CREATE
CONNECT
LIVE
inspire

End-to-end optimized image compression for machines

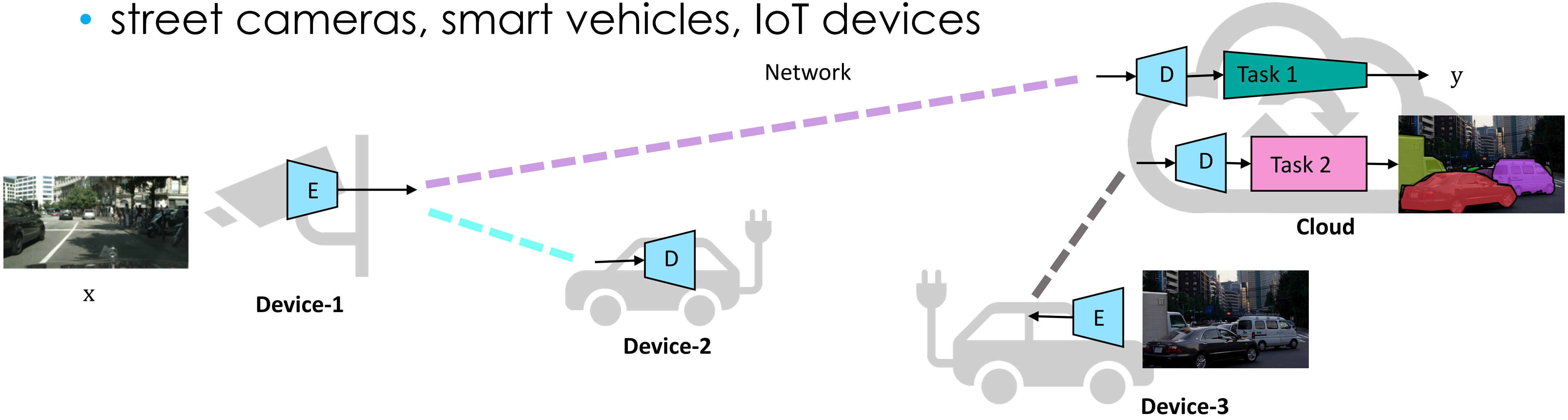
A study on object detection

Lahiru D. Chamain, Fabien Racapé, Jean Bégaint, Akshay Pushparaja and Simon Feltman

interdigital™

Networked AI applications

- **Images & videos are captured by embedded devices**
  - street cameras, smart vehicles, IoT devices

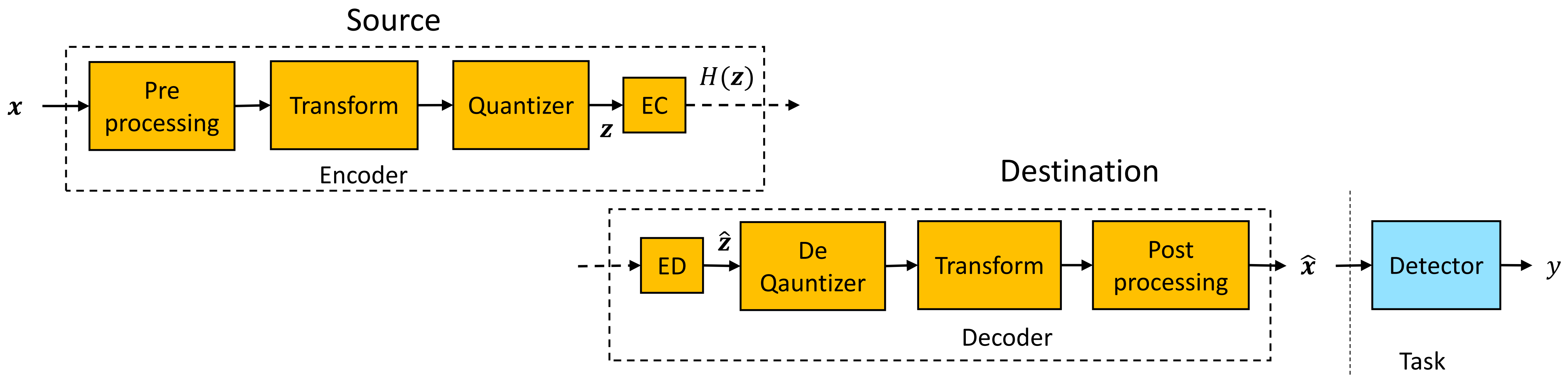


- **Content is analyzed mostly by machines for**
  - Object detection, segmentation, classification, tracking
  - Sometimes visualization

interdigital

Codecs for visualization #1

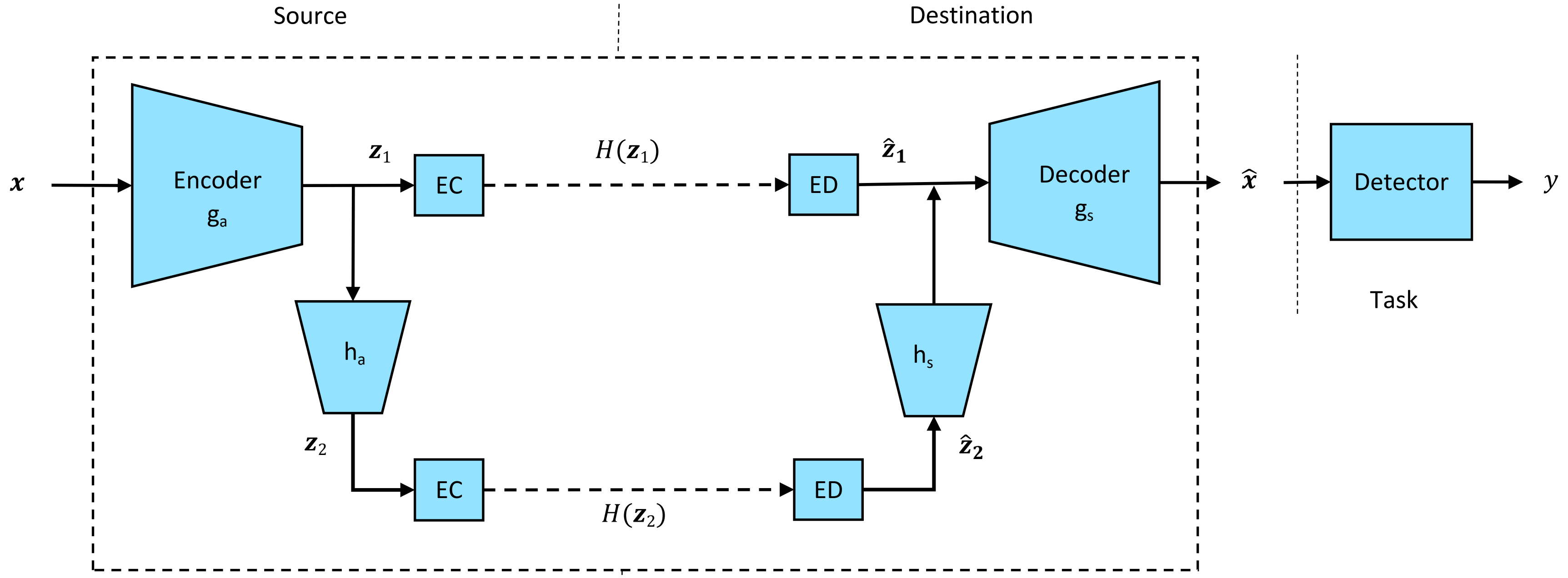
- Existing codecs are optimized for visualization
- Conventional codecs (JPEG, J2k, AVC/H.264, HEVC/H.265, AV1, VVC)



