SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

MULTI-TASK IMAGE AND VIDEO COMPRESSION

Ivan V. Bajić

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

IEEE ICIP 2022 Tutorial

ICIP2022 BORDEAUX

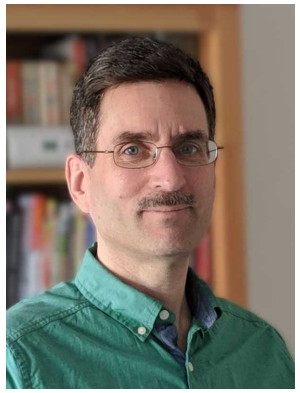
SPECIAL THANKS

People @ SFU Multimedia Lab (multimedia.fas.sfu.ca) whose work contributed to this tutorial – 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 and background

- What is multi-task compression?
- History and applications

Part 1 – Theory

- Review of information theory: mutual information, data processing inequality, RD function
- Bounds on feature compressibility
- Bit allocation in multi-task coding

Part 2 – Current practice

- Multi-task image coding
- Multi-task video coding
- Privacy

Part 3 – Standardization

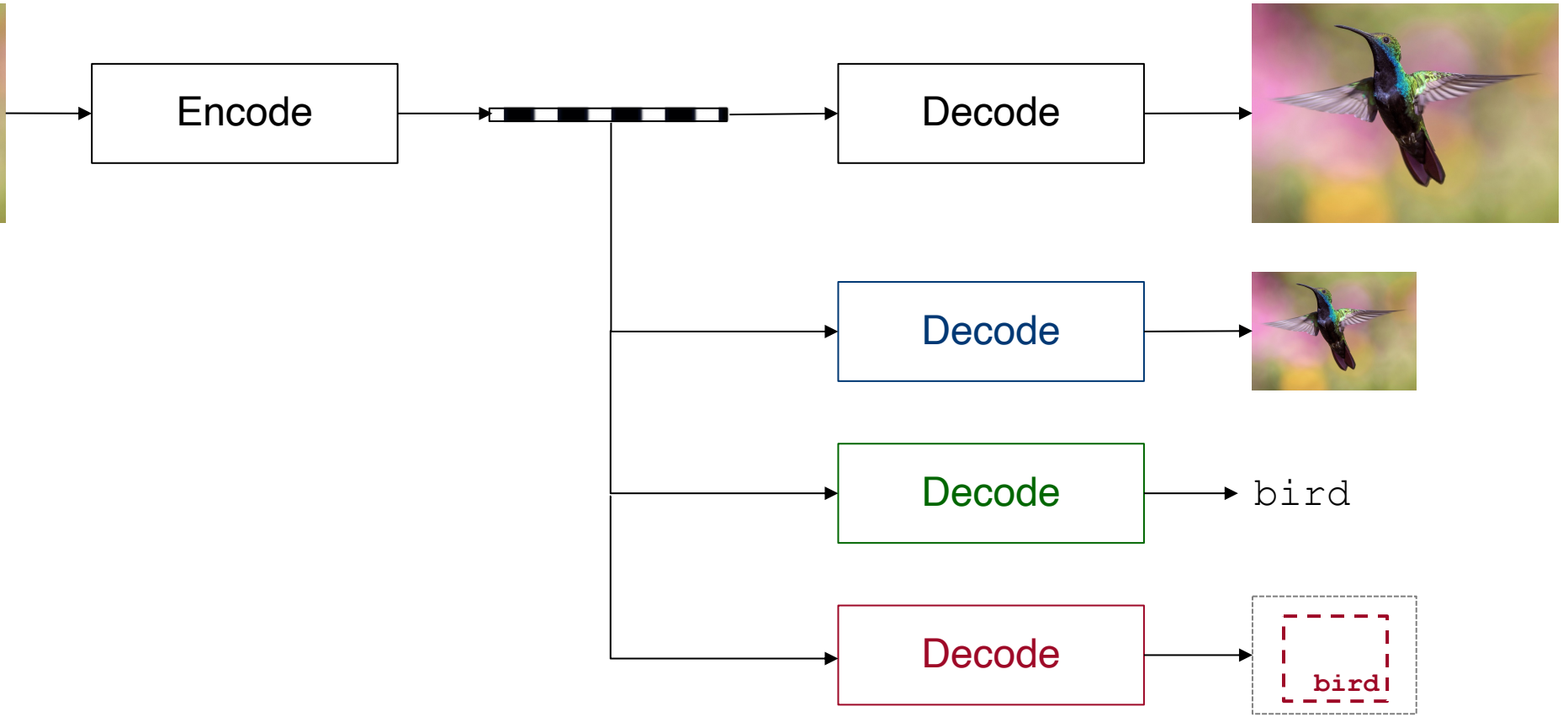
- JPEG AI
- MPEG-VCM (Video Coding for Machines)

Introduction and background

WHAT IS MULTI-TASK COMPRESSION?

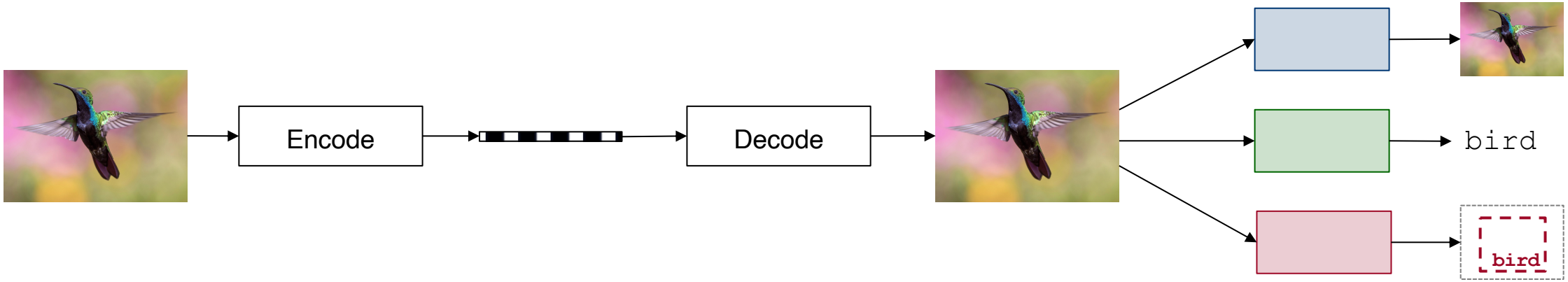
Default task: reconstruct input at the same resolution, bit-depth, etc.

This photo by Unknown Author is licensed under CC BY



WHAT IS MULTI-TASK COMPRESSION?

Why not single-task compression + multi-task post processing / analysis?



Key potential benefits of multi-task compression:

- Reduced complexity: task-specific decoding may be simpler than default task decoding + post-processing / analysis
- Avoiding input reconstruction: reduce memory requirements, improve privacy
- Lower bitrate for most tasks

HISTORY

Multiple research streams related to multi-task compression

- Scalable coding
  - Encode the source image/video to allow multiple decoding options
  - Support different quality levels, resolutions, frame rates, …
- Compressed-domain analysis
  - Start with conventional or scalable bitstream
  - Decode as needed for the task(s) without reconstructing the input

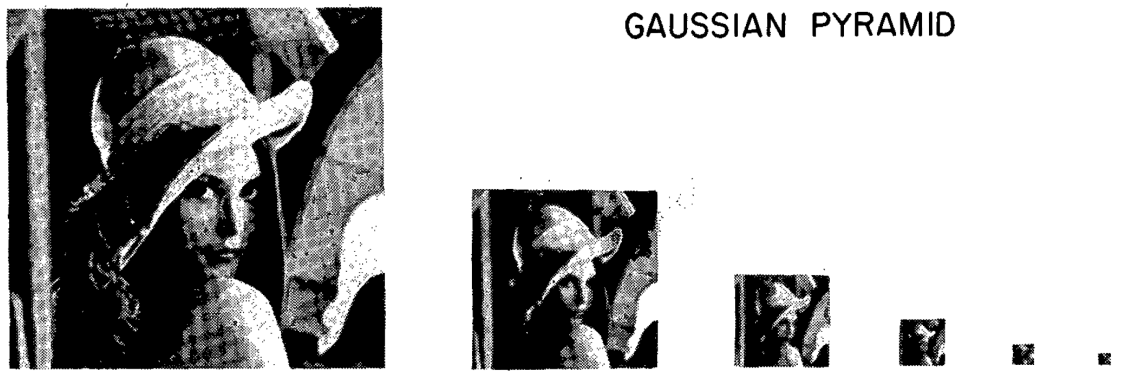
HISTORY: SCALABLE CODING

IEEE TRANSACTIONS ON COMMUNICATIONS, VOL. COM-31, NO. 4, APRIL 1983

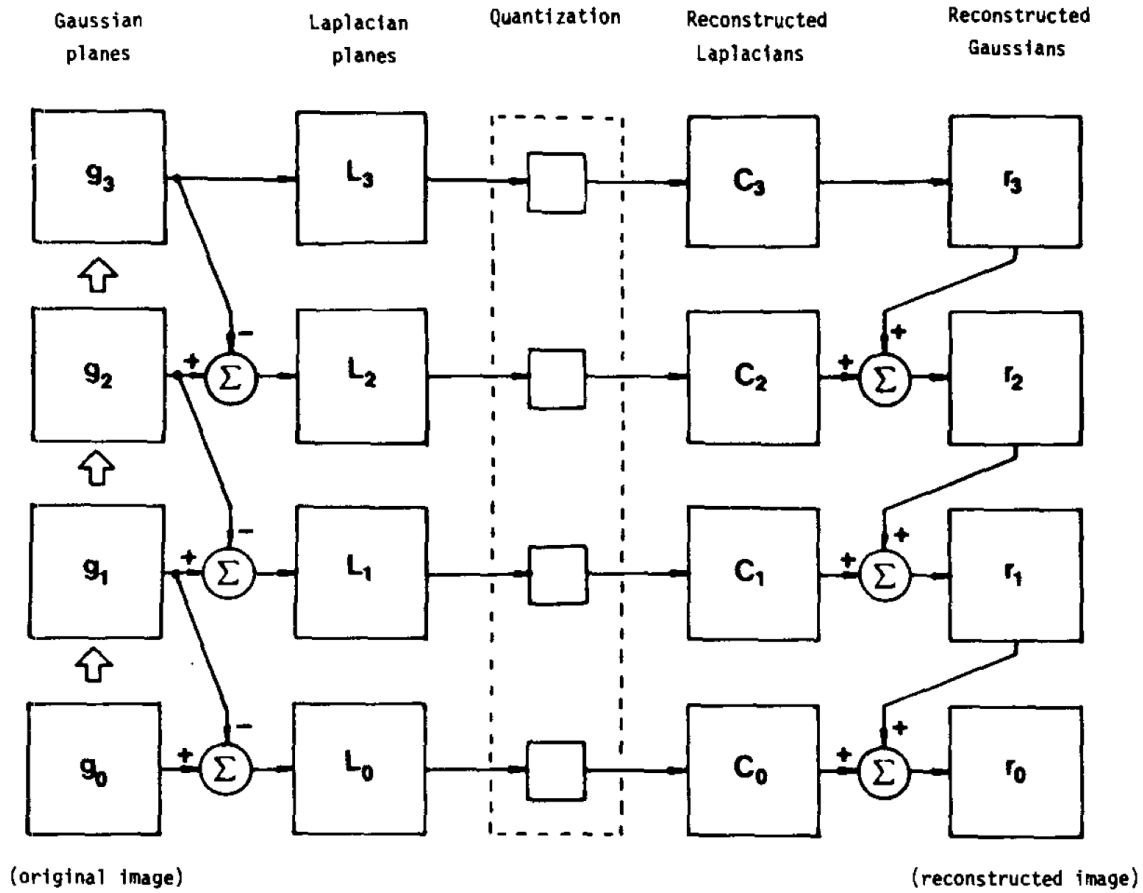
The Laplacian Pyramid as a Compact Image Code

PETER J. BURT, MEMBER, IEEE, AND EDWARD H. ADELSON

GAUSSIAN PYRAMID



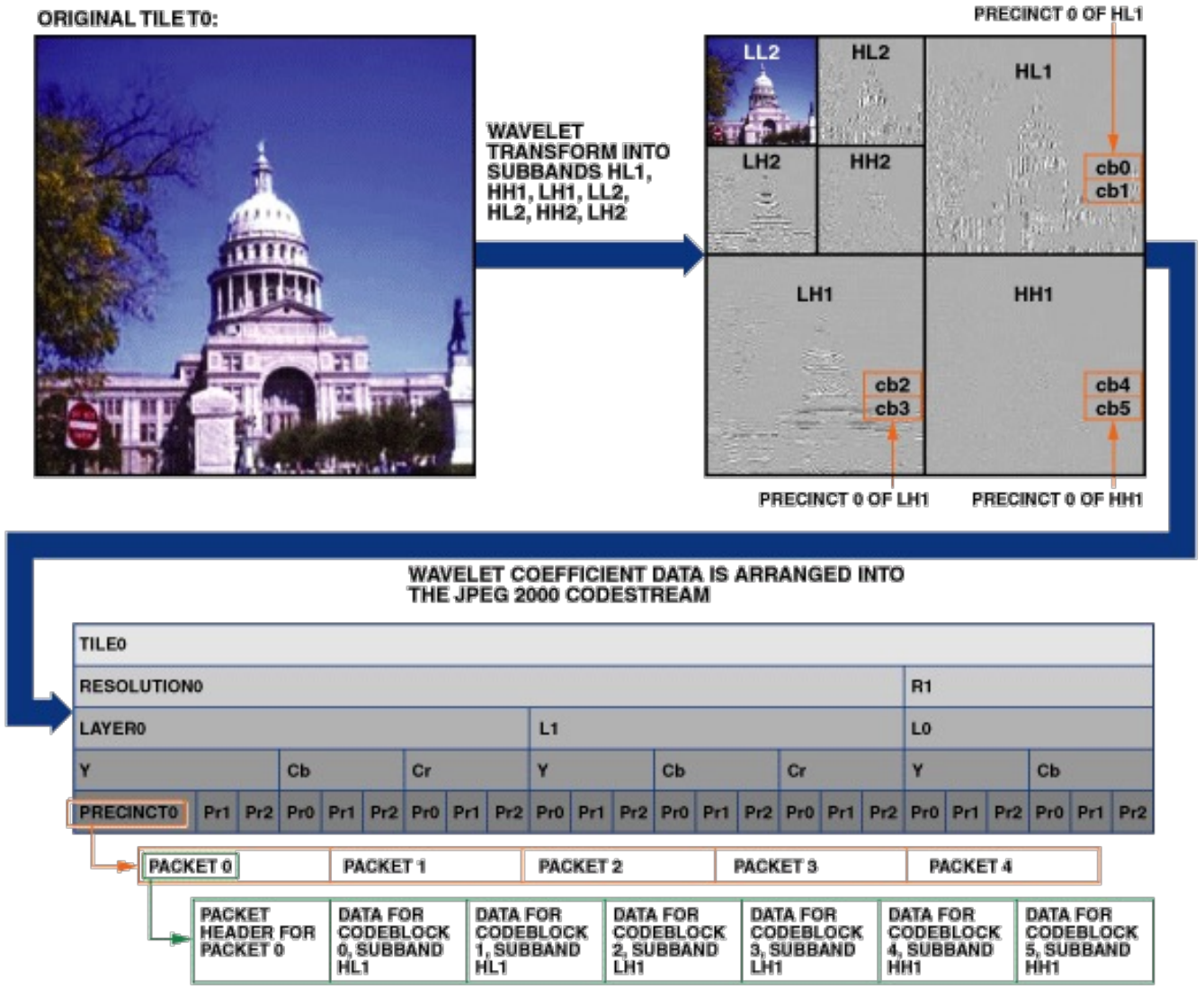
0 1 2 3 4 5



HISTORY: SCALABLE CODING

Example: JPEG 2000

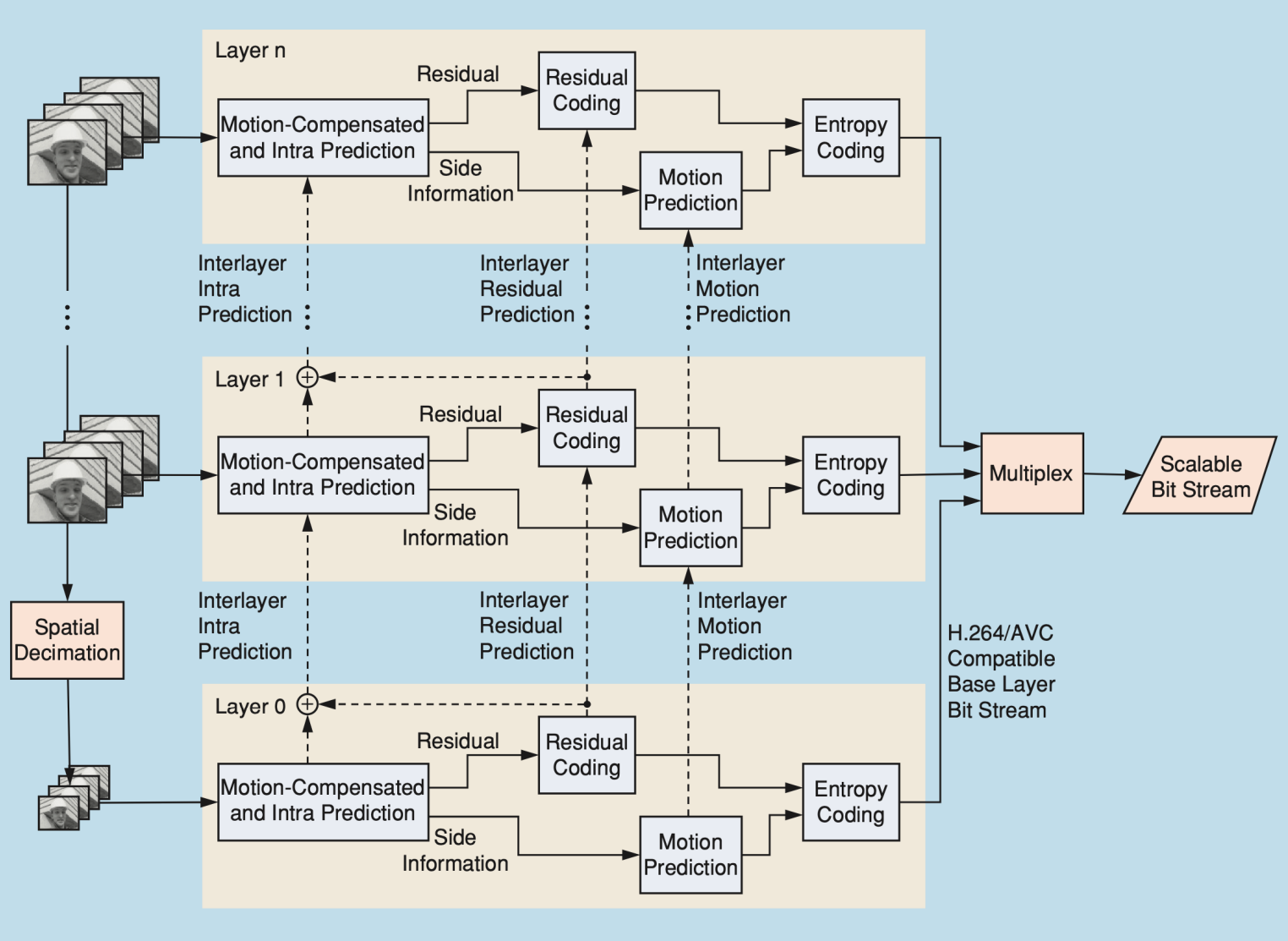
- Subband/wavelet transform
  - More efficient than Laplacian pyramid
  - Supports resolution scalability
- Also supported:
  - Quality scalability
  - Region-of-Interest (RoI) coding



ISO/IEC IS 15444-X and ITU-T T.8XX, JPEG 2000 image coding system

C. Bako, "JPEG 2000 Image Compression," *Analog Dialogue 38-09*, September 2004. https://www.analog.com/en/analog-dialogue/articles/jpeg-2000-image-compression.html

HISTORY: SCALABLE CODING

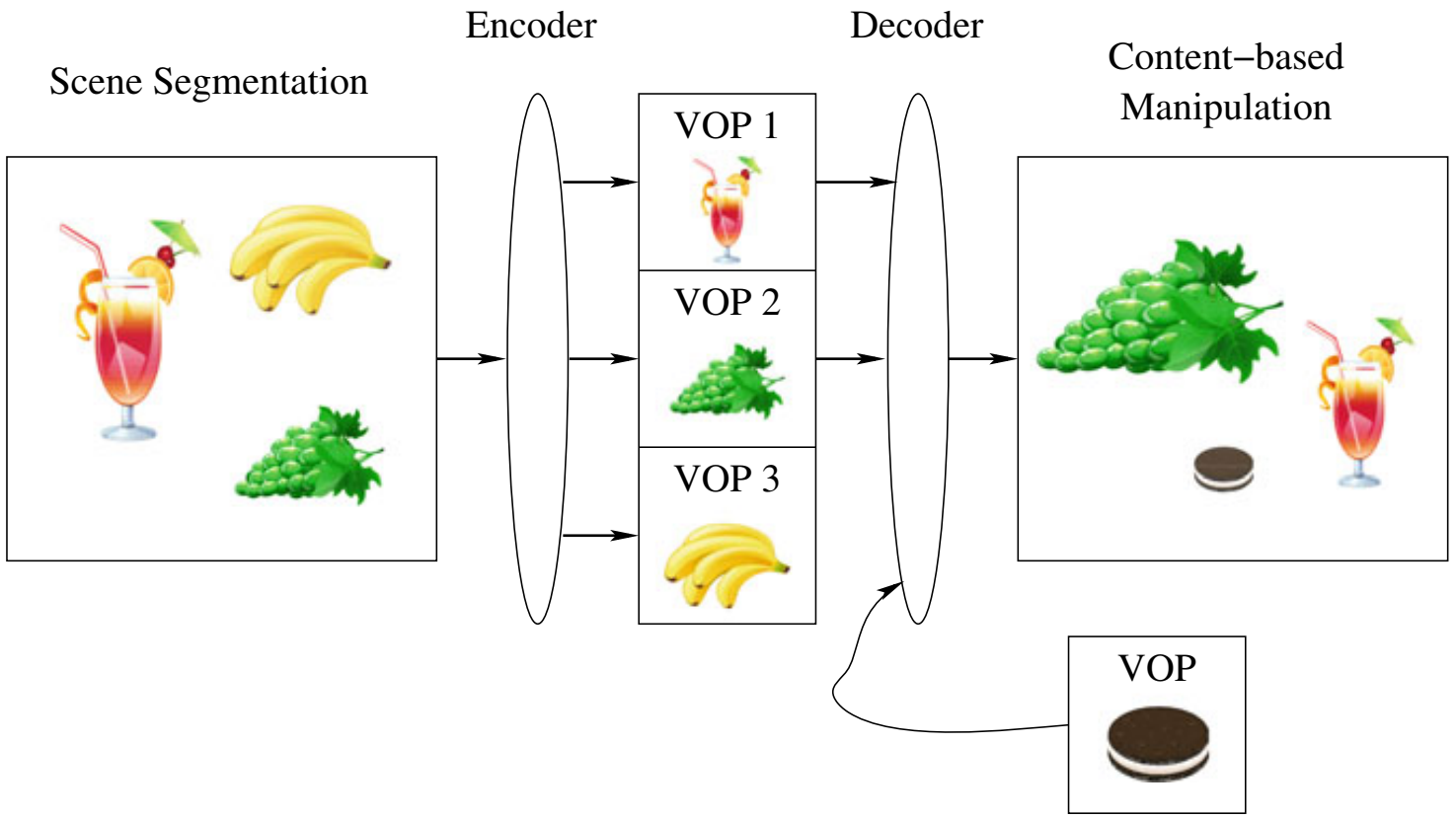


Example: scalable extension of H.264/AVC

- Base layer: lowest resolution / quality / frame rate
- Enhancement layers for higher resolutions / qualities / frame rates
- Only decode parts of the bitstream needed for the particular rendering

H. Schwarz and M. Wien, "The Scalable Video Coding Extension of the H.264/AVC Standard [Standards in a Nutshell]," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 135-141, March 2008.

HISTORY: SCALABLE CODING



Example: object-based coding in MPEG-4

- Objects encoded into VOPs
- Can be combined into a composite scene
- Multiple versions of the scene can be decoded from the same bitstream

ISO/IEC JTC 1/SC29/WG11, ISO/IEC 14496 – Coding of audio-visual objects

Z. N. Li, M. S. Drew, and J. Liu, *Fundamentals of Multimedia*, 3rd Ed., Springer, 2021.

HISTORY: SCALABLE CODING

Summary

- Scalable coding is a form of multi-task coding
- However, tasks considered so far are related to rendering – resolution, frame rate, quality, compositing
- No analysis tasks
  - Object-based coding relies on external analysis to tell it what the objects are

HISTORY: COMPRESSED-DOMAIN ANALYSIS

What is "compressed domain"?

Prediction / Transform → Q → Entropy coding → Entropy decoding → IQ → Inv. Pred. / Transform

Compressed domain?

