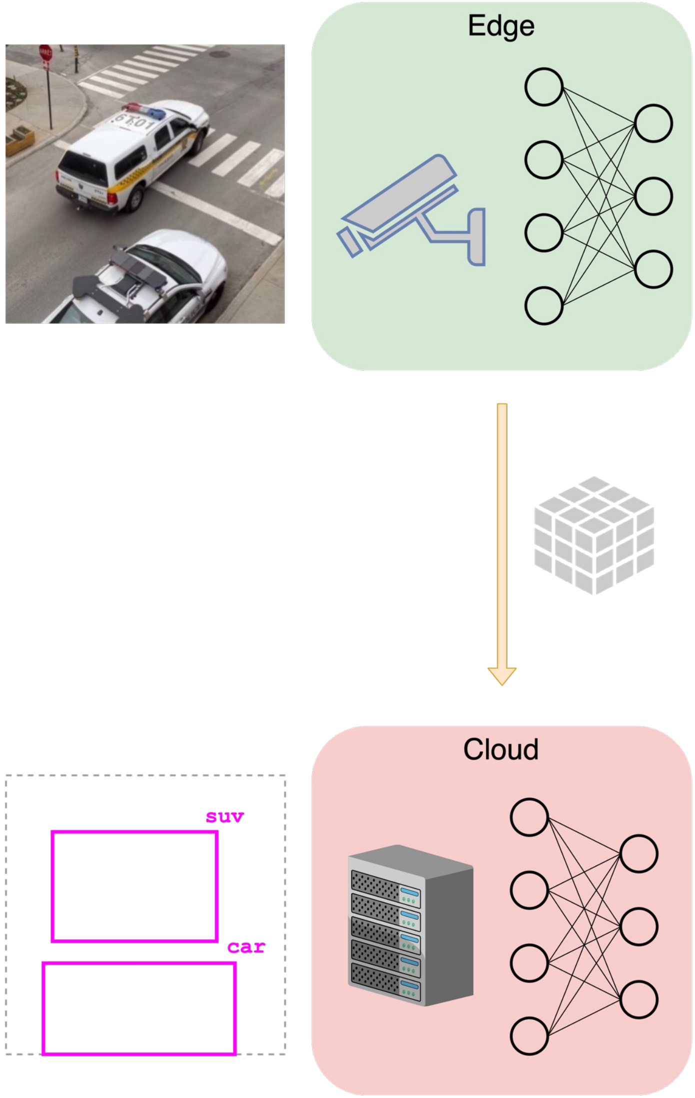
COLLABORATIVE INTELLIGENCE

(Edge-cloud) collaborative intelligence

- Between cloud-only and edge-only extremes
- Part of “intelligence” at the edge, other part at the cloud
- Features sent to the cloud, task(s) completed there
- Able to address privacy concerns
- Able to scale to available resources

Challenges:

- Design criteria?
- Bitrate?



Y. Lou et al., "Front-end smart visual sensing and back-end intelligent analysis: A unified infrastructure for economizing the visual system of city brain," IEEE JSAC, vol. 37, no. 7, pp. 1489-1503, July 2019.

I. V. Bajić, W. Lin and Y. Tian, "Collaborative intelligence: Challenges and opportunities," Proc. ICASSP, 2021, pp. 8493-8497

Part 1

Coding for machines

multimedia laboratory
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

LIMITS OF CODING FOR MACHINES

Can coding for machines be more efficient than conventional coding (for humans)?

Recall the data processing inequality (DPI)

- If $X \rightarrow Y \rightarrow Z$ is a Markov chain, then

$$I(X;Y) \geq I(X;Z)$$

- Downstream variable (Z) has no more information about input (X) than an upstream variable (Y)

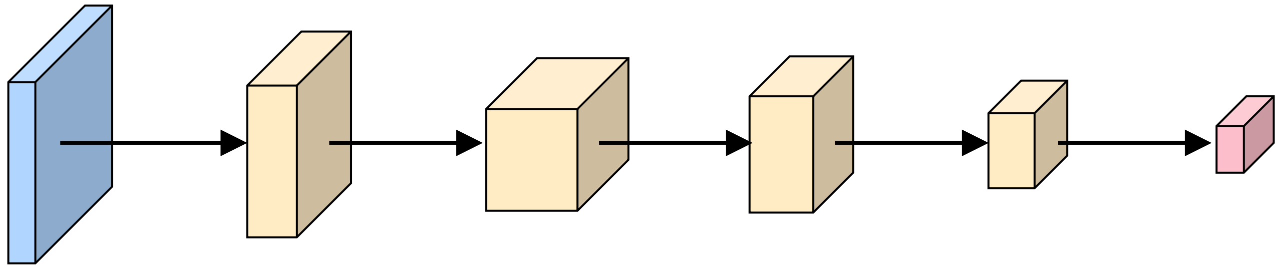
- Extended version of DPI: if $X \rightarrow Y \rightarrow Z \rightarrow W$ is a Markov chain, then

$$I(Y;Z) \geq I(X;W)$$

T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley, 2006.
R. W. Yeung, *A First Course in Information Theory*, Springer, 2006.

NEURAL NETWORK LAYERS FORM MARKOV CHAINS

- $\mathcal{Y}_i$ = output of the i-th layer in a feedforward neural network



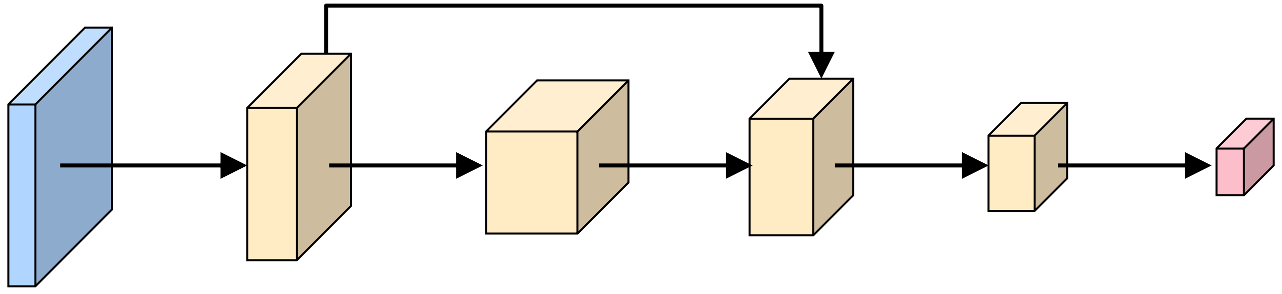
(input) X $\mathcal{Y}_1$ $\mathcal{Y}_2$ $\mathcal{Y}_3$ $\mathcal{Y}_4$ T (output)

- $X \rightarrow \mathcal{Y}_1 \rightarrow \mathcal{Y}_2 \rightarrow \mathcal{Y}_3 \rightarrow \mathcal{Y}_4 \rightarrow T$ is a Markov chain
  - So is any chain $X \rightarrow \mathcal{Y}_i \rightarrow \mathcal{Y}_j \rightarrow T$ for $i < j$
  - True for dense layers, convolutional layers, pooling layers, etc.

N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” Proc. IEEE Information Theory Workshop (ITW), Mar. 2015.

NEURAL NETWORK LAYERS FORM MARKOV CHAINS

- What about skip connections?



- $X \rightarrow \mathcal{Y}_1 \rightarrow \mathcal{Y}_2 \rightarrow \mathcal{Y}_3$ is **not** a Markov chain
  - $\mathcal{Y}_3$ depends on both $\mathcal{Y}_2$ and $\mathcal{Y}_1$, not just $\mathcal{Y}_2$
  - However, $X \rightarrow \mathcal{Y}_1 \rightarrow \mathcal{Y}_3$ is a Markov chain
  - Markovity still holds “across” skip connections, but not “under” them

FEATURE COMPRESSIBILITY

Claim: In a non-generative feedforward neural network, intermediate features are at least as compressible as the network's input.

- Let X be the input, $\mathcal{Y} = \{\mathcal{Y}_i\}$ be a set of some intermediate layer outputs (features)
- Using DPI it can be shown

$$H(\mathcal{Y}) \leq H(X) \qquad \text{and} \qquad R_{\mathcal{Y}}(D) \leq R_X(D)$$

where distortion D is measured at the network's output

- By extension

Claim: Deeper layers are at least as compressible as the shallower layers.

$$H(\mathcal{Y}_i) \leq H\left(\mathcal{Y}_j\right) \qquad \text{and} \qquad R_{\mathcal{Y}_i}(D) \leq R_{\mathcal{Y}_j}(D) \qquad \text{for } i > j$$

H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Processing, vol. 31, pp. 2739-2754, 2022.

FEATURE COMPRESSIBILITY

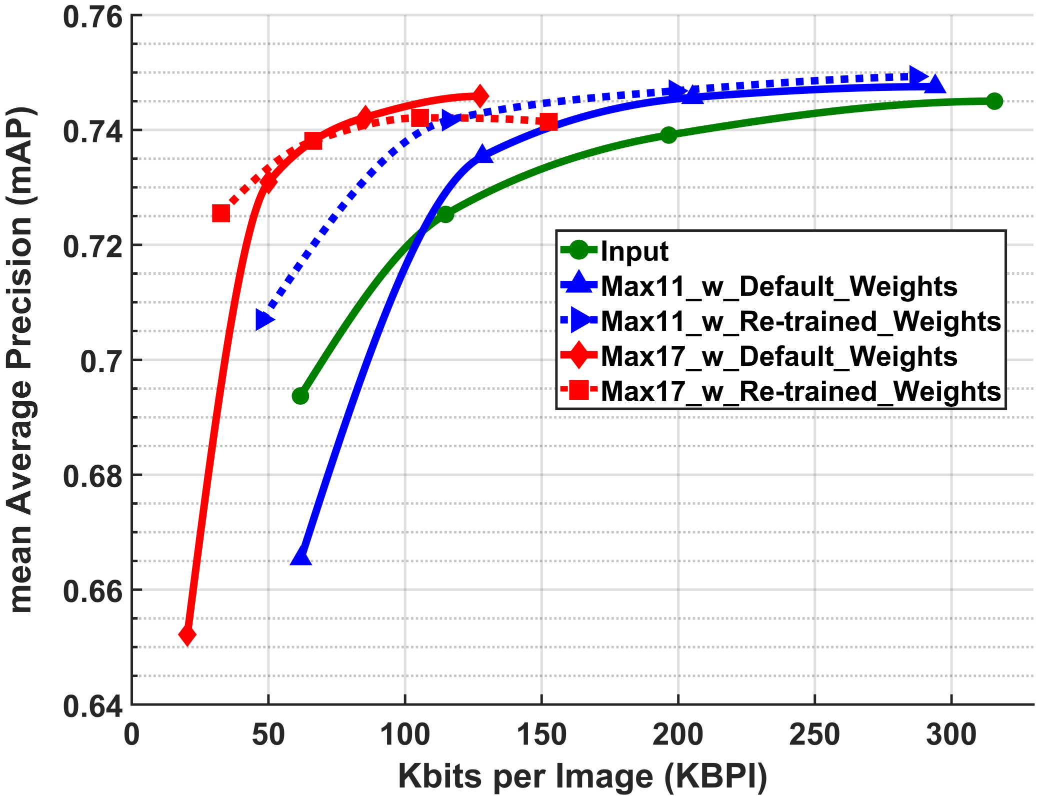
- Great news for collaborative intelligence and coding for machines!
    - Can do better than cloud-only approach (conventional coding)
- Further, the theory suggests the following design principle:
    - *Compress the deepest layer that complexity constraints will allow on the edge device*
- However:
    - Theory talks about limits; practical codecs might be far from those limits
    - Theory shows what is possible, but not exactly how to get there
    - Ideal for grant proposals 😀

