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 in a nutshell

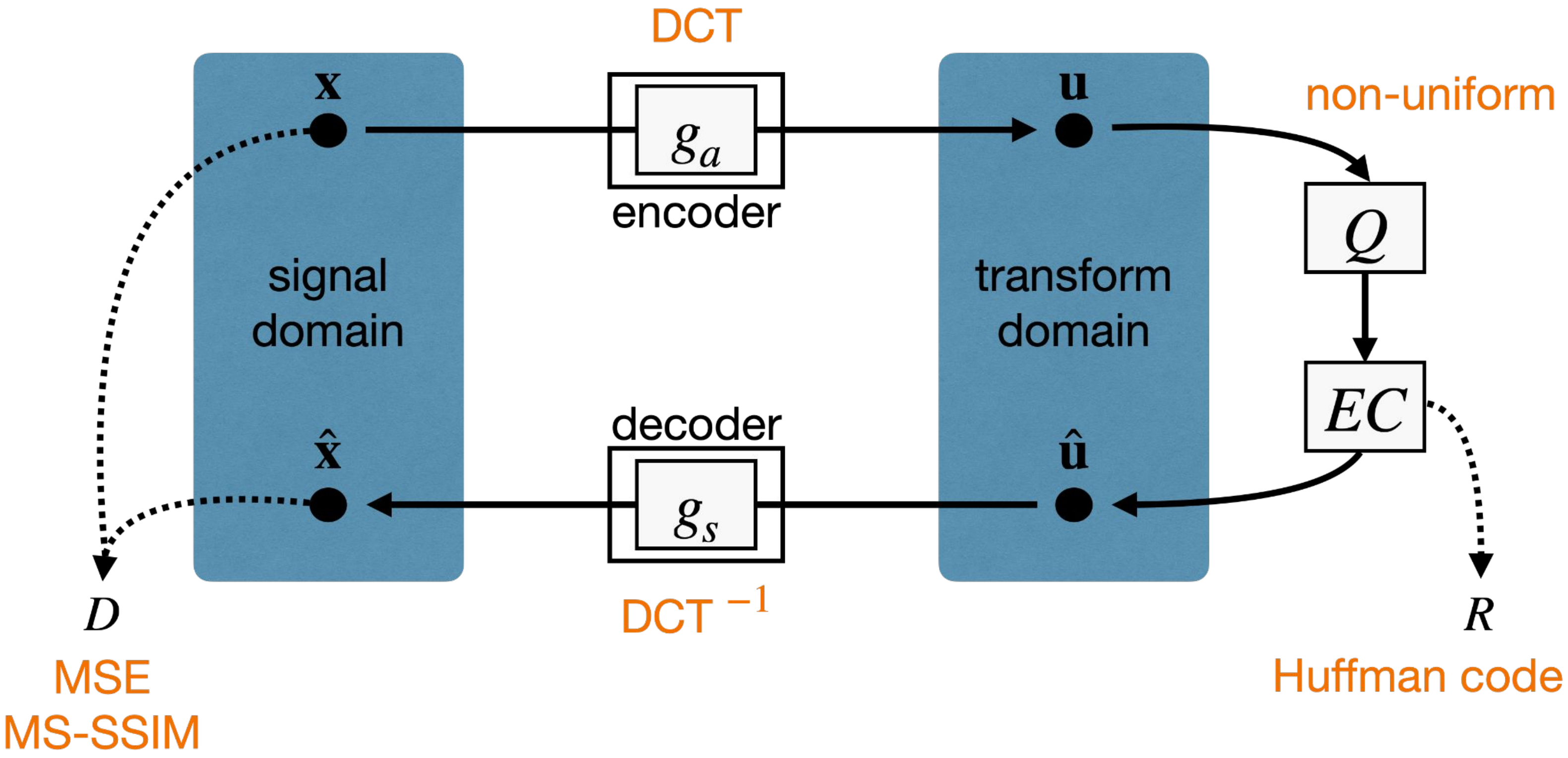


Figure adapted from [J. Ballé et al.], "End-to-end optimized image compression," ICLR, 2017.