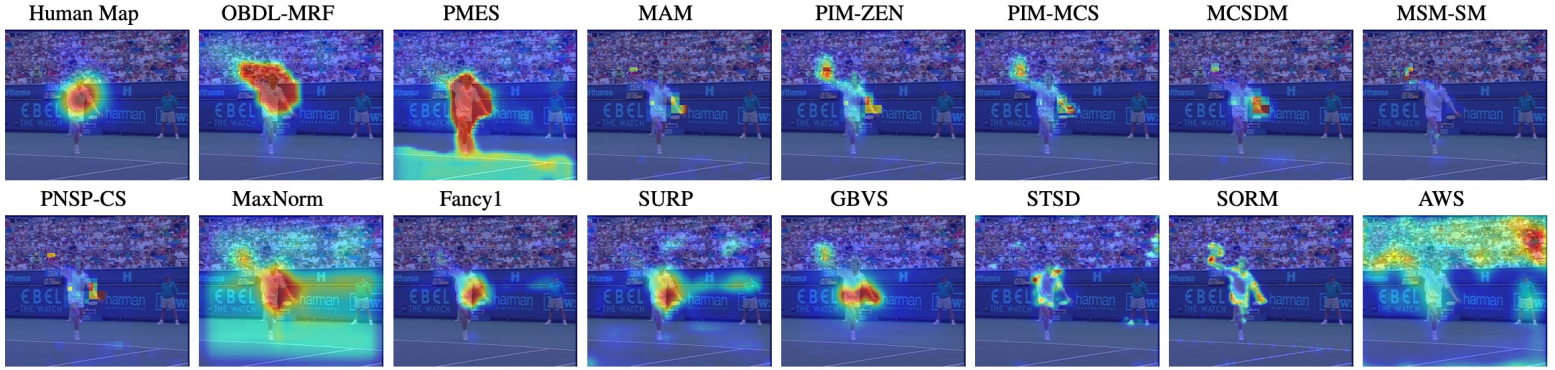
Compressed domain?

Compressed domain

HISTORY: COMPRESSED-DOMAIN ANALYSIS

“True” compressed-domain analysis

- Analyze compressed bitstream without entropy decoding
- Difficult - very few papers on this topic
  - Compressed bitstream looks like iid binary noise
- Possible to do some inference if auxiliary information is available, or if the bitstream has some special structure
  - Example: saliency estimation in H.264/AVC bitstreams

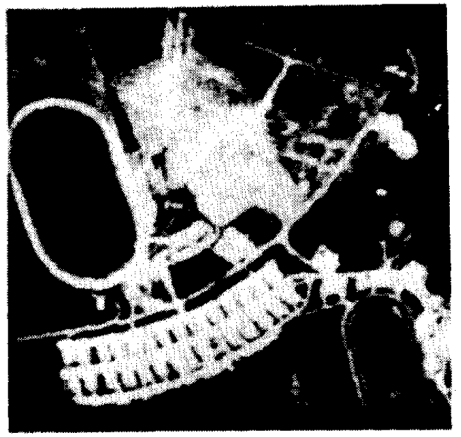


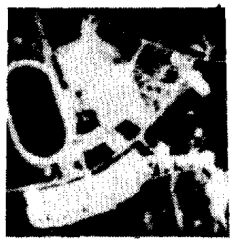
S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajić and Y. Shan, "How many bits does it take for a stimulus to be salient?" CVPR 2015, pp. 5501-5510

HISTORY: COMPRESSED-DOMAIN ANALYSIS

Transform-domain analysis

- Much easier – relationship between pixel and transform domain tractable
- Many papers on this topic, earliest dating back to 1970's!





HIERARCHIAL SEARCH FOR IMAGE MATCHING

E. L. Hall
Dept. of Electrical Engineering
University of Tennessee
Knoxville, Tennessee

R. Y. Wong
Dept. of Computer Science
University of Southern California
Los Angeles, California

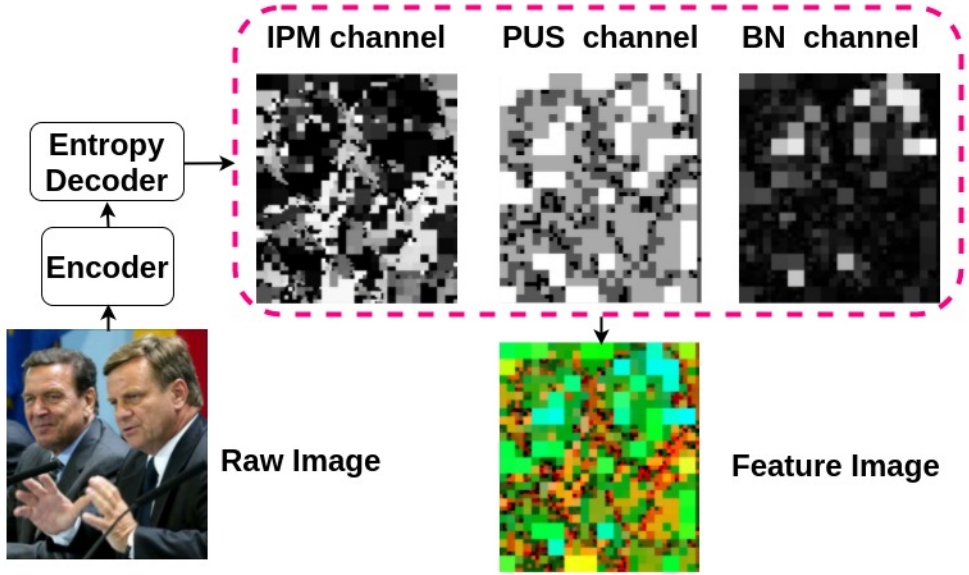
Lt. J. Rouge
U. S. Air Force Space and Missile
Systems Organization
El Segundo, California

E. L. Hall, L. J. Rouge and R. Y. Wong, "Hierarchical search for image matching," Proc. IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes, 1976, pp. 791-796

HISTORY: COMPRESSED-DOMAIN ANALYSIS

Transform-domain image analysis

- Feature extraction (e.g., SIFT)
- Indexing, search and retrieval
- Image classification
- Object detection
- Face detection
- …



S.-F. Chang, "Compressed-domain techniques for image/video indexing and manipulation," Proc. ICIP, 1995, pp. 314-317
D. G. Lowe, “Distinctive image features from scale-Invariant keypoints,” Int. Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, Nov. 2004
K. M. Au, N. F. Law, and W. C. Siu, “Unified feature analysis in JPEG and JPEG 2000-compressed domains,” Patt. Recog., vol. 40, no. 7, pp. 2049–2062, Jul. 2007
S. R. Alvar, H. Choi and I. V. Bajic, "Can you find a face in a HEVC bitstream?," Proc. ICASSP, 2018, pp. 1288-1292

HISTORY: COMPRESSED-DOMAIN ANALYSIS

Transform-domain video analysis

- Global motion estimation
- Object/motion segmentation
- Object tracking
- Action recognition
- Vehicle counting
- …

R. V. Babu, M. Tom, and P. Wadekar, “A survey on compressed domain video analysis techniques,” Multimed. Tools Appl., vol. 75, no. 2, pp. 1043–1078, Jan. 2016.

A. Smolic, M. Hoeynck and J.-R. Ohm, "Low-complexity global motion estimation from P-frame motion vectors for MPEG-7 applications," Proc. ICIP, 2000, pp. 271-274

V. Mezaris, I. Kompatsiaris, N. V. Boulgouris and M. G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," IEEE Trans. Circ. Syst. Video Technol., vol. 14, no. 5, pp. 606-621, May 2004

S. H. Khatoonabadi and I. V. Bajic, "Video object tracking in the compressed domain using spatio-temporal Markov random fields," IEEE Trans. Image Processing, vol. 22, no. 1, pp. 300-313, Jan. 2013.

C. Yeo, P. Ahammad, K. Ramchandran and S. S. Sastry, "High-speed action recognition and localization in compressed domain videos," IEEE Trans. Circuits and Systems for Video Technology, vol. 18, no. 8, pp. 1006-1015, Aug. 2008.

X. Liu, Z. Wang, J. Feng and H. Xi, "Highway vehicle counting in compressed domain," IEEE CVPR, 2016, pp. 3016-3024

HISTORY: COMPRESSED-DOMAIN ANALYSIS

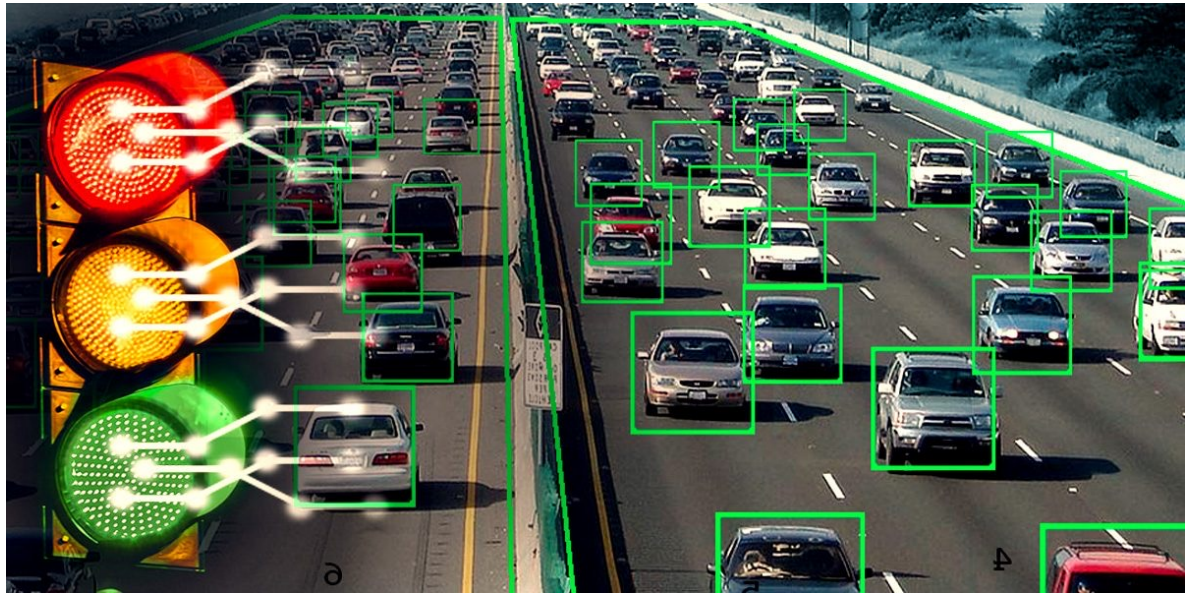
Summary

- Little work on “true” compressed-domain (entropy-coded data) analysis
- A lot of work on transform-domain analysis
  - Traditional computer vision
    - Although those transforms are not necessarily the ones used in conventional codecs
  - “Compressed vision”
    - Using elements / features found in compressed bitstreams: transform coefficients, prediction modes, motion vectors, …

APPLICATIONS

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

APPLICATIONS

Autonomous driving

- Cameras (and other sensors) mounted on the vehicle to help understand and navigate its surroundings
- Detecting vehicles, bikes, pedestrians, traffic lights and signs, speed bumps, etc.
- Lots of data, high energy usage:
  Estimated ~ 2 kWh for on-board processing of sensor data (2.5 kWh in cities) – may want to offload
- Tasks:
  - Object detection and tracking
  - Object motion prediction
  - Human viewing (occasionally)

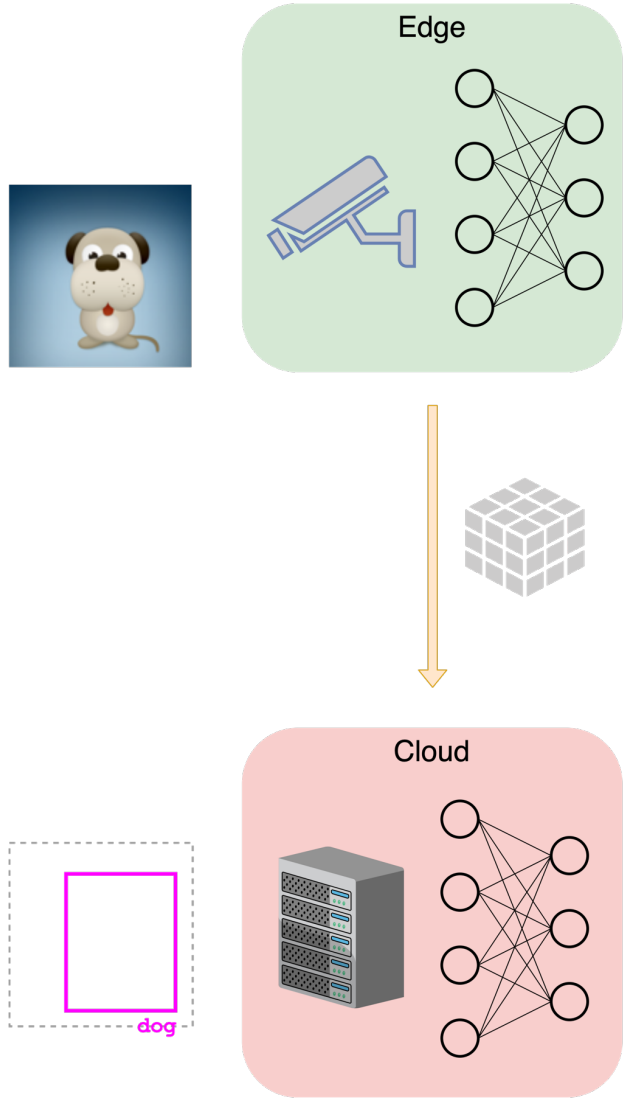
aarp.org

D. Richart, Autonomous Cars' Big Problem: The energy consumption of edge processing reduces a car's mileage with up to 30%, May 2019.
https://medium.com/@teraki/energy-consumption-required-by-edge-computing-reduces-a-autonomous-cars-mileage-with-up-to-30-46b6764ea1b7

APPLICATIONS

(Edge-cloud) collaborative intelligence

- Covers the spectrum 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
- Tasks:
  - Any machine vision task
  - Human viewing



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

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
- MPEG-VCM (Video Coding for Machines) is a related, broader standardization effort

[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.

[3] S. Ma, X. Zhang, S. Wang, X. Zhang, C. Jia and S. Wang, "Joint feature and texture coding: Toward smart video representation via front-end intelligence," IEEE Trans. Circuits and Systems for Video Technology, vol. 29, no. 10, pp. 3095-3105, Oct. 2019.

LOOKING TO THE FUTURE

- State-of-the-art performance on most vision tasks is currently achieved by Deep Neural Networks (DNNs)
- Even on the default task – input reconstruction – DNN-based coding provides state-of-the-art performance for image compression (though not yet for video)

  ⇒ DNNs provide a good unified framework for multi-task compression

J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” ICLR 2018.
Z. Cheng, H. Sun, M. Takeuchi and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," CVPR 2020, pp. 7936-7945
Y. -H. Ho, C. -C. Chan, W. -H. Peng, H. -M. Hang and M. Domański, “ANFIC: Image compression using augmented normalizing flows,” IEEE Open Journal of Circuits and Systems, vol. 2, pp. 613-626, 2021.
B. Li, J. Liang and J. Han, "Variable-rate deep image compression with vision transformers," IEEE Access, vol. 10, pp. 50323-50334, 2022.
Z. Guo, Z. Zhang, R. Feng and Z. Chen, "Causal contextual prediction for learned image compression," IEEE Trans. Circuits and Systems for Video Technology, vol. 32, no. 4, pp. 2329-2341, April 2022.
F. Brand, K. Fischer, A. Kopte, M. Windsheimer and A. Kaup, “RDONet: Rate-distortion optimized learned image compression with variable depth,” CVPRW 2022, pp. 1758-1762
W. Duan, K. Lin, C. Jia, X. Zhang, S. Ma and W. Gao, “End-to-end image compression via attention-guided information-preserving module,” ICME 2022

Questions?

Part 1

Theory

Entropy

- Let X be a discrete random variable taking on values x in some sample space $\mathcal{X}$
- The entropy of X (in bits) is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(X = x) \cdot \log_2 p(X = x)$$

- Entropy is a measure of uncertainty (randomness)
- Entropy is the limit of lossless compressibility
- Examples:
  - Fair coin: $\mathcal{X} = \{\text{Heads}, \text{Tails}\}, \quad p(X = \text{Heads}) = p(X = \text{Tails}) = 1/2, \quad H(X) = 1$ bit
  - Fair die: $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}, \quad p(X = 1) = \cdots = p(X = 6) = 1/6, \quad H(X) = \log_2 6 = 2.58$ bits

T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley, 2006.

Mutual information

- Let X and Y be discrete random variables taking on values in sample spaces $\mathcal{X}$ and $\mathcal{Y}$
- The mutual information (MI) between X and Y (in bits) is defined as

$$I(X;Y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p((X,Y)=(x,y)) \cdot \log_2 \frac{p((X,Y)=(x,y))}{p(X=x)\cdot p(Y=y)}$$

- MI is a measure of statistical dependence (linear or nonlinear) between X and Y
- MI is the amount of information that X carries about Y, and vice versa
- Examples:
  - X and Y independent $\Leftrightarrow$ $I(X;Y) = 0$
  - $I(X;X) = H(X)$: mutual information between X and itself is its own entropy

T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley, 2006.

Markov chain

- A sequence of random variables $X \rightarrow Y \rightarrow Z$ is a Markov chain if Z is conditionally independent of X, given Y

always

$$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|y,x)$$
$$= p(x) \cdot p(y|x) \cdot p(z|y)$$

if Markov chain

- If Z is a function of Y, i.e., $Z = f(Y)$, then $X \rightarrow Y \rightarrow Z$ is a Markov chain
  - Since Z is computed from Y, it does not depend on X (when Y is given)

T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley, 2006.

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)

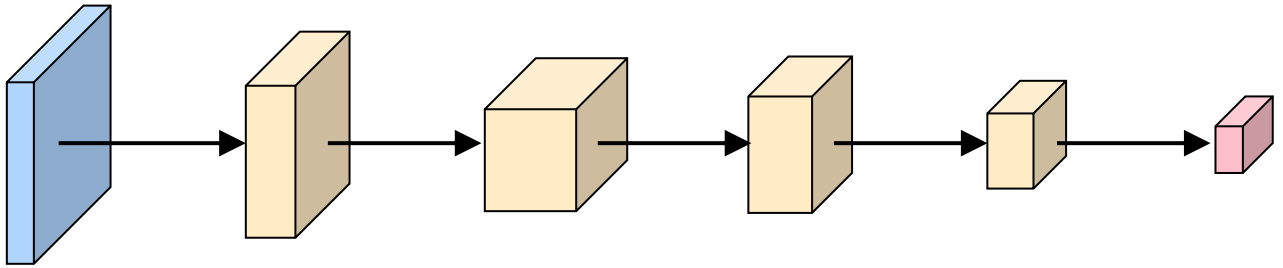
- Extended version of DPI: if $X \rightarrow Y \rightarrow Z \rightarrow W$ is a Markov chain, then

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

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

- Y_i = output of the i-th layer in a feedforward neural network

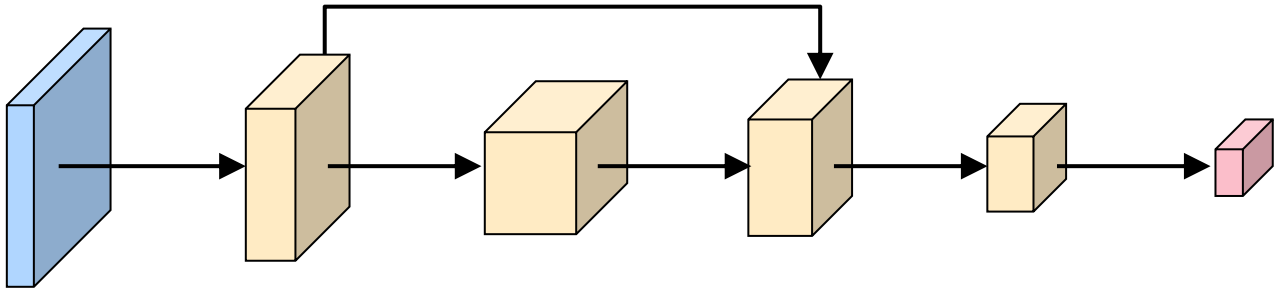


- $X \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_4 \rightarrow T$ is a Markov chain
  - So is any chain $X \rightarrow Y_i \rightarrow Y_j \rightarrow T$ for $i < j$
  - True for dense layers, convolutional layers, pooling layers, etc.

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?



- $X \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3$ is **not** a Markov chain
  - Y_3 depends on both Y_2 and Y_1, not just Y_2
  - However, $X \rightarrow Y_1 \rightarrow Y_3$ is a Markov chain
  - Markovity still holds "across" skip connections, but not "under" them

LOSSLESS FEATURE COMPRESSIBILITY

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

Proof (sketch):

- Let $\mathcal{Y} = \{\mathcal{Y}_i : 1 \leq i \leq L\}$ be a set of some intermediate layer outputs (features)
- Decompose mutual information between input X and $\mathcal{Y}$ as

$$I(X;\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y} \mid X)$$
$$= H(\mathcal{Y})$$

$H(\mathcal{Y} \mid X)$: 0, because $\mathcal{Y}$ is a function of X

- Note that $X \rightarrow X \rightarrow \mathcal{Y}$ is a Markov chain and apply DPI

$$H(X) = I(X;X) \geq I(X;\mathcal{Y}) = H(\mathcal{Y})$$

- So, $H(\mathcal{Y})$ is no larger than $H(X)$ $\Longrightarrow$ features $\mathcal{Y}$ at least as compressible (losslessly) as input X

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

LOSSLESS FEATURE COMPRESSIBILITY

- Intermediate features being more compressible than the input is good news!
- But lossless compressibility is very limiting
  - Lossy compression gives much higher compression ratios
  - Practical image and video codecs mostly lossy
  - Can we extend this result to lossy compression?

Rate-distortion function

- Let X be a random variable and $\hat{X}$ be its "quantized" version according to some conditional probability distribution $p(\hat{x} \mid x)$
- Let $d(\hat{x}, x)$ be a distortion metric – how much $\hat{x}$ differs from x
- For a given distortion level D, define set $\mathcal{P}_X(D)$ of conditional distributions as

$$\mathcal{P}_X(D) = \left\{ p(\hat{x} \mid x) \;:\; \underbrace{\sum_{x,\hat{x}} p(x) \cdot p(\hat{x} \mid x) \cdot d(\hat{x}, x)}_{\mathbb{E}[d(\hat{X}, X)]} \leq D \right\}$$

- Rate-distortion (RD) function for X is given by

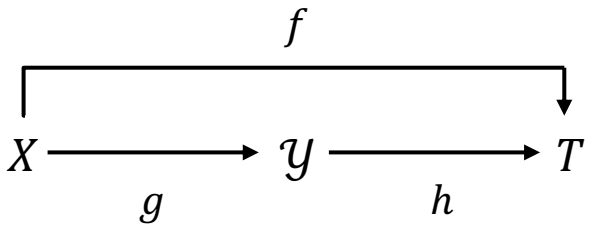
$$R_X(D) = \min_{p(\hat{x} \mid x) \in \mathcal{P}_X(D)} I(X; \hat{X})$$

- $R_X(D)$ is the minimum rate (in bits) at which you can encode X without incurring distortion $> D$

T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley, 2006.

LOSSY FEATURE COMPRESSIBILITY

- In order to use RD theory in our case, we need some modifications



- When we compress input X, we care about what happens to the output T

$$\mathcal{P}_X(D) = \{p(\hat{x} \mid x) \quad : \mathbb{E}[d(f(\hat{X}), f(X))] \leq D\}$$

- Similarly, when we compress features Y, we care about what happens to the output T

$$\mathcal{P}_Y(D) = \{p(\hat{y} \mid y) \quad : \mathbb{E}[d(h(\hat{Y}), h(Y))] \leq D\}$$

LOSSY FEATURE COMPRESSIBILITY

- We can now define the RD function for the input

$$R_X(D) = \min_{p(\hat{x} \mid x) \in \mathcal{P}_X(D)} I(X; \hat{X})$$

and the RD function for the features

$$R_Y(D) = \min_{p(\hat{y} \mid y) \in \mathcal{P}_Y(D)} I(Y; \hat{Y})$$

- In both cases, distortion is measured at the output of the network
- Distortion metric can be any metric appropriate for the network's task, e.g.
  - Mean Squared Error for regression tasks
  - Cross-entropy or accuracy for classification tasks
  - …

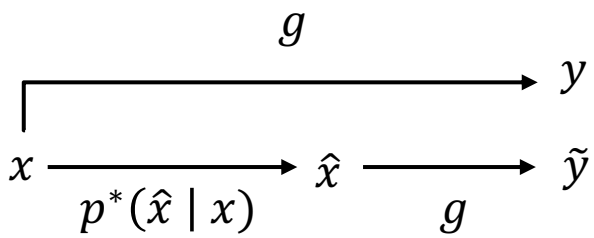
LOSSY FEATURE COMPRESSIBILITY

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

$$R_y(D) \leq R_x(D)$$

Proof (sketch):

- Let D be given and let $p^*(\hat{x} \mid x)$ be optimal for input compression (achieves $R_X(D)$)
- Draw inputs $X \sim p(x)$ and process each input x in two ways as follows



- For each x, obtain y and $\tilde{y}$
- Define $q(\tilde{y} \mid y)$ by pairing up y and $\tilde{y}$

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

Proof (sketch, continued):

- Show $q(\tilde{y}|\, y) \in \mathcal{P}_y(D)$, i.e., satisfies distortion constraint for D
  - Easy to show because $q(\tilde{y}|\, y)$ is derived from $p^*(\hat{x} \mid x) \in R_X(D)$, which satisfies distortion constraint for D
- Apply DPI to Markov chain $\tilde{Y} \rightarrow \hat{X} \rightarrow X \rightarrow Y$ to show

$$I(Y;\tilde{Y}) \leq I(X;\hat{X})$$

- When $p^*(\hat{x} \mid x)$ is used to generate $\hat{X}$, the above inequality becomes

$$I(Y;\tilde{Y}) \leq R_X(D)$$

- So we have found one distribution $q(\tilde{y}|\, y) \in \mathcal{P}_y(D)$ that achieves $I(Y;\tilde{Y})$ below $R_X(D)$. Therefore

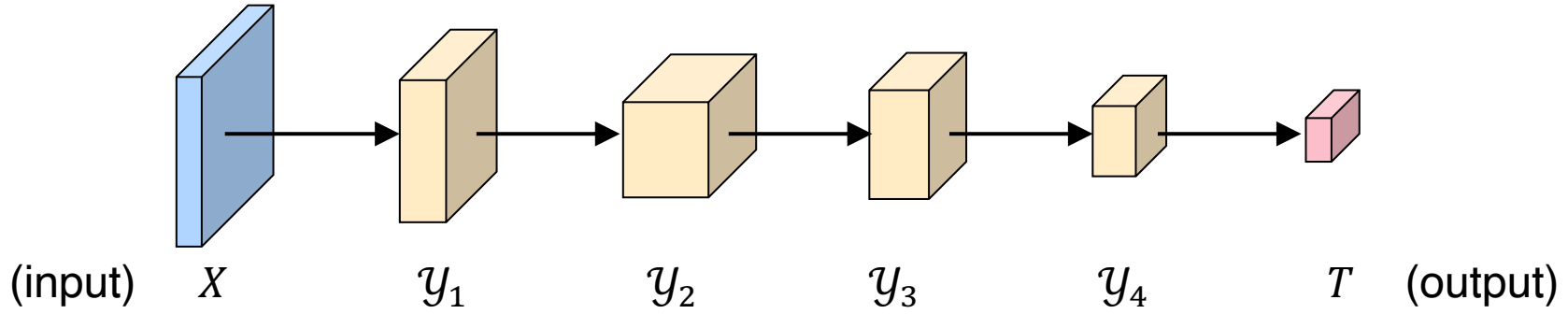
$$R_Y(D) = \min_{p(\hat{y} \mid y) \in \mathcal{P}_y(D)} I(Y;\hat{Y}) \leq R_X(D)$$

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

DEEPER MEANS MORE COMPRESSIBLE

Claim: In a non-generative feedforward neural network, deeper layers are more compressible.

