Practical image compression with deep neural networks

Luis Herranz

Computer Vision Center
Universitat Autònoma de Barcelona

March 2022

Universitat Autònoma de Barcelona
MINISTERIO DE CIENCIA E INNOVACIÓN
UNIÓN EUROPEA
FONDO SOCIAL EUROPEO
El FSE invierte en tu futuro
AGENCIA ESTATAL DE INVESTIGACIÓN

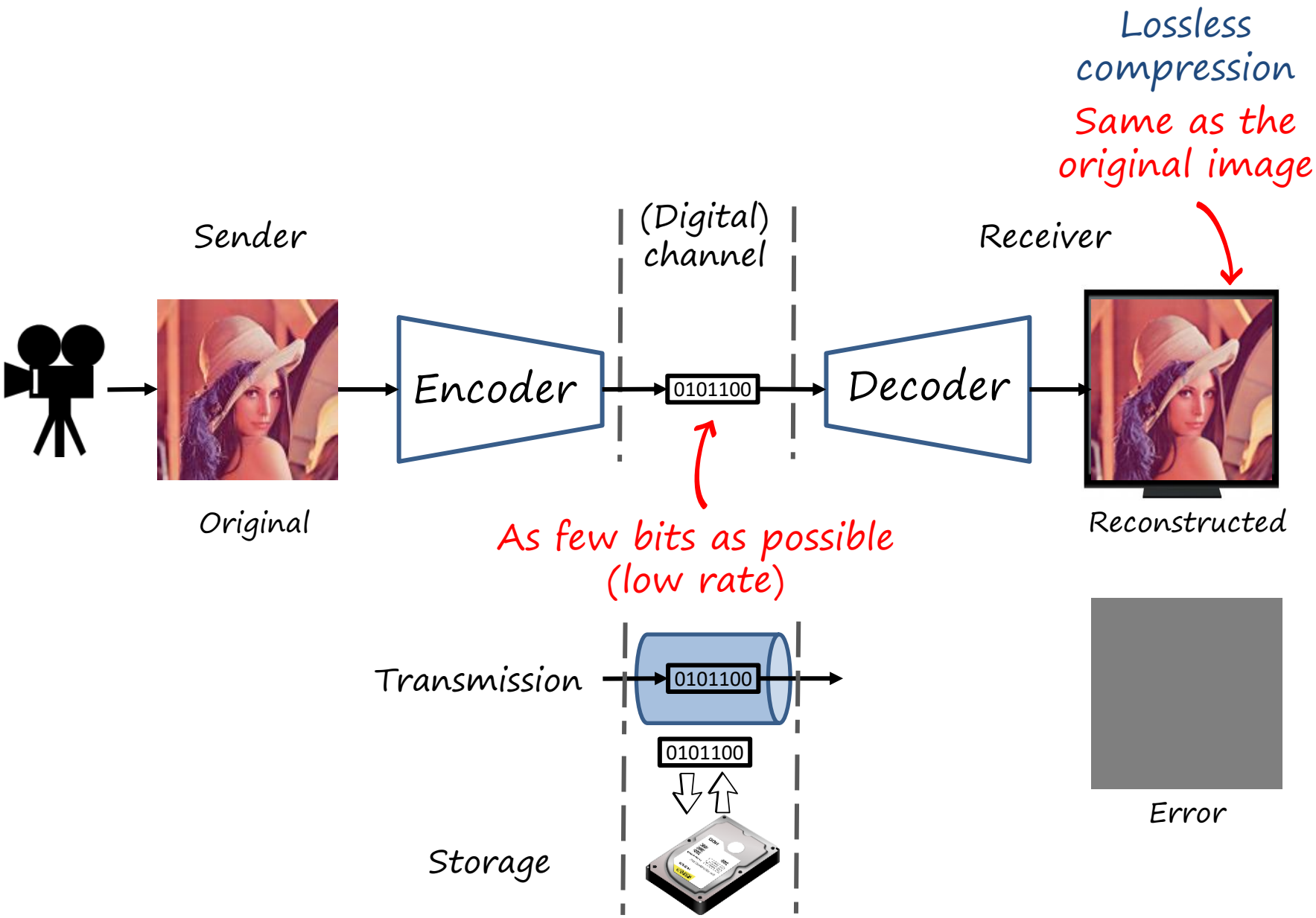
Outline

- Introduction: image coding
- Compression with neural networks
- Towards practical image compression