EXAMPLE OF FEATURE COMPRESSIBILITY

Results on YOLOv2 object detector

- Features compressed by BPG (HEVC-Intra)
- Part of VOC2007 dataset for testing
- Images from VOC2007 and VOC2012 for re-training to account for quantization
- Bit savings of up to 60% at equivalent accuracy without re-training
- Bit savings of 70% with re-training

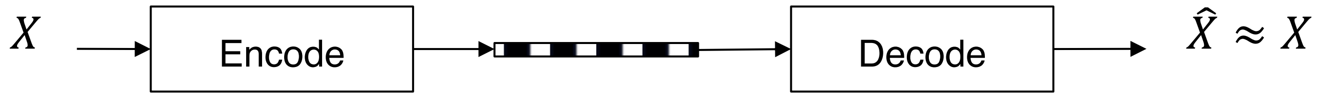
Split at	Default weights	Re-trained weights
max_11	−6.09%	**−45.23%**
max_17	−60.30%	**−70.30%**



H. Choi and I. V. Bajić, "Deep feature compression for collaborative object detection," Proc. IEEE ICIP, Oct. 2018.

GENERALIZED CODEC

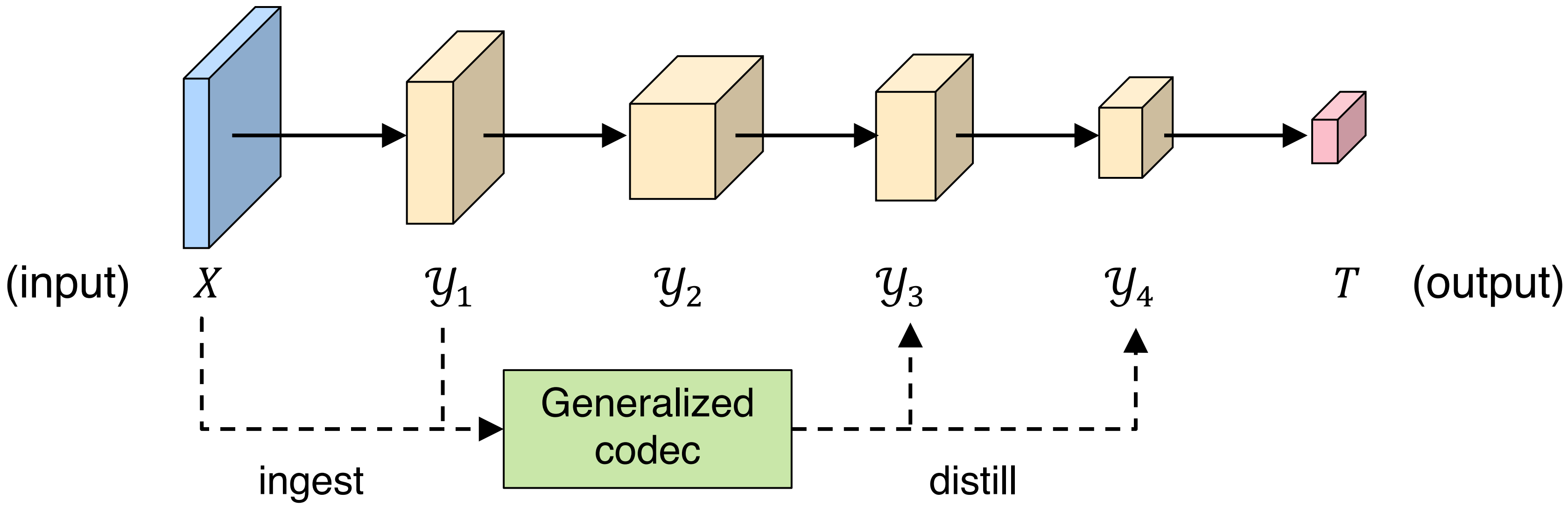
- Until now, we considered conventional coding approaches, which operate as autoencoders



- These could be applied to input coding, or feature coding

- But generalized codecs could be more useful for coding for machines

Generalized codec could ingest input or features, and output ("distill") other features



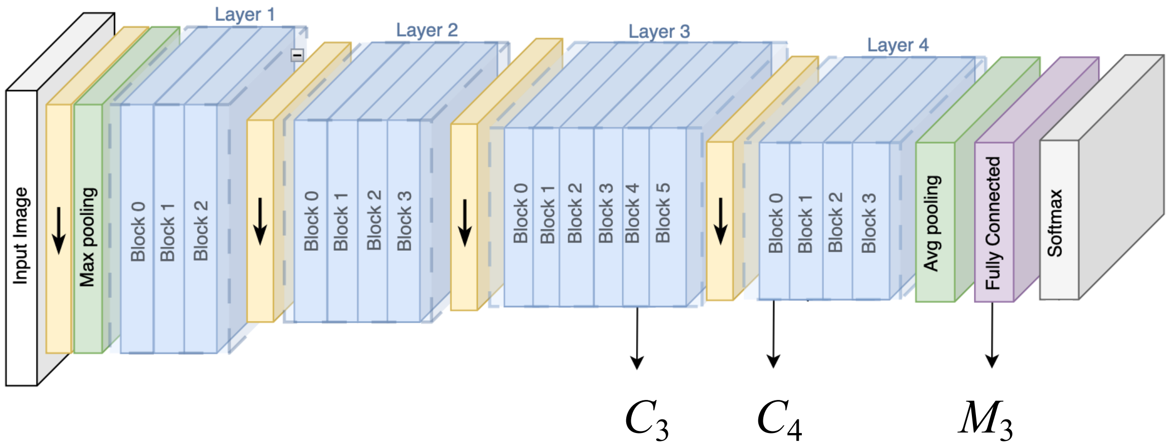
Claims: 1) For a given distillation point, all ingestion points have the same RD bound

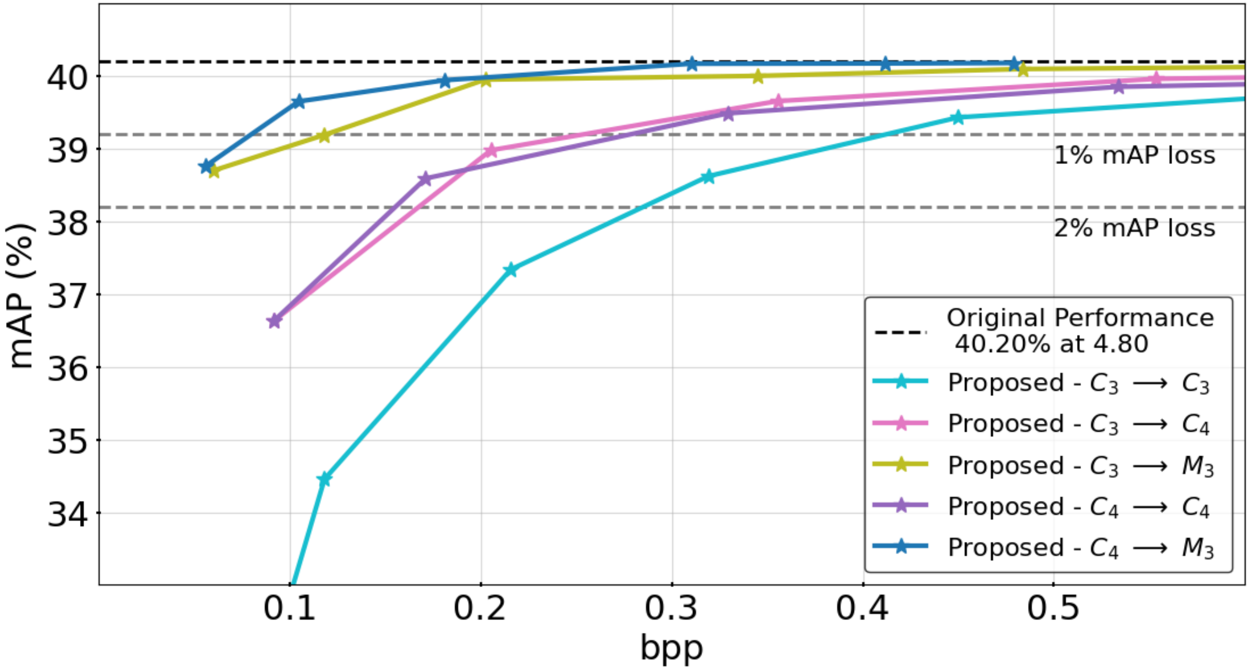
2) For a given ingestion point, deeper distillation points have better RD bounds

A. Harell, A. de Andrade, and I. V. Bajić, "Rate-distortion in image coding for machines," PCS 2022. arXiv:2209.11694
A. Harell et al., "Rate-distortion theory in coding for machines and its applications," arXiv:2305.17295

DISTILLATION USING A GENERALIZED CODEC – SOME RESULTS

Object detection with Faster R-CNN (ResNet50 backbone)





A. Harell et al., "Rate-distortion theory in coding for machines and its applications," arXiv:2305.17295

WHY CODECS FOR MACHINES (CAN) DO BETTER?