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



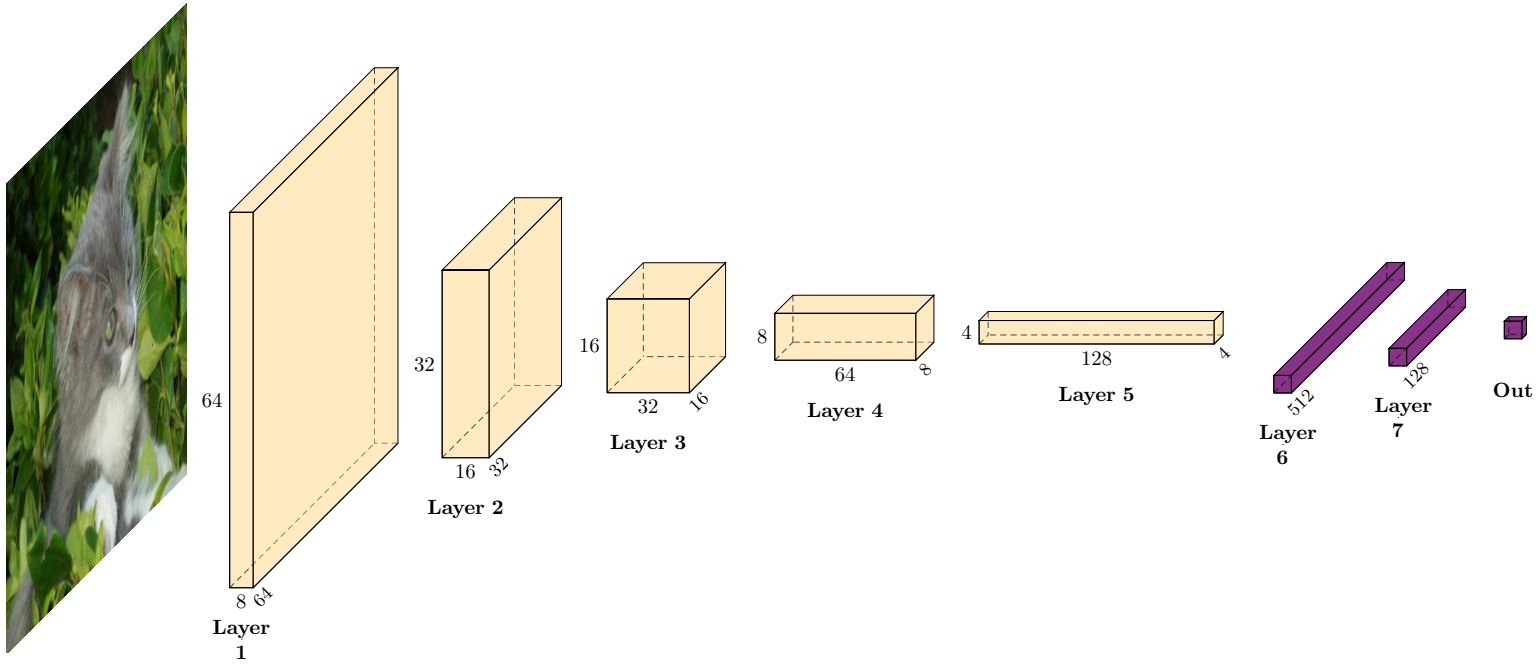
Proof: Follows from previous proofs by replacing X with $\mathcal{Y}_j$ and $\mathcal{Y}$ with $\mathcal{Y}_i$

SUMMARY OF FEATURE COMPRESSIBILITY

- Theory shows that intermediate features are at least as compressible as the network's input
- This is true for any non-generative feedforward network:
  - Regardless of what its task is (T can be any task)
  - Regardless of how many tasks there are (T can be a composite task)
- However:
  - Theory talks about limits; practical codecs might be far from those limits
  - Theory shows what is possible, but not how to get there
  - Ideal for grant proposals 😀

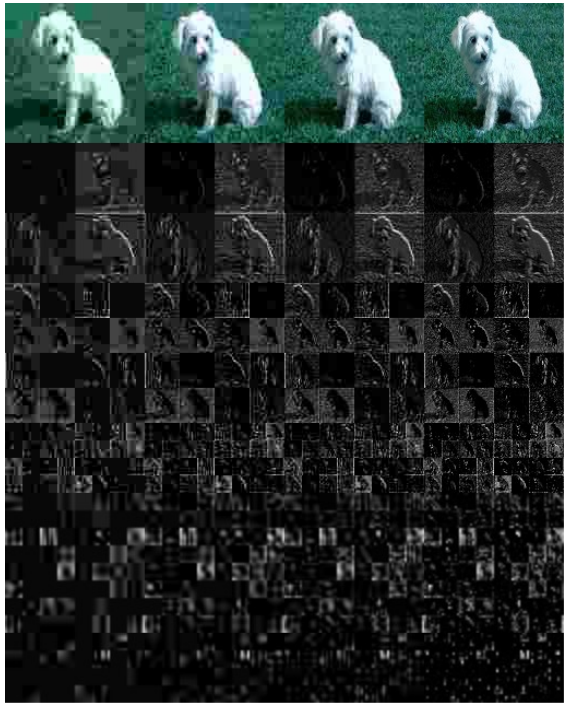
- What can we expect from practical (i.e., non-optimal) codecs?

TOY EXAMPLE OF FEATURE COMPRESSIBILITY

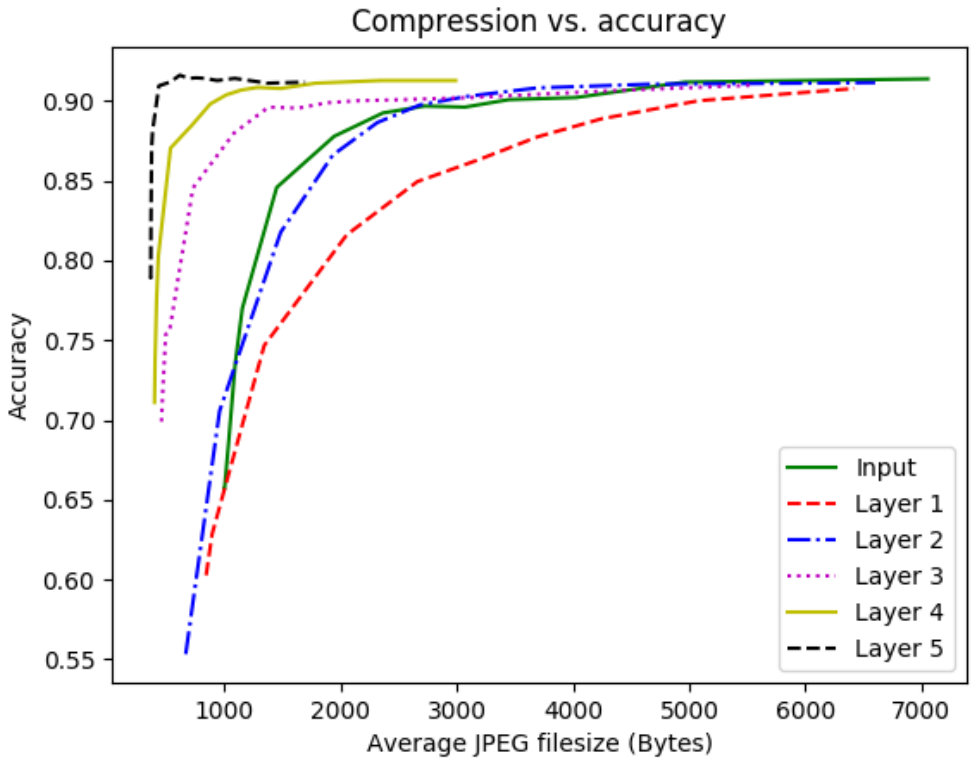


- A simple convolutional neural network (CNN) for cats vs. dogs classification
- Trained on Kaggle's cats vs. dogs dataset
- Goal: compare input compression vs. feature compression in terms of resulting classification accuracy

TOY EXAMPLE OF FEATURE COMPRESSIBILITY



Features tiled into an image and compressed using JPEG



Feature compression better than input compression starting with layer 3 – why?

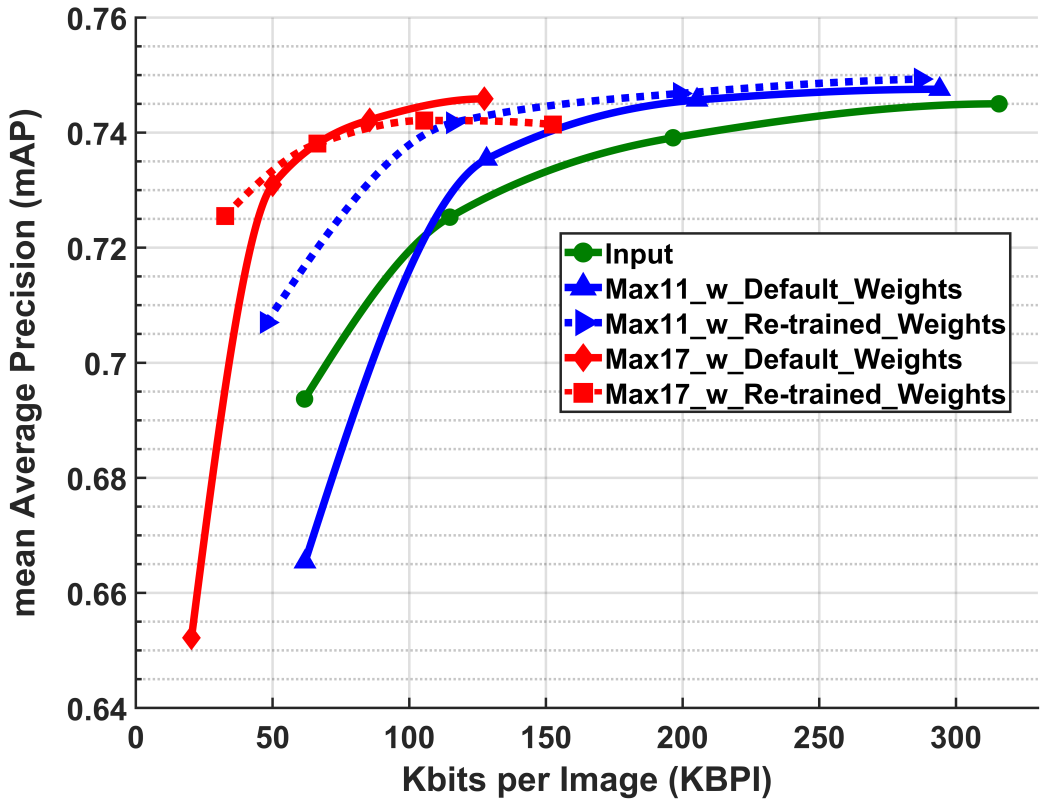
If we had an optimal encoder, this would already happen at layer 1

ANOTHER 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

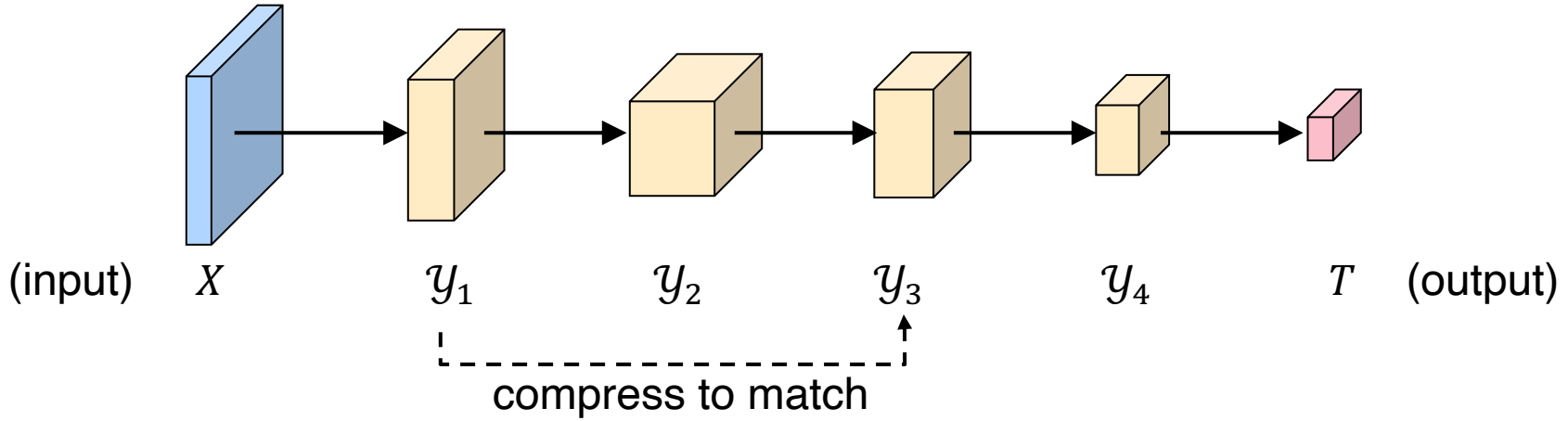
Split at	Default weights	Re-trained weights
max_11	−6.09%	**−45.23%**
max_17	−60.30%	**−70.30%**



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

DISTILLATION

- Based on the results so far, it seems one needs to move the compression point in order to achieve gains – this makes encoder more complicated
- But there is another way, via “distillation” – no need to move the compression point



Claim: Under certain conditions, compressing to match (“distill”) deeper layers is better.

A. Harell, A. de Andrade, and I. V. Bajić, ”Rate-distortion in image coding for machines," PCS 2022. arXiv:2209.11694

CODING FOR MULTIPLE TASKS

- Different tasks have different distortion metrics
- Need to define task importance
  - Need to be careful about scale of different distortion metrics
- Need to allocate bits appropriately

- One way to bring task distortions to a common scale

accuracy w/o compression

accuracy after compression

$$D_i = \frac{|\overline{A_i} - A_i|}{\overline{A_i}} \cdot 100$$

% change in task accuracy due to compression

- A_i could be mAP, IoU, Jaccard index, MSE, PSNR, …

S. R. Alvar and I. V. Bajić, "Pareto-optimal bit allocation for collaborative intelligence," IEEE Trans. Image Processing, vol. 30, Feb. 2021.

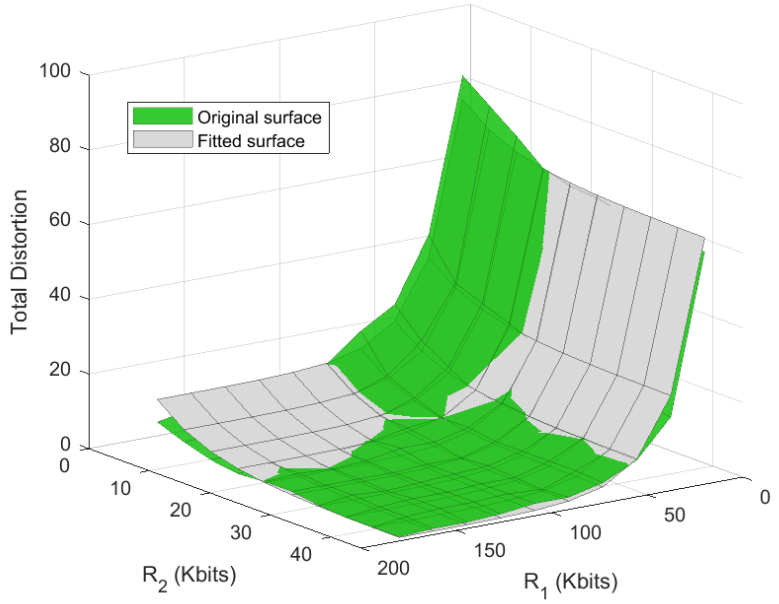
CODING FOR MULTIPLE TASKS

- Tractable rate-distortion (RD) model

$$D_i(R_1, \ldots, R_N) \approx \gamma_i + \sum_{j=1}^{N} \alpha_{i,j} 2^{-\beta_{i,j} R_j}$$

where R_j is the rate of the j-th coding unit

- Benefits of this RD model:
  - "Makes sense" – distortion reduces exponentially with rates
  - Fits the data: $R^2 > 0.94$ in all our tests
  - Tractable – distortion is convex and monotonically decreasing with rate



S. R. Alvar and I. V. Bajić, "Pareto-optimal bit allocation for collaborative intelligence," IEEE Trans. Image Processing, vol. 30, Feb. 2021.

CODING FOR MULTIPLE TASKS

Claim: Let $w_i \geq 0, \sum_{i=1}^{N} w_j = 1$ be the relative importance of task $i \in \{1, \ldots, T\}$, so that the total distortion over all tasks is

$$D_t(R_1, \ldots, R_N) = \sum_{i=1}^{T} w_i \cdot D_i(R_1, \ldots, R_N) \approx \gamma + \sum_{j=1}^{N} \alpha_i 2^{-\beta_j R_j}$$

Then the optimal bit allocation to minimize $D_t(R_1, \ldots, R_N)$ subject to $\sum_{j=1}^{N} R_j \leq R_t$ is

$$R_j^* = \frac{1}{\beta_j}\left[\log_2\{(\ln 2)\alpha_j\beta_j\} - \log_2\lambda\right]^+$$

where $[x]^+ = \max(0, x)$ and λ is the Lagrange multiplier.

Proof: Via Karush-Kuhn-Tucker (KKT) conditions.

S. R. Alvar and I. V. Bajić, "Pareto-optimal bit allocation for collaborative intelligence," IEEE Trans. Image Processing, vol. 30, Feb. 2021.

CODING FOR MULTIPLE TASKS

- In the previous result we relied on task weights w_i to convert a multi-objective problem into a single-objective problem
- But what if task importance is not know in advance?
- General problem is multi-objective optimization:

$$\text{minimize } \{D_1(R_1, \ldots, R_N), \ldots, D_T(R_1, \ldots, R_N)\}$$

$$\text{subject to } \sum_{j=1}^{N} R_j \leq R_t$$

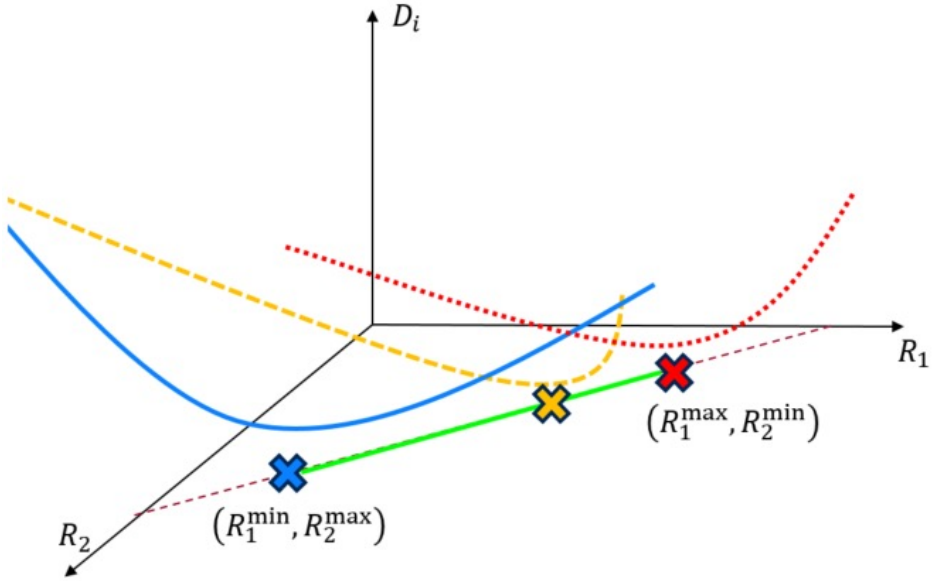
- Can be solved numerically
- Because of convexity, it can also be solved analytically in the case of two coding units ($N = 2$) and any number of tasks T

S. R. Alvar and I. V. Bajić, "Pareto-optimal bit allocation for collaborative intelligence," IEEE Trans. Image Processing, vol. 30, Feb. 2021.

CODING FOR MULTIPLE TASKS

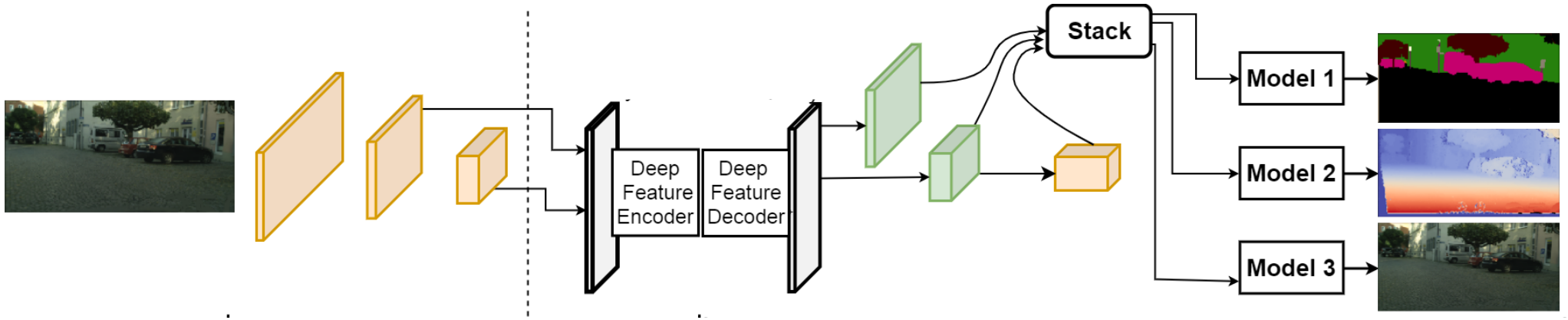
Claim: Let (R_1^i, R_2^i) be the rates on the line $R_1 + R_2 = R_t$ that minimize $D_i(R_1, R_2)$, and let $R_1^{\max} = \max\{R_1^i\}$, $R_1^{\min} = \min\{R_1^i\}$, $R_2^{\max} = 1 - R_1^{\min}$, and $R_2^{\min} = 1 - R_1^{\max}$. Then any point on the line $R_1 + R_2 = R_t$ between $(R_1^{\min}, R_2^{\max})$ and $(R_1^{\max}, R_2^{\min})$ is Pareto-optimal, and there are no Pareto-optimal solutions outside of this line segment.

Proof: Follows from the properties of distortion surfaces.

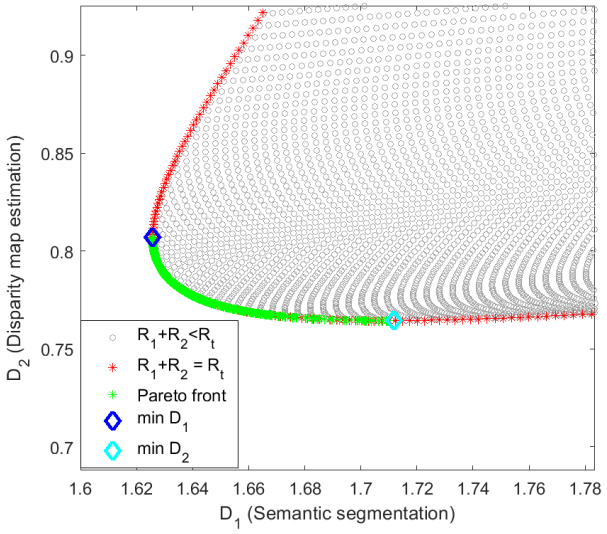


S. R. Alvar and I. V. Bajić, "Pareto-optimal bit allocation for collaborative intelligence," IEEE Trans. Image Processing, vol. 30, Feb. 2021.