Nonlinear Transform Coding

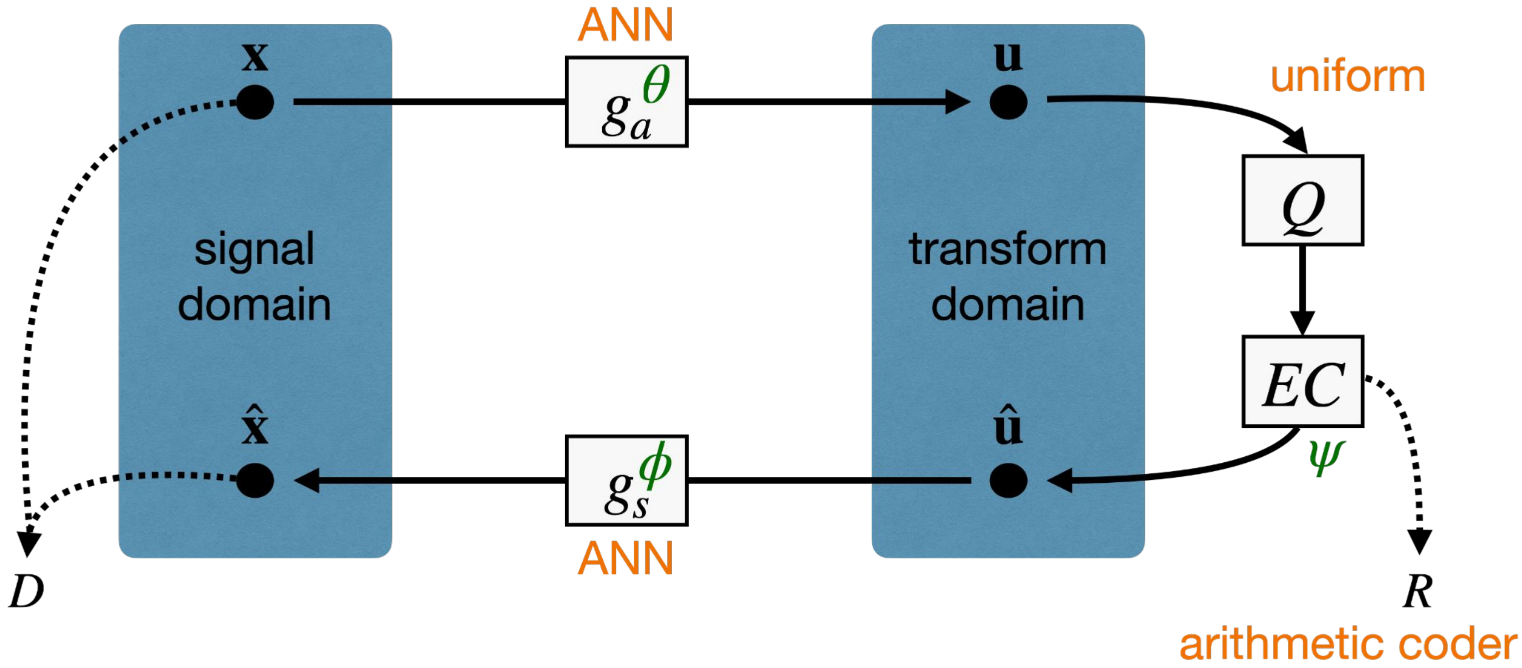
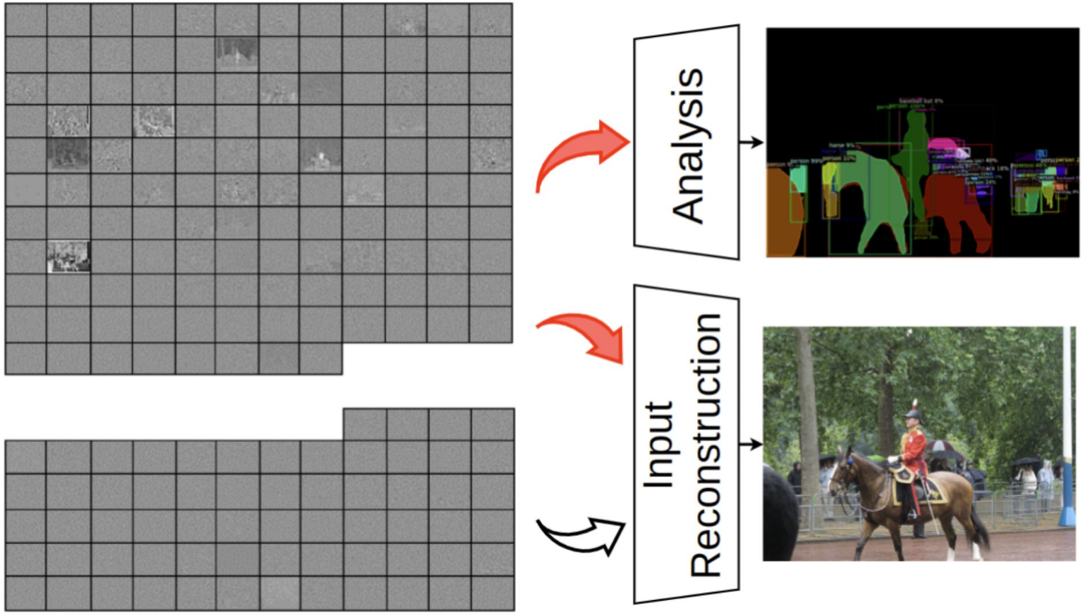


Figure adapted from [J. Ballé et al.], "End-to-end optimized image compression," ICLR, 2017.