- Engineered blocks, no backpropagation

Codecs for visualization #2

- **ANN-based codecs:**
  - Scale hyperprior [1], L3C [2]
  - learned encoder, decoder and entropy modelling



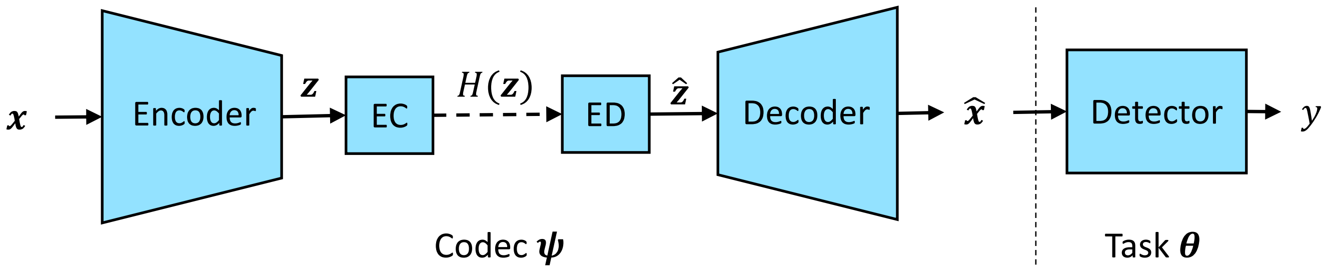
- **Can be optimized end-to-end for any differentiable task**

[1] Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). "Variational image compression with a scale hyperprior.", ICLR 2018.
[2] Mentzer, Fabian, et al. "Practical full resolution learned lossless image compression." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Motivation

- **Existing codecs are optimized for visualization:**
  - Rate-Distortion criterion, usually PSNR, sometimes MS-SSIM
- **Not optimal for rate-task accuracy**
  - ANN-based codecs can be optimized end-to-end for rate-task accuracy
- **Goal:**
  - to provide a benchmark for rate-task accuracy performance improvements with end-to-end training

Studied configurations



scale-hyperprior [1]

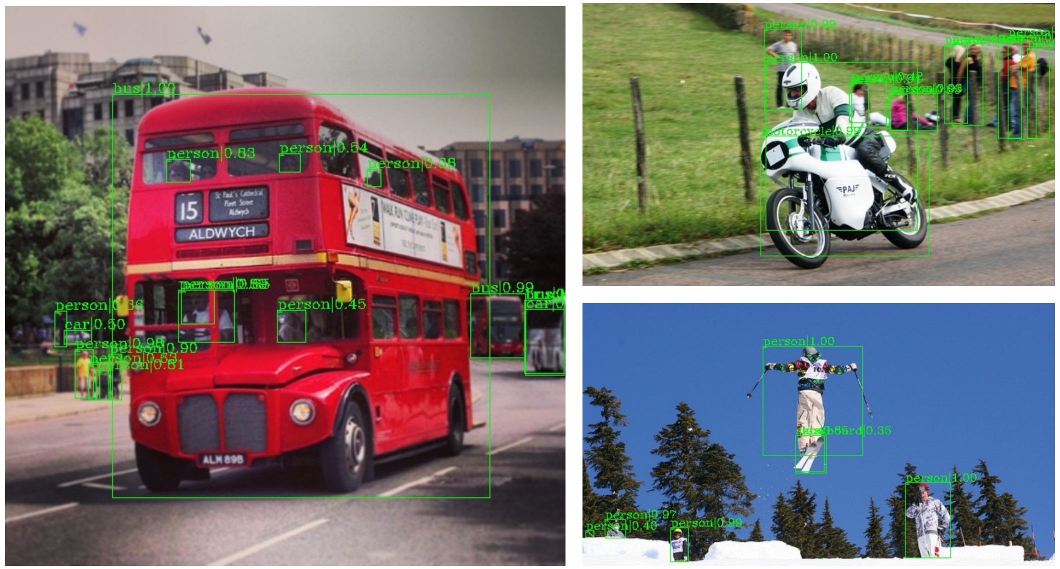
Object detection

1. **Baseline inference**
2. **Task model fine-tuning (T-FT)**
3. **Codec fine-tuning (C-FT)**
4. **Joint end-to-end fine-tuning (J-FT)**

[1] Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). "Variational image compression with a scale hyperprior.", ICLR 2018.