CODING FOR MULTIPLE TASKS

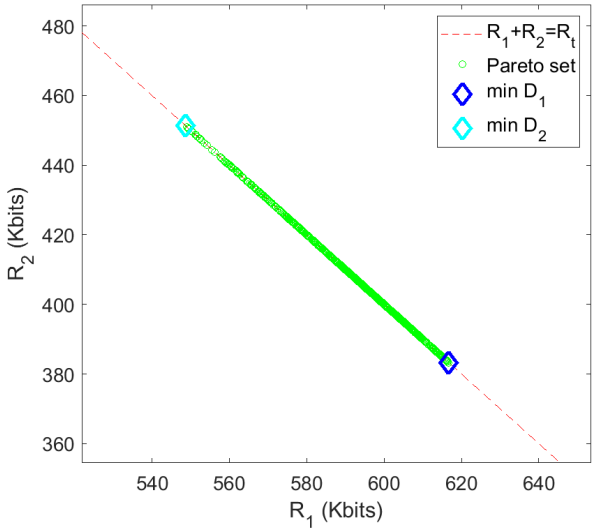


Pareto front



Pareto set

(rates that achieve the Pareto front)



S. R. Alvar and I. V. Bajić, "Pareto-optimal bit allocation for collaborative intelligence," IEEE Trans. Image Processing, vol. 30,Feb. 2021.

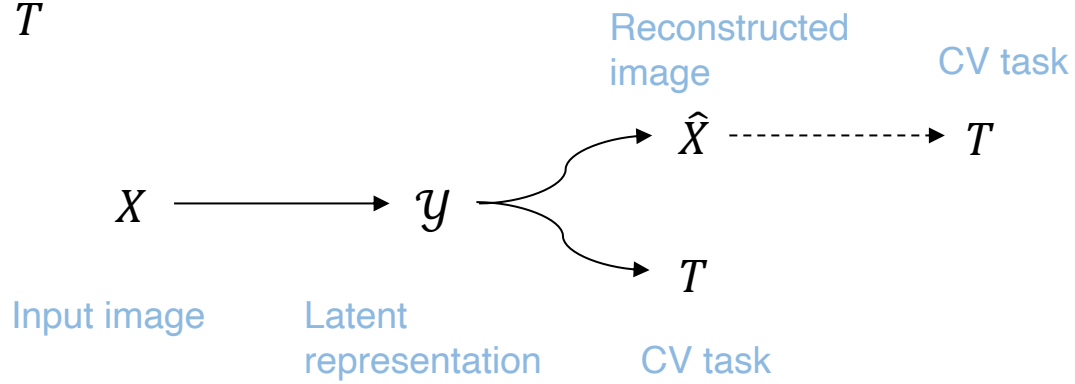
Questions?

Part 2

Current practice

LATENT SPACE SCALABILITY

- The tasks often include input image reconstruction ($\hat{X}$) and/or some computer vision (CV) inference tasks T



- In the discussion so far, it seems that all features supported all tasks; but a better design is possible
- CV inference can also be obtained from $\hat{X}$ (common in practice)
- Data processing inequality (DPI) applied to $Y \rightarrow \hat{X} \rightarrow T$:

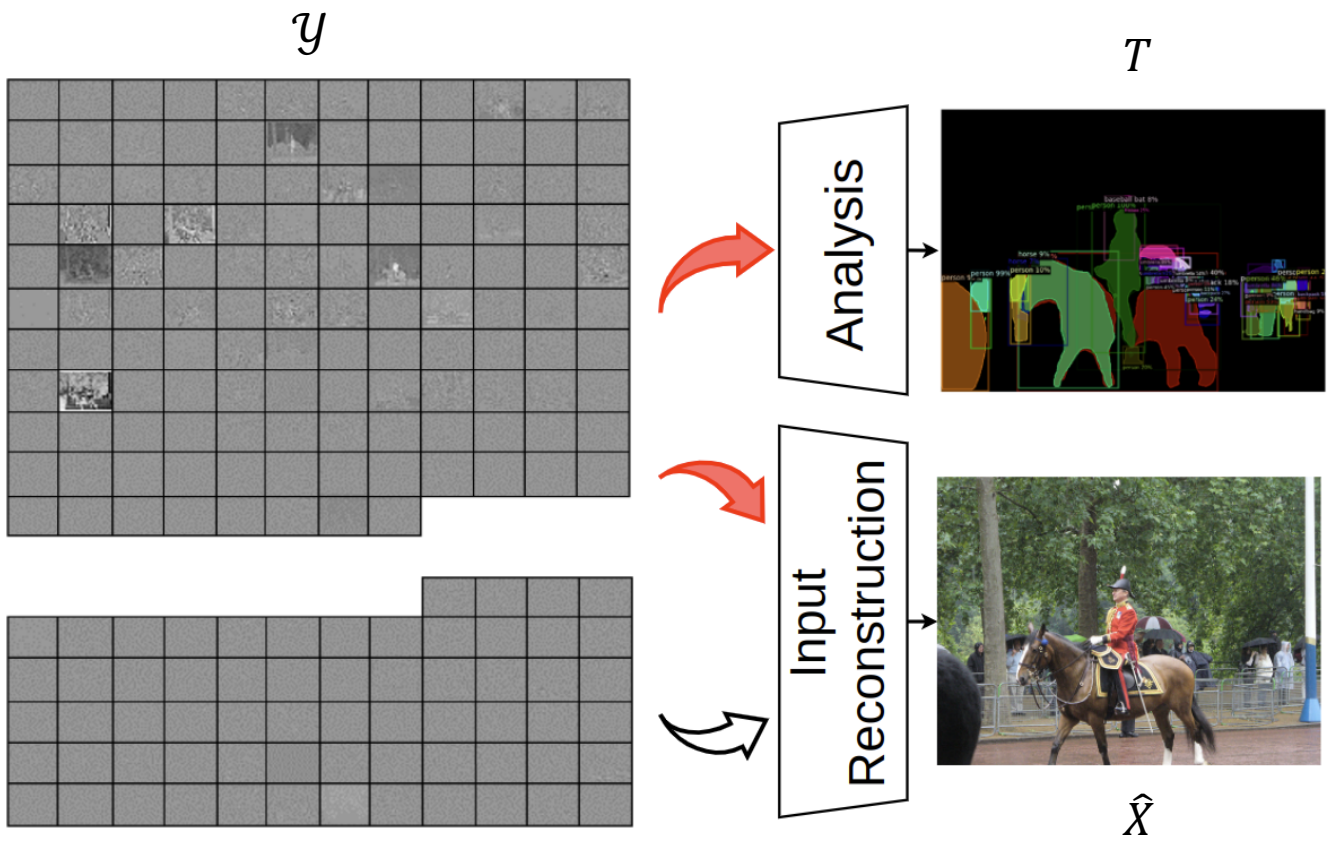
$$I(Y;\hat{X}) \geq I(Y;T)$$

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

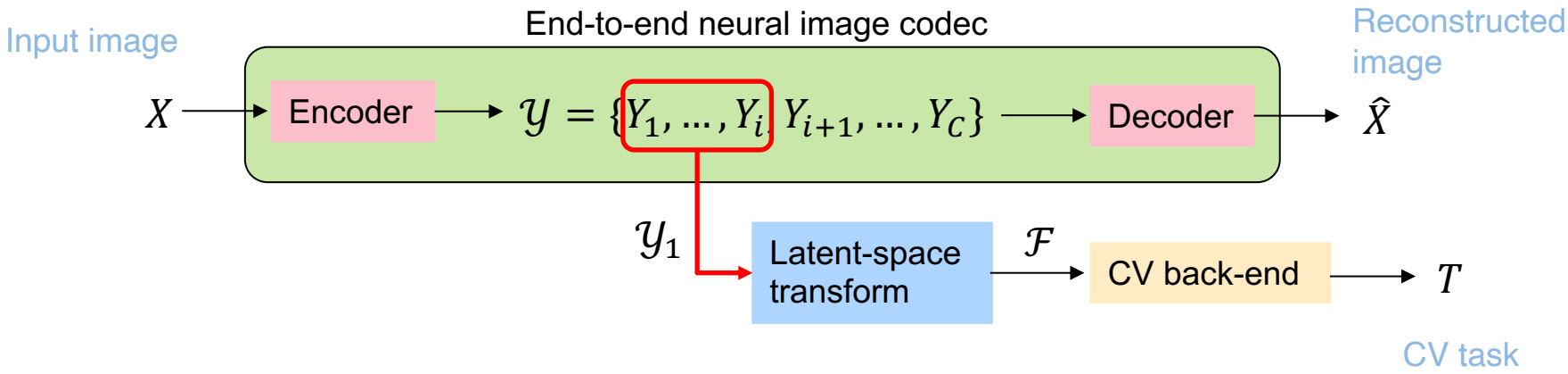
$$I(\mathcal{Y};\hat{X}) \geq I(\mathcal{Y};T)$$

- 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

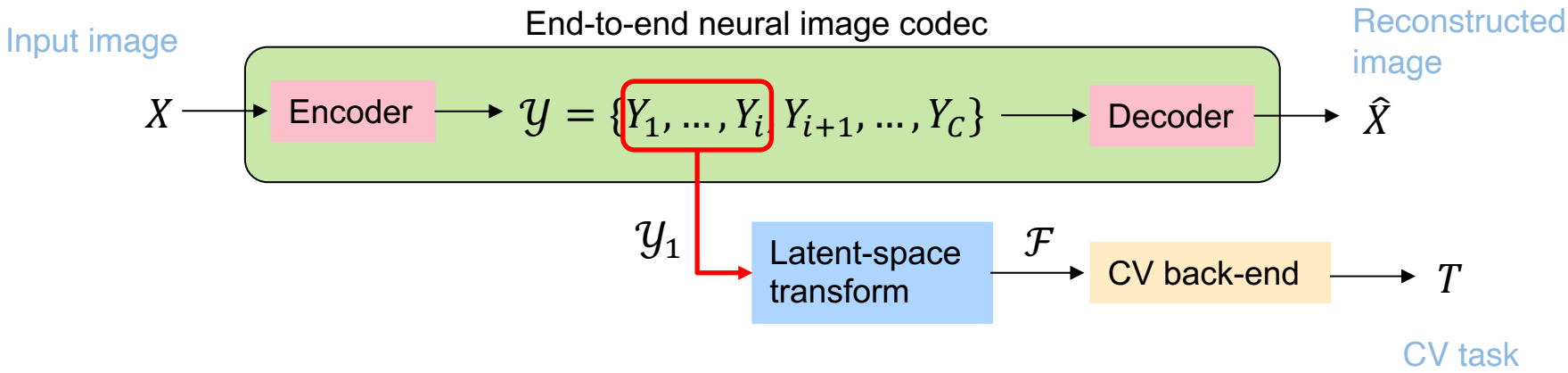


Example 2-layer scalable system:

- 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:

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

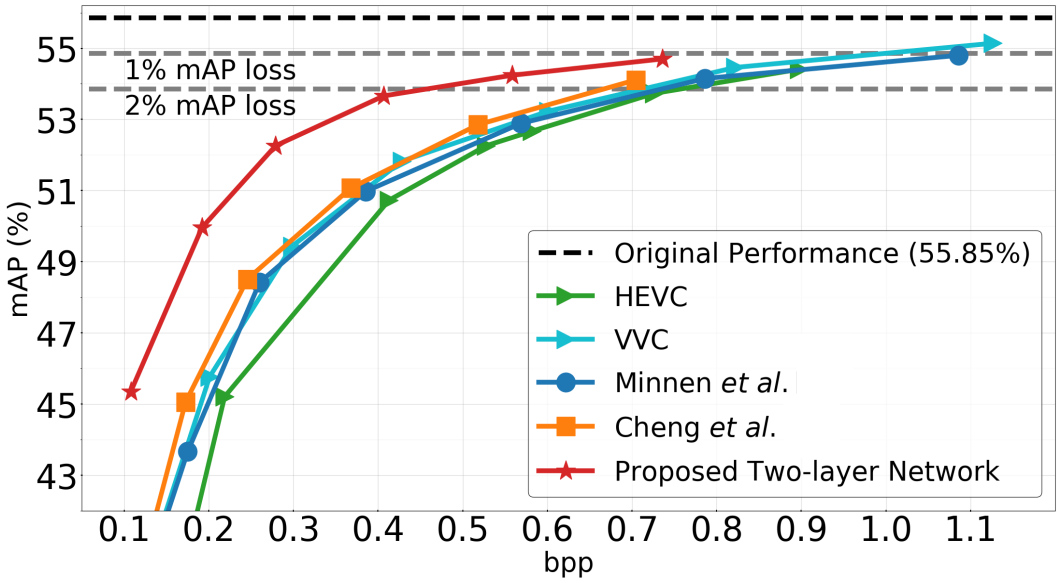
- R is the rate estimate [2]
- Distortion D composed of input reconstruction $\mathrm{MSE}\left(X, \hat{X}\right)$ and CV feature reconstruction $\mathrm{MSE}\left(\mathcal{F}, \hat{\mathcal{F}}\right)$
- Since $\mathrm{MSE}\left(\mathcal{F}, \hat{\mathcal{F}}\right)$ depends only on $\mathcal{Y}_1$ (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

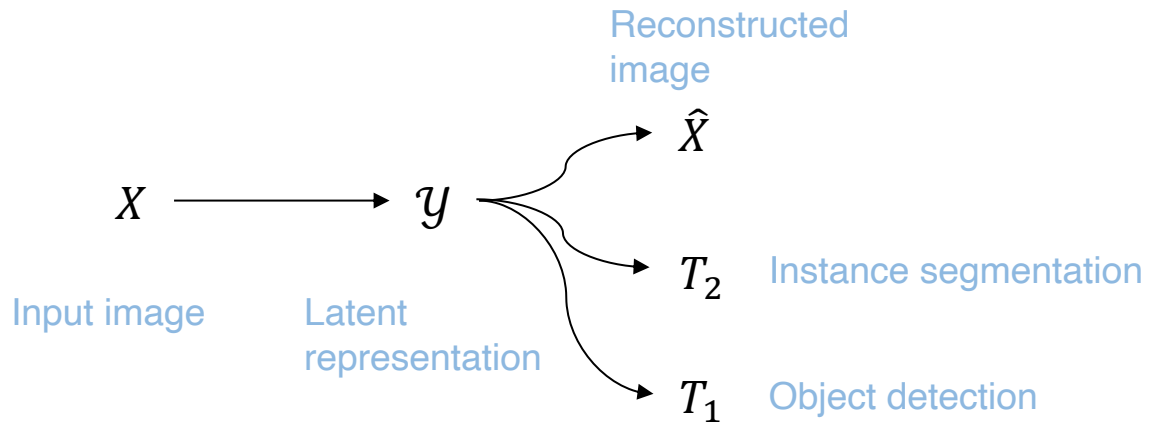


	Two-layer Network	
Benchmarks	BD-Bitrate	BD-mAP
VVC	–39.8	2.79
HEVC	–47.9	4.55
Minnen *et al.*	–41.3	3.26
Cheng *et al.*	–37.4	2.89

[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.

LATENT SPACE SCALABILITY

Three tasks

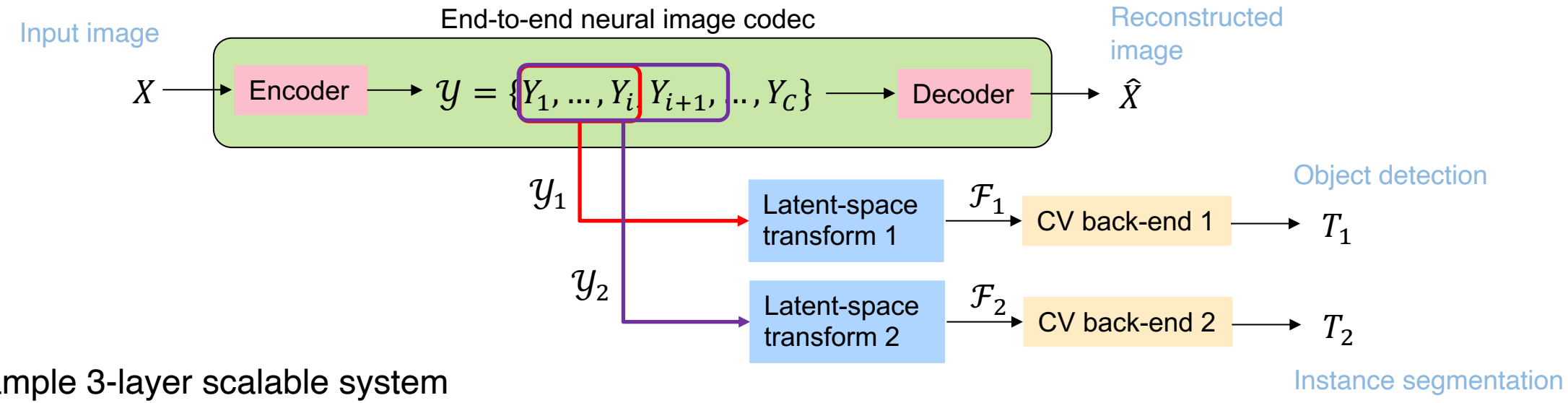


$$X \longrightarrow Y \longrightarrow \hat{X} \longrightarrow T_2 \longrightarrow T_1$$

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

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

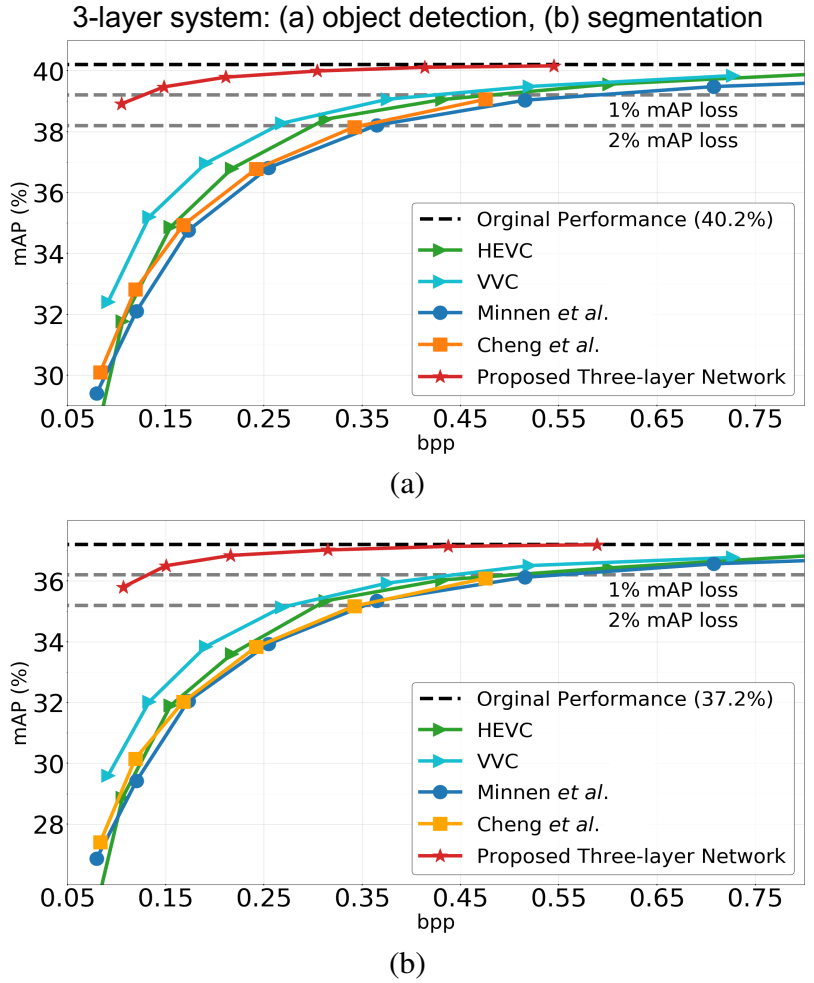
- End-to-end image codec backbone [2]
- CV task 1: object detection using Detectron [3] Faster RCNN
- CV task 2: instance segmentation using Detectron [3] Mask RCNN
  - Object detection $\subset$ semantic segmentation $\Longrightarrow$ $y_1 \subset y_2$

[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

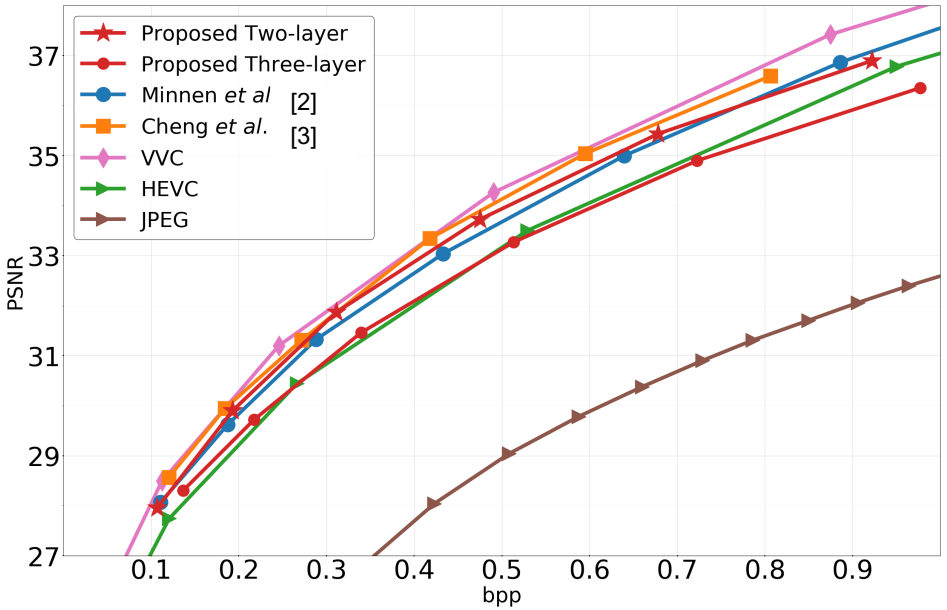
- Detection and segmentation experiments on COCO
- Again, Performance much better than compressing input directly:
  - 71 – 78% bit savings compared to state-of-the-art image codecs
  - 2.3 – 3.5% more accurate detection at the same bit rate

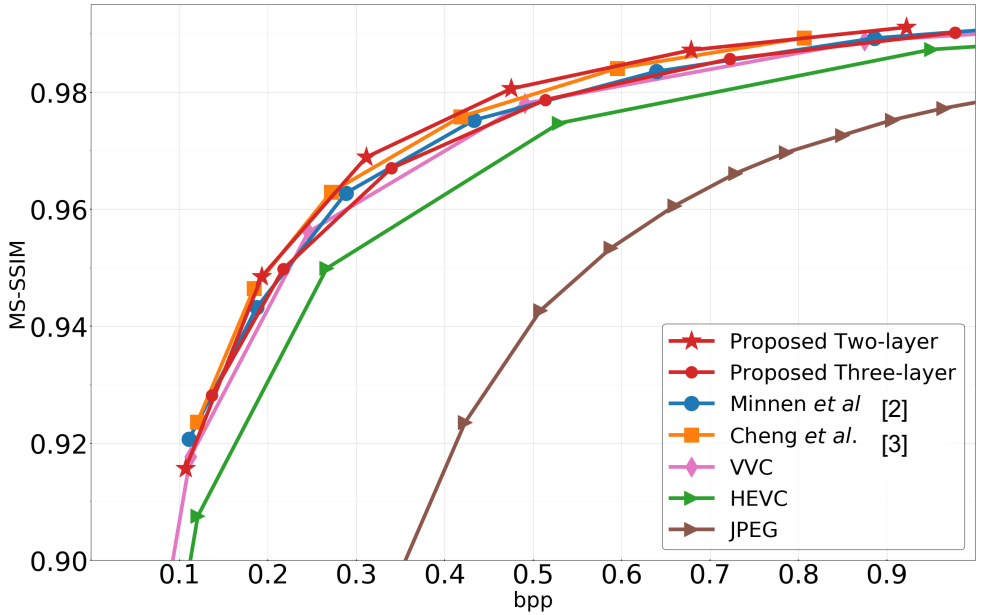
	Three-layer Network			
	Object Detection		Segmentation	
Benchmarks	BD-Bitrate	BD-mAP	BD-Bitrate	BD-mAP
VVC	–73.2	2.33	–71.2	2.34
HEVC	–73.2	3.05	–74.7	2.96
Minnen *et al.*	–78.7	3.73	–77.2	3.38
Cheng *et al.*	–76.6	3.62	–75.4	3.49



[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.