Multi-Task Image Coding

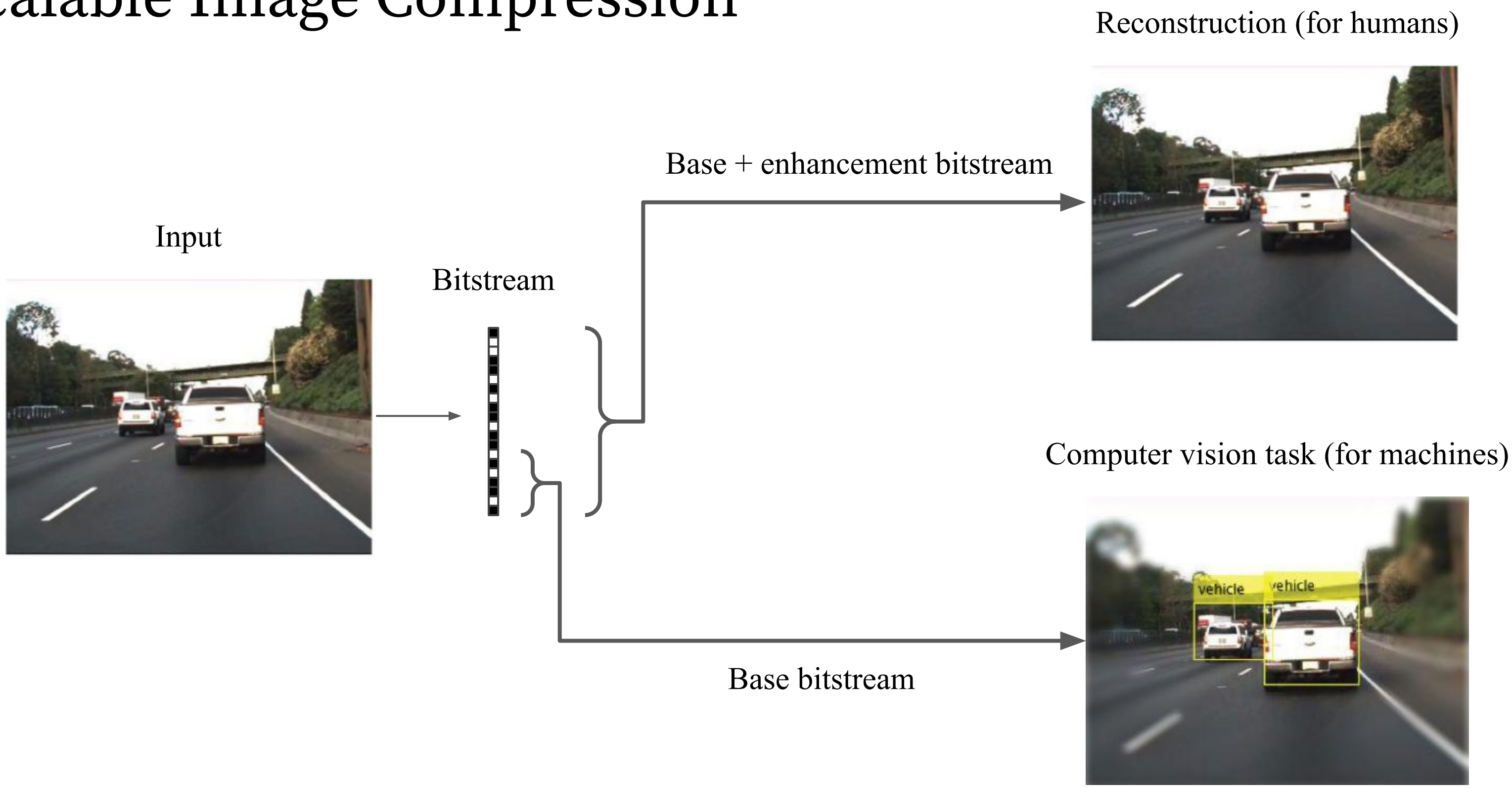
Split transform domain latent space for machine analytics. “**V**ideo **C**oding for **M**achines” (VCM).



⇒ Improvement in bitrate for analytics tasks without reduction in accuracy.

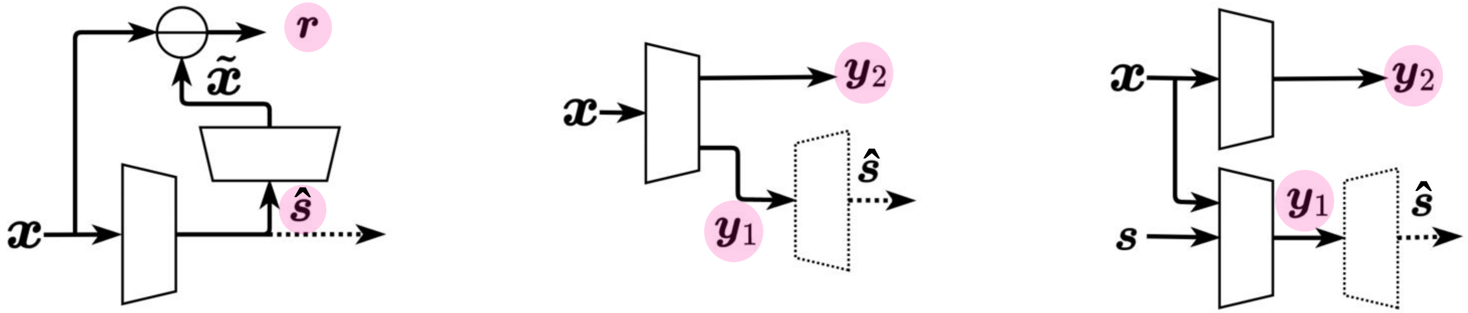
Figure courtesy of [H. Choi et al.], “Scalable Image Coding for Humans and Machines,” *IEEE Transactions on Image Processing*, 2022.

Scalable Image Compression



Prior Work

Chamain et al. | Choi et al. | Proposed



(a) "Enhancement" is residual error in reconstructing from "base".

(b) "Base" and "enhancement" obtained from same transform on x.

(c) "Base" from x, s and "enhancement" only from x.

Transmitted bitstreams are highlighted.

[Yan et al.], "End-to-end optimized image compression for machines, a study," *DCC,* 2021.
[Choi et al.], "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing,* 2022.

Prior Work: Choi et al.

