Learned Disentangled Latent Representations for Scalable Image Coding for Humans and Machines

Ezgi Ozyilkan[†,*,¶], Mateen Ulhaq[‡,*,¶], Hyomin Choi[*], Fabien Racapé[*]

¶ Joint first authors.

† Dept. of Electrical and Computer Engineering, New York University
‡ School of Engineering Science, Simon Fraser University
∗ InterDigital – Emerging Technologies Lab

ezgi.ozyilkan@nyu.edu, mulhaq@sfu.ca,
{hyomin.choi, fabien.racape}@interdigital.com

This work was done while E. Ozyilkan and M. Ulhaq were interns at InterDigital.

Contents

Traditional Transform Coding: JPEG

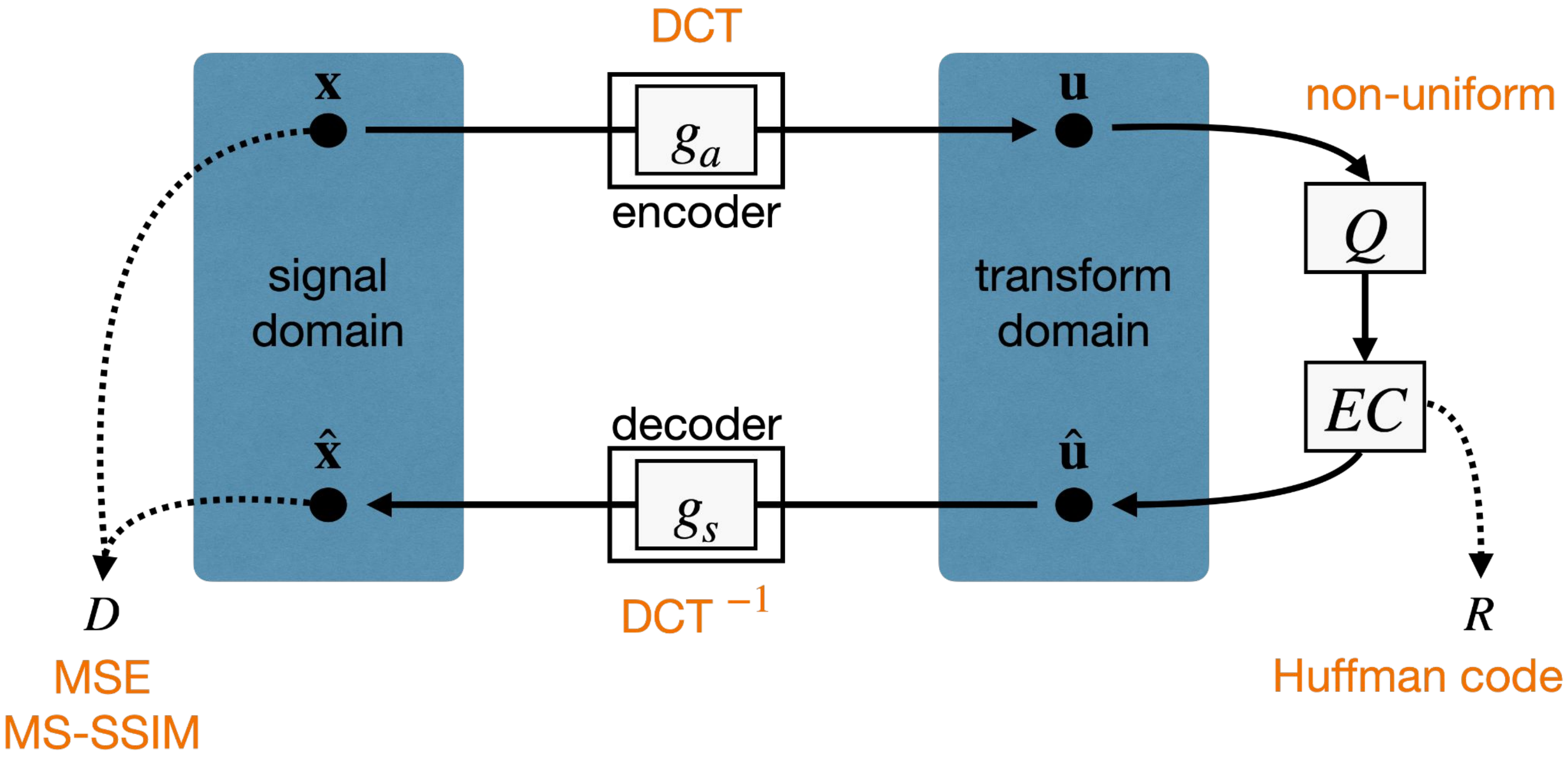


Figure adapted from [Ballé et al.], "End-to-end optimized image compression," *ICLR,* 2017.