LATENT SPACE SCALABILITY





Results on the Kodak dataset

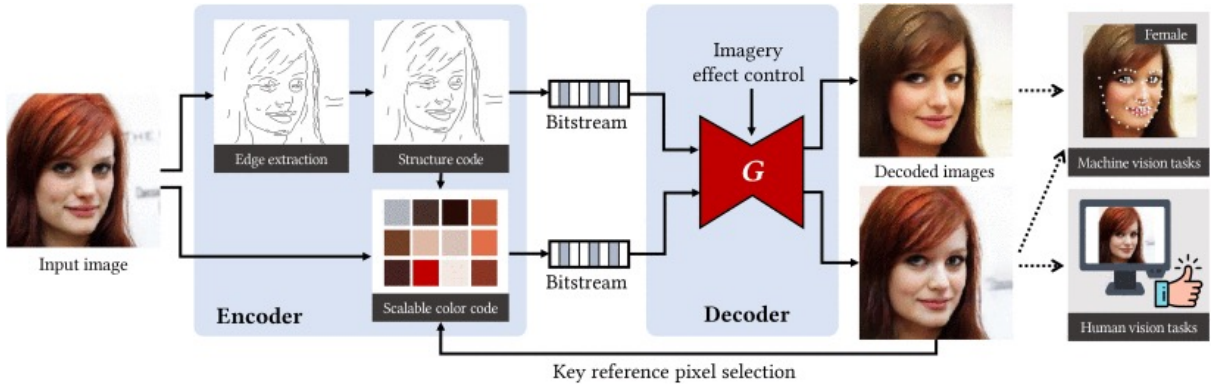
- Proposed scalable codec comparable to state-of-the-art on input reconstruction
- 10 – 20% degradation by adding a scalability layer (2 → 3), in line with earlier work on scalable video coding

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

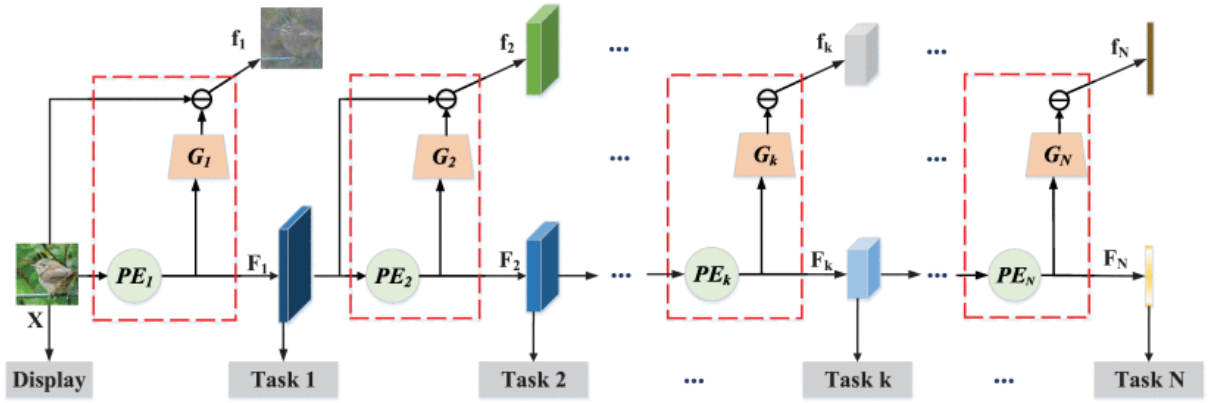
[1] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," IEEE TIP, 2022.
[2] D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," NeurIPS, 2018.
[3] Z. Cheng et al., "Learned image compression with discretized gaussian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

OTHER HUMAN-MACHINE IMAGE CODING SYSTEMS

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



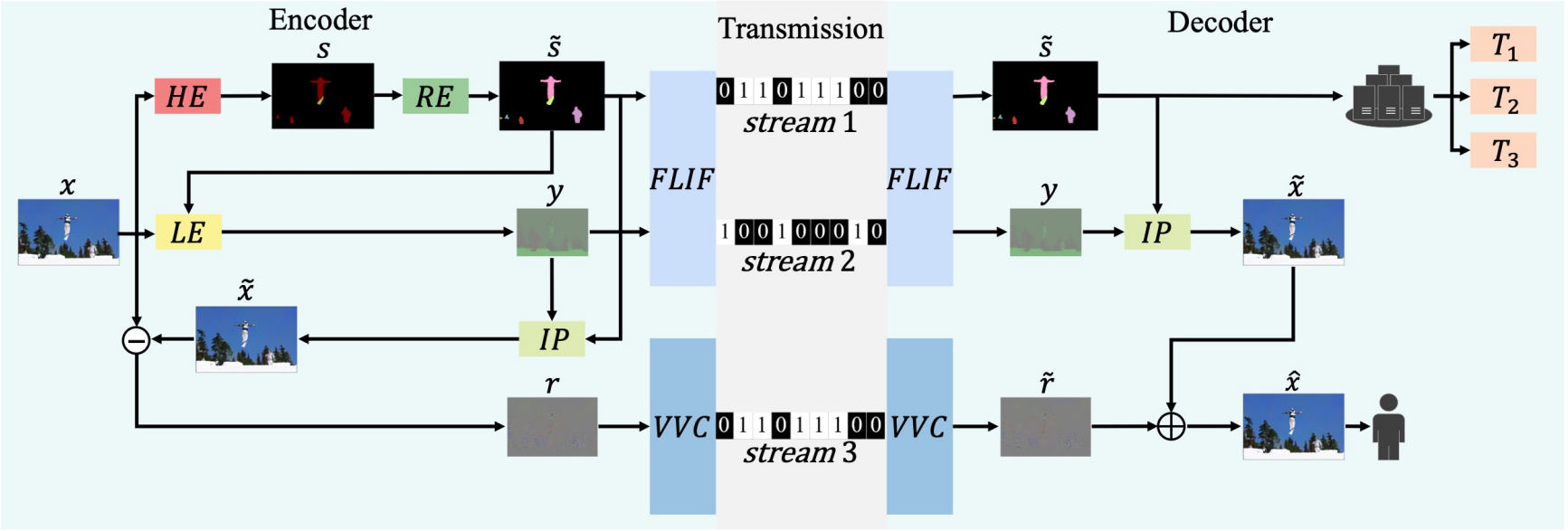
- Semantic-to-signal-scalable coding [2]
  - Base: deepest feature
  - Enhancements: information lost when going layer to layer



[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. On 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.

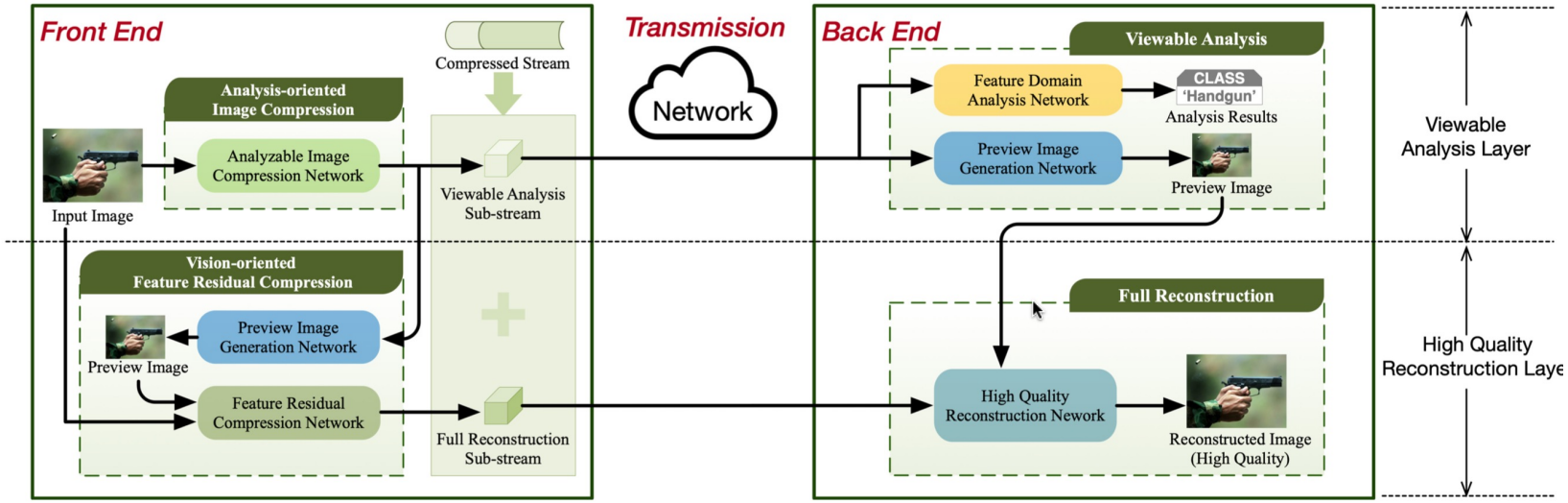
OTHER HUMAN-MACHINE IMAGE CODING SYSTEMS



- 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.

OTHER HUMAN-MACHINE IMAGE CODING SYSTEMS



- Human-machine coding for IoT [1]
  - Base: classification + preview
  - Enhancement: reconstruction residual
- A few other approaches [2, 3]

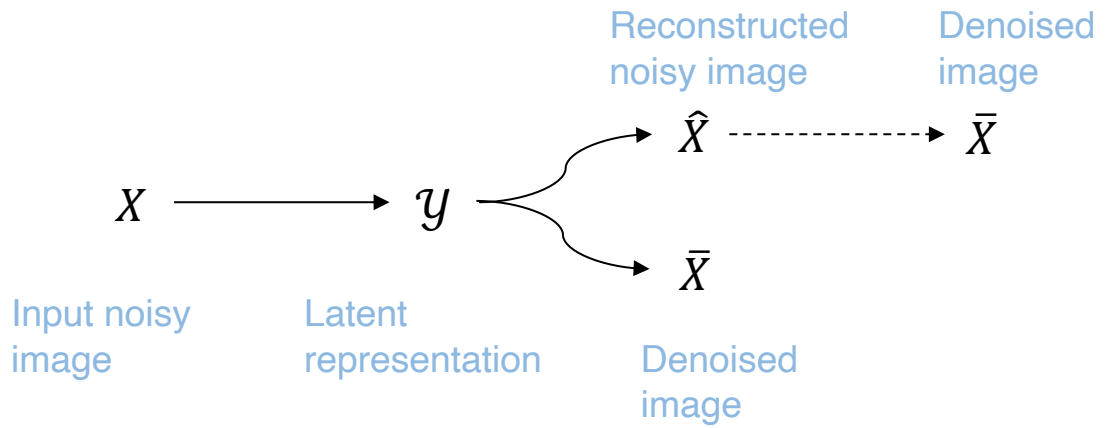
[1] Z. Wang, F. Li, J. Xu and P. C. Cosman, "Human–machine interaction-oriented image coding for resource-constrained visual monitoring in IoT," IEEE Internet of Things Journal, vol. 9, no. 17, pp. 16181-16195, 1 Sept. 2022.

[2] N. Patwa, N. Ahuja, S. Somayazulu, O. Tickoo, S. Varadarajan and S. Koolagudi, "Semantic-preserving image compression," Proc. ICIP, 2020, pp. 1281-1285

[3] M. Wang, Z. Zhang, J. Li, M. Ma and X. Fan, "Deep joint source-channel coding for multi-task network," IEEE Signal Processing Letters, vol. 28, pp. 1973-1977, 2021.

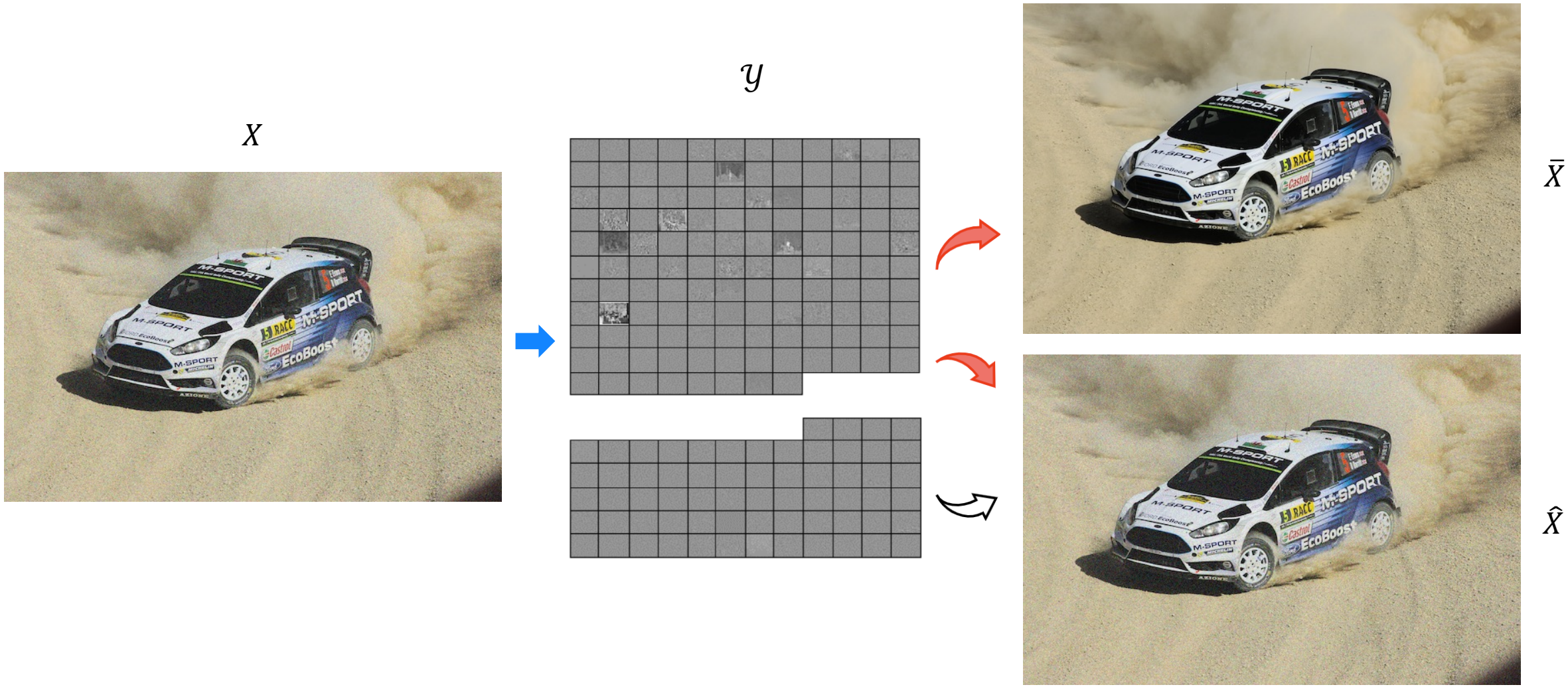
Compressed-domain denoising

- One of the scenarios in the JPEG AI call for proposals
- Provide both the denoised image and noisy image from compressed representation



- Data processing inequality (DPI) applied to $y \rightarrow \hat{X} \rightarrow \bar{X}$: $I(y;\hat{X}) \geq I(y;\bar{X})$
- Problem can be solved by latent-space scalability
  - Information needed for $\bar{X}$ is a subset of that needed for $\hat{X}$

SCALABLE LATENT SPACE FOR DENOISING

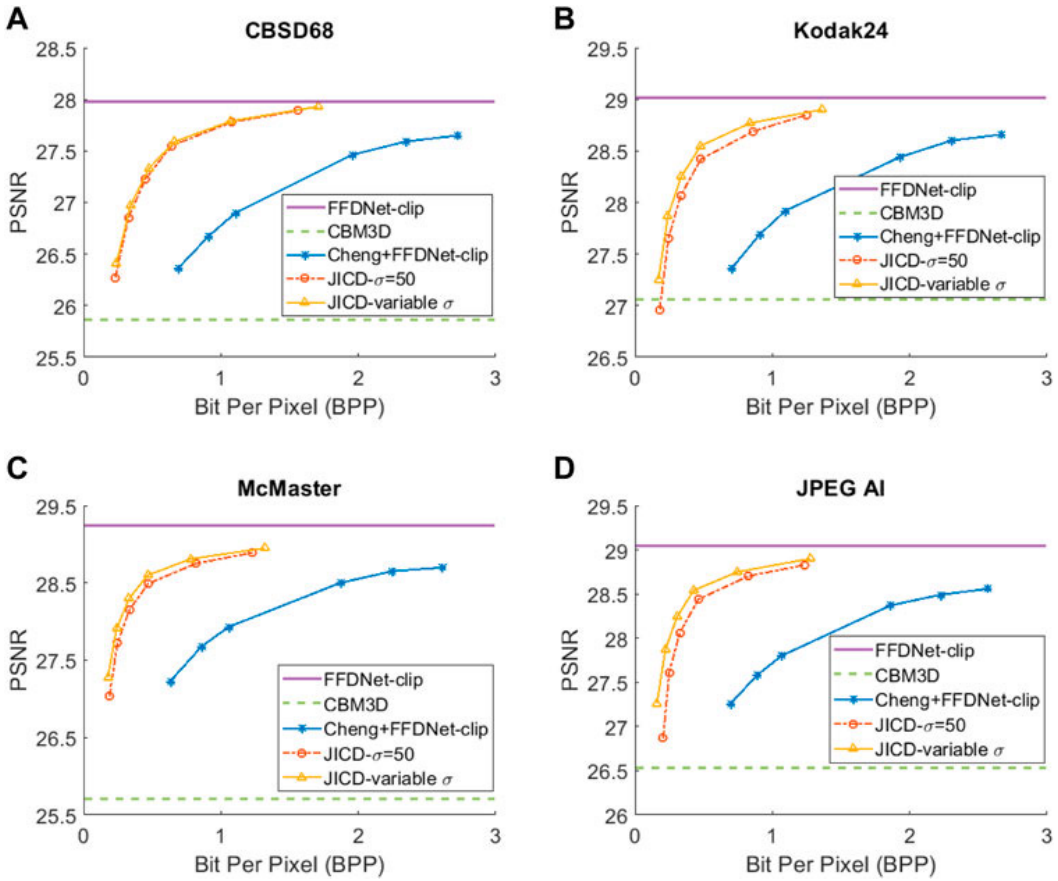


S. R. Alvar, M. Ulhaq, H. Choi, and I. V. Bajić, "Joint image compression and denoising via latent-space scalability," Front. Signal Process., 2022.

LATENT SPACE SCALABILITY FOR DENOISING

Experimental setup

- Six models trained using the Cheng2020 backbone [2], tested on four other datasets
- System trained on CLIC dataset with additive Gaussian noise $\sigma \in \{15, 25, 50\}$
- Compared against CBM3D [3] and FFD-Net [4]
- In terms of AWGN denoising performance, on large noise, better than CBM3D without compression

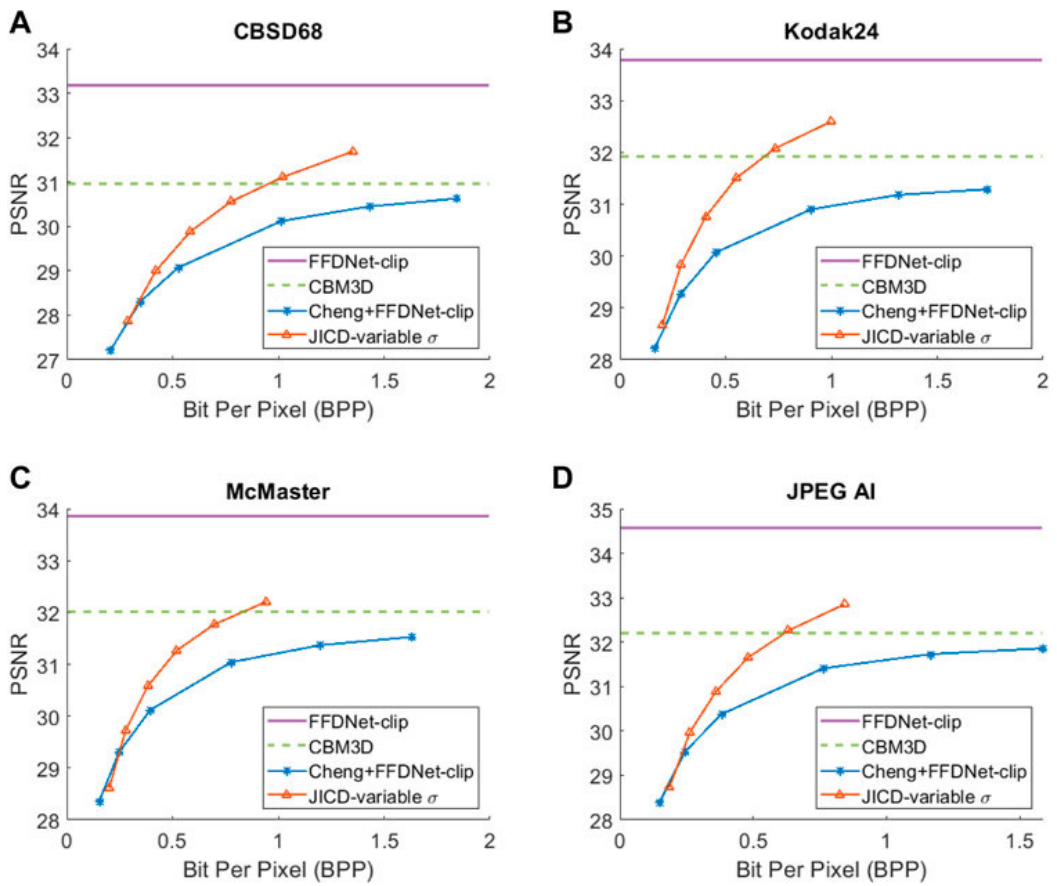


[1] S. R. Alvar, M. Ulhaq, H. Choi, and I. V. Bajić, "Joint image compression and denoising via latent-space scalability," Front. Signal Process., 2022.
[2] Z. Cheng et al., "Learned image compression with discretized gaussian mixture likelihoods and attention modules," Proc. IEEE CVPR, 2020.
[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," IEEE Trans. Image Process. 2007, pp. 2080–2095
[4] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," IEEE Trans. Image Process., 2018, pp. 4608–4622

LATENT SPACE SCALABILITY FOR DENOISING

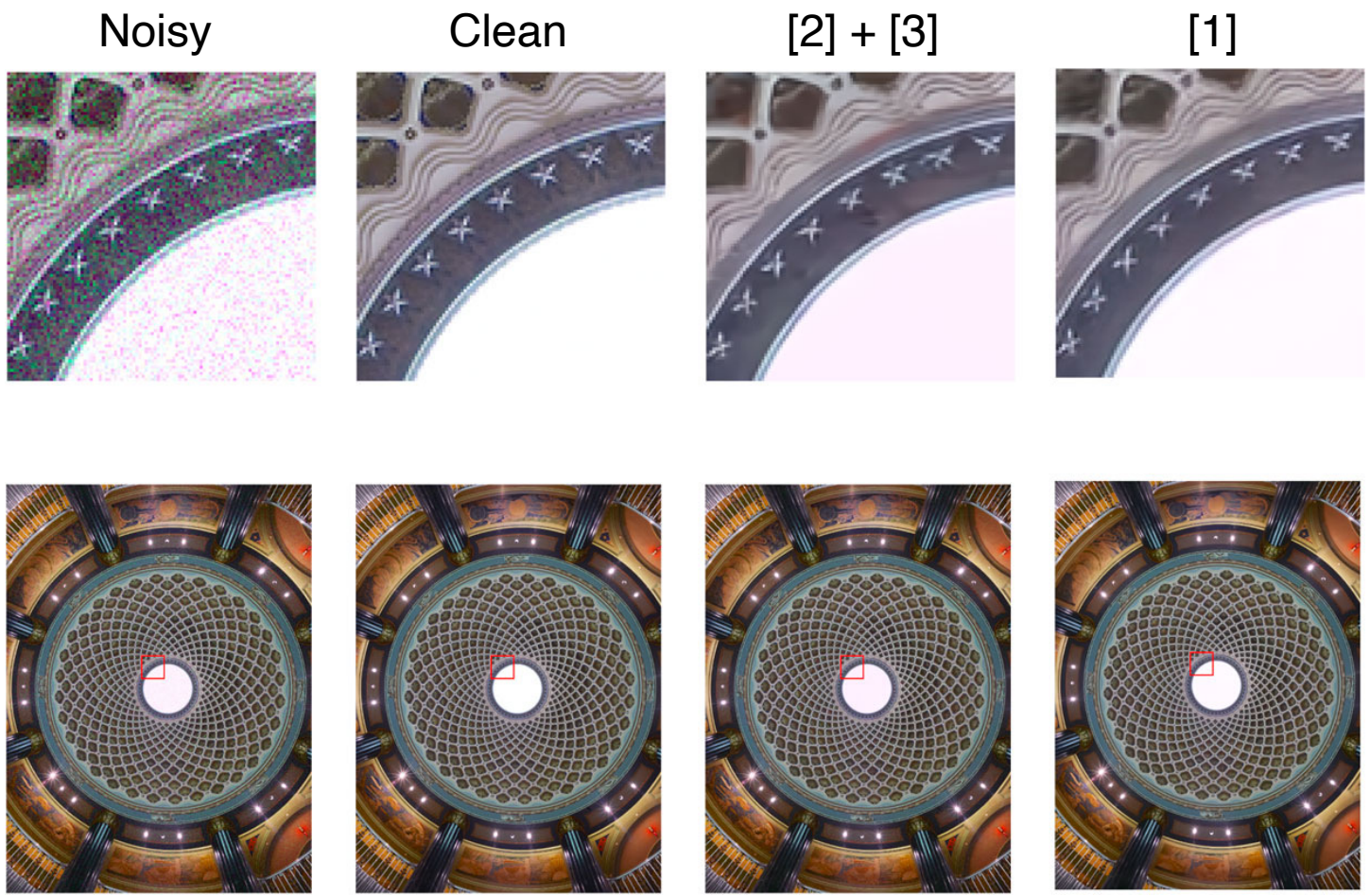
Unseen noise removal

- Tested on Poissonian-Gaussian noise model [2] that wasn't used in training
- Noise generator [3] with parameters fitted on the SIDD [4] dataset was used
  - Same noise generator was used in JPEG AI evaluation
- Surpasses CBM3D at bitrates around 1 bpp and higher



[1] S. R. Alvar, M. Ulhaq, H. Choi, and I. V. Bajić, "Joint image compression and denoising via latent-space scalability," Front. Signal Process., 2022.
[2] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data," IEEE Trans. Image Process., 2008, pp. 1737–1754
[3] S. Ranjbar Alvar and I. V. Bajić, "Practical noise simulation for RGB images," arXiv preprint arXiv:2201.12773, 2022.
[4] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," CVPR 2018.