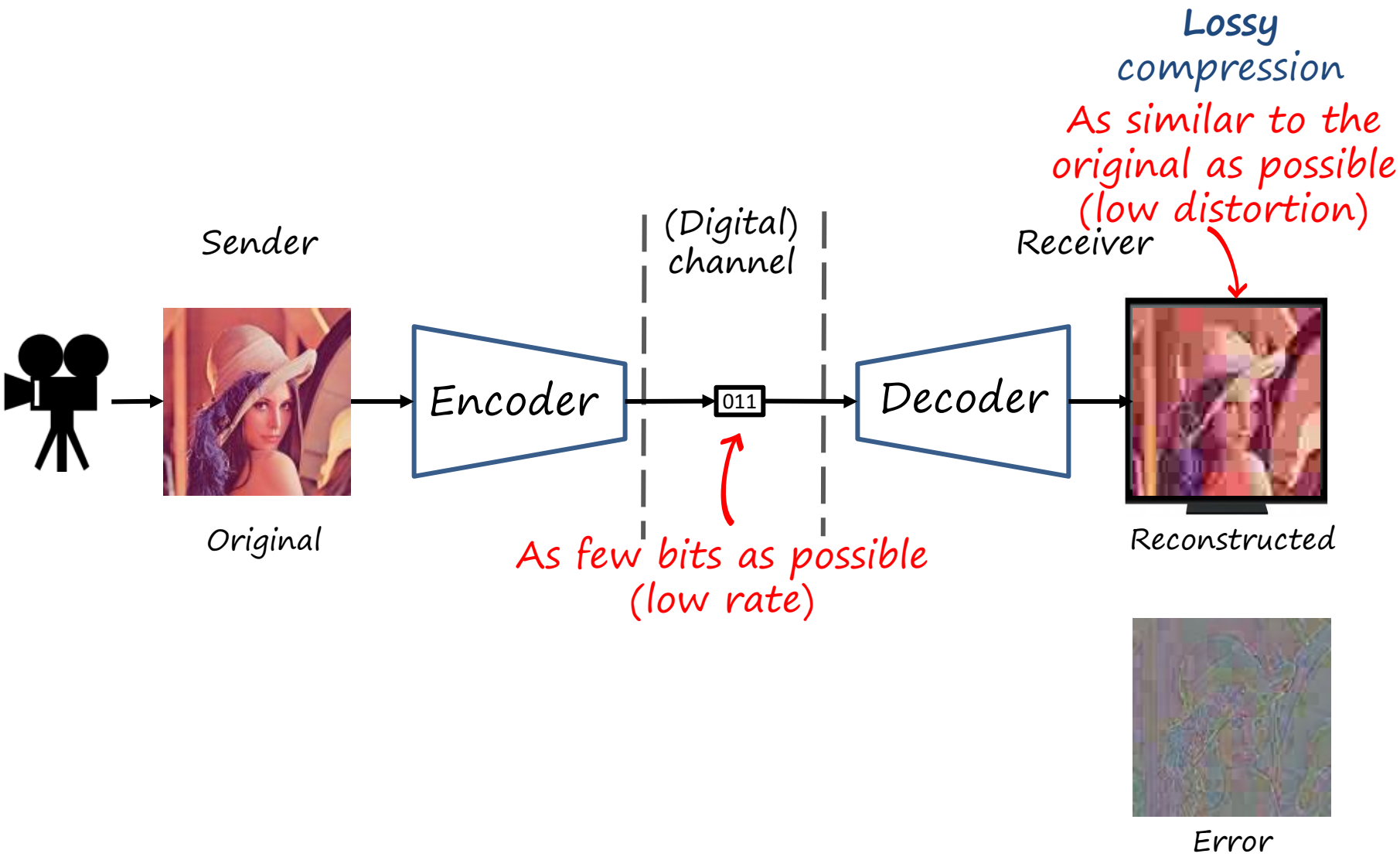
Outline

- Introduction: image coding
- Compression with neural networks
- Towards practical image compression