Nonlinear Transform Coding

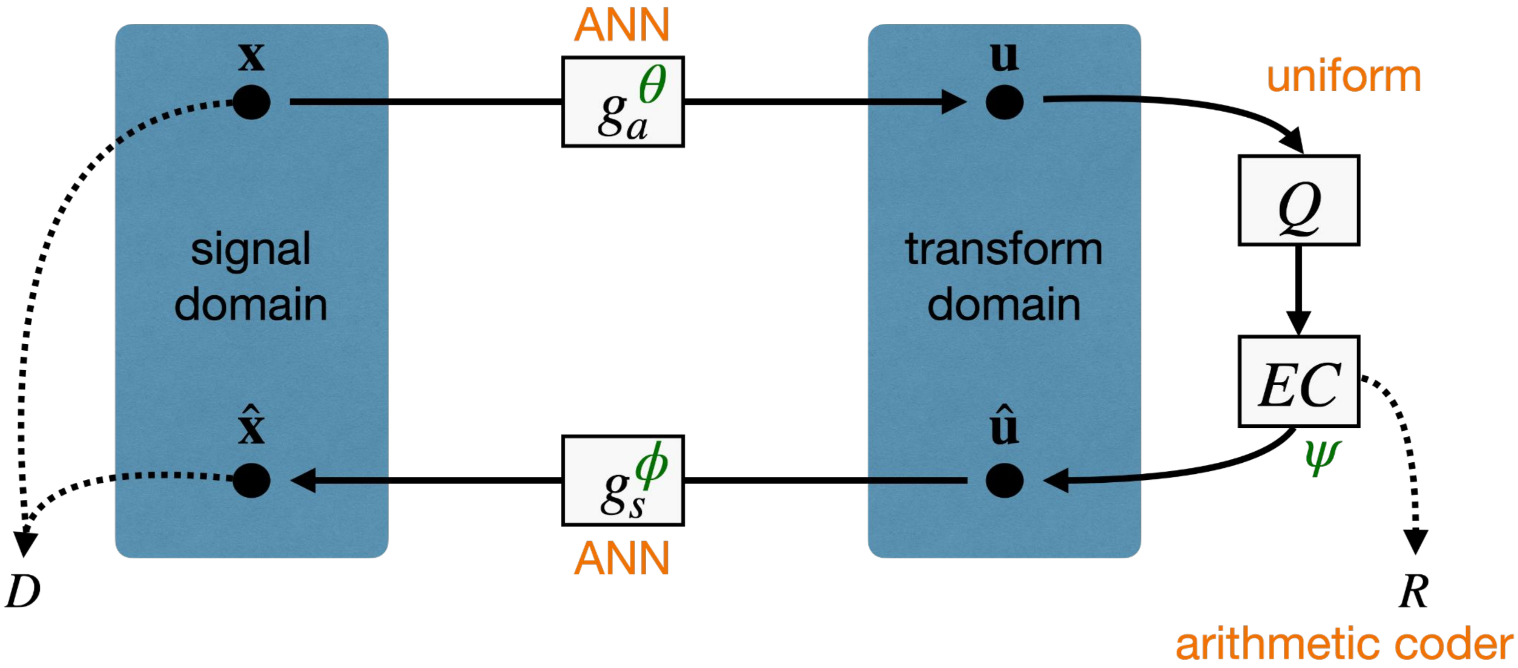
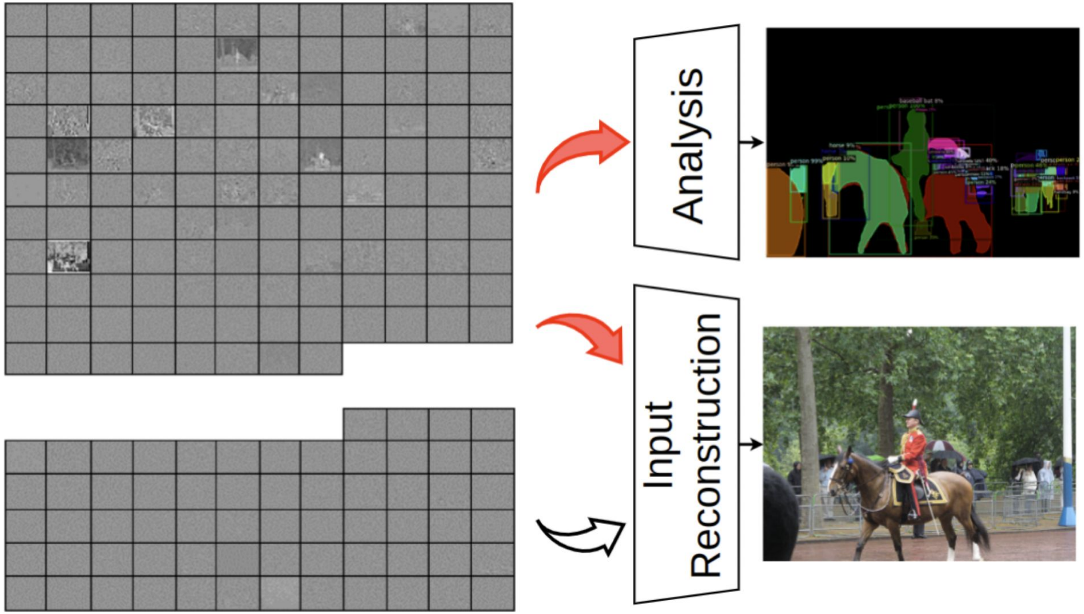