- Conventional codecs remove statistical redundancy (which may manifest itself in space, time, frequency, latent space, etc.)
    - They operate as autoencoders
- Codecs for machines, in addition to removing statistical redundancy, also remove **irrelevant information**
    - They can operate as generalized codecs

- Example: for the purpose of monitoring traffic, the tree is irrelevant (but not redundant)

SOME EXAMPLES OF CODING FOR MACHINES

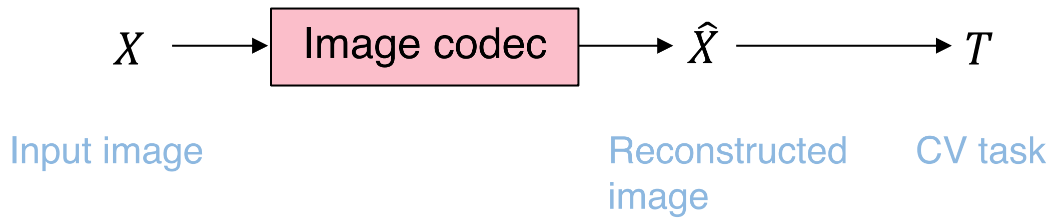
Reference	Computer vision task(s)
H. Choi and I. V. Bajić, "High efficiency compression for object detection," Proc. IEEE ICASSP, 2018, pp. 1729-1796.	Object detection
H. Choi and I. V. Bajić, "Near-lossless deep feature compression for collaborative intelligence," Proc. IEEE MMSP, 2018.	Object detection
H. Choi and I. V. Bajić, "Deep feature compression for collaborative object detection," Proc. IEEE ICIP, 2018, pp. 3743-3747.	Object detection
N. Patwa, N. Ahuja, S. Somayazulu, O. Tickoo, S. Varadarajan and S. Koolagudi, "Semantic-Preserving Image Compression," Proc. IEEE ICIP, 2020, pp. 1281-1285.	Image classification
Y. Matsubara, R. Yang, M. Levorato and S. Mandt, "Supervised Compression for Resource-Constrained Edge Computing Systems," Proc. IEEE/CVF WACV, 2022, pp. 923-933.	Image classification Object detection Object segmentation
Z. Yuan, S. Rawlekar, S. Garg, E. Erkip and Y. Wang, "Feature Compression for Rate Constrained Object Detection on the Edge," Proc. IEEE MIPR, 2022.	Object detection
Z. Duan and F. Zhu, "Efficient Feature Compression for Edge-Cloud Systems," Proc. PCS, 2022, pp. 187-191	Image classification
Z. Zhang and Y. Liu, "Side Information Driven Image Coding for Machines," Proc. PCS, 2022, pp. 193-197	Image classification
K. Fischer, F. Brand and A. Kaup, "Boosting Neural Image Compression for Machines Using Latent Space Masking," IEEE Trans Circuits Syst. Video Technol., 2022, Early Access.	Semantic segmentation

Part 2

Coding for humans and machines

CODING FOR HUMANS AND MACHINES

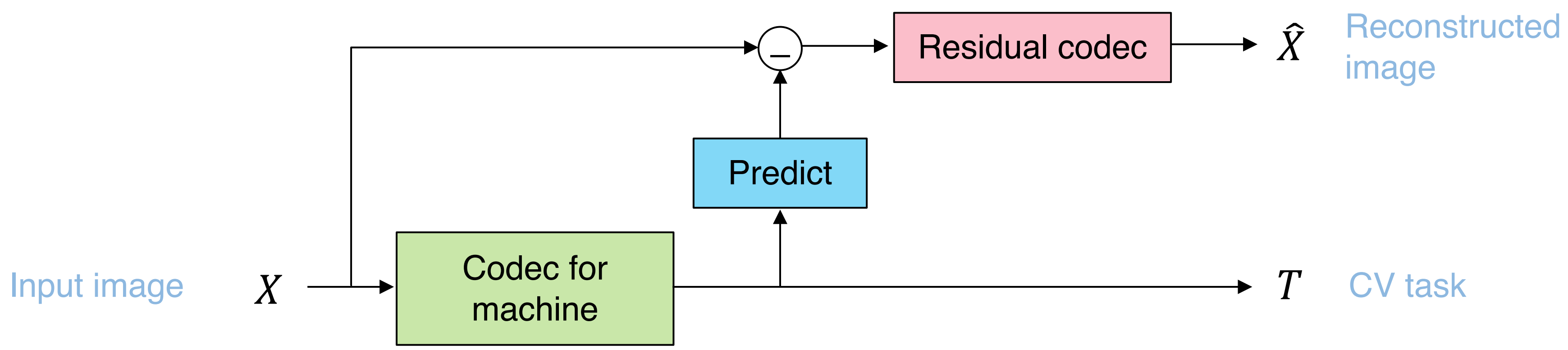
- To support human viewing, input image reconstruction ($\hat{X}$) is needed in addition to computer vision (CV) task(s) T
- A possible solution: reconstruct $\hat{X}$ first, then feed it to a CV model



- Challenges:
  - $\hat{X}$ has to be good for both human viewing and subsequent CV analysis task
  - Bitrate dominated by input reconstruction, which is higher than bitrate for CV analysis; if human viewing is needed only occasionally, this is wasteful

CODING FOR HUMANS AND MACHINES

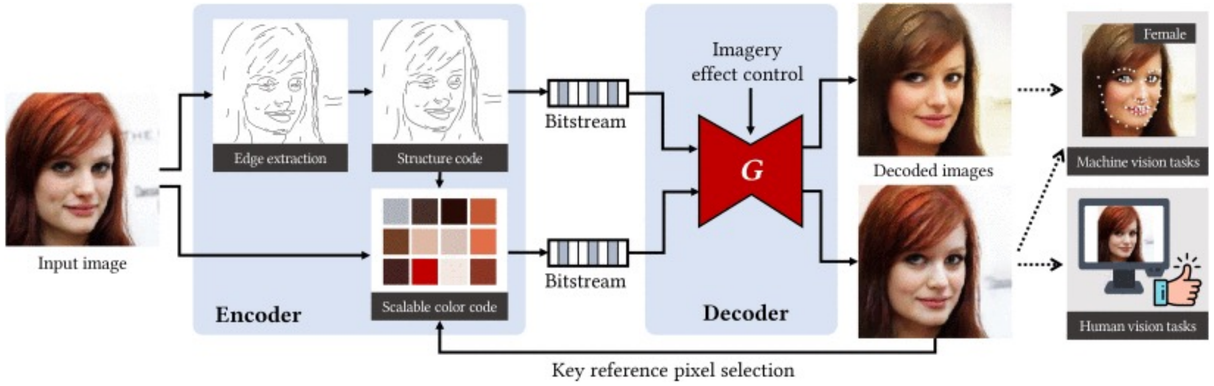
- Better solution: perform CV analysis first, input reconstruction if needed



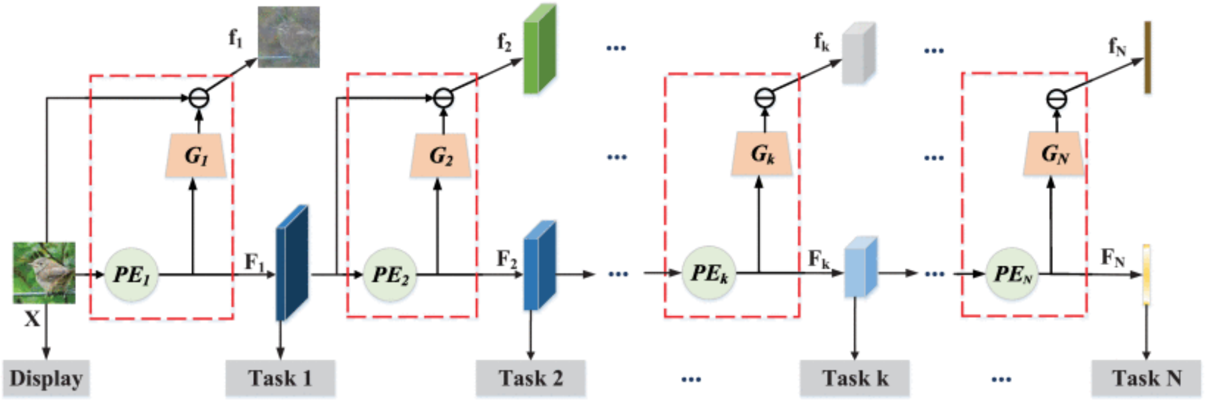
- Advantage: operates at CV task rate when human viewing not needed
- Challenges:
  - Predictor design
  - Residual codec design

EXAMPLES OF SCALABLE HUMAN-MACHINE CODING SYSTEMS

- Scalable face image coding [1]
  - Base: facial landmark keypoints
  - Enhancement: color and texture info
  - Uses generative face decoder

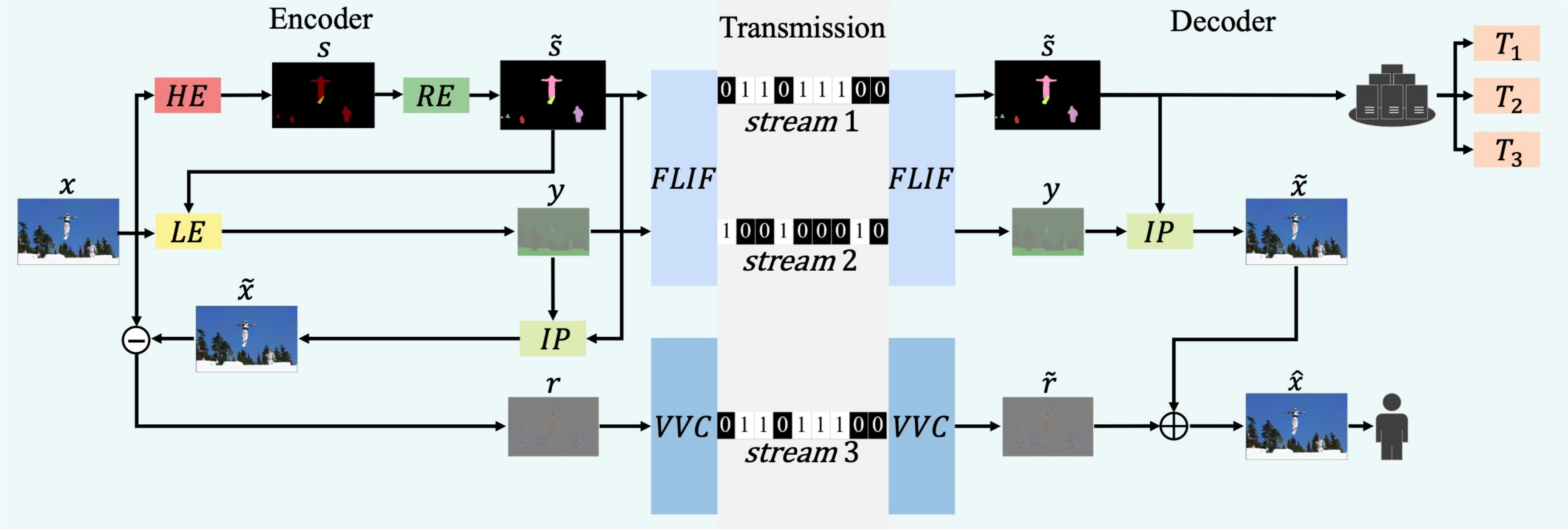


- Semantic-to-signal-scalable coding [2]
  - Base: deepest feature
  - Enhancements: information lost when going layer to layer



[1] S. Yang, Y. Hu, W. Yang, L. -Y. Duan and J. Liu, "Towards coding for human and machine vision: Scalable face image coding," IEEE Trans. Multimedia, vol. 23, pp. 2957-2971, 2021.

[2] N. Yan, C. Gao, D. Liu, H. Li, L. Li and F. Wu, "SSSIC: Semantics-to-signal scalable image coding with learned structural representations," IEEE Trans. Image Processing, vol. 30, pp. 8939-8954, 2021.