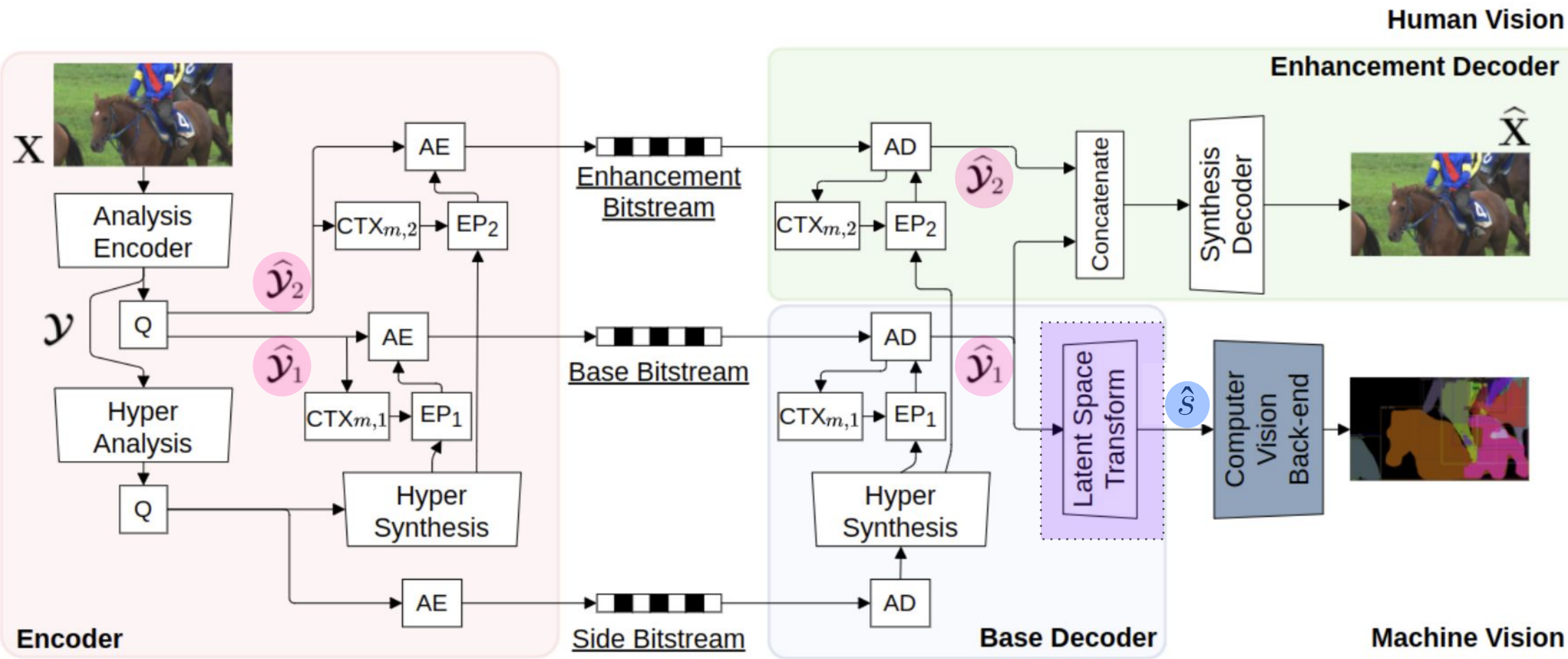


Figure adapted from [H. Choi et al.], "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing*, 2022.

interdigital.

Idea: Learned Disentangled Latent Spaces

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

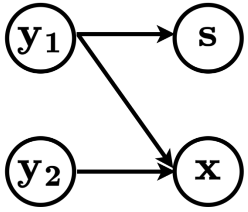
Proposed approach is based on variational inference.

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

The data likelihood is given by integrating:

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

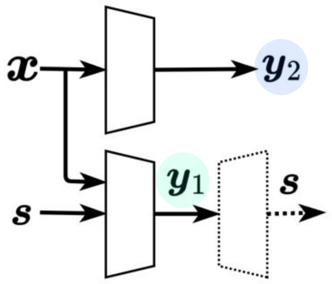
Unfortunately, intractable!



graphical model

Overcoming Intractability

Introduce variational posterior.



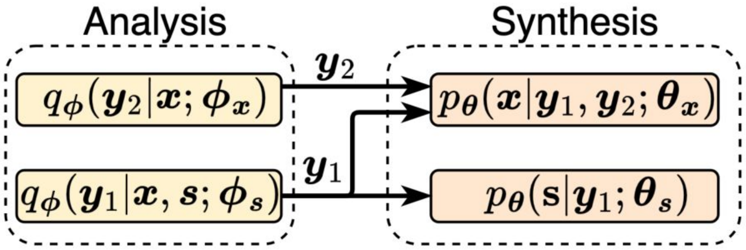
$$q_\phi(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{x}, \boldsymbol{s}) = \underbrace{q_\phi(\boldsymbol{y}_1 \mid \boldsymbol{x}, \boldsymbol{s})}_{\substack{\boldsymbol{y}_1 \text{ derived} \\ \text{from } \boldsymbol{x}, \boldsymbol{s}}} \underbrace{q_\phi(\boldsymbol{y}_2 \mid \boldsymbol{x})}_{\substack{\boldsymbol{y}_2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

Impose above factorization by system model.