LATENT SPACE SCALABILITY FOR DENOISING



[1] S. R. Alvar, M. Ulhaq, H. Choi, and I. V. Bajić, “Joint image compression and denoising via latent-space scalability,” Front. Signal Process., 2022.
[2] Z. Cheng et al., “Learned image compression with discretized gaussian mixture likelihoods and attention modules,” Proc. IEEE CVPR, 2020.
[3] K. Zhang, W. Zuo, and L. Zhang, “FFDNet: Toward a fast and flexible solution for CNN-based image denoising,” IEEE Trans. Image Process., 2018, pp. 4608–4622

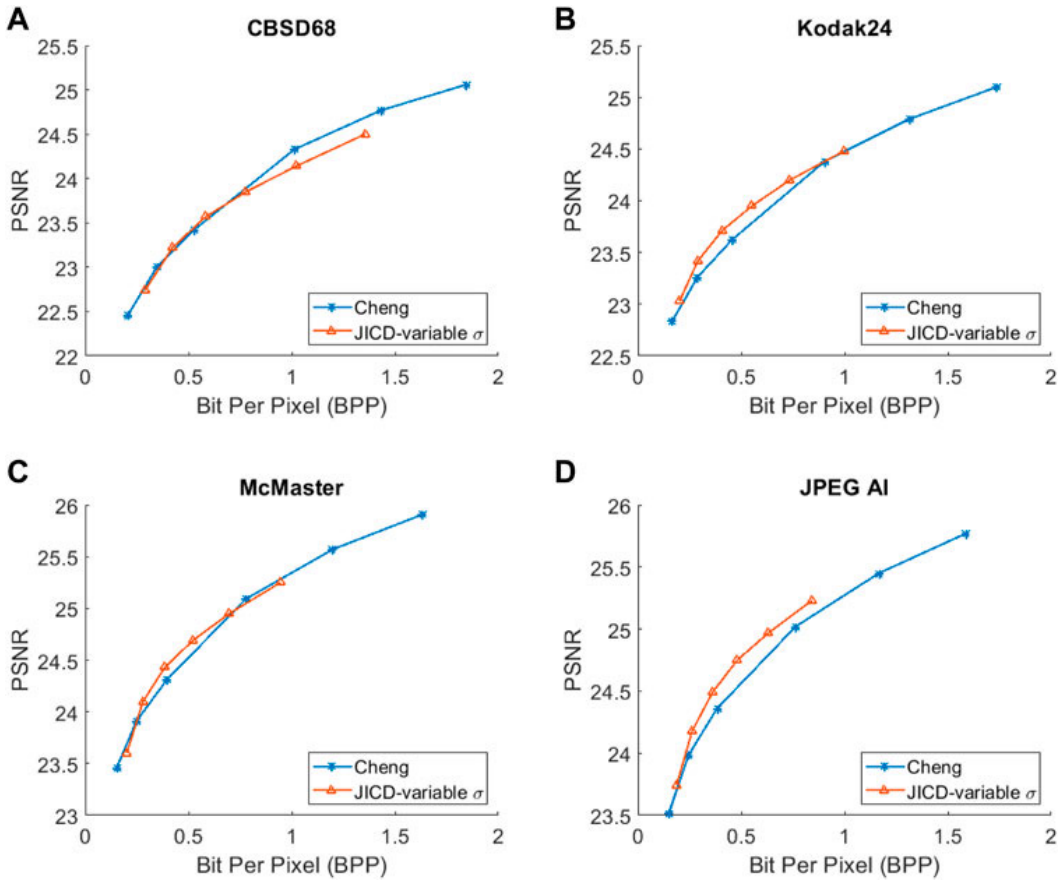
LATENT SPACE SCALABILITY FOR DENOISING

Noisy input reconstruction

- Whole latent space used
- Slightly worse than [2] on CBSD68, better on other datasets

BD-rate results

Noise type	Model	CBSD68	Kodak24	McMaster	JPEG AI
Practical noise simulator	variable σ	5.50%	−11.74%	−3.97%	−13.49%

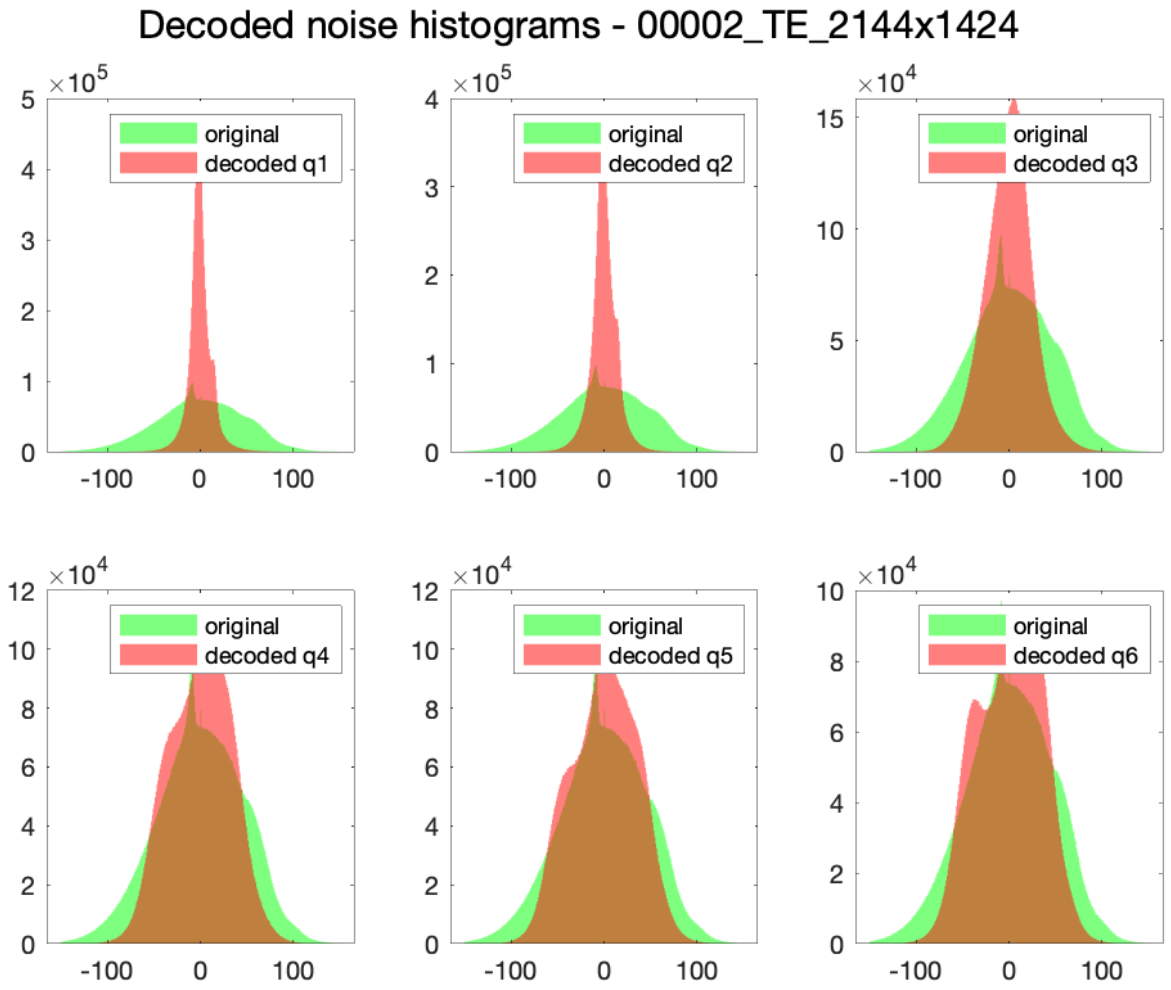


[1] S. R. Alvar, M. Ulhaq, H. Choi, and I. V. Bajić, "Joint image compression and denoising via latent-space scalability," Front. Signal Process., 2022.
[2] Z. Cheng et al., "Learned image compression with discretized gaussian mixture likelihoods and attention modules," Proc. IEEE CVPR, 2020.

LATENT SPACE SCALABILITY FOR DENOISING

Reconstructed noise distribution

- Compare distribution of input noise vs. distribution of reconstructed noise
- Example: one image from the JPEG AI test set, Gaussian noise with $\sigma = 50$
- At low bitrates, only low-variance reconstructed noise can be supported
- As the bitrate increases, reconstructed noise distribution better matches the input noise distribution



H. Choi, M. Ulhaq, S. R. Alvar, and I. V. Bajić, "Latent-space scalability for compressed domain denoising ," ISO/IEC JTC 1/SC29/WG1 M93047, Oct. 2021.

MULTI-TASK IMAGE COMPRESSION

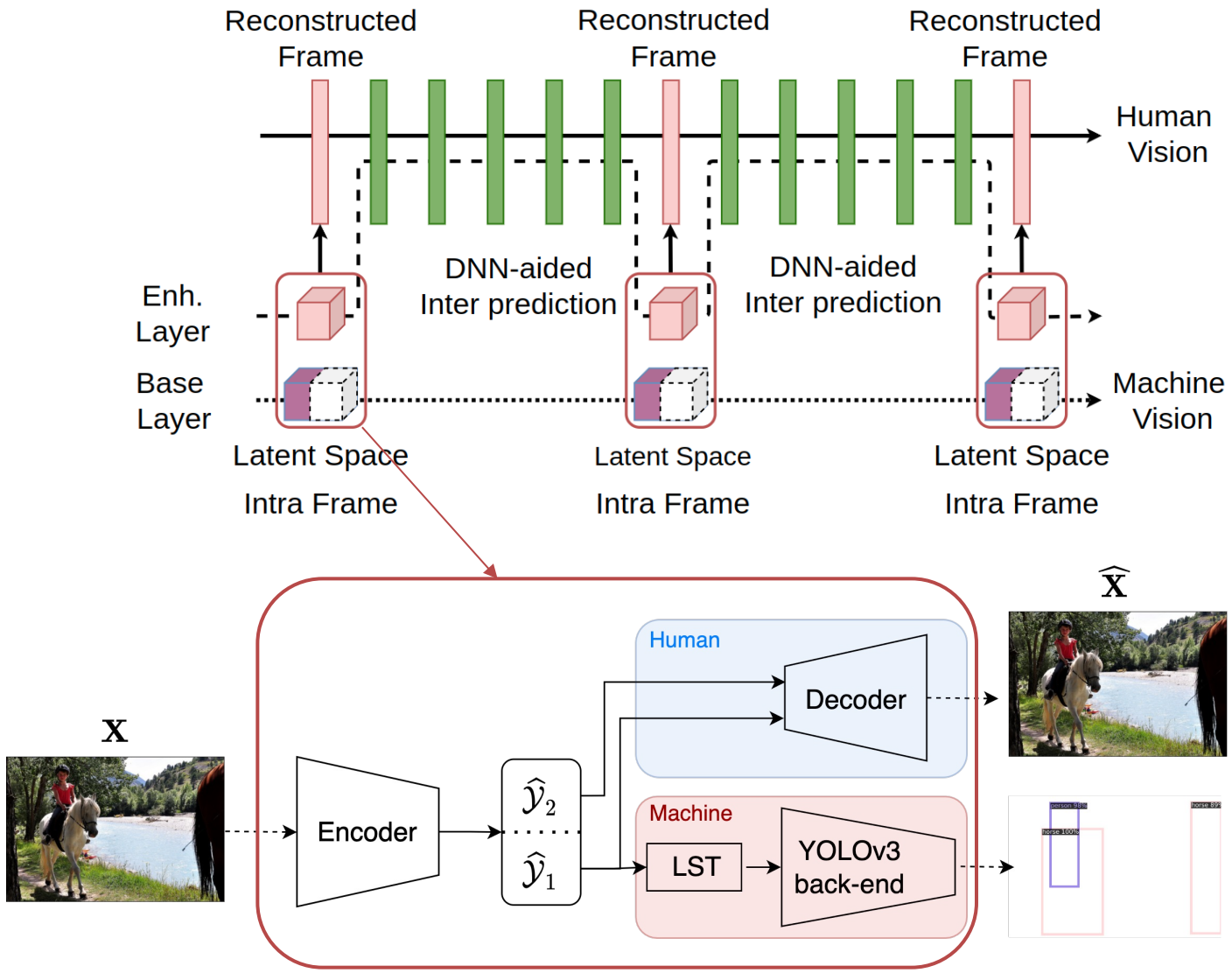
Summary

- Already a number of papers in the literature describing multi-task image compression
- Base task: computer vision
  - Usually classification, sometimes object detection and/or segmentation
- Additional tasks: computer or human vision
- Computer vision 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

MULTI-TASK VIDEO COMPRESSION

Example of a scalable 2-task video compression system

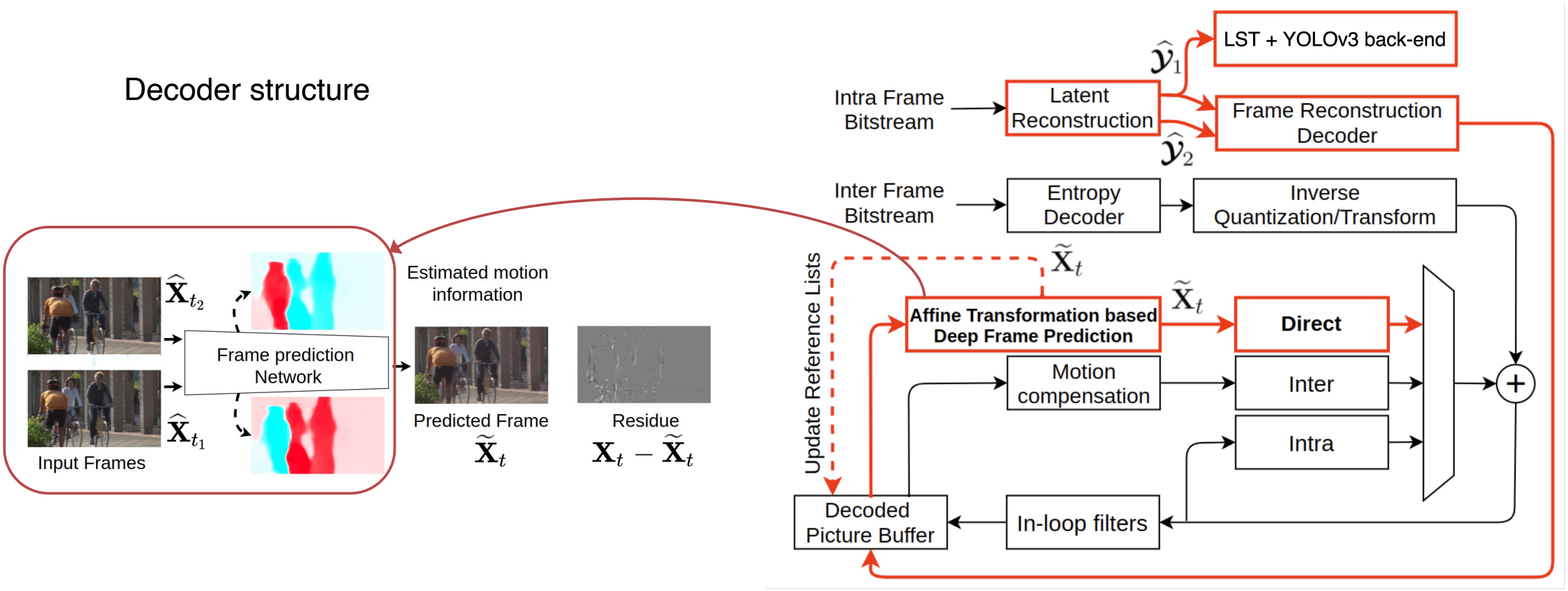
- Base layer: object detection
- Enhancement layer: input reconstruction
- Intra frames coded using the scalable human-machine image codec presented earlier
- Inter frames coded using DNN-aided HEVC pipeline



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

MULTI-TASK VIDEO COMPRESSION

Decoder structure



H. Choi and I. V. Bajić, "Scalable video coding for humans and machines," Proc. IEEE MMSP, 2022.
H. Choi and I. V. Bajić, "Affine transformation-based deep frame prediction," IEEE Trans. Image Processing, vol. 30, pp. 3321-3334, Feb. 2021.

MULTI-TASK VIDEO COMPRESSION

All Intra (detection [2] & reconstruction)

Class	Sequence	HEVC (HM-16.20) Machine Vision BD-rate- mAP	HEVC (HM-16.20) Human Vision BD-rate- PSNR	HEVC (HM-16.20) Human Vision BD-rate- MS-SSIM	VVC (VTM-10.0) Machine Vision BD-rate- mAP	VVC (VTM-10.0) Human Vision BD-rate- PSNR	VVC (VTM-10.0) Human Vision BD-rate- MS-SSIM
A	PeopleOnStreet	**-37.17%**	8.55%	**-22.93%**	**-29.52%**	36.47%	**-6.34%**
	Traffic	33.82%	16.80%	**-20.72%**	61.09%	44.38%	**-4.09%**
	Average	**-1.68%**	12.67%	**-21.83%**	15.78%	40.42%	**-5.21%**
B	BQTerrace	16.37%	29.84%	**-18.33%**	**-2.26%**	73.32%	7.84%
	BasketballDrive	**-49.91%**	24.57%	**-13.63%**	**-47.16%**	64.10%	9.47%
	Cactus	**-30.68%**	20.79%	**-19.18%**	**-46.64%**	55.70%	2.28%
	Kimono	**-75.00%**	1.37%	**-15.72%**	**-70.98%**	24.91%	0.74%
	ParkScene	**-35.81%**	14.63%	**-16.45%**	**-20.30%**	40.05%	**-0.63%**
	Average	**-35.01%**	18.24%	**-16.66%**	**-37.47%**	51.62%	3.94%
C	BQMall	**-51.04%**	1.07%	**-20.80%**	**-51.96%**	31.80%	0.95%
	BasketballDrill	**-37.45%**	0.62%	**-22.76%**	**-46.88%**	46.70%	5.09%
	PartyScene	**-8.01%**	15.60%	**-12.54%**	**-12.25%**	43.87%	5.33%
	RaceHorses	27.07%	8.49%	**-11.43%**	**-36.60%**	38.90%	8.37%
	Average	**-17.36%**	6.44%	**-16.88%**	**-36.92%**	40.32%	4.94%
D	BQSquare	**-6.51%**	7.39%	**-25.10%**	**-15.38%**	32.52%	**-10.52%**
	BasketballPass	**-57.82%**	**-2.33%**	**-16.14%**	**-55.58%**	29.18%	6.82%
	BlowingBubbles	**-15.49%**	1.08%	**-15.26%**	**-2.86%**	30.57%	5.72%
	RaceHorses	21.69%	**-4.15%**	**-11.10%**	**-22.45%**	27.46%	11.82%
	Average	**-14.53%**	0.50%	**-16.90%**	**-24.07%**	29.93%	3.46%
E	Johnny	116.35%	7.87%	**-19.50%**	86.62%	47.54%	7.45%
	KristenAndSara	**-39.08%**	7.48%	**-29.17%**	**-8.03%**	42.40%	**-8.88%**
	Average	38.64%	6.21%	**-24.90%**	39.29%	41.19%	**-2.60%**
Avg. (A - D)		**-20.40%**	9.62%	**-17.47%**	**-26.65%**	41.33%	2.86%
Avg. (A - E)		**-13.45%**	9.05%	**-18.71%**	**-18.89%**	41.31%	1.95%

Random Access (reconstruction only)

Class	Sequence	HEVC (HM-16.20) BD-rate (PSNR)	HEVC (HM-16.20) BD-rate (MS-SSIM)	VVC (VTM-10.0) BD-rate (PSNR)	VVC (VTM-10.0) BD-rate (MS-SSIM)
A	PeopleOnStreet	**-1.27%**	**-12.15%**	20.82%	9.41%
	Traffic	21.88%	8.90%	48.65%	33.31%
	Average	10.30%	**-1.63%**	34.74%	21.36%
B	BQTerrace	21.70%	3.32%	55.15%	32.94%
	BasketballDrive	5.85%	**-2.02%**	42.65%	31.89%
	Cactus	16.54%	**-1.89%**	49.58%	27.42%
	Kimono	0.50%	**-9.96%**	29.06%	14.88%
	ParkScene	14.13%	0.86%	39.48%	23.98%
	Average	11.74%	**-1.94%**	43.18%	26.22%
C	BQMall	3.14%	**-9.64%**	40.89%	22.20%
	BasketballDrill	10.91%	**-4.05%**	56.60%	54.33%
	PartyScene	12.99%	**-0.45%**	43.24%	24.76%
	RaceHorses	4.23%	**-1.58%**	37.94%	31.42%
	Average	7.82%	**-3.93%**	44.67%	33.18%
D	BQSquare	7.38%	**-9.49%**	50.49%	19.02%
	BasketballPass	**-2.86%**	**-9.68%**	36.77%	23.01%
	BlowingBubbles	4.18%	**-6.94%**	39.37%	21.03%
	RaceHorses	**-2.71%**	**-4.75%**	38.38%	31.18%
	Average	1.50%	**-7.71%**	41.25%	23.56%
E	FourPeople	11.52%	**-11.51%**	45.47%	13.16%
	Johnny	17.84%	**-2.49%**	62.58%	32.28%
	KristenAndSara	14.26%	**-16.50%**	53.67%	11.36%
	Average	14.54%	**-10.17%**	53.90%	18.94%
Avg. (A - D)		7.77%	**-3.97%**	41.94%	26.72%
Avg. (A - E)		8.90%	**-5.00%**	43.93%	25.42%

[1] H. Choi and I. V. Bajić, "Scalable video coding for humans and machines," Proc. IEEE MMSP, 2022.

[2] H. Choi, E. Hosseini, S. R. Alvar, R. A. Cohen, and I. V. Bajić, "A dataset of labelled objects on raw video sequences," Data in Brief, vol. 34, article no. 106701, Feb. 2021.

MULTI-TASK VIDEO COMPRESSION

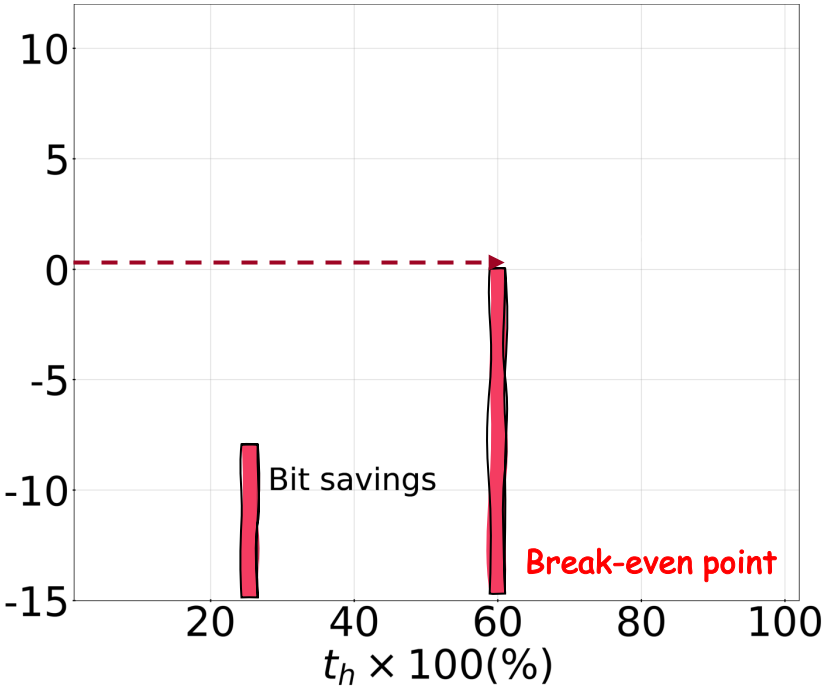
Break even point

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

$$(1 - t_h) \cdot 0.8655 + t_h \cdot 1.0905 \leq 1$$

frac. time machine vision

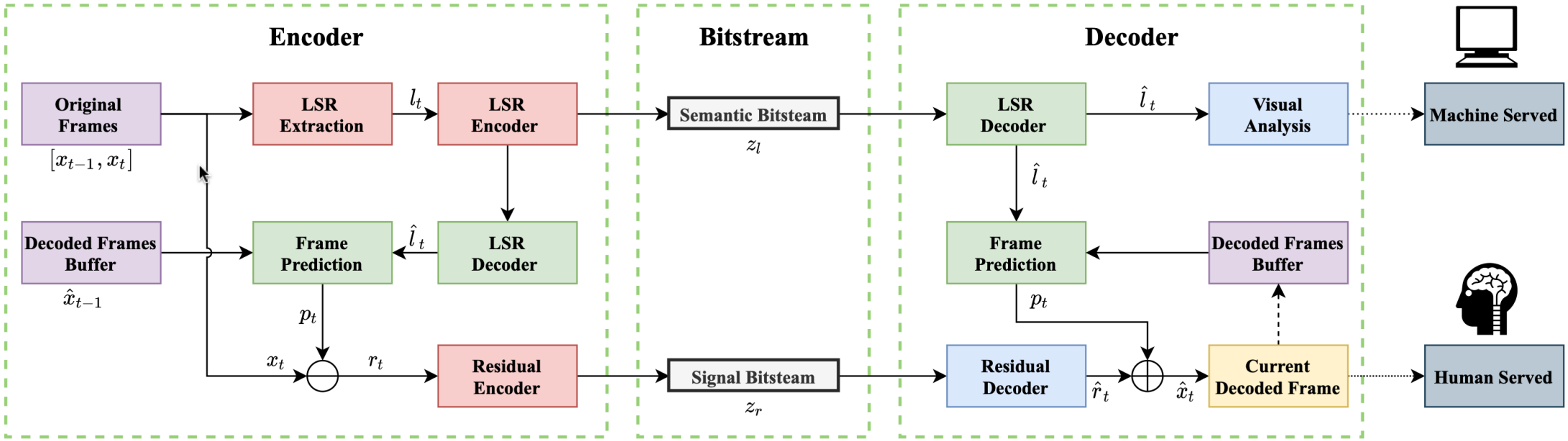
frac. time human vision



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.