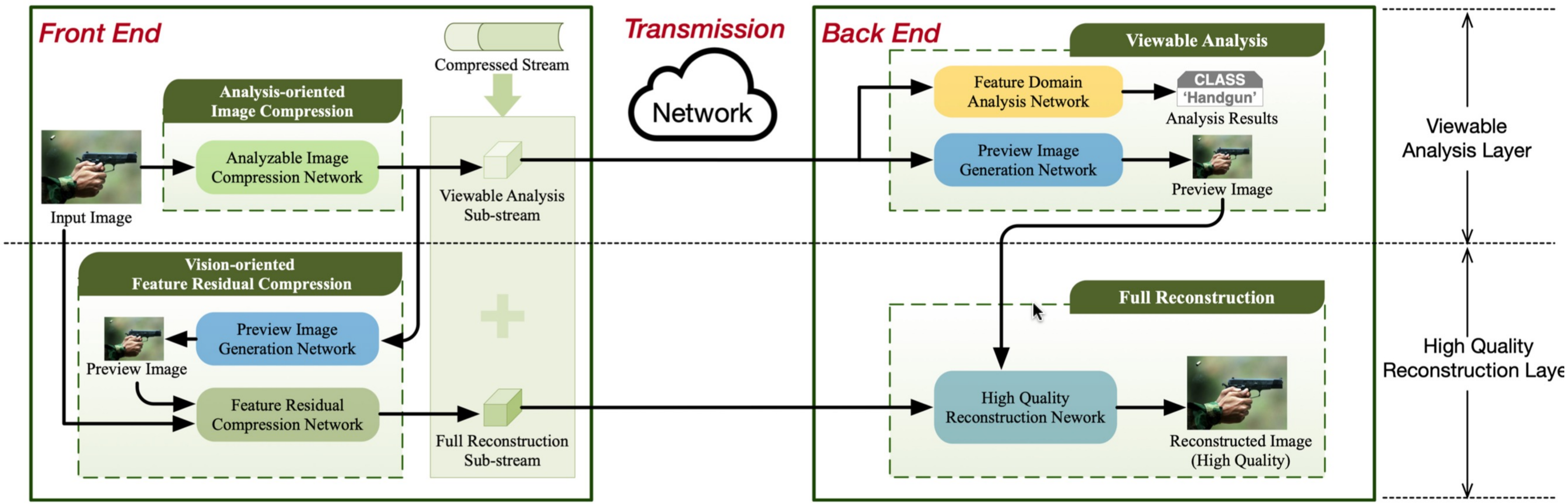


⊖ Element-wise subtraction ⊕ Element-wise addition

- Scalable human-machine coding using conventional encoders
  - Base: segmentation information
  - First enhancement: preview
  - Second enhancement: reconstruction residual

S. Chen, J. Jin, L. Meng, W. Lin, Z. Chen, T.-S. Chang, Z. Li, H. Zhang, "A new image codec paradigm for human and machine uses," arXiv preprint arXiv:2112.10071, Dec. 2021.

EXAMPLES OF SCALABLE HUMAN-MACHINE CODING SYSTEMS

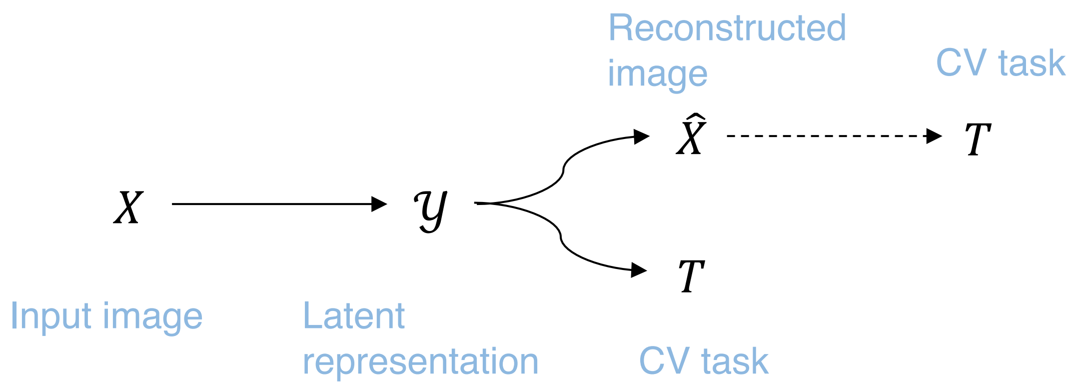


- Human-machine coding for IoT
  - Base: classification + preview
  - Enhancement: reconstruction residual

Z. Wang, F. Li, J. Xu and P. C. Cosman, "Human–machine interaction-oriented image coding for resource-constrained visual monitoring in IoT," IEEE Internet of Things Journal, vol. 9, no. 17, pp. 16181-16195, 1 Sept. 2022.

LATENT SPACE SCALABILITY FOR HUMAN-MACHINE CODING

- Structured latent space to support input reconstruction ($\hat{X}$) and CV tasks (T) ***efficiently***



- CV analysis can also be obtained from $\hat{X}$
- Data processing inequality (DPI) applied to $\mathcal{Y} \rightarrow \hat{X} \rightarrow T$:

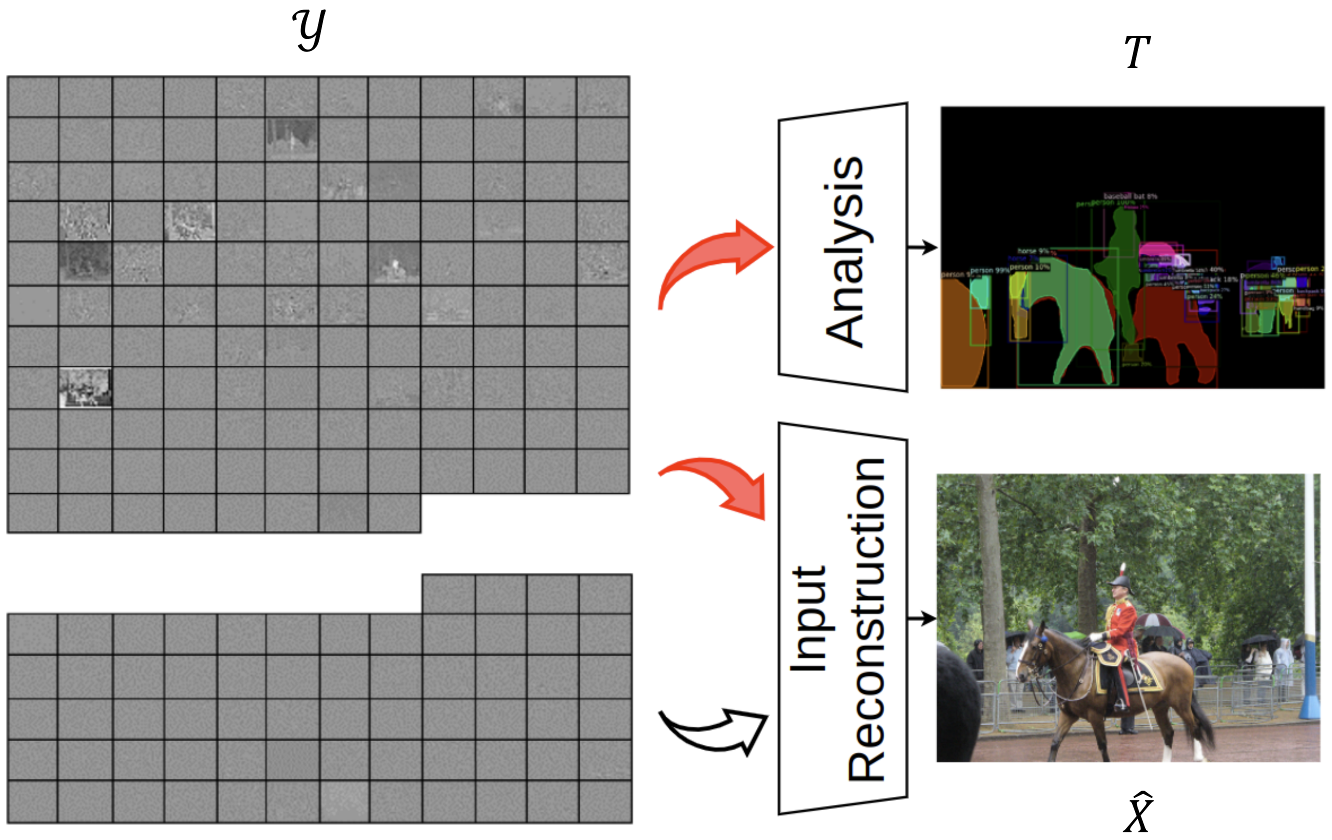
$$I(\mathcal{Y};\hat{X}) \geq I(\mathcal{Y};T)$$

H. Choi and I. V. Bajić, “Latent-space scalability for multi-task collaborative intelligence,” Proc. IEEE ICIP, pp. 3562-3566, Sep. 2021.
H. Choi and I. V. Bajić, “Scalable image coding for humans and machines,” IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.

LATENT SPACE SCALABILITY FOR HUMAN-MACHINE CODING

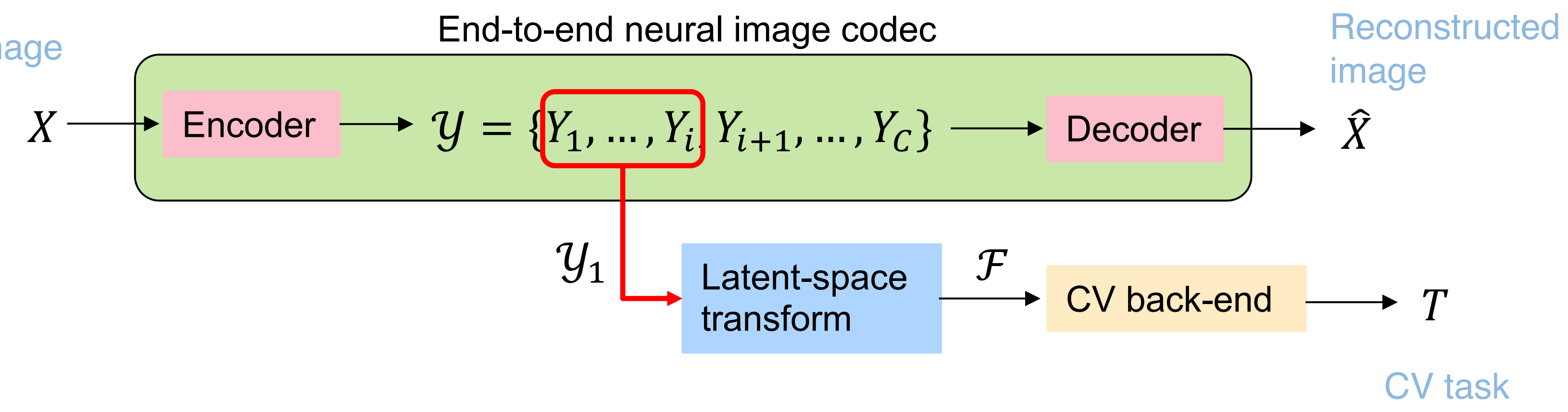
$$I(\mathcal{Y};\hat{X}) \geq I(\mathcal{Y};T)$$

- Latent space $\mathcal{Y}$ contains less information about CV task T than about input reconstruction $\hat{X}$
- Dedicate a subset of $\mathcal{Y}$ to T, all of it to $\hat{X}$
- When only T is needed, decode only a subset of $\mathcal{Y}$



H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.