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

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

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



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

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

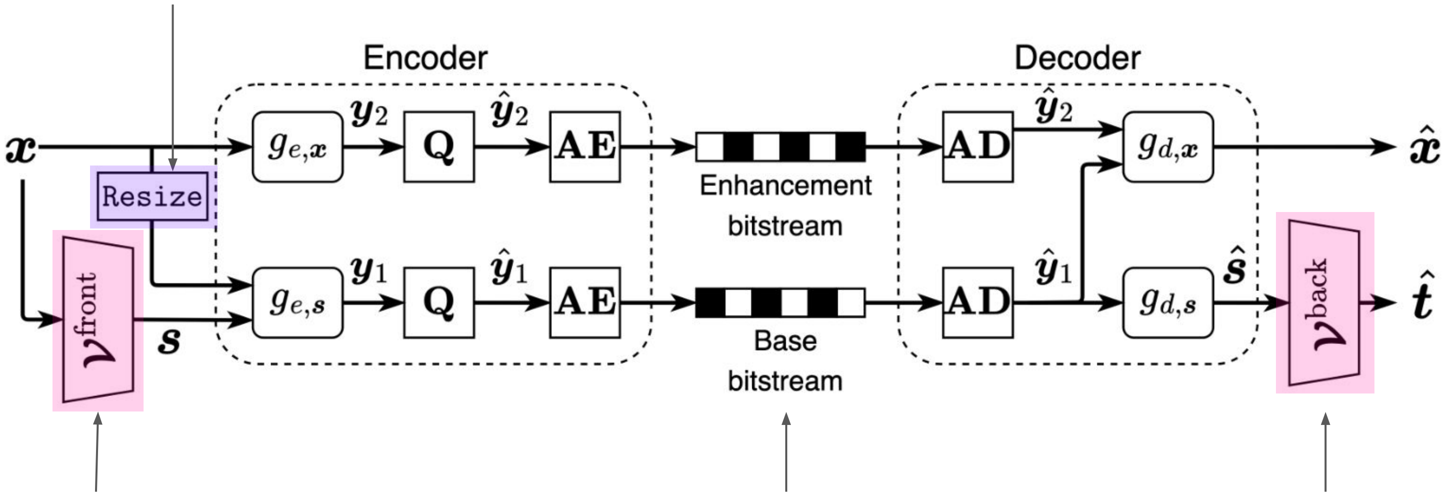
$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}\sim p(\boldsymbol{x},\boldsymbol{s})}\mathbb{E}_{\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2\sim q_\phi}\Bigg[\Bigg(\overbrace{\log q_\phi(\tilde{\boldsymbol{y}}_1 \mid \boldsymbol{x},\boldsymbol{s};\phi_{\boldsymbol{s}})}^{0} + \overbrace{\log q_\phi(\tilde{\boldsymbol{y}}_2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

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

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

Proposed Architecture

“Resize” to match latent dimensions.

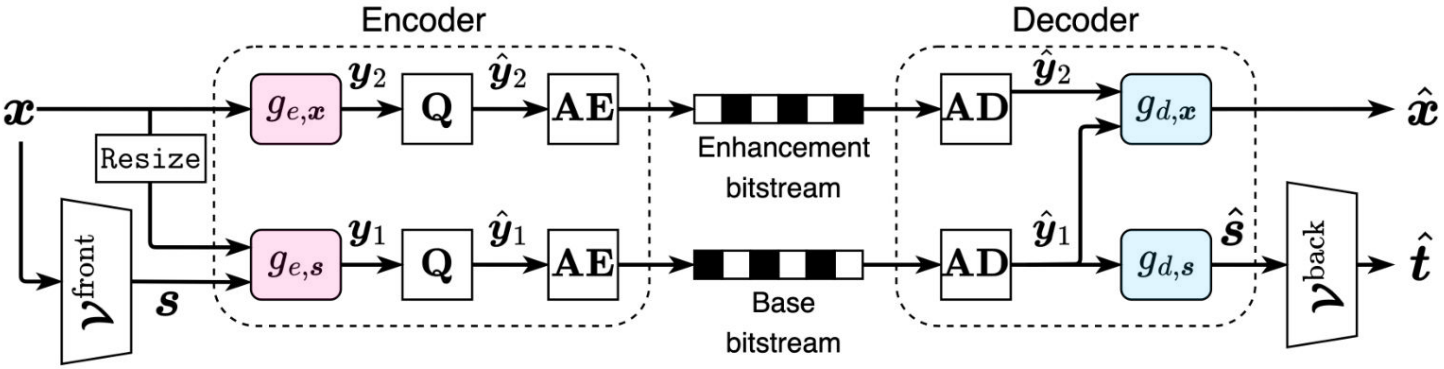


Features are generated from “front” half of task model.

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

Features are fed into “back” half of task model.

Proposed Architecture



	Encoder				Decoder			
	$g_{e,\boldsymbol{s}}$		$g_{e,\boldsymbol{x}}$		$g_{d,\boldsymbol{s}}$		$g_{d,\boldsymbol{x}}$	
No.	Layer	In/Out	Layer	In/Out	Layer	In/Out	Layer	In/Out
1	conv5s1	$C_{\boldsymbol{s}}+3/N$	conv5s2	$3/N$	deconv5s1	M_1/N	deconv5s2	M/N
2	conv5s1	N/N	conv5s2	N/N	deconv5s1	N/N	deconv5s2	N/N
3	conv5s2	N/M_1	conv5s2	N/N	deconv5s2	$N/C_{\boldsymbol{s}}$	deconv5s2	N/N
4			conv5s2	N/M_2			deconv5s2	$N/3$