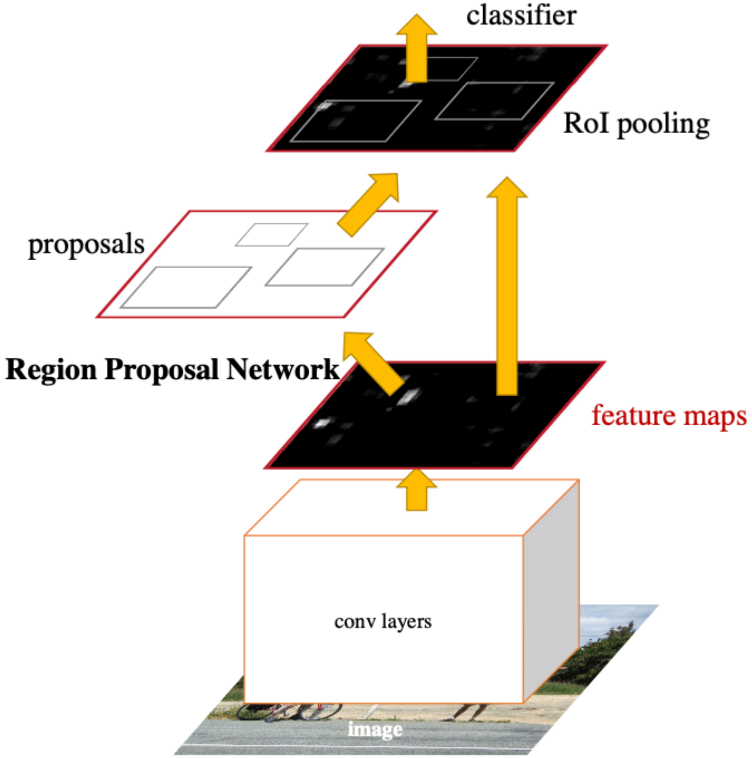
Object detection , dataset: COCO 2017

- **Common Objects in Context**
  - >180,000 training | 5,000 validation images
- **Metric: mAP**



https://cocodataset.org/#home

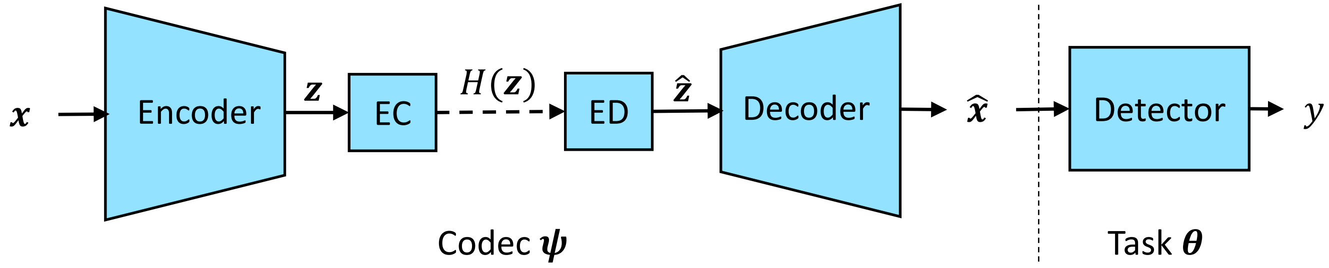
Faster R-CNN (Res-50 backbone)[1]



[1] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99)

Baseline: Inference (off-the-shelf)

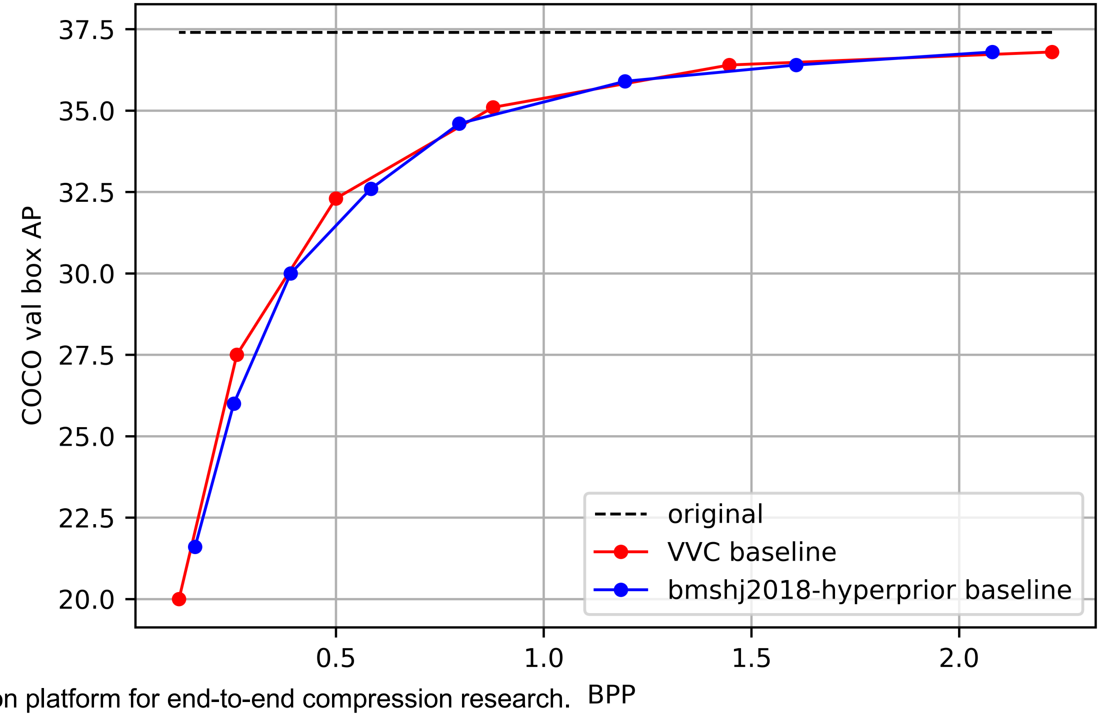
- **Codecs:**
  - Scale hyperpriors (learned) [1]
  - Versatile Video Coding (VVC)