LATENT SPACE SCALABILITY FOR HUMAN-MACHINE CODING

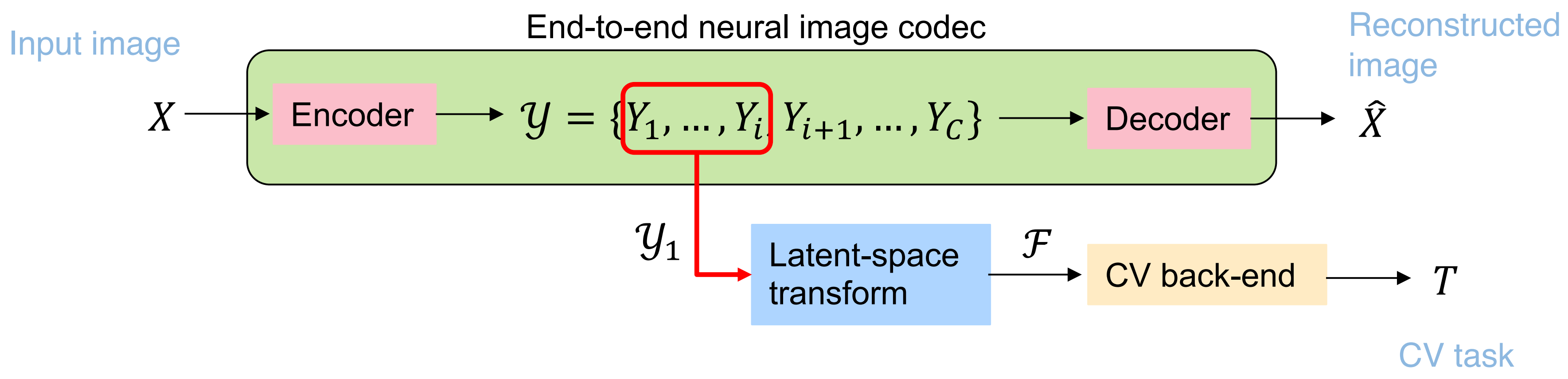


Example 2-layer scalable system:

- End-to-end image codec backbone [2]
- Subset of latent space ($\mathcal{Y}_1$) needs to be transformed into the latent space $\mathcal{F}$ of the CV back-end
  - Need latent-space transform (another neural network)
- CV back-end (for object detection) is YOLOv3 [3] starting at layer 13

[1] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.
[2] Z. Cheng et al., "Learned image compression with discretized gaussian mixture likelihoods and attention modules," Proc. IEEE CVPR, 2020.
[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, Apr. 2018.

LATENT SPACE SCALABILITY



- Loss function:

$$\mathcal{L} = R + \lambda \cdot \underbrace{\left[\mathrm{MSE}(X, \hat{X}) + \gamma \cdot \mathrm{MSE}(\mathcal{F}, \hat{\mathcal{F}})\right]}_{D}$$

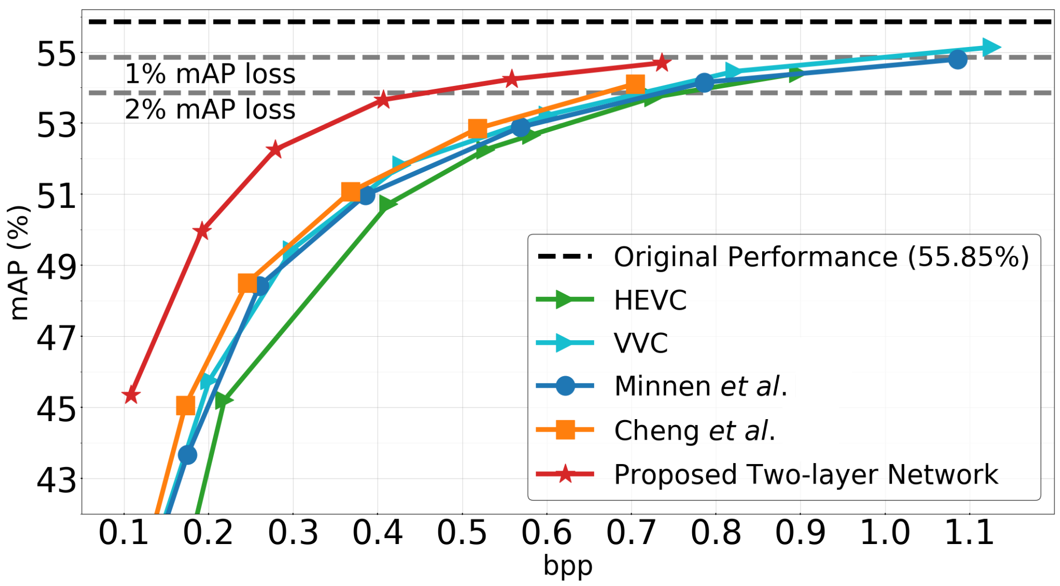
- R is the rate estimate [2]
- Distortion D composed of input reconstruction $\mathrm{MSE}(X, \hat{X})$ and CV feature reconstruction $\mathrm{MSE}(\mathcal{F}, \hat{\mathcal{F}})$
- Since $\mathrm{MSE}(\mathcal{F}, \hat{\mathcal{F}})$ depends only on $\mathcal{Y}_1$ (and not on $\mathcal{Y} \backslash \mathcal{Y}_1$), CV-relevant information is steered to $\mathcal{Y}_1$

[1]. H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.
[2]. D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," NeurIPS, 2018.

LATENT SPACE SCALABILITY

- Object detection experiments on the COCO dataset
- Performance much better than compressing input directly:
  - 37 – 48% bit savings compared to state-of-the-art image codecs
  - 2.8 – 4.5% more accurate detection at the same bit rate
  - Reason: not all pixel details are needed for object detection

2-layer system: object detection + input reconstruction



	Two-layer Network	
Benchmarks	BD-Bitrate	BD-mAP
VVC	–39.8	2.79
HEVC	–47.9	4.55
Minnen *et al.*	–41.3	3.26
Cheng *et al.*	–37.4	2.89

[1] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.
[2] Z. Cheng et al., "Learned image compression with discretized gaussian mixture likelihoods and attention modules," Proc. IEEE CVPR, 2020.
[3] D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," NeurIPS, 2018.

Three tasks

$X \longrightarrow \mathcal{Y}$

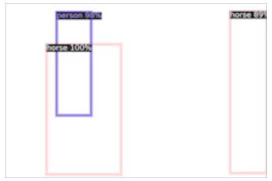
Input image — Latent representation

$\mathcal{Y} \rightarrow \hat{X}$ Reconstructed image

$\mathcal{Y} \rightarrow T_2$ Instance segmentation

$\mathcal{Y} \rightarrow T_1$ Object detection



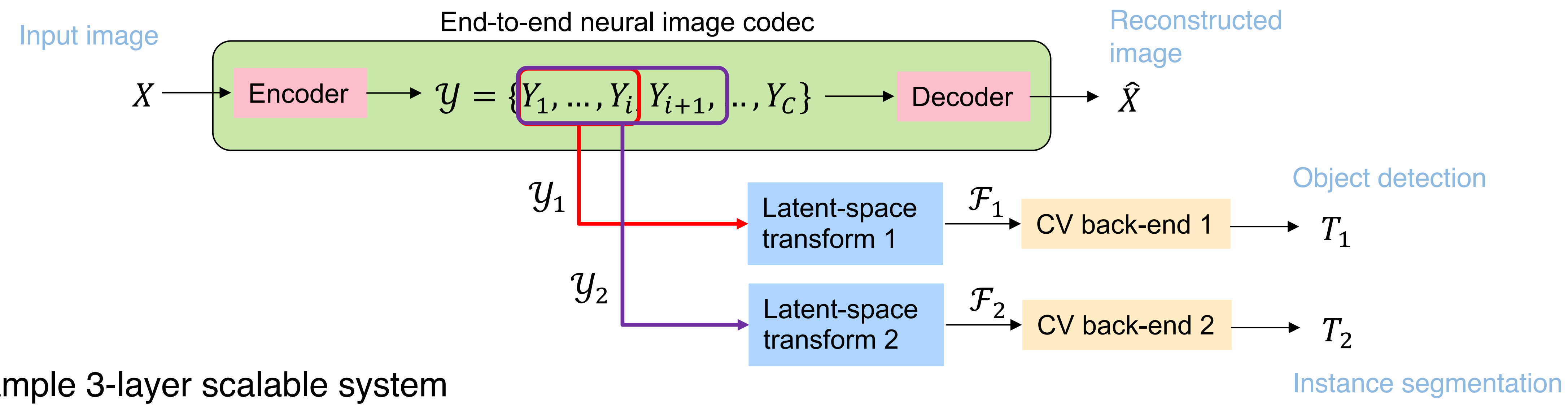



$$X \longrightarrow \mathcal{Y} \longrightarrow \hat{X} \longrightarrow T_2 \longrightarrow T_1$$

$$I(\mathcal{Y};\hat{X}) \geq I(\mathcal{Y};T_2) \geq I(\mathcal{Y};T_1)$$

H. Choi and I. V. Bajić, “Scalable image coding for humans and machines,” IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.

LATENT SPACE SCALABILITY



Example 3-layer scalable system

- End-to-end image codec backbone [2]
- CV task 1: object detection using Detectron [3] Faster RCNN
- CV task 2: instance segmentation using Detectron [3] Mask RCNN
  - Object detection ⊂ semantic segmentation $\Rightarrow$ $\mathcal{Y}_1 \subset \mathcal{Y}_2$

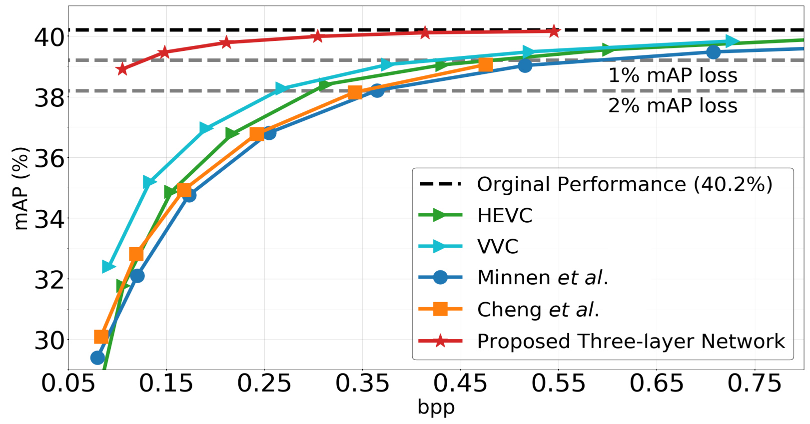
[1] H. Choi and I. V. Bajić, “Scalable image coding for humans and machines,” IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.
[2] Z. Cheng et al., “Learned image compression with discretized gaussian mixture likelihoods and attention modules,” Proc. IEEE CVPR, 2020.
[3] R. Girshick et al., “Detectron,” https://github.com/facebookresearch/detectron, 2018.