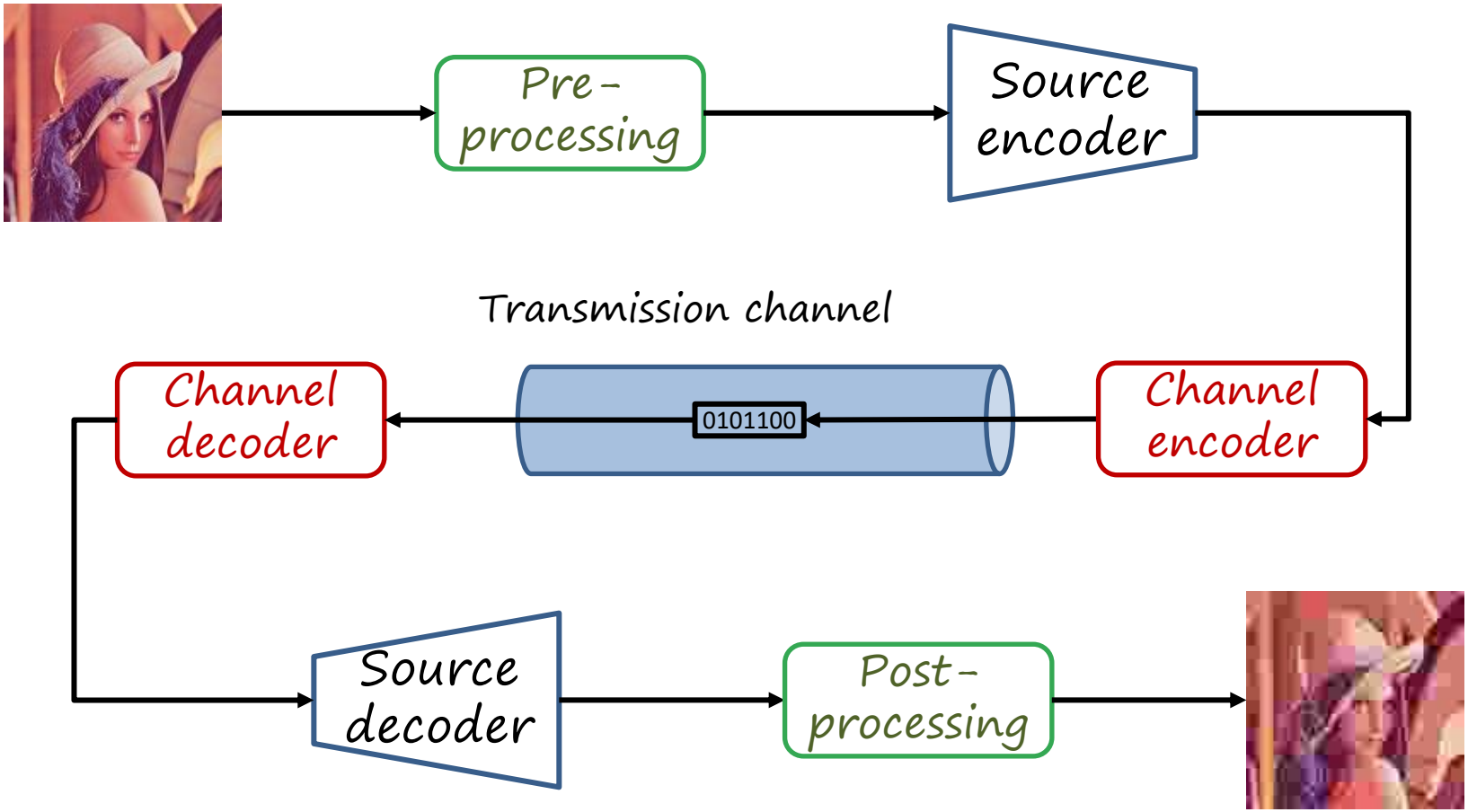
The visual communication problem