Experimental Setup

\# of channels for $\boldsymbol{y}_1$

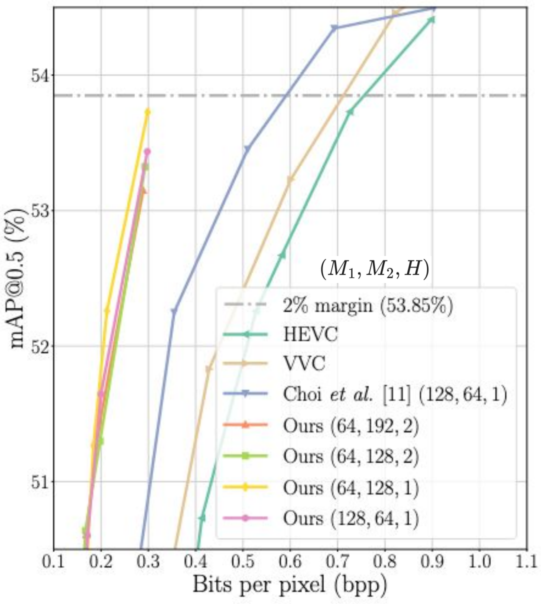
\# of channels for $\boldsymbol{y}_2$

\# of hyperprior blocks

- Various architecture configurations for the tuple (M_1, M_2, H).
- Train on Vimeo-90K dataset with "distortion" computed using mean-squared error (MSE).
- Evaluate object detection on COCO 2014 validation dataset using mAP (IoU=0.5).
- 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 model (*without* PixelCNN-style autoregression) from Choi et al.

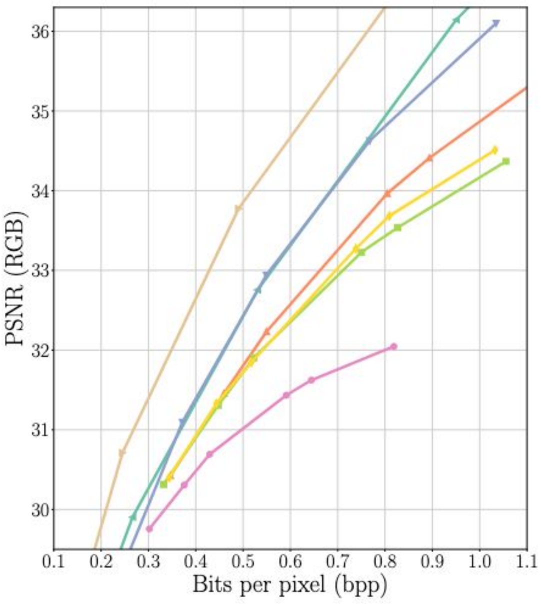
[HEVC] http://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.20+SCM-8.8/
[VVC] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-12.3/
[Vimeo-90K] Xue et al. "Video Enhancement with Task-Oriented Flow," *IJCV*, 2019.
[COCO 2014] T.-Y. Lin et al., "Microsoft COCO: Common objects 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.
[H. Choi et al.] "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing*, 2022.
[PixelCNN] Oord et al., "Pixel Recurrent Neural Networks," *PMLR*, 2016.

Performance Across Various Metrics



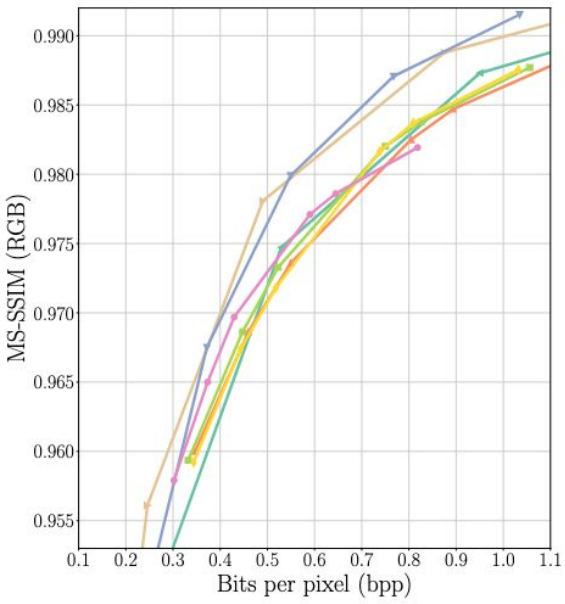
(a)

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



(b)