- **Scale hyperpriors:**
  - Optimized for MSE

- **Faster R-CNN: off-the-shelf**

- **Scale hyperpriors: as good as VVC**



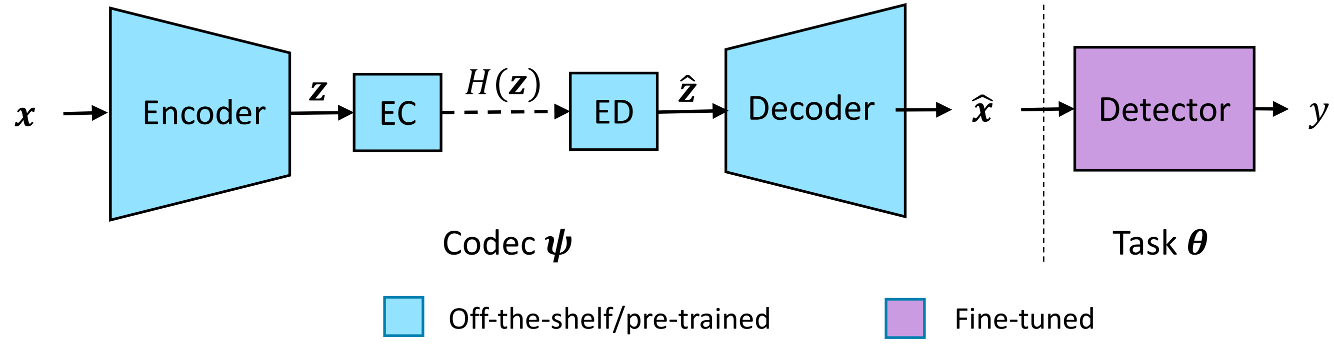
[1] Bégaint, J., Racapé, F., Feltman, S., & Pushparaja, A. (2020). CompressAI: a PyTorch library and evaluation platform for end-to-end compression research.

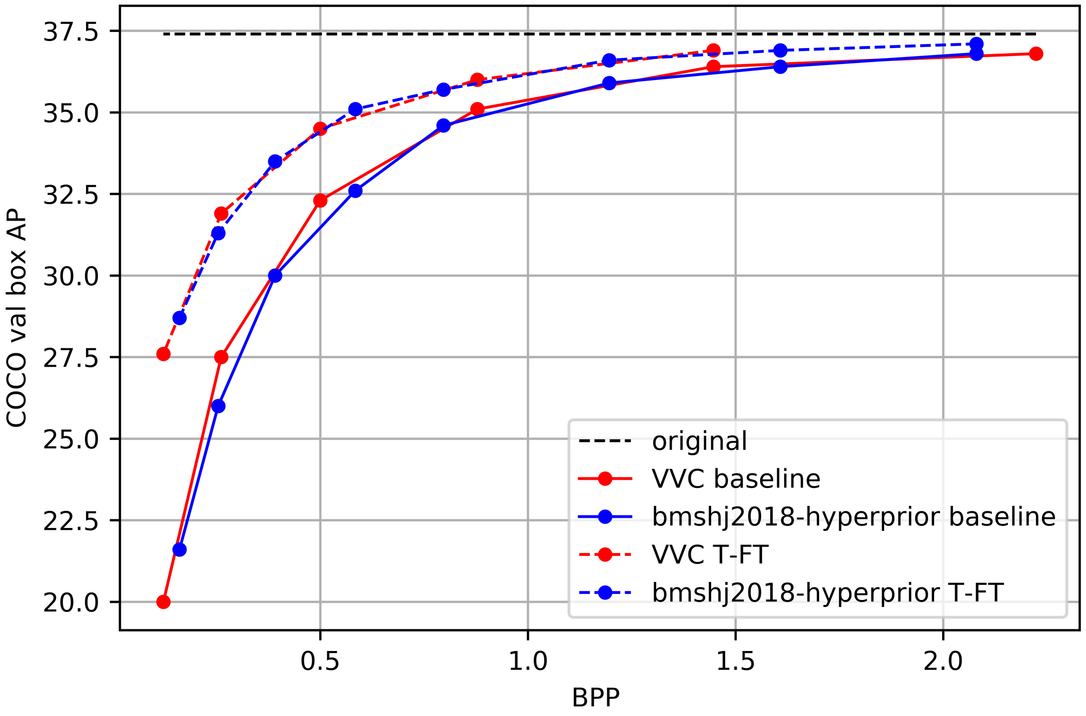
Task model fine-tuning (T-FT)