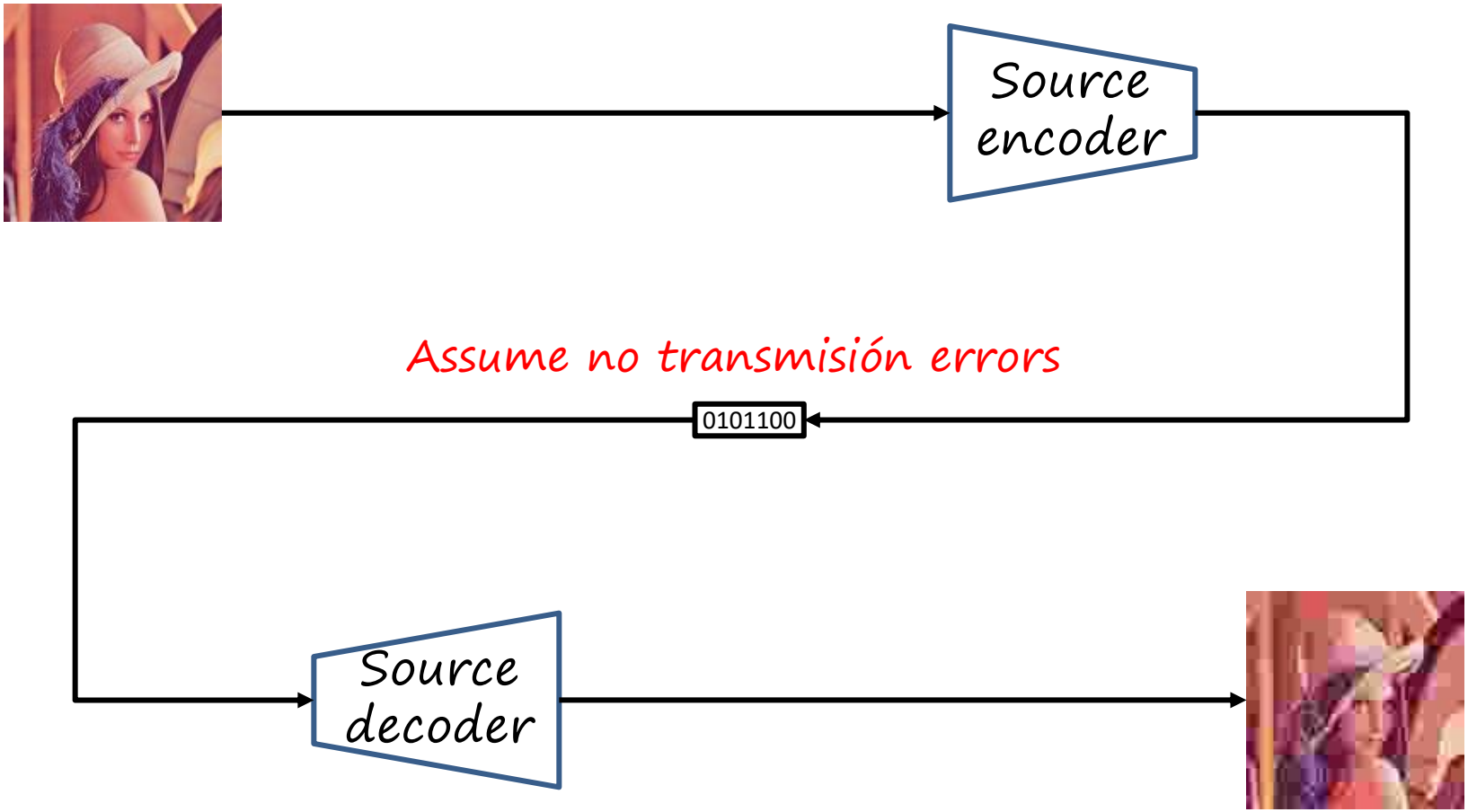
The visual communication problem