Input reconstruction on Kodak
PSNR vs. bpp

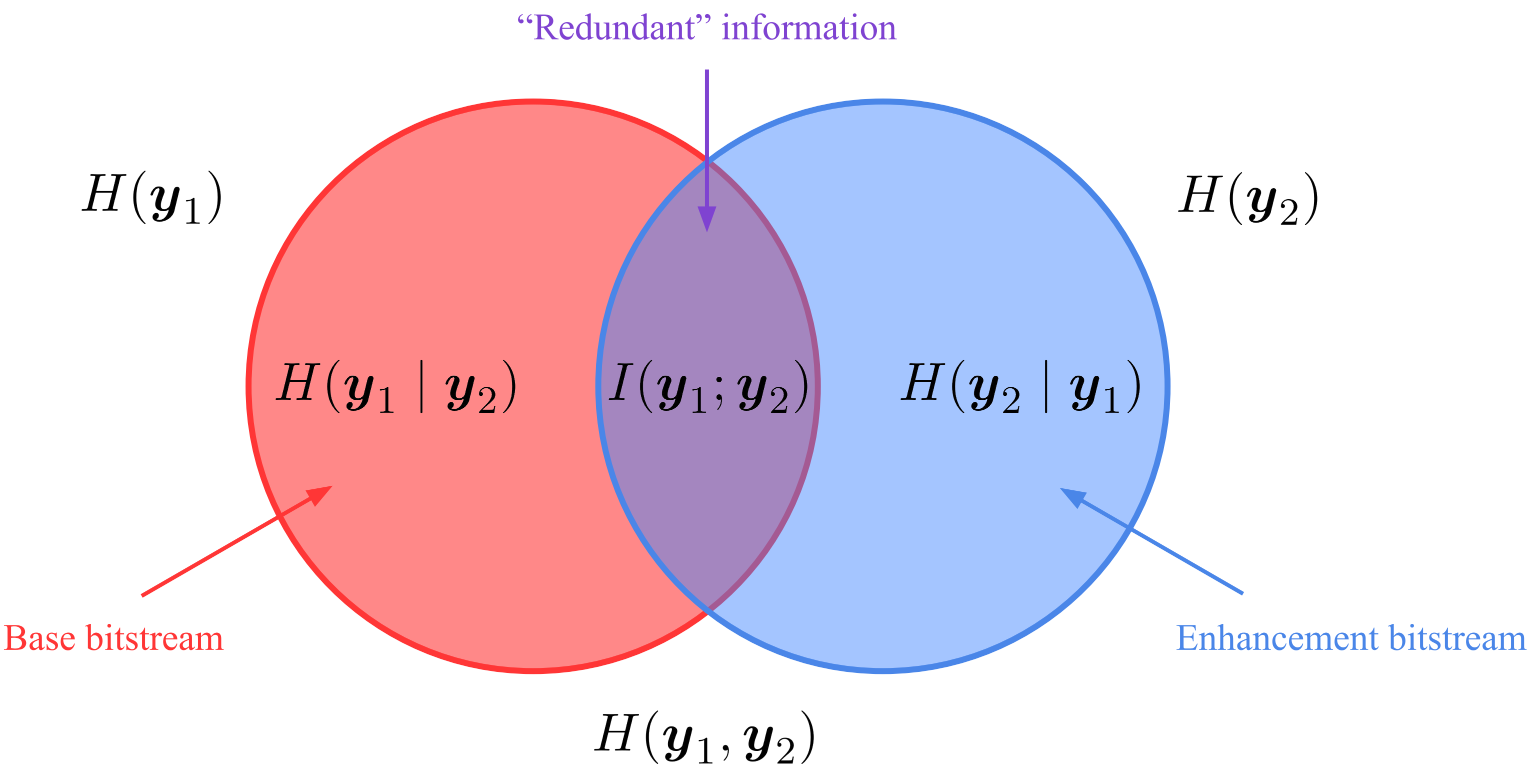


(c)

Input reconstruction on Kodak
MS-SSIM vs. bpp

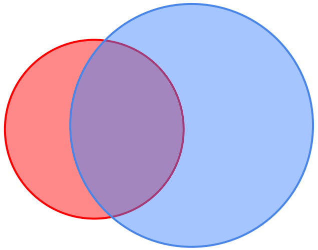
Baseline accuracy of YOLOv3 on COCO 2014 val, including JPEG-compressed images, is 55.85% mAP at 4.80 bpp.

Quick Recap of Entropy and Mutual Information



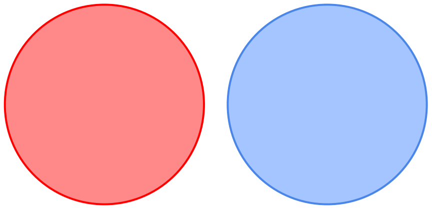
Disentanglement

(i)



Redundancy in bitstreams
Shaded area = 1
Practical rate cost = 1.5

(ii)



Fully disentangled
Shaded area = 1
Practical rate cost = 1

Redundancy $\propto I(\boldsymbol{y}_1; \boldsymbol{y}_2)$

Shaded area $= H(\boldsymbol{y}_1, \boldsymbol{y}_2)$

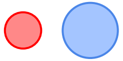
Practical rate cost $= H(\boldsymbol{y}_1) + H(\boldsymbol{y}_2)$

Redundancy

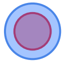
Definition: $$\operatorname{Rdn}(\boldsymbol{y}_i \mid \boldsymbol{y}_j) \triangleq \frac{I(\boldsymbol{y}_i;\boldsymbol{y}_j)}{H(\boldsymbol{y}_i)} = 1 - \frac{H(\boldsymbol{y}_i|\boldsymbol{y}_j)}{H(\boldsymbol{y}_i)}$$

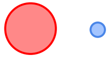
	Base	Enhancement
Codec	$H(\boldsymbol{y}_1)$	$H(\boldsymbol{y}_2)$
Ours	0.3	0.7
Choi et al.	0.6	0.05

Codec entropy rates (in bits per pixel) measured at 2% loss threshold in mAP.