- **Fixed off-the-shelf codecs**
- **Detector fine-tuning:**

$$Loss = E[loss_{task}(y, y_{gt})]$$

- **Similar improvement for both codecs**





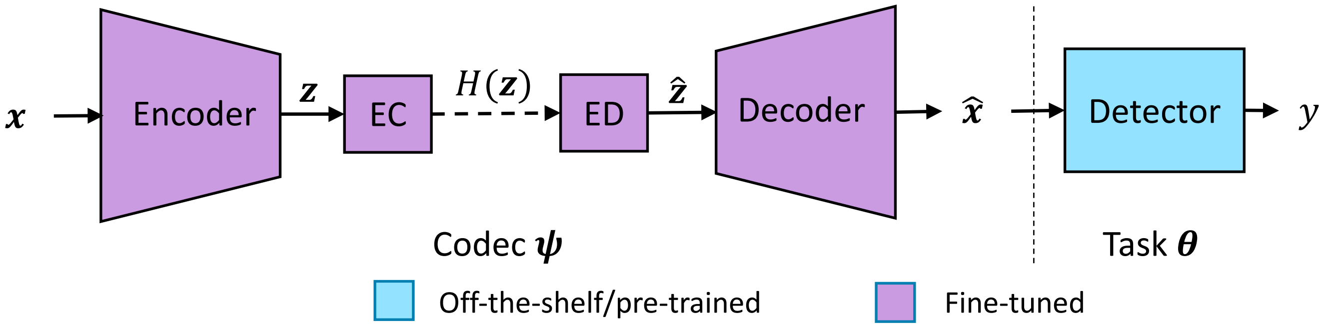
interdigital

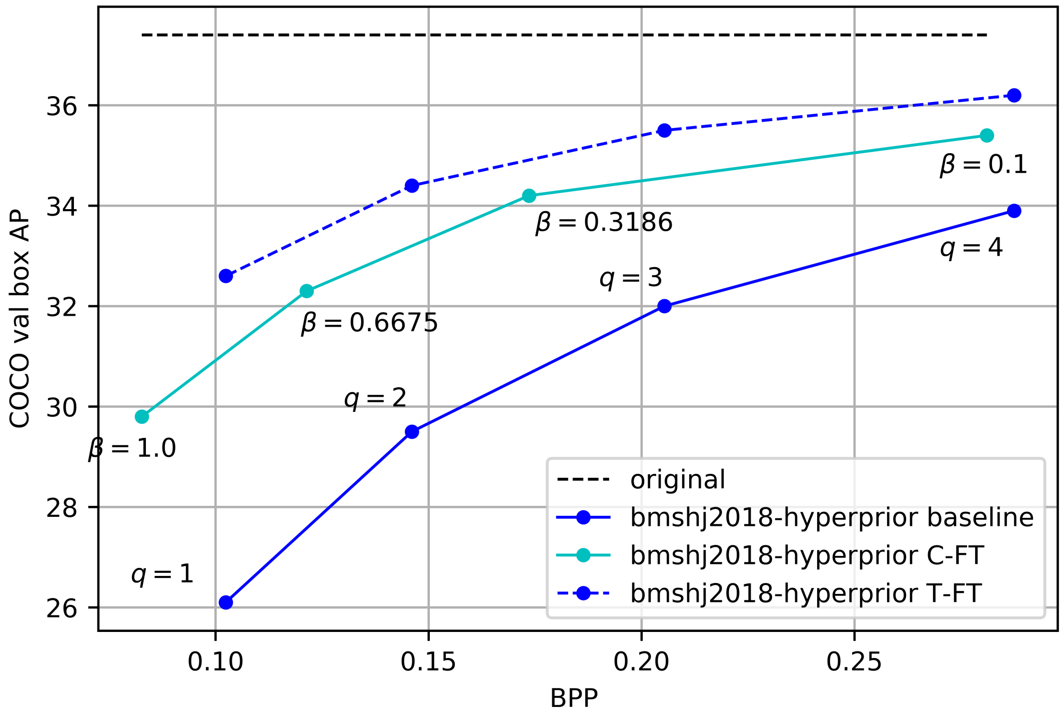
Codec fine-tuning (C-FT)

- **Fixed off-the-shelf Detector**
- **Codec finetuned to minimize:**

$$Loss = E[loss_{task}(y, y_{gt})] + \beta E[H(\boldsymbol{z})]$$

- **β : control parameter**
- **Codec learns to drop irrelevant features for detection**
- **Rate-accuracy is lower compared to Task-FT**





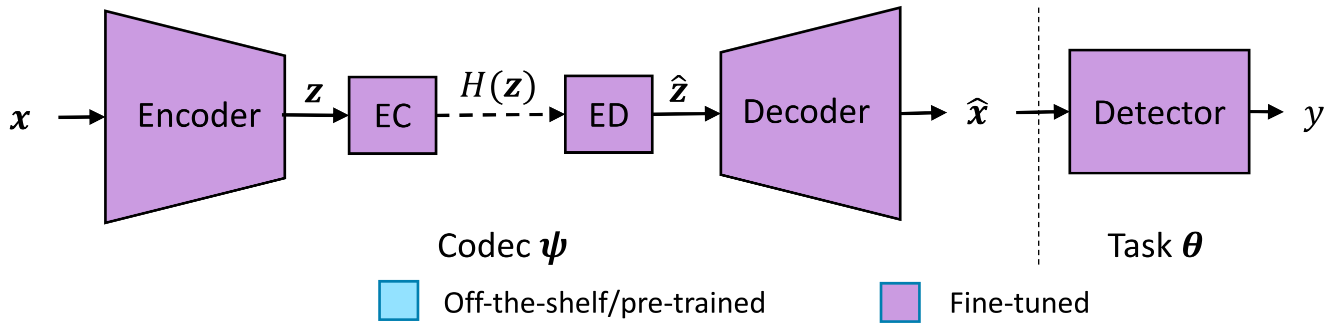
interdigital

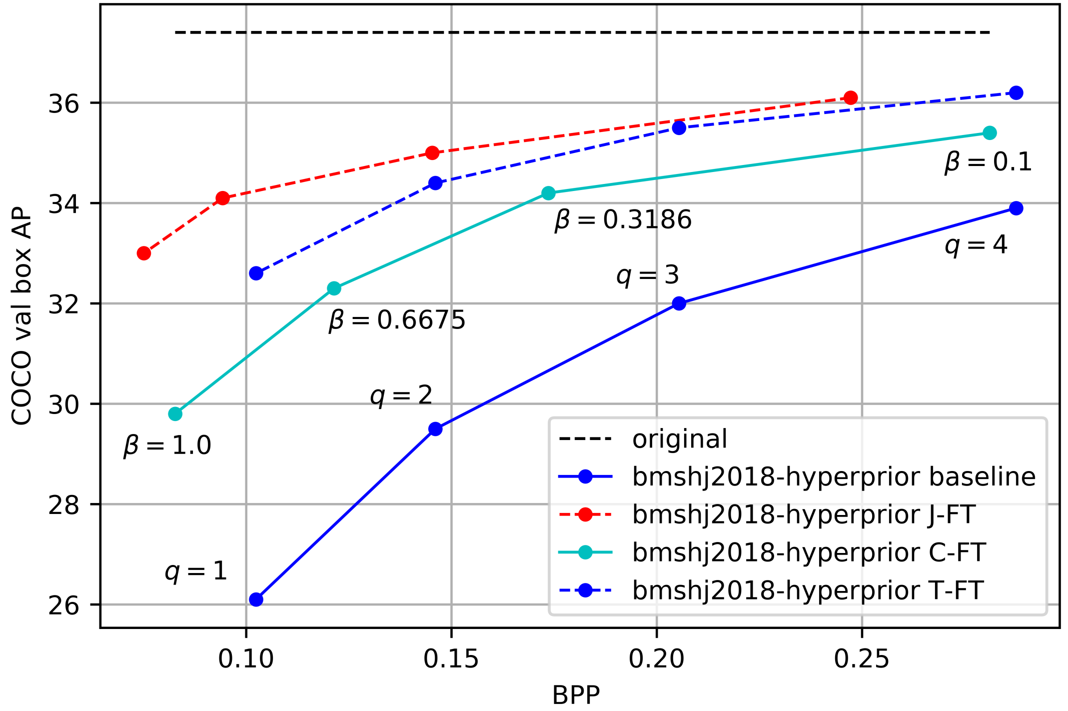
Joint end-to-end fine-tuning (J-FT)