Figure adapted from [Ballé et al.], "End-to-end optimized image compression," *ICLR*, 2017.

Multi-Task Image Coding

Split transform domain latent space for machine analytics. $\propto$ MPEG's "**V**ideo **C**oding for **M**achines"



$\Rightarrow$ Improvement in bitrate for {**reconstruction**, **analytics**} while not sacrificing {**distortion**, **accuracy**}.

Figure courtesy of [Choi et al.], "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing*, 2022.

Scalable Image Coding

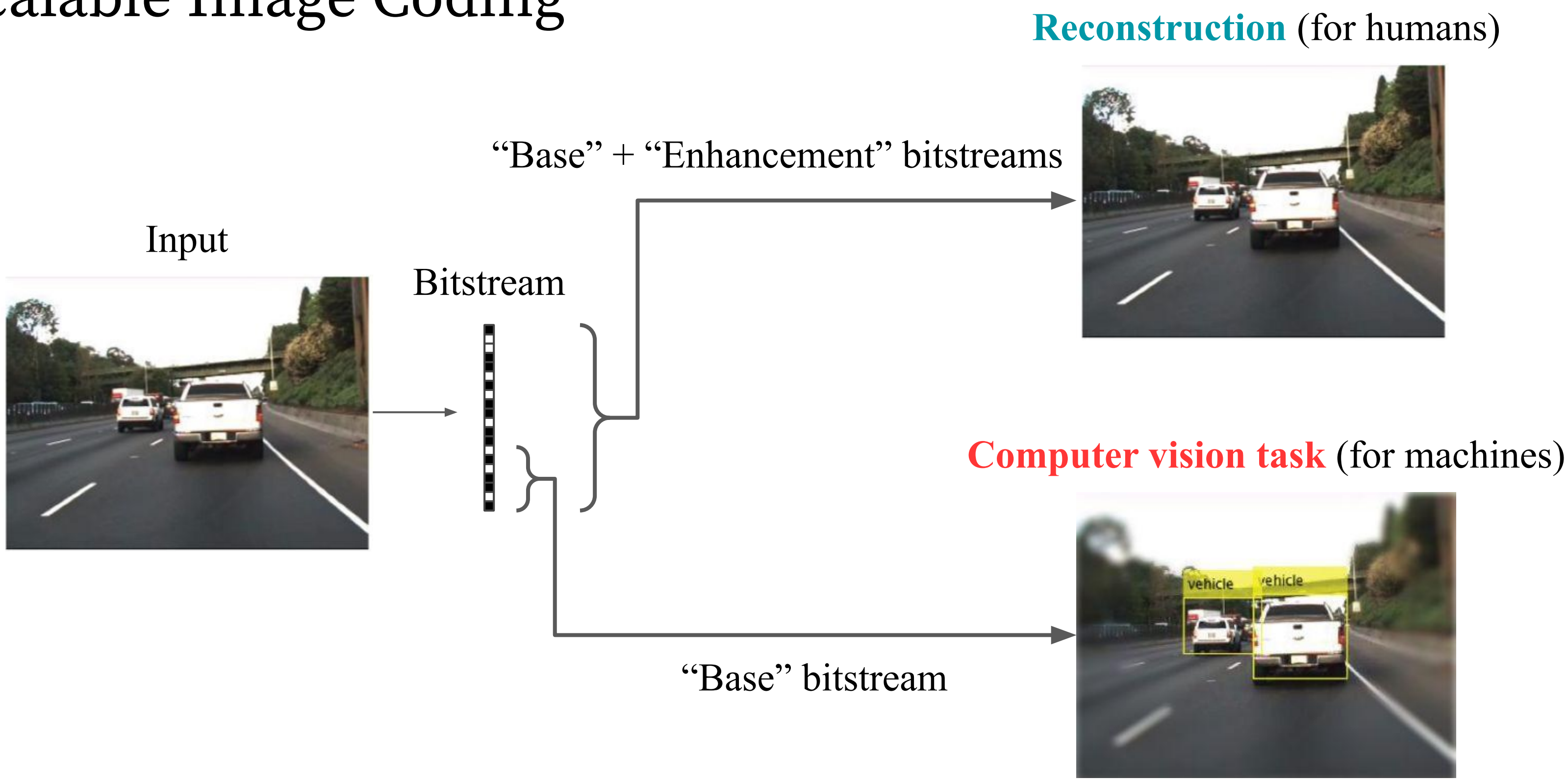
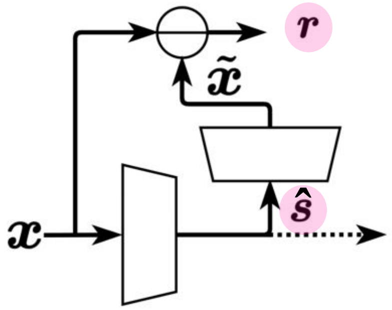