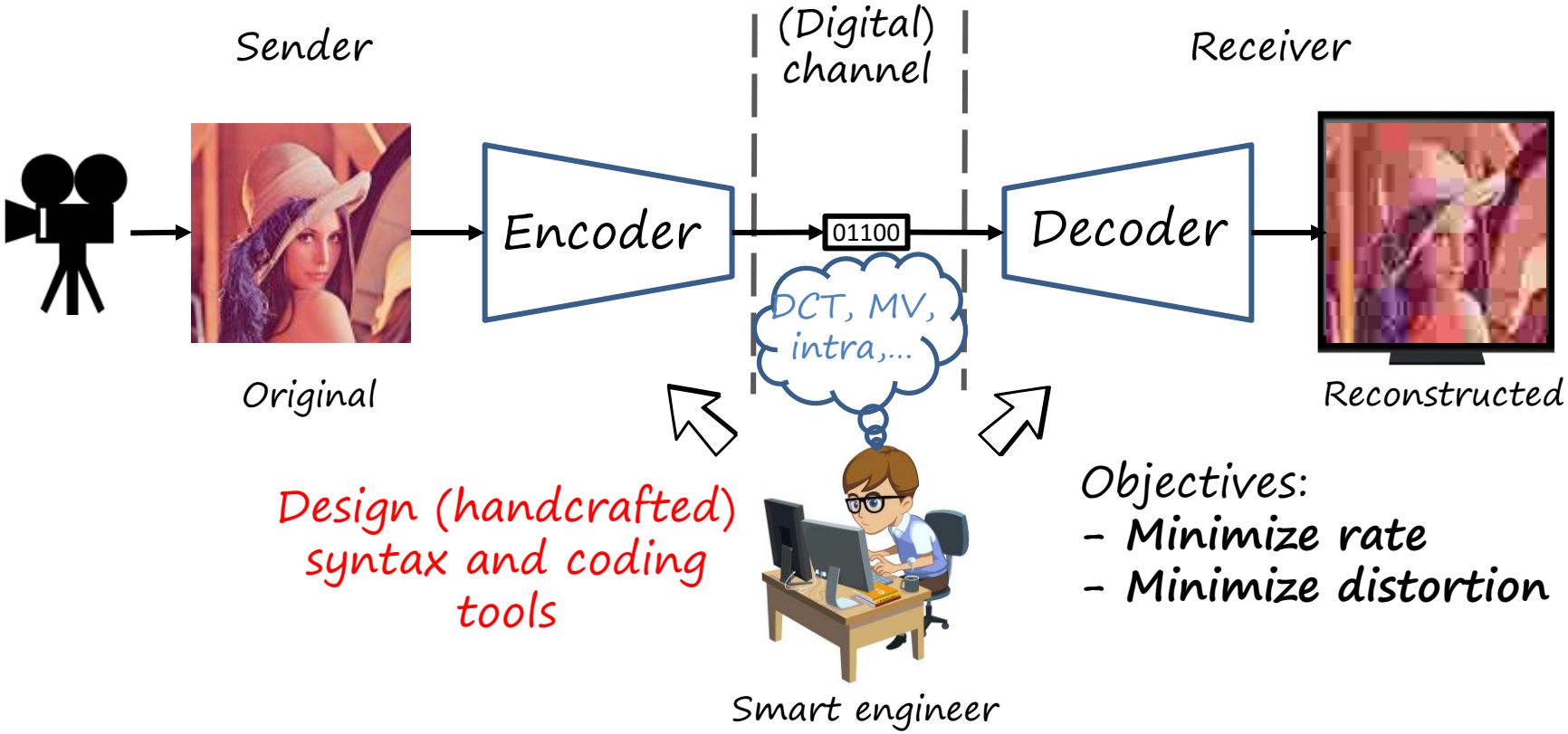
Pre/post-processing, source coding and channel coding