- Training codec + task
- Minimize to find:

$$\boldsymbol{\psi}^{\beta}, \boldsymbol{\theta}^{\beta} = \arg min \ E[loss_{task}(y, y_{gt})] + \beta . E[H(\boldsymbol{z})]$$

- High rate-accuracy at lower BPPs





interdigital

Visual comparison #1

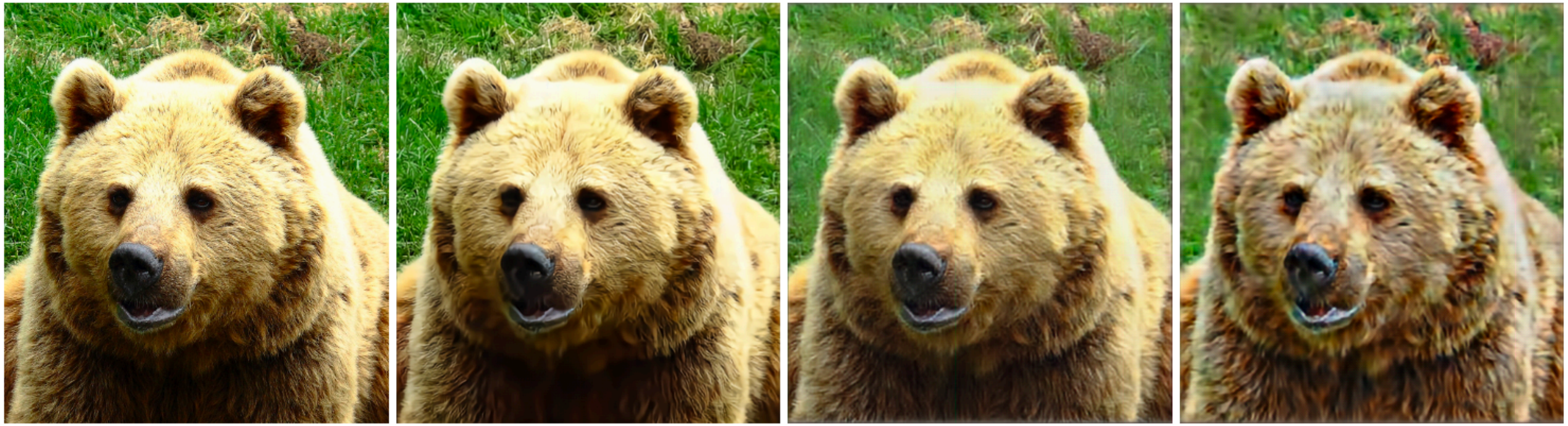
- **Same bit-rate: 0.1 Bpp**

Original	T-FT	C-FT	J-FT
box mAP	32.6	29.8	34.1 (**+4.6%**)
PSNR (dB)	30.49	17.70	16.31
bpp	0.1024	0.0827	0.0943