LATENT SPACE SCALABILITY

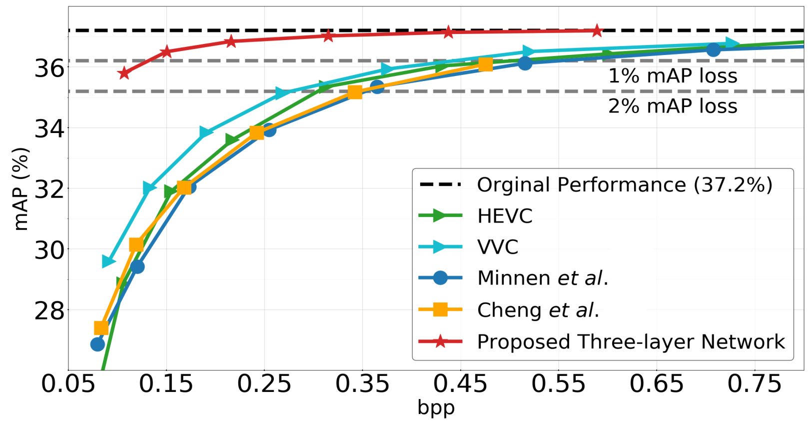
- Detection and segmentation experiments on COCO
- Again, Performance much better than compressing input directly:
  - 71 – 78% bit savings compared to state-of-the-art image codecs
  - 2.3 – 3.5% more accurate detection at the same bit rate

	Three-layer Network			
	Object Detection		Segmentation	
Benchmarks	BD-Bitrate	BD-mAP	BD-Bitrate	BD-mAP
VVC	–73.2	2.33	–71.2	2.34
HEVC	–73.2	3.05	–74.7	2.96
Minnen *et al.*	–78.7	3.73	–77.2	3.38
Cheng *et al.*	–76.6	3.62	–75.4	3.49

3-layer system: (a) object detection, (b) segmentation

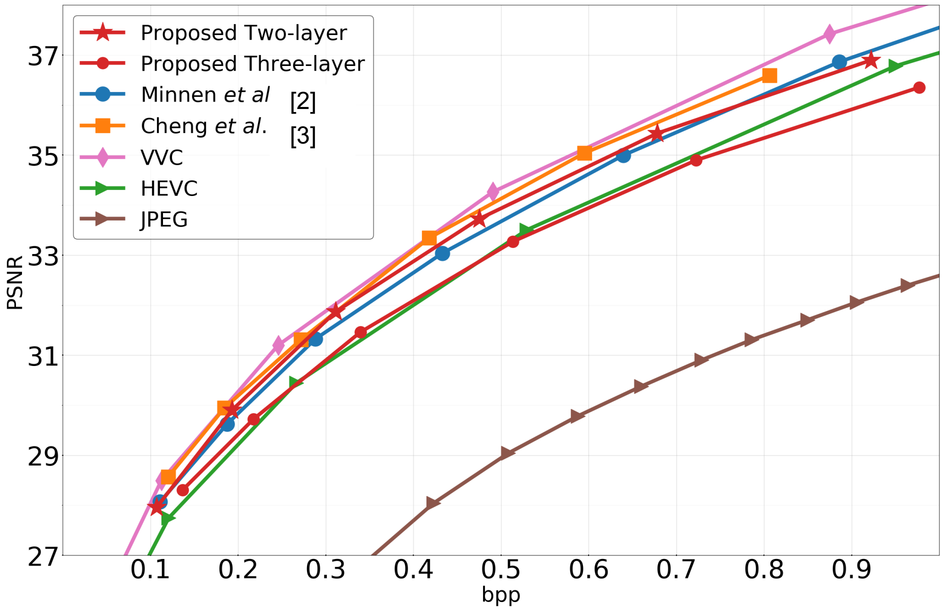


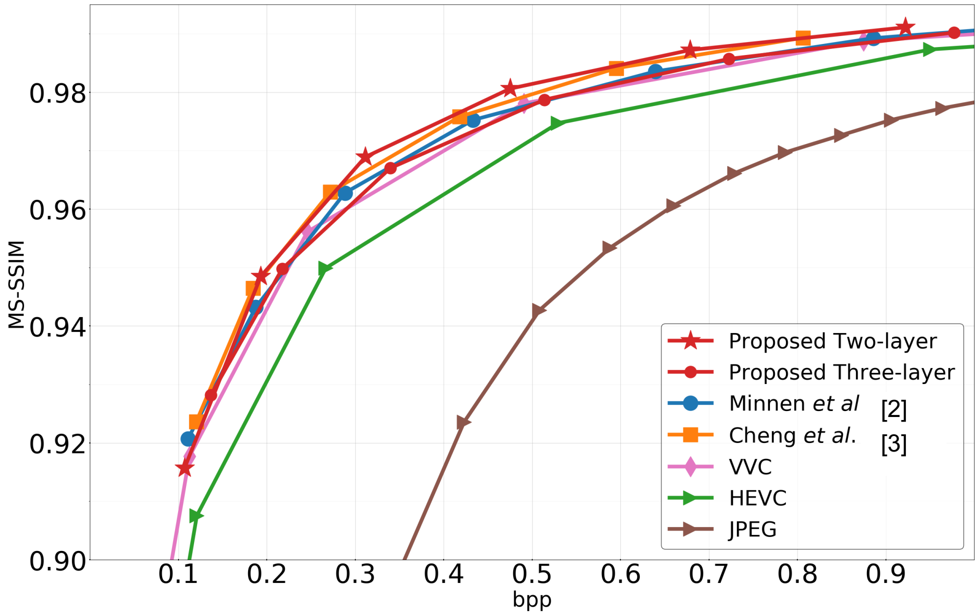
(a)



(b)

[1] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Processing, pp. 2739-2754, Mar. 2022.
[2] Z. Cheng et al., "Learned image compression with discretized gaussian mixture likelihoods and attention modules," Proc. IEEE CVPR, 2020.
[3] D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," NeurIPS, 2018.





Results on the Kodak dataset

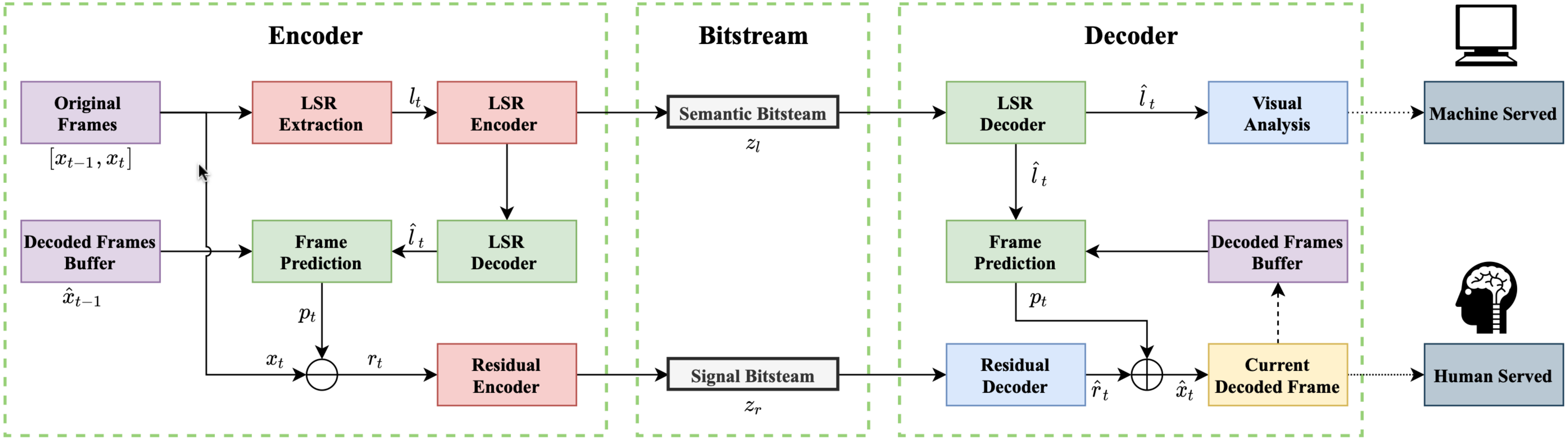
- Proposed scalable codec comparable to state-of-the-art on input reconstruction
- 10 – 20% degradation by adding a scalability layer (2 → 3), in line with earlier work on scalable video coding

Benchmarks	Proposed methods			
	Two-layer Network		Three-layer Network	
	BD-Bitrate (PSNR)	BD-Bitrate (MS-SSIM)	BD-Bitrate (PSNR)	BD-Bitrate (MS-SSIM)
VVC	10.17	–7.83	30.43	2.14
HEVC	–14.27	–26.15	1.38	–17.96
JPEG	–63.99	–63.99	–57.25	–57.84
[2]	–3.58	–7.83	14.02	2.06
[3]	4.49	–1.90	24.24	9.55
Two-layer Network	-		18.84	11.95

[1] H. Choi and I. V. Bajić, “Scalable image coding for humans and machines,” IEEE TIP, 2022.
[2] D. Minnen, J. Balle, and G. D. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” NeurIPS, 2018.
[3] Z. Cheng et al., “Learned image compression with discretized gaussian mixture likelihoods and attention modules,” Proc. IEEE CVPR, 2020.

Summary

- Already a number of studies in the literature describing human-machine image coding
- Base task: computer vision
  - Usually classification, object detection and/or segmentation
- Enhancement task(s): computer or human vision
- CV tasks require fewer bits than input reconstruction
  - Practically demonstrated in many cases
  - Theoretical justification
  - Still a ways to go:
    - ImageNet classification requires $\log_2 1000 \approx 10$ bits ≈ 0.0002 bpp for a 224×224 image; best currently available feature coding systems require > 0.01 bpp to maintain accuracy