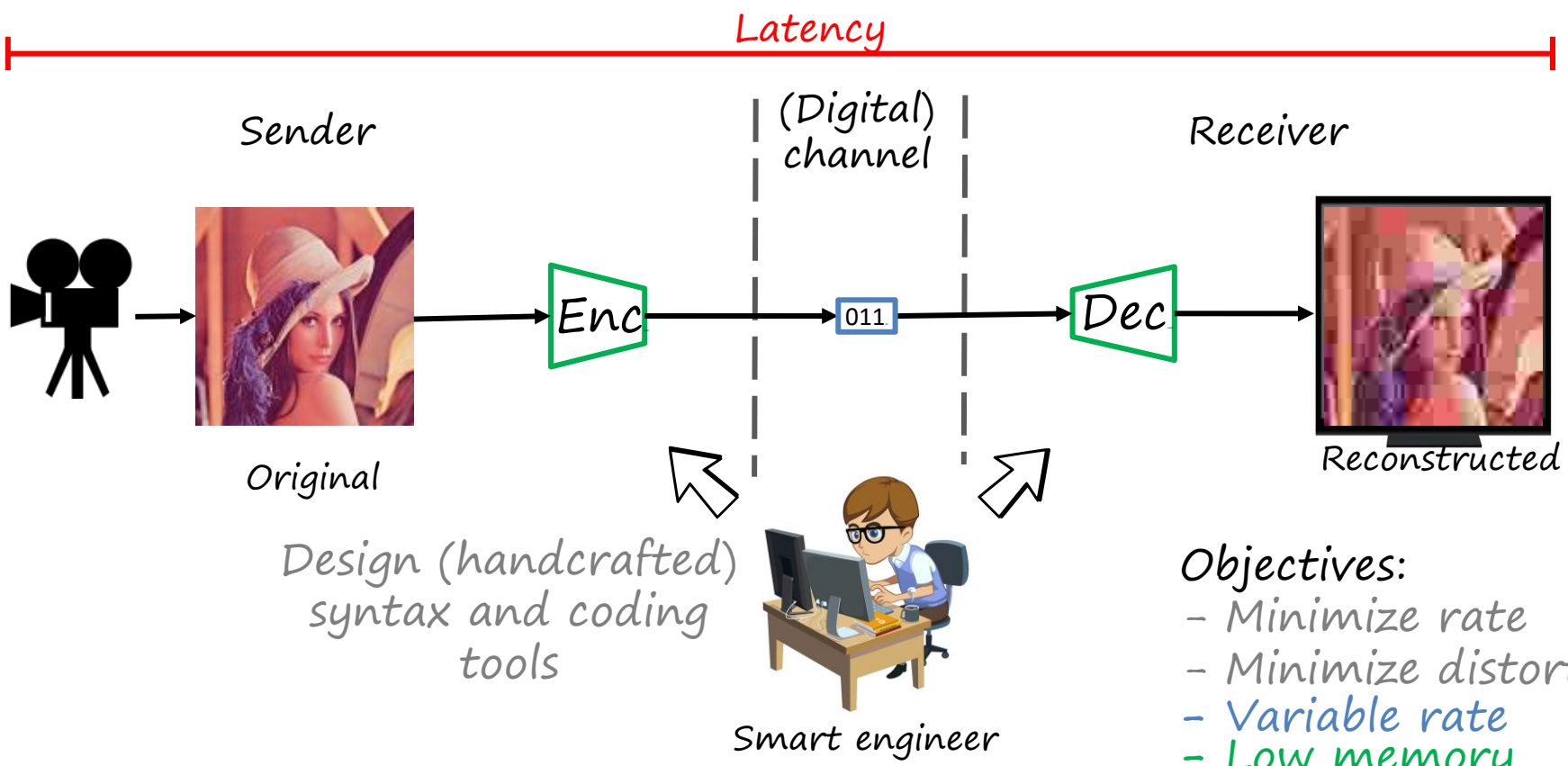
Source coding only



Developing traditional image/video codecs