Image taken from https://www.mathworks.com/discovery/object-detection.html

Prior Work

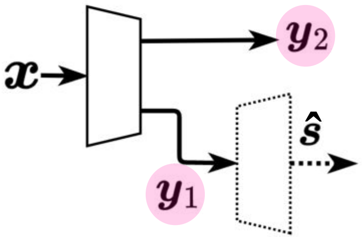
Yan et al.



(a)

"Enhancement" is residual error in reconstructing from "base".

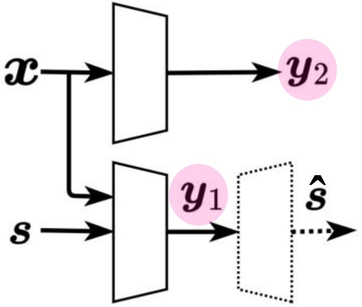
Choi et al.



(b)

"Base" and "enhancement" obtained from same transform on x.

<u>Proposed</u>



(c)

"Base" from x, s and "enhancement" only from x.

Transmitted bitstreams are highlighted.

[Yan et al.], "SSSIC: Semantics-to-signal scalable image coding with learned structural representations," *IEEE Transactions on Image Processing,* 2021.

[Choi et al.], "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing,* 2022.

Proposed Scalable Image Coding Framework

Idea: Learned Disentangled Latent Spaces

Motivation is to have little (or none!) excess rate: $I(\boldsymbol{y}_1; \boldsymbol{y}_2) \approx 0\,.$

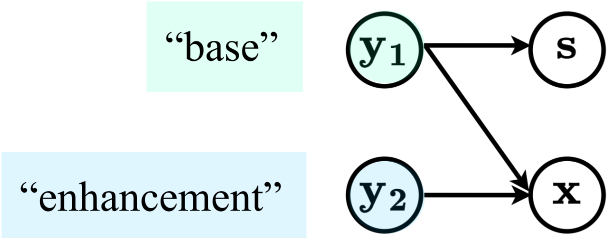
Proposed approach is based on variational inference.

$$
\begin{aligned}
p_\theta(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{y}_1, \boldsymbol{y}_2) &= p(\boldsymbol{y}_1)\, p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1)\, p_\theta(\boldsymbol{x} \mid \boldsymbol{y}_1, \boldsymbol{y}_2)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{x}) && \text{by chain rule} \\
&= p(\boldsymbol{y}_1)\, p(\boldsymbol{y}_2)\, p_\theta(\boldsymbol{x} \mid \boldsymbol{y}_1, \boldsymbol{y}_2)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{y}_1) && \text{since } \boldsymbol{y}_1 \perp\!\!\!\perp \boldsymbol{y}_2 \\
& && \text{and } (\boldsymbol{s} \perp\!\!\!\perp \boldsymbol{y}_2) \mid \boldsymbol{y}_1
\end{aligned}
$$

The data likelihood is given by integrating:

$$
p_\theta(\boldsymbol{x}, \boldsymbol{s}) = \iint p_\theta(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{y}_1, \boldsymbol{y}_2)\; d\boldsymbol{y}_1 d\boldsymbol{y}_2
$$

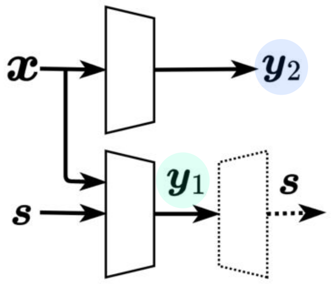
Unfortunately, intractable!!



graphical model

Overcoming Intractability

Introduce approximate posterior.