HMFVC

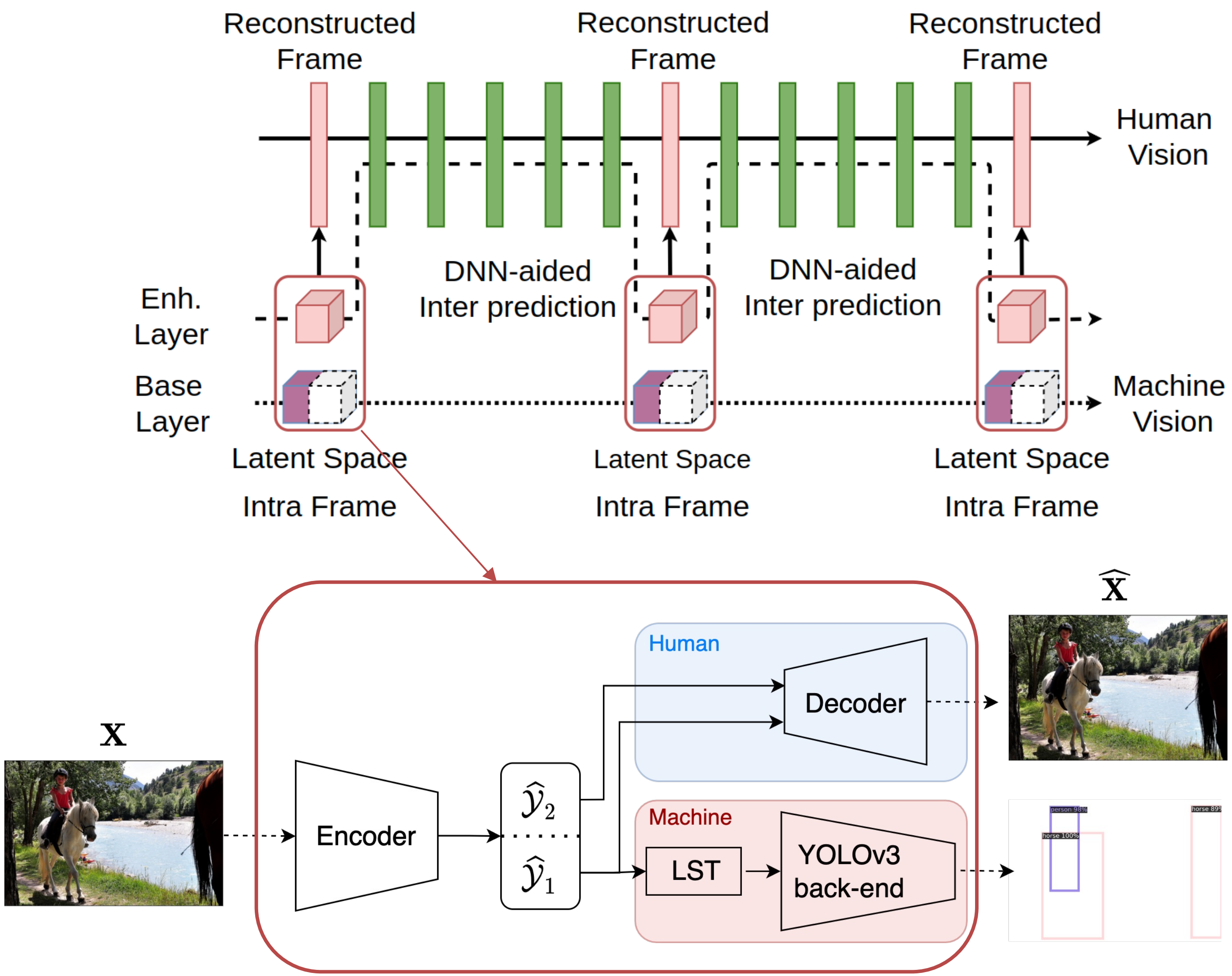
- Base layer: action recognition or object detection
- Enhancement: input reconstruction

Z. Huang, C. Jia, S. Wang, and S. Ma, "HMFVC: A human-machine friendly video compression scheme," IEEE Trans. Circ. Syst. Video Technol., Early Access, 2022.

HUMAN-MACHINE VIDEO CODING

Example of a scalable 2-task video compression system

- Base layer: object detection
- Enhancement layer: input reconstruction
- Intra frames coded using the scalable human-machine image codec presented earlier
- Inter frames coded using DNN-aided HEVC pipeline

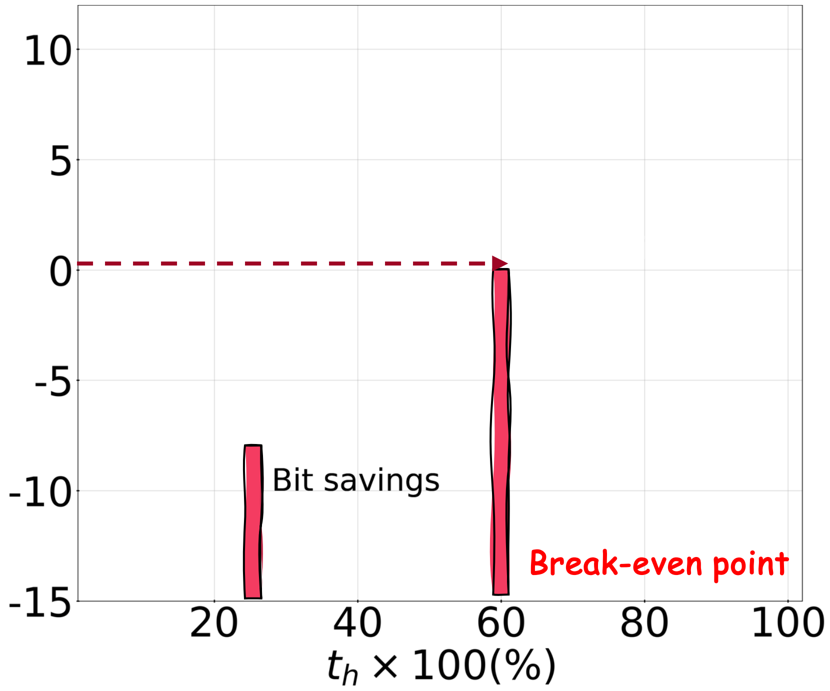


H. Choi and I. V. Bajić, “Scalable video coding for humans and machines,” Proc. IEEE MMSP, 2022.

Break even point

Benchmark		HEVC (HM-16.20)		
		Machine Vision	Human Vision	
		BD-rate-		
Class	Sequence	mAP	PSNR	MS-SSIM
Avg. (A - D)		**-20.40%**	9.62%	**-17.47%**
Avg. (A - E)		**-13.45%**	9.05%	**-18.71%**

$$\underbrace{(1 - t_h)}_{\text{frac. time machine vision}} \cdot 0.8655 + \underbrace{t_h}_{\text{frac. time human vision}} \cdot 1.0905 \leq 1$$



vs. HEVC		vs. VVC	
PSNR	MS-SSIM	PSNR	MS-SSIM
59.8%	100%	31.4%	90.7%

H. Choi and I. V. Bajić, “Scalable video coding for humans and machines,” Proc. IEEE MMSP, 2022.

INFORMATION DIAGRAM

Input image (lossless representation)

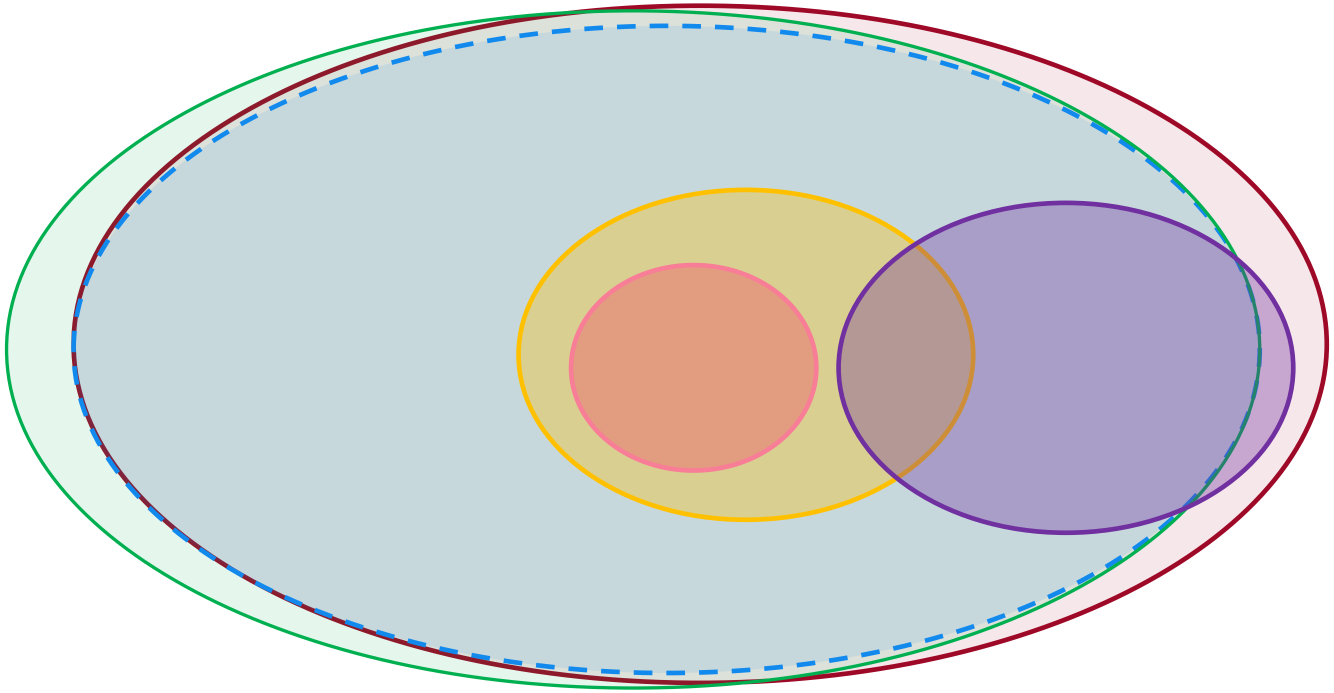
Visually lossless representation

Perceptually-optimized representation

Machine task 1

Machine task 2

Machine task 3



Part 3

Standardization

EXISTING STANDARDS

Compact Descriptors for Visual Search (CDVS) [1]

- For image-related vision tasks, especially search and retrieval
- Handcrafted features: SIFT and Fisher Vectors

Compact Descriptors for Video Analysis (CDVA) [2]

- For video-related vision tasks, especially search and retrieval
- Also considered learnt features

[1] L. -Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, "Overview of the MPEG-CDVS standard," IEEE Trans. Image Processing, vol. 25, no. 1, pp. 179-194, Jan. 2016.
[2] L. -Y. Duan, Y. Lou, Y. Bai, T. Huang, W. Gao, V. Chandrasekhar, J. Lin, S. Wang, and A. C. Kot, "Compact descriptors for video analysis: The emerging MPEG standard," IEEE MultiMedia, vol. 26, no. 2, pp. 44-54, 1 April-June 2019.

JPEG AI

- Scope

  "*The scope of the JPEG AI is the creation of a learning-based image coding standard offering a **single-stream, compact** compressed domain representation, targeting both **human visualization**, with significant compression efficiency improvement over image coding standards in common use at equivalent subjective quality, and effective performance for **image processing and computer vision tasks**, with the goal of supporting a **royalty-free baseline**.*" [JPEG AI White Paper, 2021]

- Difference from earlier image coding standards
  - Learning-based
  - Support for image processing and computer vision tasks (besides default input reconstruction)