Visual comparison #2

- **Same accuracy**

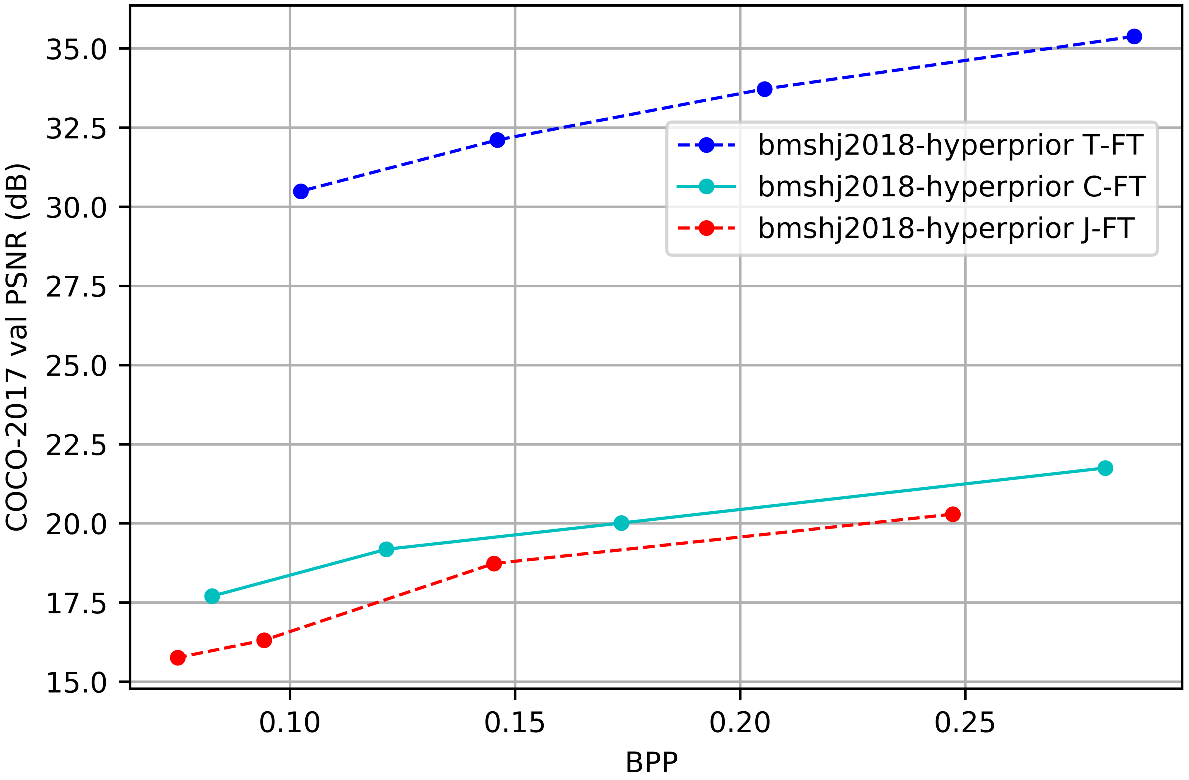
Original T-FT C-FT J-FT



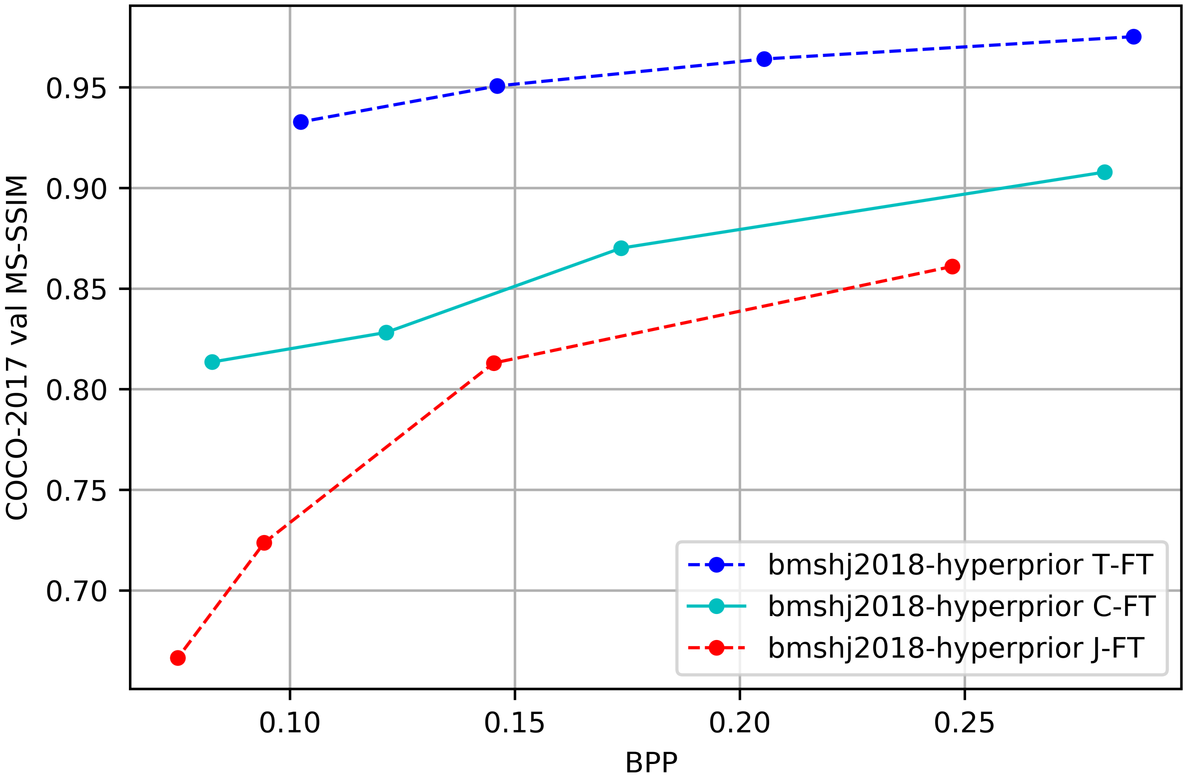
box mAP	34.4	34.2	34.1
PSNR (dB)	32.11	20.01	16.31
bpp	0.1461 (+**54.9%**)	0.1736 (+**84.1%**)	0.0943