<u>Proposed</u>

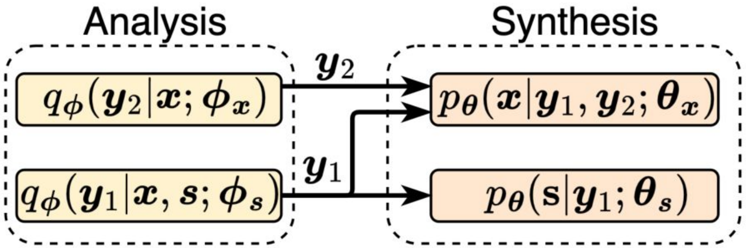
$$q_\phi(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{x}, \boldsymbol{s}) = \underbrace{q_\phi(\boldsymbol{y}_1 \mid \boldsymbol{x}, \boldsymbol{s})}_{\substack{\boldsymbol{y}_1 \text{ derived} \\ \text{from } \boldsymbol{x}, \boldsymbol{s}}} \underbrace{q_\phi(\boldsymbol{y}_2 \mid \boldsymbol{x})}_{\substack{\boldsymbol{y}_2 \text{ derived} \\ \text{from } \boldsymbol{x}}}$$

Impose above factorization by system model.

Loss function construction turns out to be similar to Ballé et al. (2018).

We seek to minimize Kullback-Leibler (KL) divergence between q_ϕ, p_θ.

[Ballé et al.], "Variational Image Compression with a Scale Hyperprior," *ICLR*, 2018.



Minimize KL between q_ϕ, p_θ over dataset of $\boldsymbol{x}, \boldsymbol{s}$:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}\sim p(\boldsymbol{x},\boldsymbol{s})}\left[D_{\mathrm{KL}}\left(q_\phi(\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2 \mid \boldsymbol{x},\boldsymbol{s}) \,\|\, p_\theta(\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2 \mid \boldsymbol{x},\boldsymbol{s})\right)\right]$$

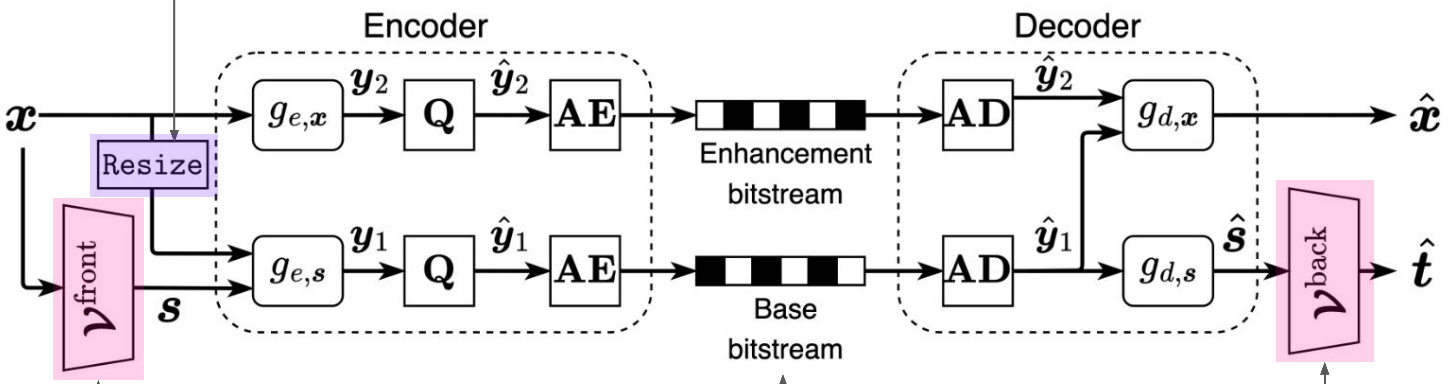
$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}\sim p(\boldsymbol{x},\boldsymbol{s})}\mathbb{E}_{\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2\sim q_\phi}\Bigg[\Big(\overbrace{\log q_\phi(\tilde{\boldsymbol{y}}_1 \mid \boldsymbol{x},\boldsymbol{s};\phi_{\boldsymbol{s}})}^{0} + \overbrace{\log q_\phi(\tilde{\boldsymbol{y}}_2 \mid \boldsymbol{x};\phi_{\boldsymbol{x}})}^{0}\Big)$$