MULTI-TASK VIDEO COMPRESSION



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.

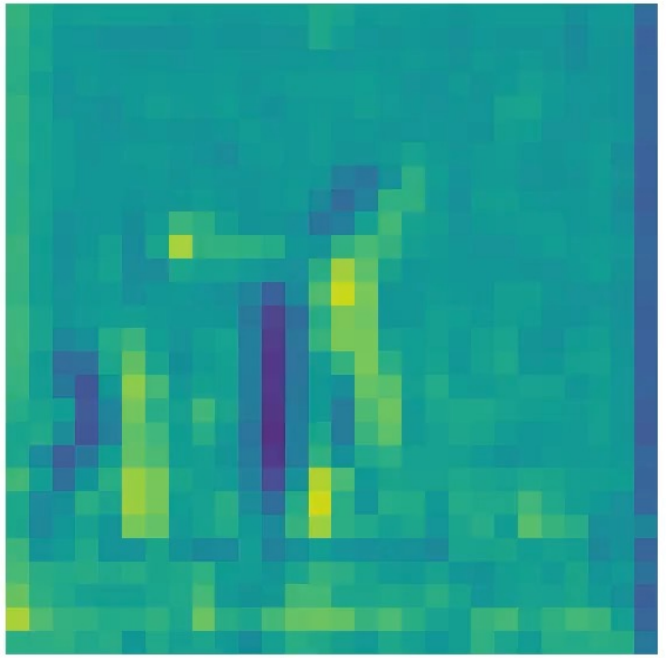
Questions?

LATENT-SPACE MOTION

What is shown in the image?

Observation:

- Input motion seems to be preserved in the latent space
- Why?

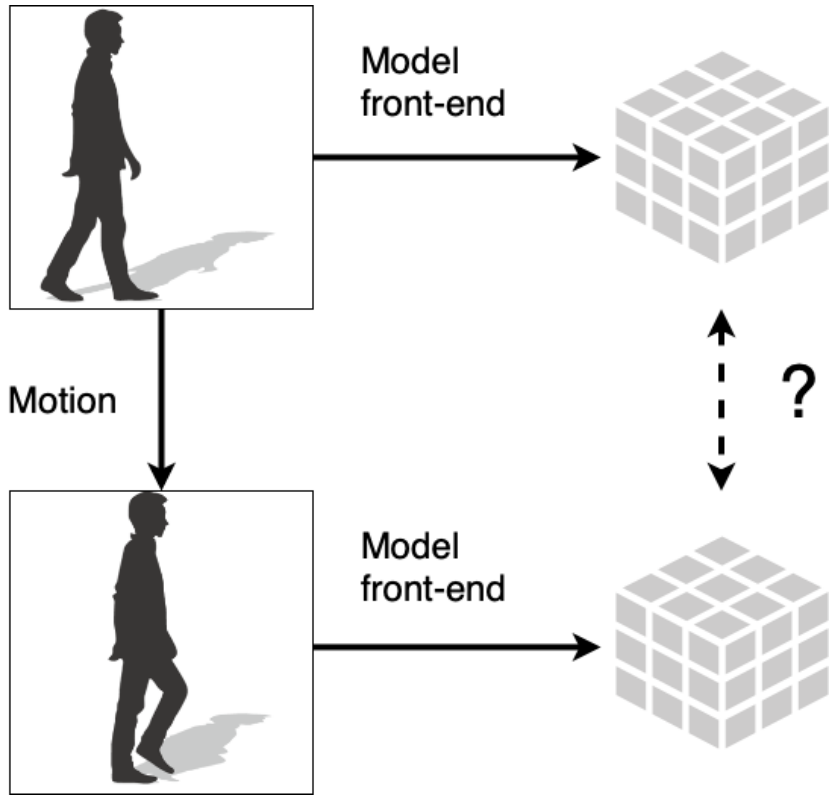


One feature tensor channel from `add_3` layer of ResNet-34

LATENT-SPACE MOTION

Understanding latent-space motion

- Consider motion in the input space between two consecutive frames
- Map each frame to the latent space via the model front—end
- What is the relationship between the corresponding feature tensors?



M. Ulhaq and I. V. Bajić, ”Latent space motion analysis for collaborative intelligence," Proc. IEEE ICASSP, pp. 8498-8502, Jun. 2021.

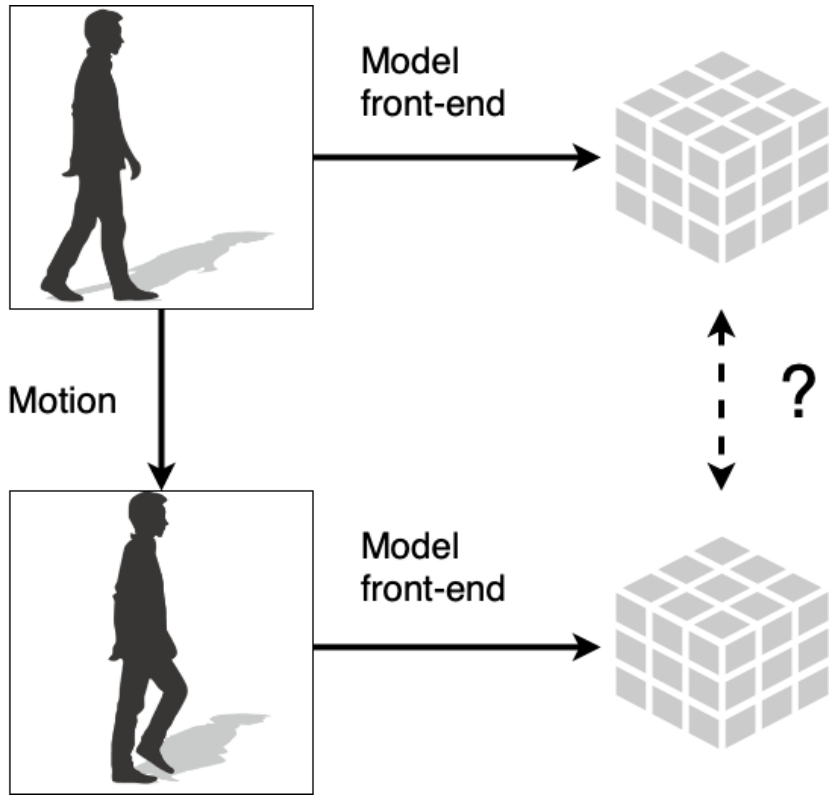
LATENT-SPACE MOTION

- A popular motion model in computer vision is “optical flow”:

$$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0$$

  - I – image intensity; t – time
  - (v_x, v_y) – optical flow

- If this model describes motion in the input space, what it its equivalent in the latent space?



M. Ulhaq and I. V. Bajić, ”Latent space motion analysis for collaborative intelligence," Proc. IEEE ICASSP, pp. 8498-8502, Jun. 2021.

LATENT-SPACE MOTION

Common operations in convolutional networks:

1. Convolution
2. Nonlinear activation
3. Batch normalization
4. Pooling
    - Max pooling
    - Mean pooling
    - Learnt pooling (strided convolution)

- Examine the effect of each of these on the optical flow PDE

M. Ulhaq and I. V. Bajić, ”Latent space motion analysis for collaborative intelligence," Proc. IEEE ICASSP, pp. 8498-8502, Jun. 2021.

LATENT-SPACE MOTION

- When input image I is convolved with kernel f, the resulting flow equation is

$$\frac{\partial}{\partial x}(f * I)u_x + \frac{\partial}{\partial y}(f * I)u_y + \frac{\partial}{\partial t}(f * I) = 0$$

where (u_x, u_y) is the flow field after convolution

- Convolution and differentiation commute:

$$f * \left(\frac{\partial I}{\partial x}u_x + \frac{\partial I}{\partial y}u_y + \frac{\partial I}{\partial t}\right) = 0$$

same flow equation as in input space $\Longrightarrow$ solution to input flow is one solution to output flow

M. Ulhaq and I. V. Bajić, "Latent space motion analysis for collaborative intelligence," Proc. IEEE ICASSP, pp. 8498-8502, Jun. 2021.

LATENT-SPACE MOTION

- When input image I passes through nonlinear activation $\sigma(\cdot)$, the resulting flow equation is

$$\frac{\partial\,\sigma(I)}{\partial x}u_x + \frac{\partial\,\sigma(I)}{\partial y}u_y + \frac{\partial\,\sigma(I)}{\partial t} = 0$$

where (u_x, u_y) is the flow field after nonlinear activation

- Using the chain rule of differentiation:

$$\sigma'(I) \cdot \left(\frac{\partial I}{\partial x}u_x + \frac{\partial I}{\partial y}u_y + \frac{\partial I}{\partial t}\right) = 0$$

same flow equation as in input space $\Longrightarrow$ solution to input flow is one solution to output flow

M. Ulhaq and I. V. Bajić, "Latent space motion analysis for collaborative intelligence," Proc. IEEE ICASSP, pp. 8498-8502, Jun. 2021.

LATENT-SPACE MOTION

Summary

- Optical flow of the input remains one (approximate) solution to the optical flow after common operations (convolution, nonlinear activation, pooling, etc.)
- Pooling with a spatial scale change causes a corresponding scale change in the optical flow
  - For example, 2×2 pooling scales the flow field by a factor of ½
- This is why input motion is approximately preserved in the latent space
- This also justifies using techniques originally developed for input-space motion (optical flow, block-based motion estimation/compensation) for feature-domain coding

M. Ulhaq and I. V. Bajić, ”Latent space motion analysis for collaborative intelligence," Proc. IEEE ICASSP, pp. 8498-8502, Jun. 2021.

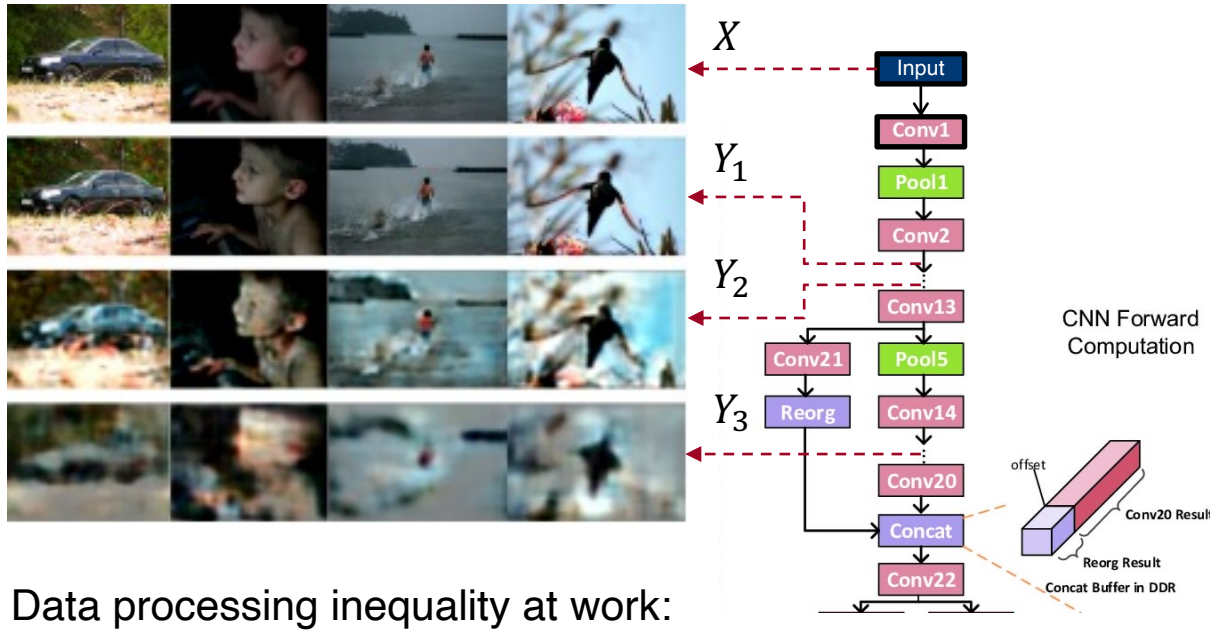
Questions?

PRIVACY

- In many multi-task systems, we code latent-space features
- Are features privacy-preserving?
- Need precise definition of privacy
- Strategies for privacy
  - Adding noise to features
  - Information-theoretic privacy
  - Resilience to model inversion attack

Not compression friendly

Input reconstruction from YOLOv2



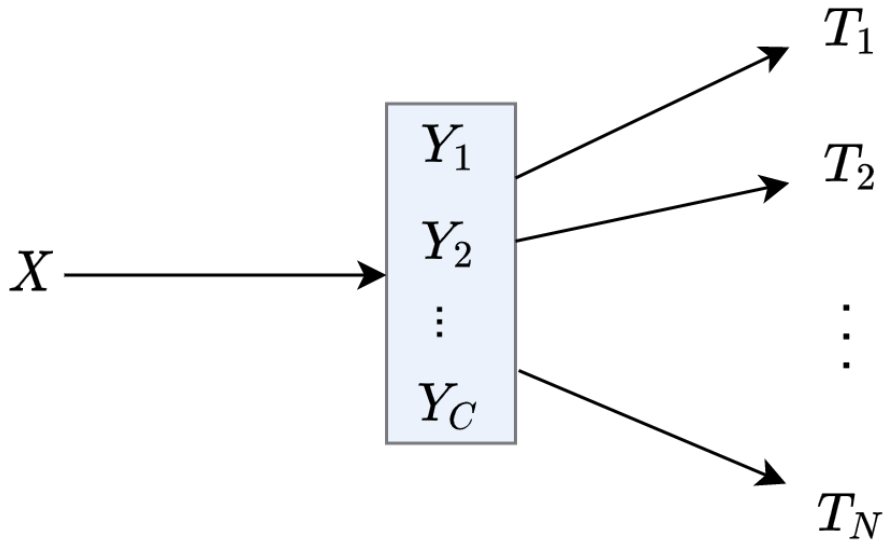
Data processing inequality at work:

$$I(X;Y_1) \geq I(X;Y_2) \geq I(X;Y_3)$$

H. Choi and I. V. Bajić, "Near-lossless deep feature compression for collaborative intelligence," Proc. IEEE MMSP, Aug. 2018.
Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," Proc. 35th Annual Computer Security Applications Conference, p. 148–162, 2019

PRIVACY FAN

- "Privacy fan" – a post-hoc information-theoretic privacy model for multi-task compression
- Start with a pre-trained model
- $Y_1, \ldots, Y_C$ - features
- $T_1, \ldots, T_N$ - tasks
- Some task outputs reveal private information (e.g. input reconstruction), some not
- Let $\mathcal{P}$ be the set of "private" tasks
- Goal: identify a set of features $\mathcal{B}$ that carry minimum information about private tasks, while providing sufficient information about non-private ones



S. R. Alvar and I. V. Bajić, "Scalable privacy in multi-task image compression," Proc. IEEE VCIP, Dec. 2021.

PRIVACY FAN

- Privacy fan formulation

$$\min_{\mathcal{B}} \sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{P}} I(Y_i;T_j)\,, \qquad \text{such that} \sum_{i\in\mathcal{B}} \sum_{j\notin\mathcal{P}} I(Y_i;T_j) \geq R$$

- Solution: define a Lagrangian $\mathcal{L}_i$ for each feature Y_i:

$$\mathcal{L}_i = \sum_{j\in\mathcal{P}} I(Y_i;T_j) - \beta \cdot \sum_{j\notin\mathcal{P}} I(Y_i;T_j)$$

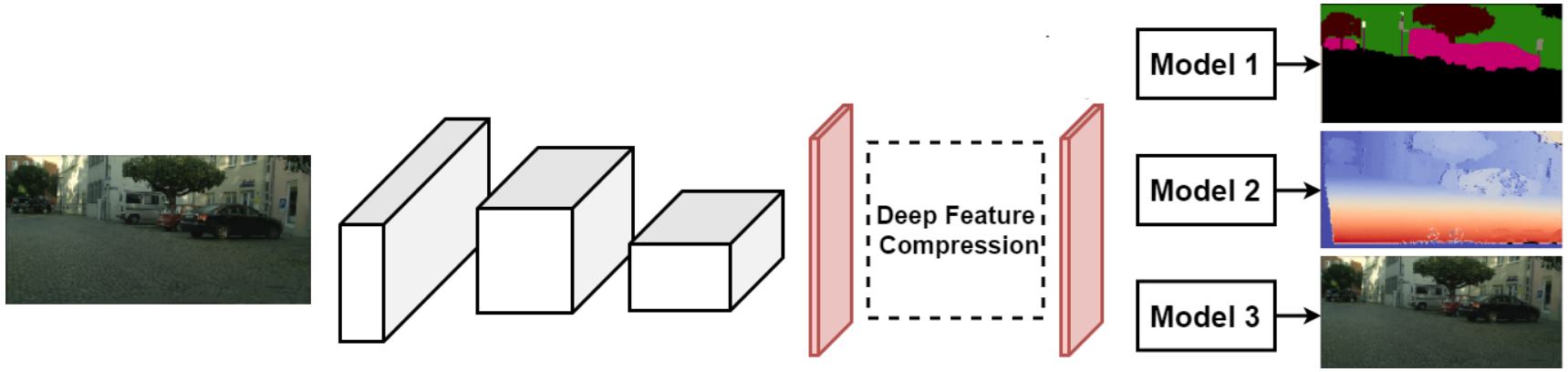
  where $\beta > 0$ is the Lagrange multiplier controlling the privacy-accuracy trade-off

  - $\mathcal{B} = \{Y_i \,:\, \mathcal{L}_i < 0\}$

- Special case, practically important: set $\mathcal{B}$ is limited to C' features: $|\mathcal{B}| \leq C'$

S. R. Alvar and I. V. Bajić, "Scalable privacy in multi-task image compression," Proc. IEEE VCIP, Dec. 2021.

SCALABLE PRIVACY



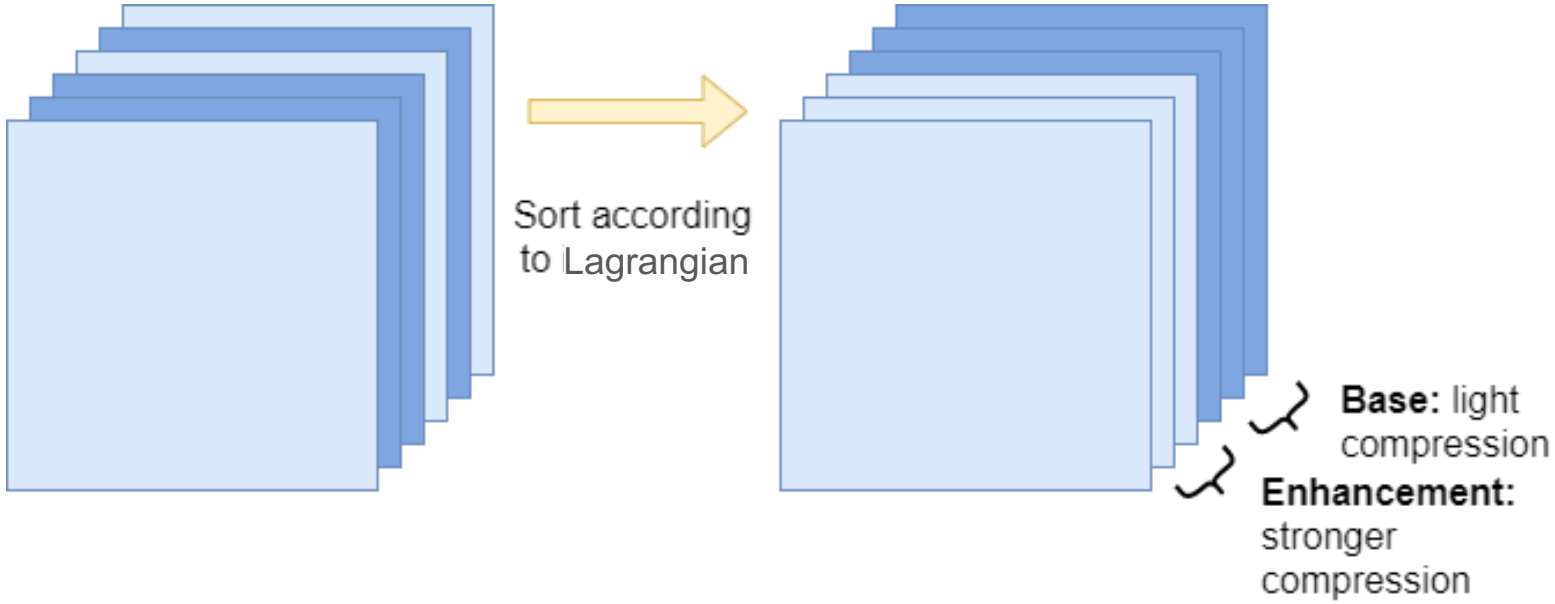
- Lagrangians:

$$\mathcal{L}_i = \underbrace{I(Y_i;T_3)}_{\text{Input reconstruction (private)}} - \beta \cdot \underbrace{[I(Y_i;T_1) + I(Y_i;T_2)]}_{\text{Segmentation and depth est. (non-private)}}$$

- Obtain set $\mathcal{B}$ by solving the privacy fan – call these "base" features

S. R. Alvar and I. V. Bajić, "Scalable privacy in multi-task image compression," Proc. IEEE VCIP, Dec. 2021.

SCALABLE PRIVACY

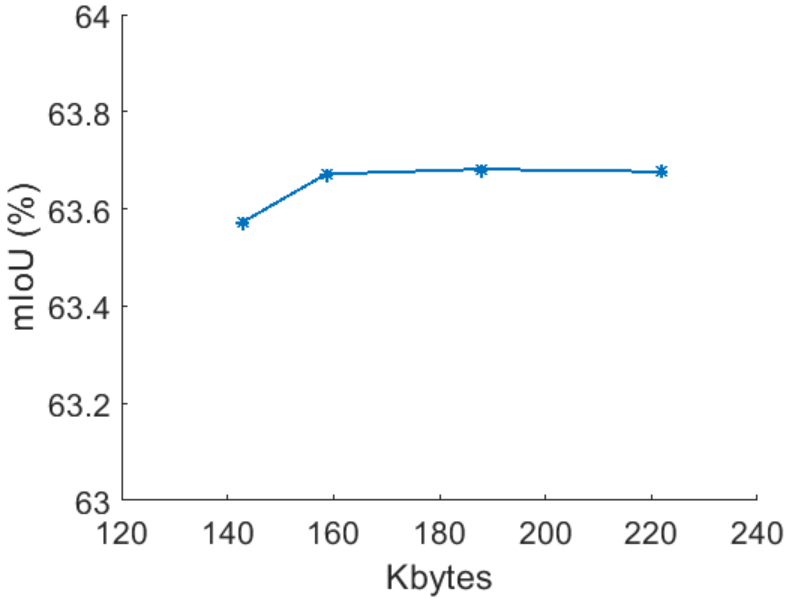


- Encode “base” features at high quality, other (“enhancement”) features at lower quality, depending on the application

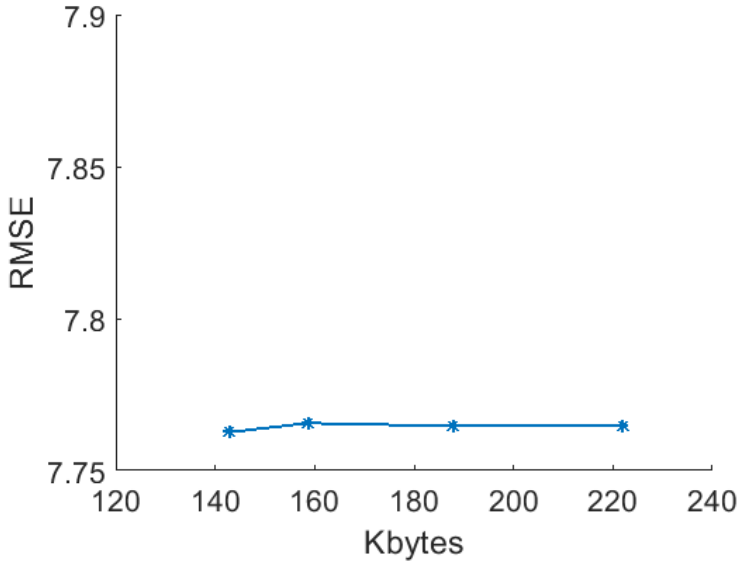
S. R. Alvar and I. V. Bajić, “Scalable privacy in multi-task image compression,” Proc. IEEE VCIP, Dec. 2021.