Synthesis quality

PSNR

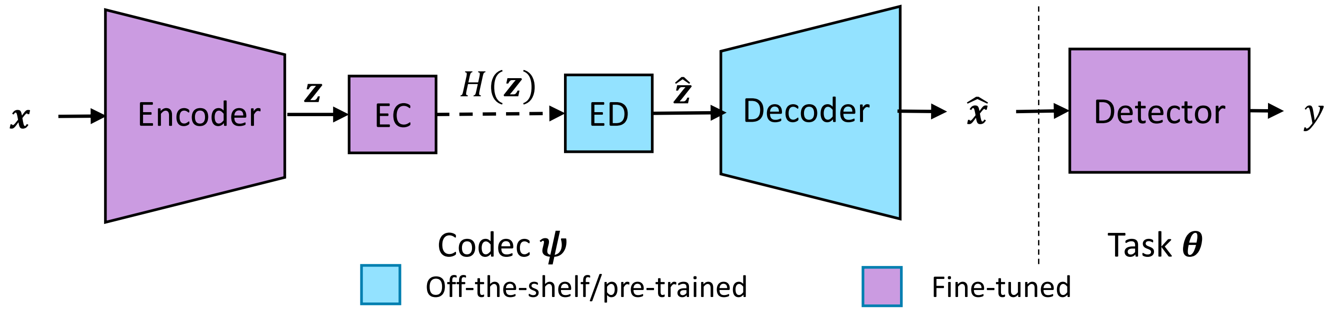


MS-SSIM



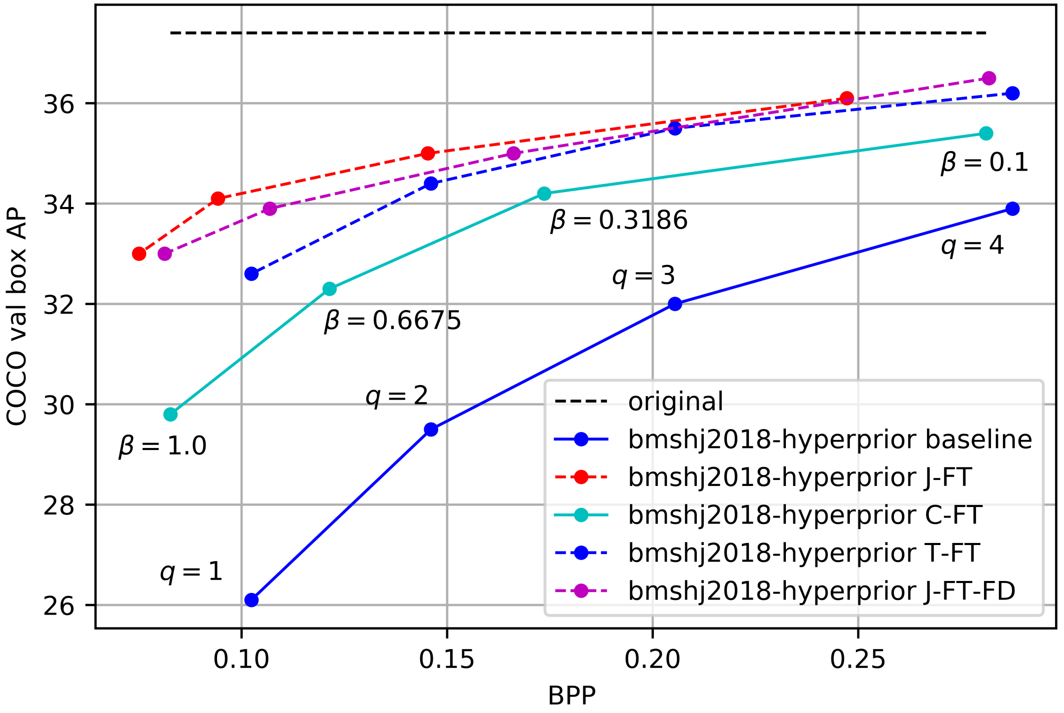
Fixed decoder #1

- **Fixed off-the-shelf decoder**
  - End devices may use a common decoder



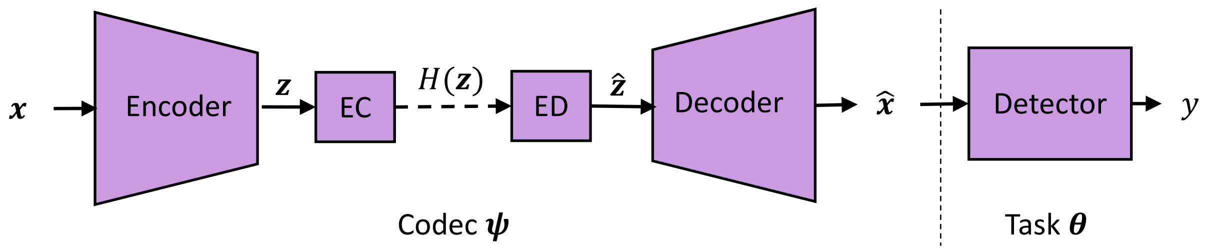
- **Better than C-FT and T-FT at lower rates**

- **Small rate-accuracy drop from J-FT**



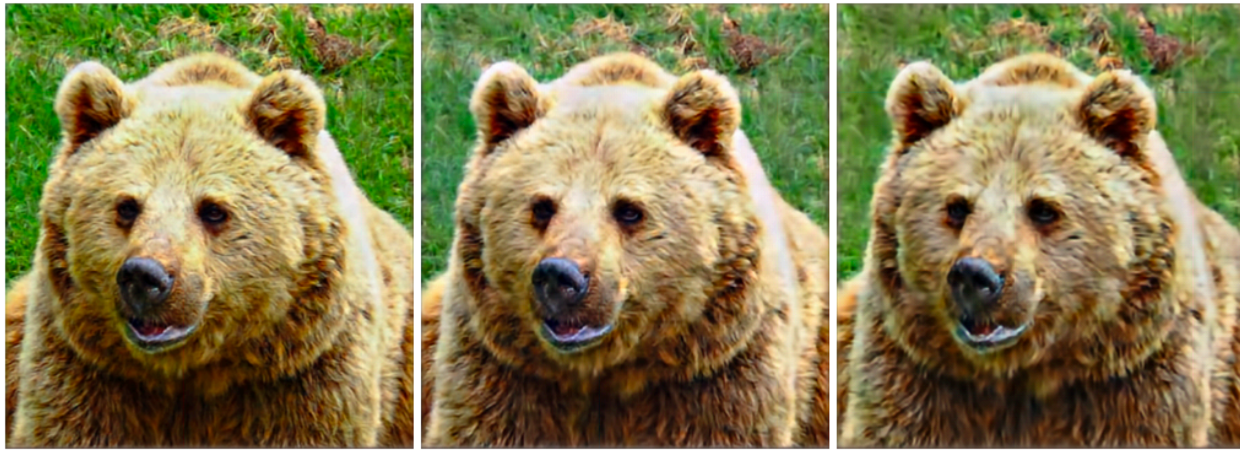
interdigital

Fixed decoder #2



End-to-end fine-tuning (J-FT)

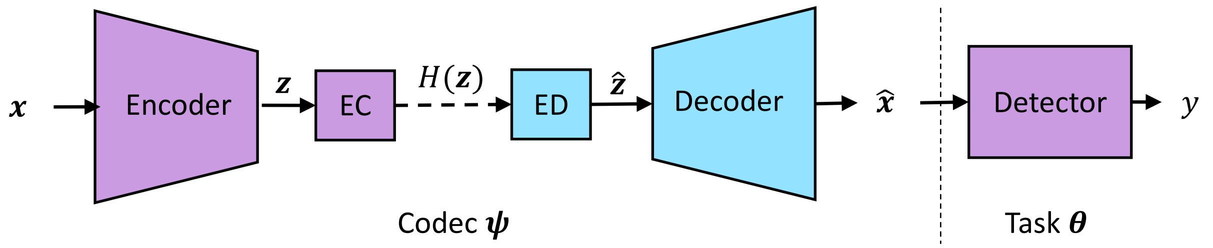
ψ_β^J



	$\beta = 0.1$	$\beta = 0.3186$	$\beta = 0.6675$
BPP	0.2472	0.1453	0.0943

Off-the-shelf/pre-trained

Fine-tuned



End-to-end fine-tuning (J-FT)
with fixed common decoder

$\psi_q^{J-\mathrm{FT}-\mathrm{FD}}$

	$\beta = 0.1$	$\beta = 0.3186$	$\beta = 0.6675$
BPP	0.2818	0.1662	0.1069

interdigital

Conclusion

- **Presented: a study on object detection related to end-to-end learning of code-task chain.**
  - Showed: Better rate-accuracy performance by jointly optimizing the codec-task model
- **Visual comparison: optimized codecs produce images with highlighted features**
- **Future work:**
  - presented Study: reuse pretrained codec, fine-tune for machine task.
  - video, feed decoded features directly to modified tasks (need to optimize enc only)

Questions and comments

hdchamain@ucdavis.edu
fabien.racape@interdigital.com

Thank you!