$$-\Big(\underbrace{\log p_\theta(\boldsymbol{x} \mid \tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2;\theta_{\boldsymbol{x}})}_{D_{\boldsymbol{x}}} + \underbrace{\log p_\theta(\boldsymbol{s} \mid \tilde{\boldsymbol{y}}_1;\theta_{\boldsymbol{s}})}_{D_{\boldsymbol{s}}} + \underbrace{\log p(\tilde{\boldsymbol{y}}_1)}_{R_{y_1}} + \underbrace{\log p(\tilde{\boldsymbol{y}}_2)}_{R_{\boldsymbol{y}_2}}\Big)\Bigg] + \text{const.}$$

$$\mathcal{L} = R_{\boldsymbol{y}_1} + R_{\boldsymbol{y}_2} + \lambda \cdot D_{\boldsymbol{x}} + \gamma \cdot D_{\boldsymbol{s}}$$

Proposed Architecture

`Resize` to match latent dimensions.



Features are generated from "front" half of task model.

Model is able to create a more task-optimized bitstreams.

Features are fed into "back" half of task model.

Experimental Setup

of channels for $\boldsymbol{y}_1$

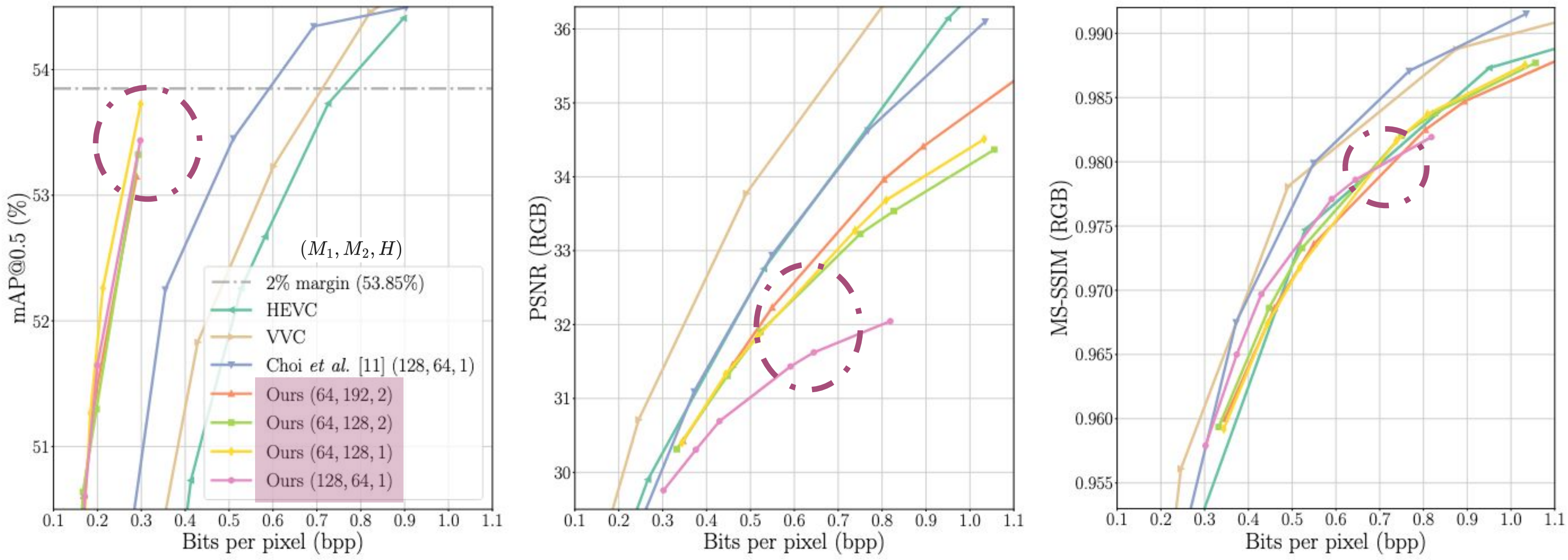
# of channels for $\boldsymbol{y}_2$

# of hyperprior blocks

- Various architecture configurations for the tuple (M_1, M_2, H).
- Train on Vimeo-90K dataset with "distortion" computed using MSE.
- Evaluate object detection on COCO 2014 validation dataset.
- Evaluate input reconstruction on Kodak dataset using MSE and MS-SSIM.
- Benchmark performance in comparison with:
  - Standard codecs such as HEVC, VVC $\Rightarrow$ do not support task-scalability!
  - Comparative scalable image compression model from Choi et al.