SCALABLE PRIVACY

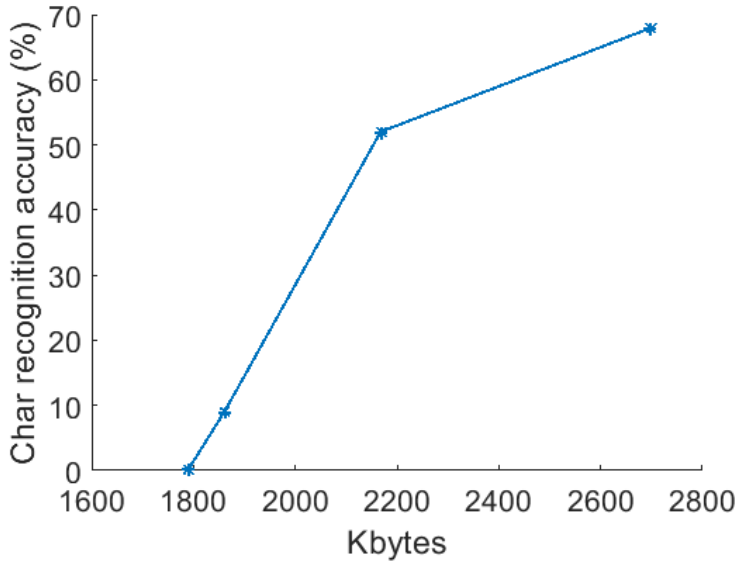
Varying the rate of enhancement layer



Semantic segmentation



Depth estimation



Character recognition

S. R. Alvar and I. V. Bajić, “Scalable privacy in multi-task image compression,” Proc. IEEE VCIP, Dec. 2021.
S. R. Alvar, K. Uyanik, and I. V. Bajić, “License plate privacy in visual analysis of traffic scenes,” to be presented at IEEE MIPR, Aug. 2022.

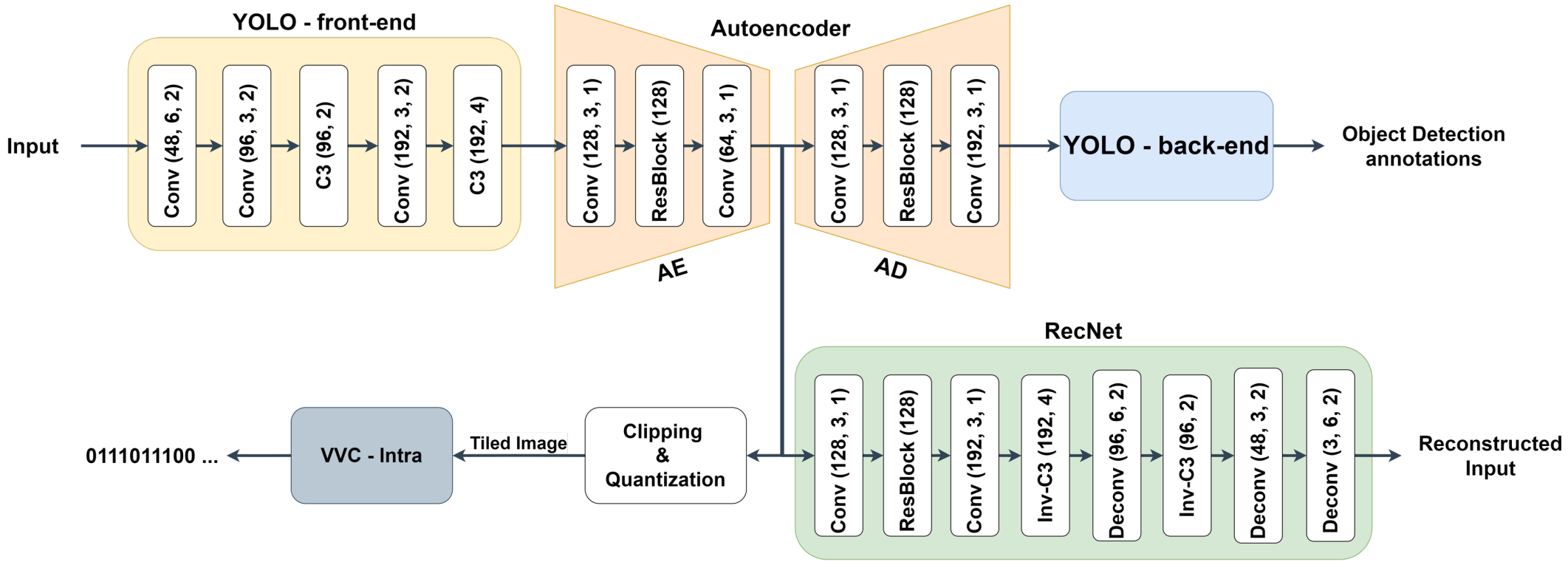
SCALABLE PRIVACY

Varying the rate of enhancement layer

- Segmentation and depth estimation accuracy approximately the same in all cases
- Character recognition accuracy increases with increasing enhancement rate

S. R. Alvar and I. V. Bajić, “Scalable privacy in multi-task image compression,” Proc. IEEE VCIP, Dec. 2021.
S. R. Alvar, K. Uyanik, and I. V. Bajić, “License plate privacy in visual analysis of traffic scenes,” Proc. IEEE MIPR, Aug. 2022.

RESISTANCE AGAINST MODEL INVERSION ATTACK



- Another approach to privacy: train autoencoder to make it more difficult to recover input image from encoded features (model inversion attack)

B. Azizian and I. V. Bajić, "Privacy-preserving feature coding for machines," Picture Coding Symposium (PCS), 2022.

RESISTANCE AGAINST MODEL INVERSION ATTACK

- Reconstruction loss

$$\mathcal{L}_{\text{rec}} = \left\|X - \hat{X}\right\|_1 + \beta \cdot \left(\left\|S_x * \left(X - \hat{X}\right)\right\|_1 + \left\|S_y * \left(X - \hat{X}\right)\right\|_1\right)$$

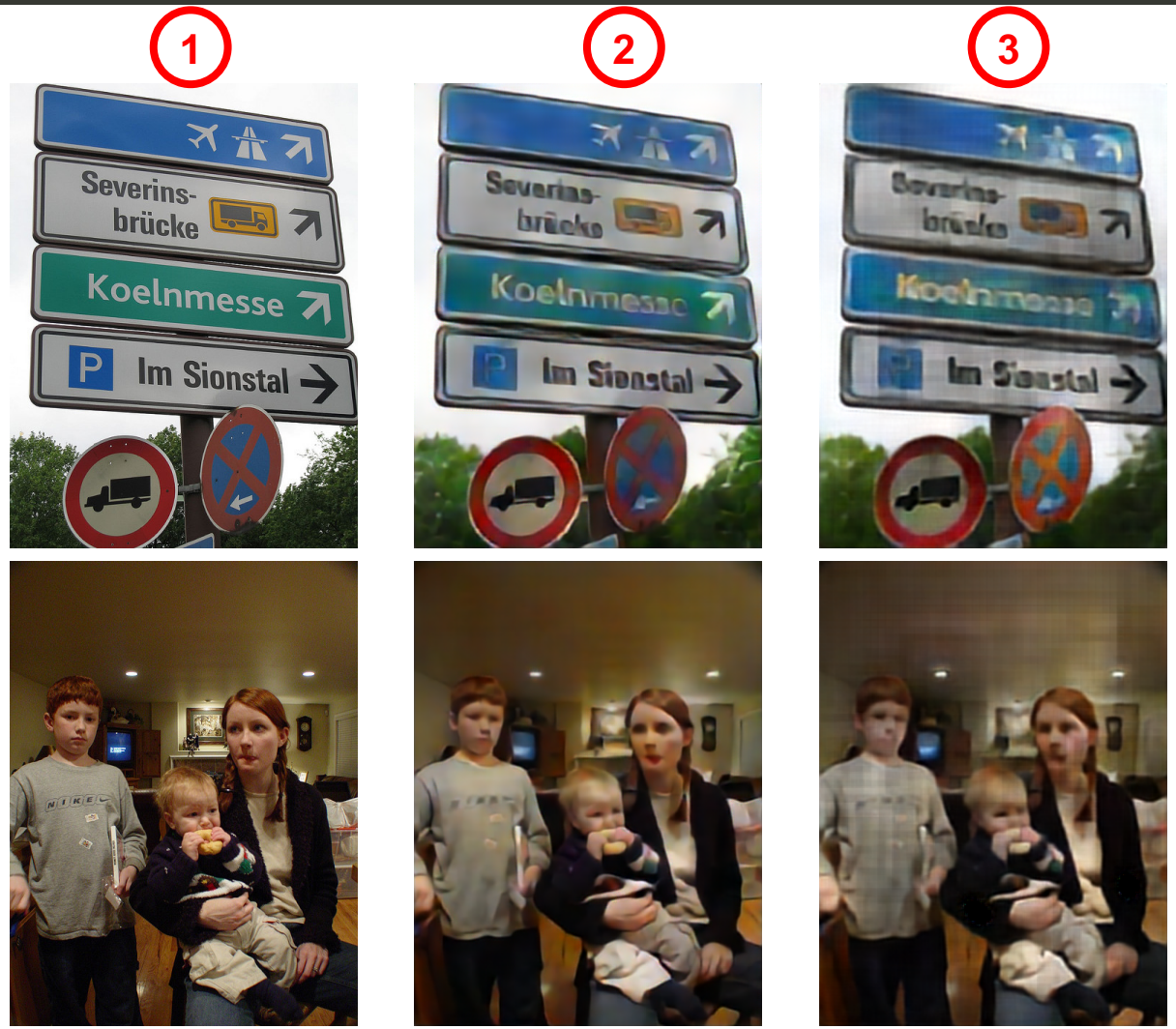
- Autoencoder's loss

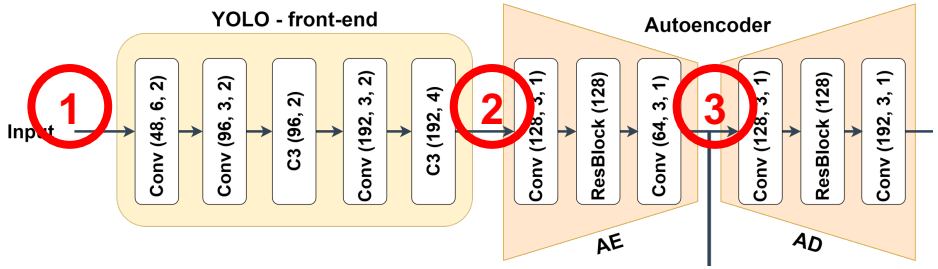
$$\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{obj}} - w \cdot \mathcal{L}_{\text{rec}}$$

- Adversarial training – alternate between:
  - Train decoder using $\mathcal{L}_{\text{rec}}$ (autoencoder frozen) – encourage decoder to be as good as it can on recovering input image, especially edges
  - Train autoencoder using $\mathcal{L}_{\text{AE}}$ (decoder frozen) – penalize encoder if decoder does a good job (reverse sign of $\mathcal{L}_{\text{rec}}$)

B. Azizian and I. V. Bajić, "Privacy-preserving feature coding for machines," Picture Coding Symposium (PCS), 2022.

RESISTANCE AGAINST MODEL INVERSION ATTACK

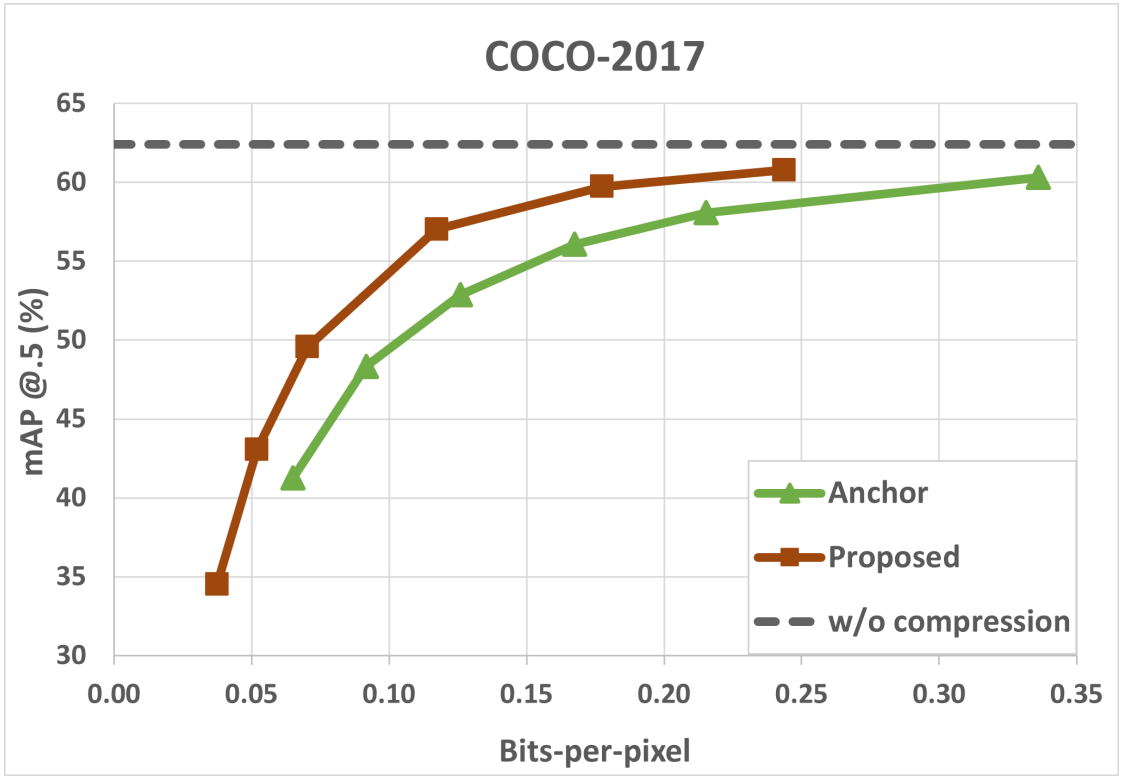




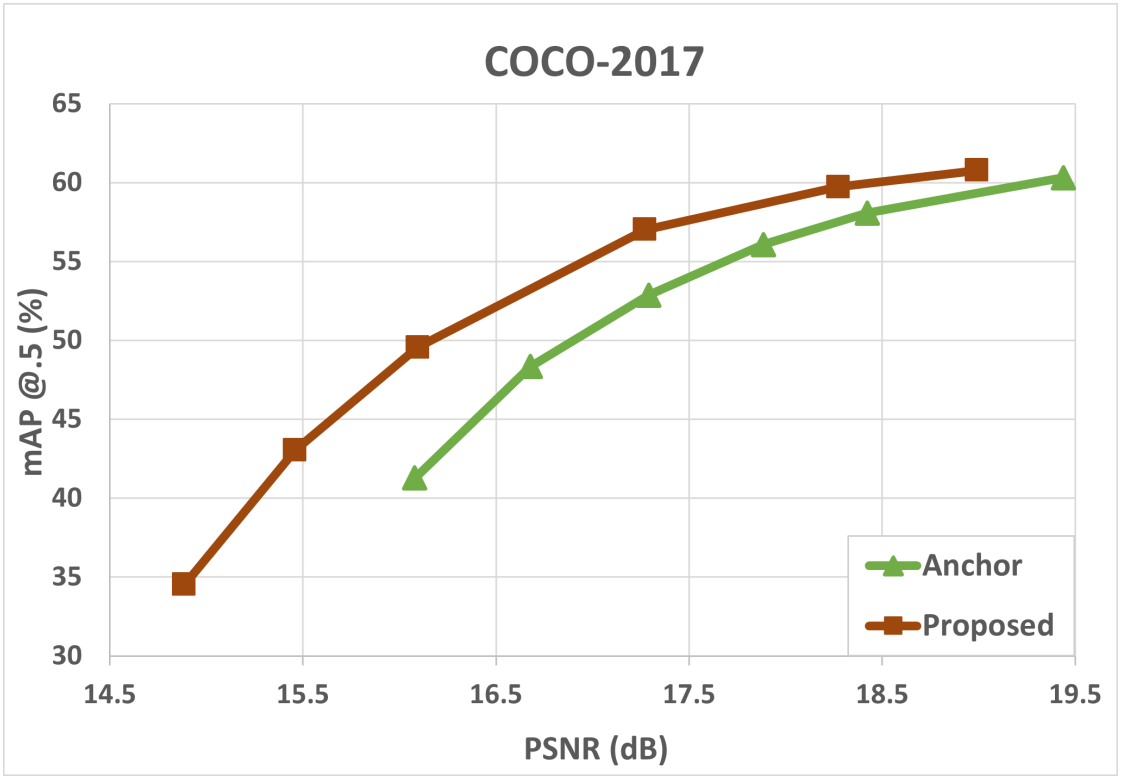
- Showing images recovered from YOLOv5's own features 2 and autoencoder's bottleneck features 3
- Details harder to distinguish in the images recovered from bottleneck features

B. Azizian and I. V. Bajić, "Privacy-preserving feature coding for machines," Picture Coding Symposium (PCS), 2022.

RESISTANCE AGAINST MODEL INVERSION ATTACK



Lower bit rate at the same accuracy (BD-rate = –31.3%)



Lower reconstruction PSNR at the same accuracy (BD-PSNR = –0.76dB)

- Anchor: YOLOv5 features compressed using VVC
- Proposed: AE bottleneck features compressed using VVC

B. Azizian and I. V. Bajić, “Privacy-preserving feature coding for machines,” Picture Coding Symposium (PCS), 2022.

Questions?

Part 3

Standardization

STANDARDIZATION

- Standards are important
  - Ensure interoperability
  - Give developers confidence that their products will have a large market
- There are several standardization activities related to multi-task compression
- We will briefly describe two:
  - JPEG AI (Joint Photographic Experts Group – Artificial Intelligence)
  - MPEG-VCM (Motion Pictures Experts Group – Video Coding for Machines)

ISO/IEC JTC 1/SC29/WG1 N90049, "White Paper on JPEG AI Scope and Framework v1.0," 2021.
W. Gao et al., “Recent standard development activities on Video Coding for Machines,” arXiv:2105.12653, May 2021.

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)

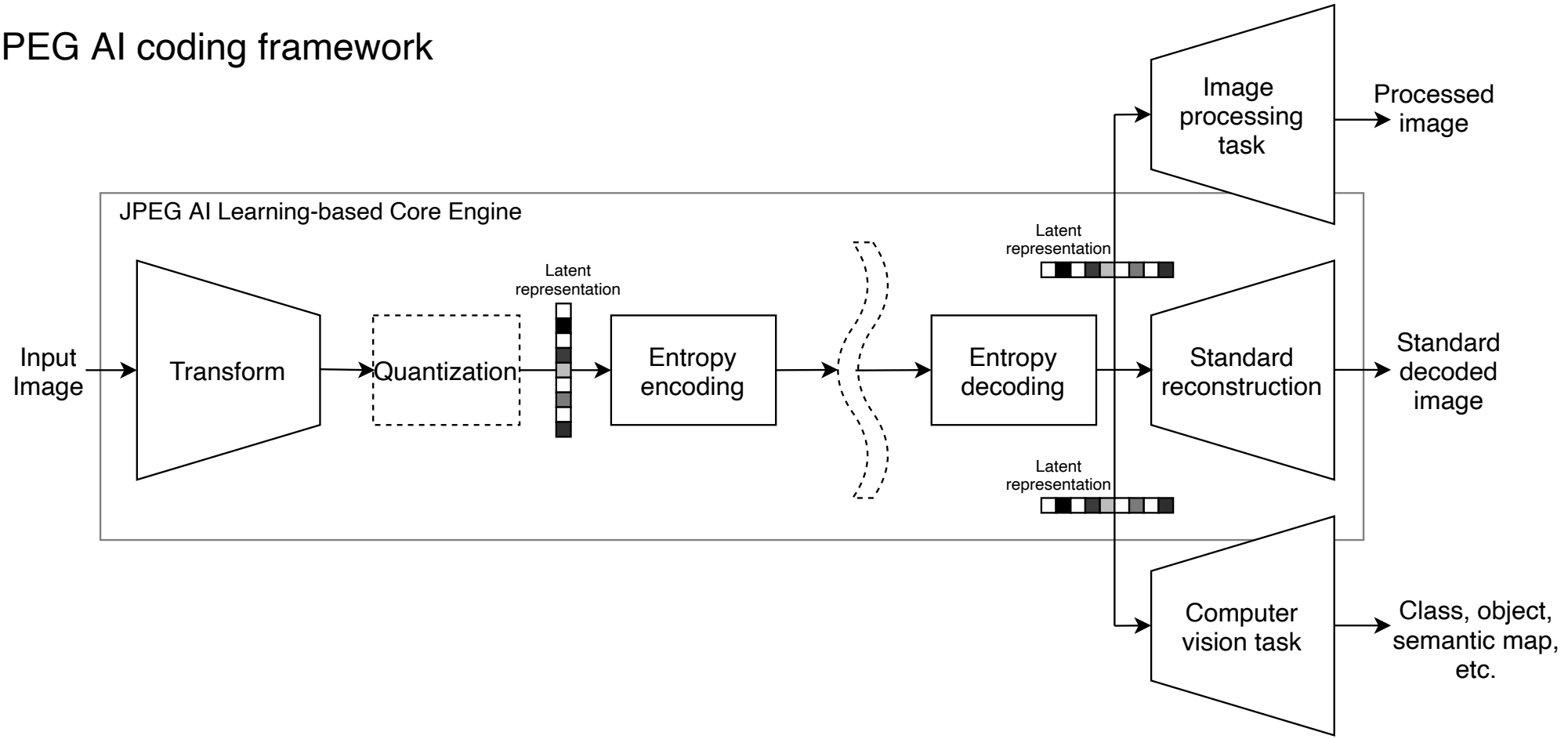
https://jpeg.org/jpegai/
ISO/IEC JTC 1/SC29/WG1 N90049, "White Paper on JPEG AI Scope and Framework v1.0," 2021.

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.

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.

JPEG AI

- Timeline
  - January 2022 – Final Call for Proposals
  - February 2022 – Proposal registration
  - April 2022 – Proposal submission
  - October 2022 – Verification Model under Consideration (VMuC)
  - ...
  - October 2023 – Draft standard
  - April 2024 – Final standard

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

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.

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

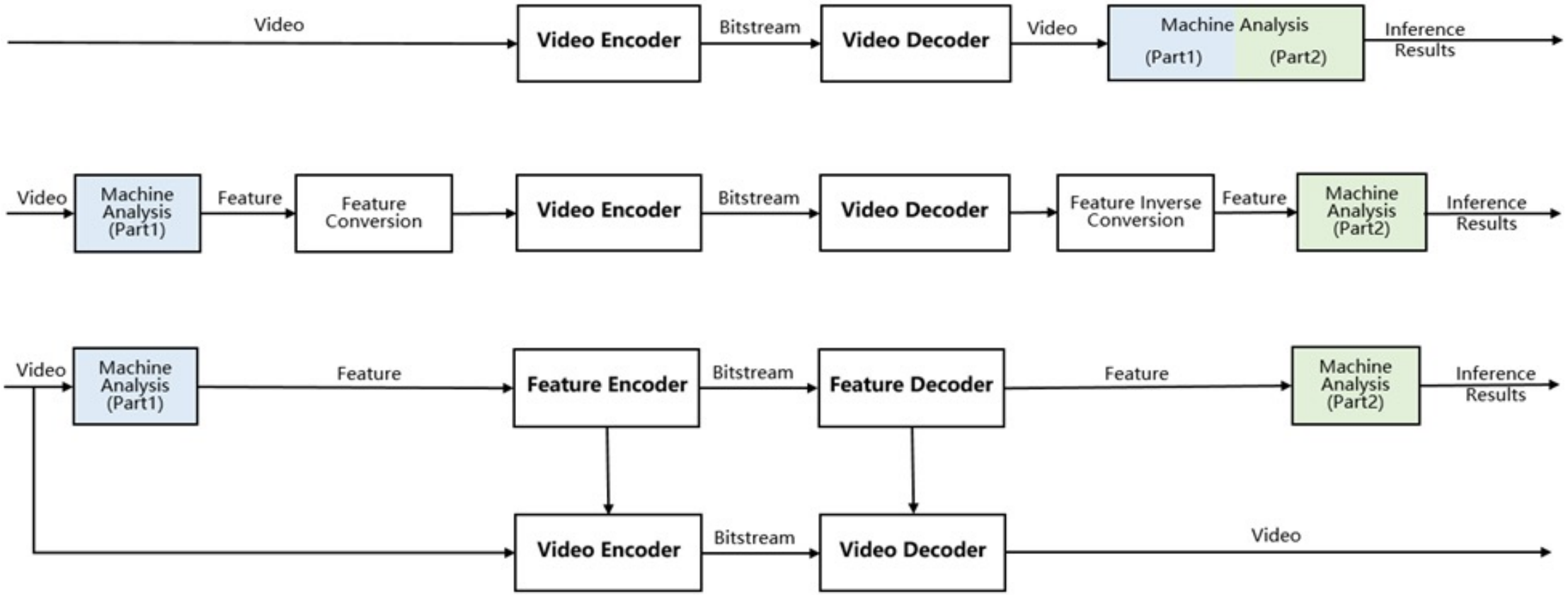
- Track 1 – Feature extraction and compression
  - Focus on machine vision
  - Call for Evidence (CfE): July 2022
  - Response to CfE: October 2022

- Track 2 – Image and video compression
  - Both human and machine vision
  - Call for Proposals (CfP): April 2022
  - Response to CfP: October 2022

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

- Multi-task compression is not new
  - But there are exciting new developments and techniques
  - Some requirements are new (e.g., lower bitrate for machine vision)
- What we have learned:
  - Features produced by neural networks are more compressible than the input
  - Learning-based techniques are good at distinguishing what is relevant for machine vision vs. other information
    - Unified framework for compression and analysis
  - Privacy is an open challenge
    - More work is needed on precise definitions and quantification of privacy in the context of multi-task compression
  - Related standardization activities: JPEG AI and MPEG-VCM

Thank you!

Questions?