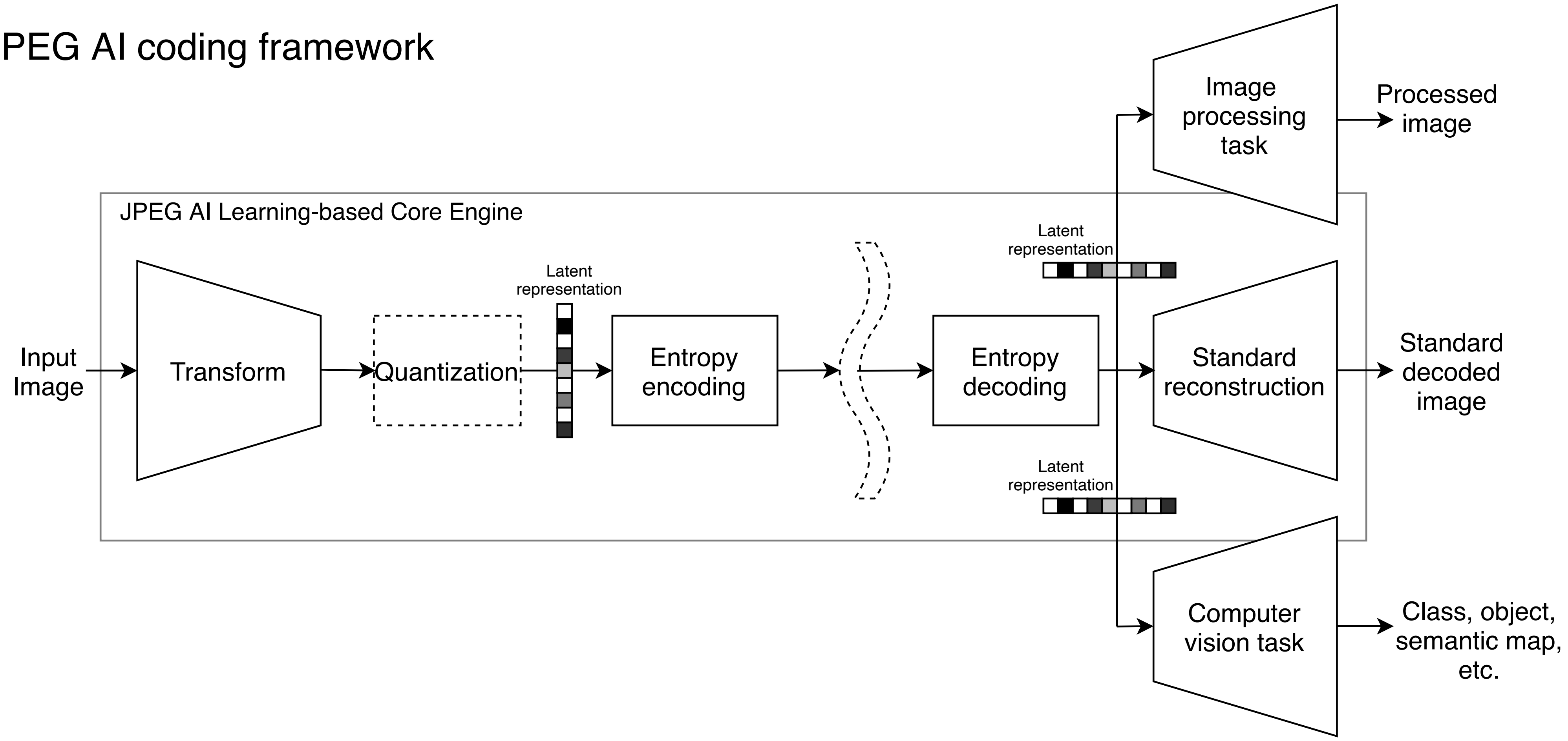
https://jpeg.org/jpegai/
ISO/IEC JTC 1/SC29/WG1 N90049, "White Paper on JPEG AI Scope and Framework v1.0," 2021.

JPEG AI

- Use cases
  - Cloud storage
  - Visual surveillance
  - Autonomous vehicles and devices
  - Image collection storage and management
  - Live monitoring of visual data
  - Media distribution
  - Television broadcast distribution and editing

ISO/IEC JTC 1/SC29/WG1 N92014, REQ "JPEG AI Second Draft Call for Proposals," 92nd Meeting, July 2021.

JPEG AI coding framework



ISO/IEC JTC 1/SC29/WG1 N92014, REQ "JPEG AI Second Draft Call for Proposals," 92nd Meeting, July 2021.

JPEG AI

- Examples of image processing tasks
  - Super-resolution
  - Denoising
  - Low-light enhancement, exposure compensation, color correction
  - Inpainting
- Examples of computer vision tasks
  - Image classification
  - Object/face detection, recognition, identification
  - Semantic segmentation
  - Event detection, action recognition

ISO/IEC JTC 1/SC29/WG1 N92014, REQ "JPEG AI Second Draft Call for Proposals," 92nd Meeting, July 2021.
ISO/IEC JTC 1/SC29/WG1 N100190, REQ " Submission Instructions for the JPEG AI Call for Proposals," 95th Meeting, April 2022.

JPEG AI

CfP results: average BD-rate over several quality metrics

TEAMID	BD-rate performance			CPU dec. time		
	J2K	HEVC	VVC	J2K	HEVC	VVC
TEAM12	**-39.3%**	**-13.2%**	**-3.1%**	601	606	484
TEAM13	**-31.5%**	**-2.1%**	**10.6%**	21	21	16
TEAM14	**-57.2%**	**-39.6%**	**-32.3%**	39	39	31
TEAM15	**-6.7%**	**33.6%**	**51.2%**	25	25	19
TEAM16	**-47.7%**	**-26.6%**	**-17.9%**	44	44	34
TEAM17	**-21.5%**	**15.4%**	**32.0%**	98	98	75
TEAM19	**-34.2%**	**-4.4%**	**8.6%**	21	21	16
TEAM21	**-33.4%**	**1.6%**	**13.8%**	153	153	118
TEAM22	**-32.6%**	**-4.9%**	**7.2%**	136	136	105
TEAM24	**-56.5%**	**-37.4%**	**-29.9%**	44	44	34

J. Ascenso, “JPEG AI Learning-based Image Compression,” Second AG4 Workshop on JPEG and MPEG Emerging Activities, Sept. 2022.

MPEG-VCM

- Scope

  "MPEG-VCM aims to define a bitstream for ***compressing video or feature extracted from video*** *that is efficient in terms of bitrate/size and can be* ***used by a network of machines after decompression*** *to perform multiple tasks without significantly degrading task performance. The decoded video or feature can be used for* ***machine consumption or hybrid machine and human consumption****.*

  The differences between VCM and video coding with deep learning are:

  1. *VCM is used for machine consumption or hybrid machine and human consumption, while current video coding aims for human consumption;*
  2. *VCM technologies could be but is not required to be based on deep learning*
  3. *VCM can achieve analysis efficiency, computational offloading and privacy protection as well as compression efficiency, while traditional video coding pursues mainly on compression efficiency. "* [VCM m57648 , 2021]

Y. Zhang et al., "[VCM] Updates to use cases and requirements for video coding for machines", m57648, July 2021.

MPEG-VCM

- Use cases
  - Surveillance
  - Intelligent transportation
  - Smart city
  - Intelligent industry
  - Intelligent content
  - Consumer electronics
  - Smart retail
  - Smart agriculture
  - Autonomous vehicles / UAV

Y. Zhang et al., "[VCM] Updates to use cases and requirements for video coding for machines", m57648, July 2021.

MPEG-VCM

- Examples of image processing tasks
  - Image/video enhancement
  - Stereo/Multiview processing
- Examples of computer vision tasks
  - Object detection, segmentation, masking, tracking, measurement
  - Event search, detection, prediction
  - Anomaly detection
  - Crowd density estimation
  - Pose estimation and tracking

Y. Zhang et al., "[VCM] Updates to use cases and requirements for video coding for machines", m57648, July 2021.
ISO/IEC JTC 1/SC 29/WG 2, "Evaluation Framework for Video Coding for Machines ," N0193, Apr. 2022.

MPEG-VCM

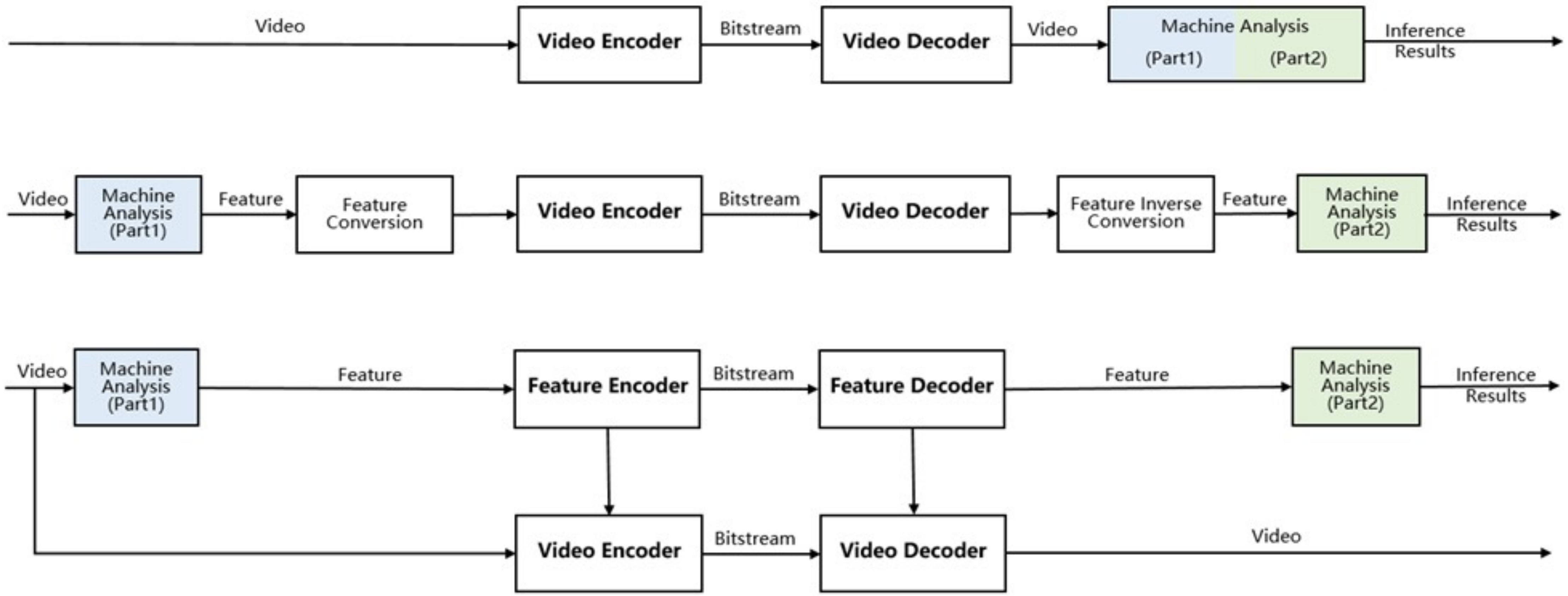
Machine vision tasks and datasets for evaluation

Machine Task	Network Architecture	Evaluation Dataset	Evaluation Metric
Object Detection	Faster R-CNN with ResNeXt-101 backbone	OpenImageV6 TVD FLIR SFU-HW-object-v1	mAP@0.5 mAP@[0.5:0.95]
Instance Segmentation	Mask R-CNN with ResNeXt-101 backbone	OpenImageV6 TVD	mAP@0.5
Object Tracking	JDE-1088x608	TVD ~~HiEve-10*~~	MOTA

S. Liu, "Updates on Video Coding for Machines," Second AG4 Workshop on JPEG and MPEG Emerging Activities, Sept. 2022.

MPEG-VCM

Coding pipelines under consideration



ISO/IEC JTC 1/SC29/WG2 N78, "Evaluation Framework for Video Coding for Machines," April 2021.

SUMMARY

- Coding for machines is an important emerging topic
  - Generalized codecs
  - Theoretical understanding based on classical RD theory + extensions
  - Already shown gains of >70% over the best image/video codecs on several tasks

- Human-machine coding (multi-task coding in general)
  - Requires extension of classical RD theory
  - Most existing work on image coding, less for video coding
  - Related standardization activities: JPEG AI and MPEG-VCM

Thank you!

Questions?