[HEVC] http://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.20+SCM-8.8/
[VVC] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-12.3/
[Vimeo-90K] Xue et al. "Video Enhancement with Task-Oriented Flow," *IJCV*, 2019.
[COCO 2014] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," 2014.
[Kodak] http://r0k.us/graphics/kodak/
[MS-SSIM] Z. Wang et al., "Multiscale structural similarity for image quality assessment," *Asilomar Conf. Signals, Systems, and Computers,* 2003.
[Choi et al.] "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing*, 2022.

Performance Across Various Metrics



(a)
Task accuracy on COCO 2014 val
mAP (IoU=0.5) vs. bpp

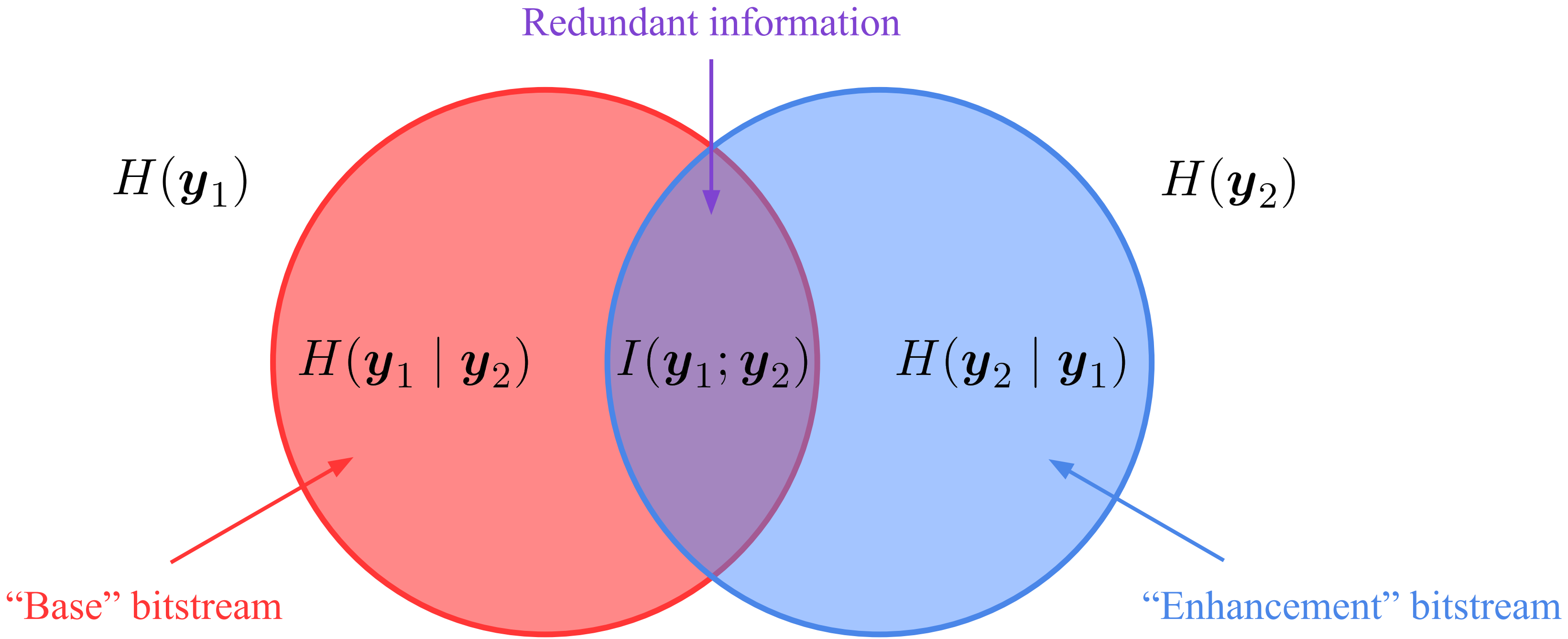
(b)
Input reconstruction on Kodak
PSNR vs. bpp

(c)
Input reconstruction on Kodak
MS-SSIM vs. bpp

Baseline accuracy of YOLOv3 on COCO 2014 val, including JPEG-compressed images, is 55.85% mAP at 4.80 bpp.