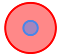


$$0 \leq \operatorname{Rdn}(\boldsymbol{y}_2 \mid \boldsymbol{y}_1) \leq 0.4$$



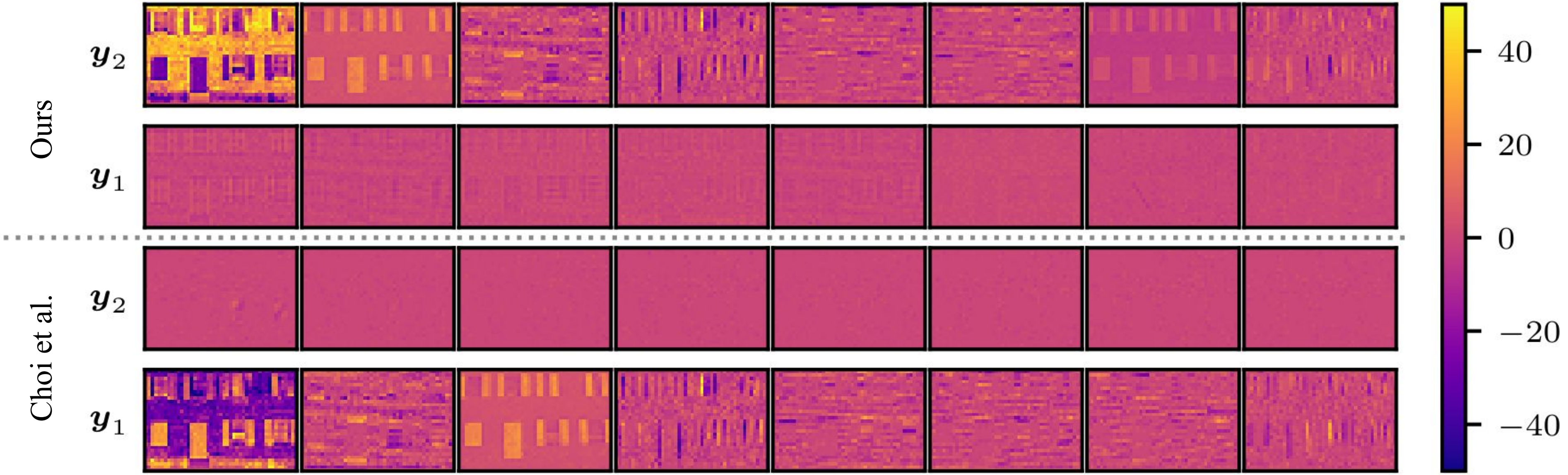


$$0 \leq \operatorname{Rdn}(\boldsymbol{y}_2 \mid \boldsymbol{y}_1) \leq 1.0$$



Bounds on redundancy in enhancement bitstream under respective entropy models.

Feature maps



top-8 channels ordered by rate

$\boldsymbol{y}_1$ = base (for machine vision)
$\boldsymbol{y}_2$ = enhancement (for humans)

Conclusion and Future Work

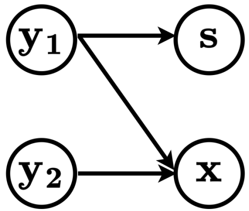
- DNN-based image codec with a new variational formulation.
  - Offers latent-space scalability for human and machine tasks.
  - New way of disentangling the learned latent representations.
- Significant bit reductions at the base layer.
- Needs further investigation about improving reconstruction quality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

Thank you

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.



Introduce variables: $y_1 \in \mathbb{R}^{d_1}$, $y_2 \in \mathbb{R}^{d_2}$

Under this model, data likelihood is given by:

$$p_\theta(x, s, y_1, y_2) = \underbrace{p(y_1)p(y_2)}_{\boxed{\text{assumption 1: } Y_1 \perp Y_2}} p_\theta(x \,|\, y_1, y_2) \underbrace{p_\theta(s \,|\, y_1)}_{\boxed{\text{assumption 2: } Y_2 - Y_1 - S}}.$$

$$\implies \quad p_\theta(x, s) = \int\!\!\int p_\theta(x, s, y_1, y_2) dy_1 dy_2.$$

<u>*intractible*</u> integral !!

Overcoming Intractability

Introduce factored variational posterior.

$$q_\phi(y_1, y_2 \,|\, x, s) = \underbrace{q_\phi(y_1 \,|\, x, s)}_{\text{extract 'common' latents from both } (X,S)} \; \underbrace{q_\phi(y_2 \,|\, x)}_{\text{extract 'enhancement' ones only from } X}$$

Impose above factorization by system model.

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

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

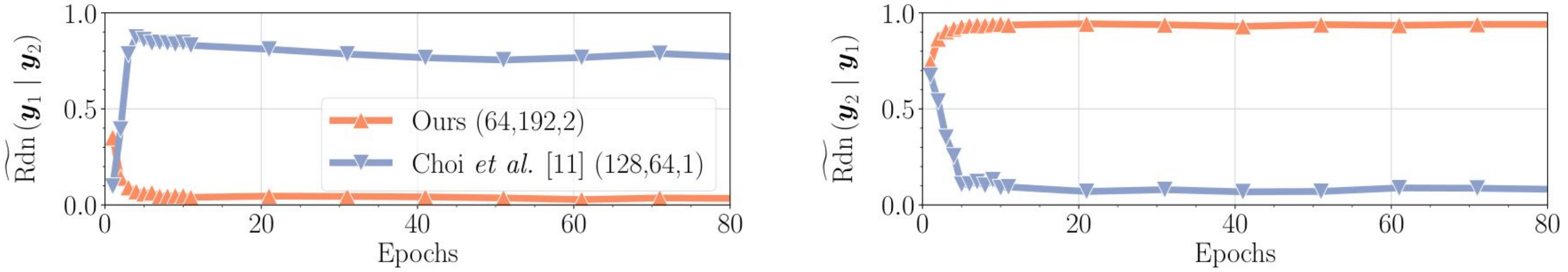
[J. Ballé et al.], "Variational Image Compression with a Scale Hyperprior," *ICLR*, 2018.

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

Clustering function



Evolution of the redundancy metrics during training.