Insights into Information Flow

Quick Recap of Entropy and Mutual Information

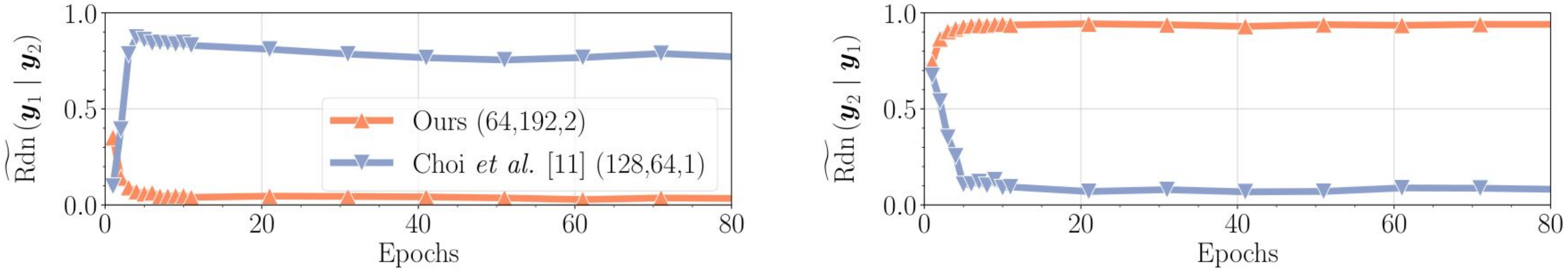


Redundancy During Training

Definition: $\mathrm{Rdn}\left(\boldsymbol{y}_i \mid \boldsymbol{y}_j\right) \triangleq \frac{I(\boldsymbol{y}_i;\boldsymbol{y}_j)}{H(\boldsymbol{y}_i)} = 1 - \frac{H(\boldsymbol{y}_i|\boldsymbol{y}_j)}{H(\boldsymbol{y}_i)}$

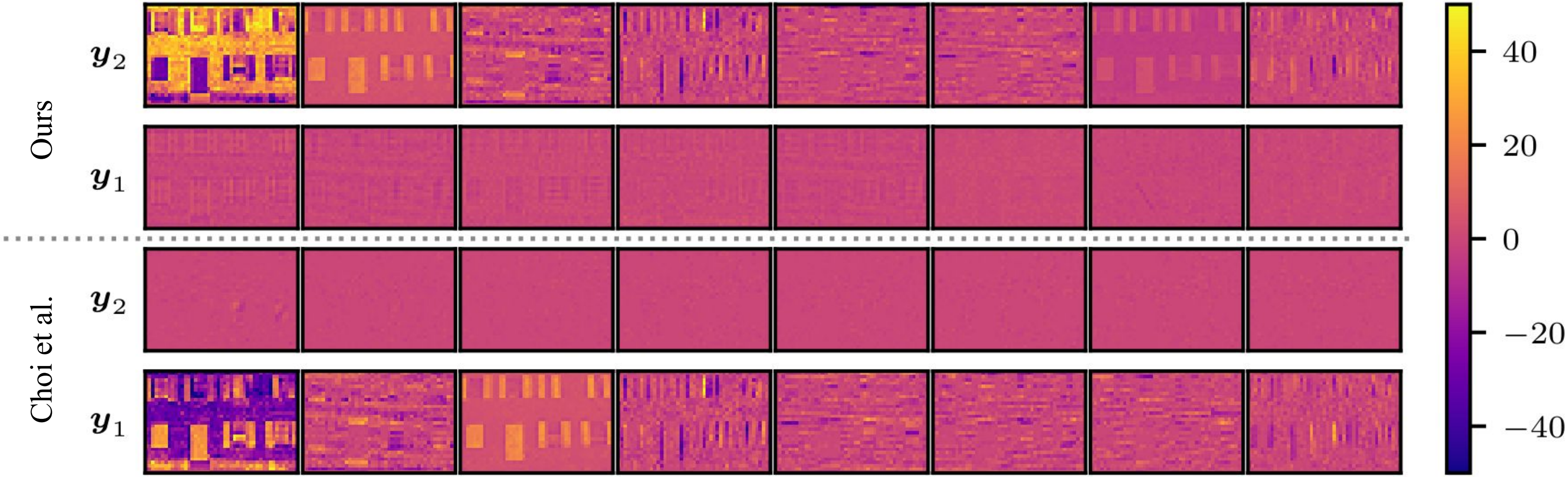
Estimate: $\widetilde{\mathrm{Rdn}}\left(\boldsymbol{y}_i \mid \boldsymbol{y}_j\right) = 1 - \frac{1}{H(\boldsymbol{y}_i)} \sum_{k\in\{1,\ldots,K\}} p(k)\, H(\bar{\boldsymbol{y}}_i \mid c(\bar{\boldsymbol{y}}_j) = k)$

Clustering function



Evolution of the redundancy metrics during training.

Feature Maps



top-8 channels ordered by rate

y_1 = "base" (for machine vision)
y_2 = "enhancement" (for humans)

Conclusion

- Learned image codec with a new formulation.
  - Offers latent-space scalability for human and machine tasks.
  - New way of disentangling the learned representations.
- Significant bit reductions at the “base” layer.
- Needs further investigation about improving reconstruction quality while maintaining the analytics performance.

Questions ?