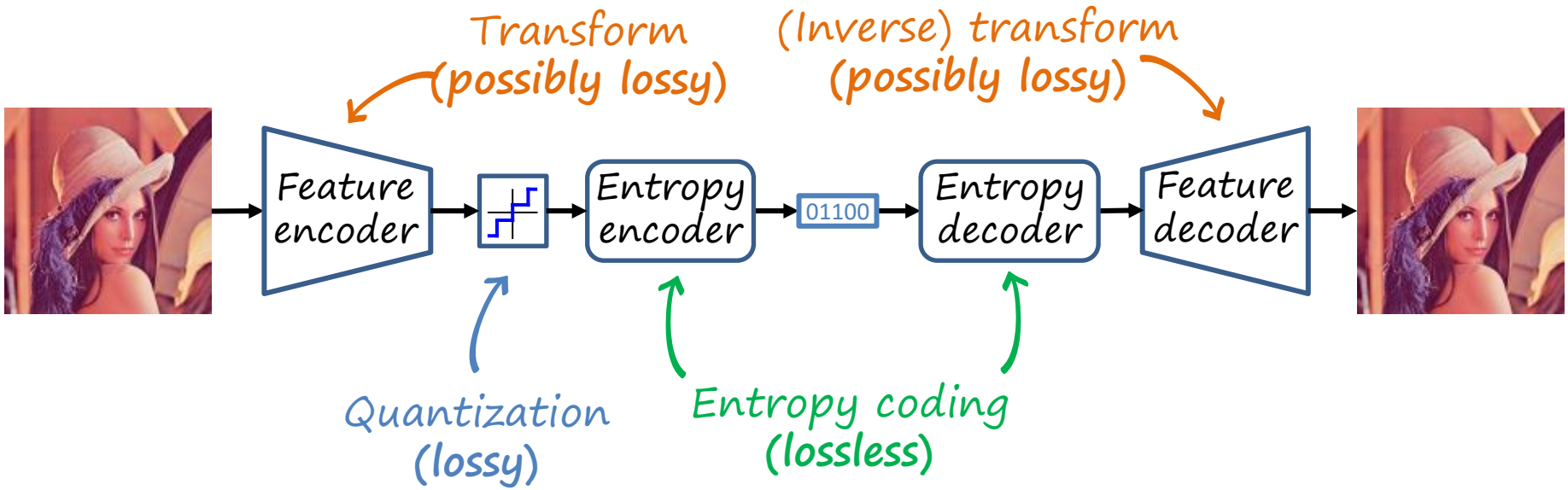
… for practical applications



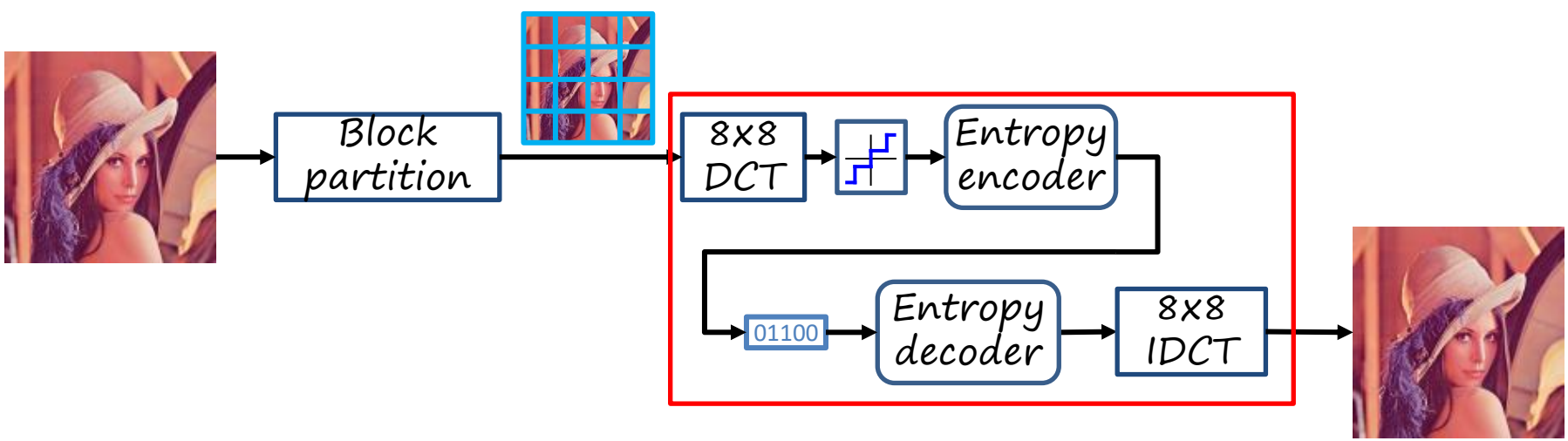
Objectives:

- Minimize rate
- Minimize distortion
- Variable rate
- Low memory
- Low computation
- Low latency
- Compatibility
- Domain-specific

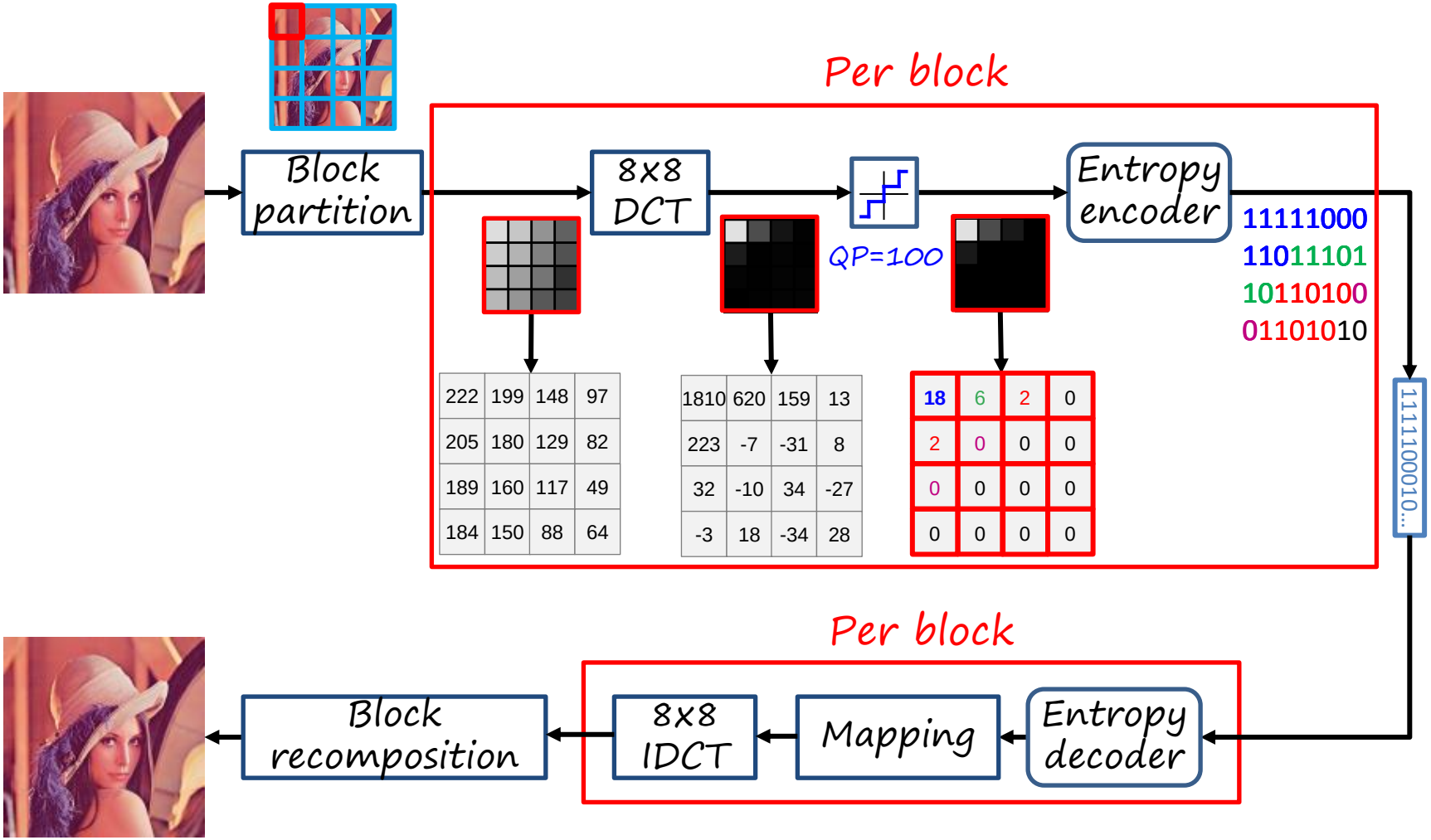
Transform coding pipeline



Example: block-based transform coding (e.g. JPEG, MPEG-2, H.264)



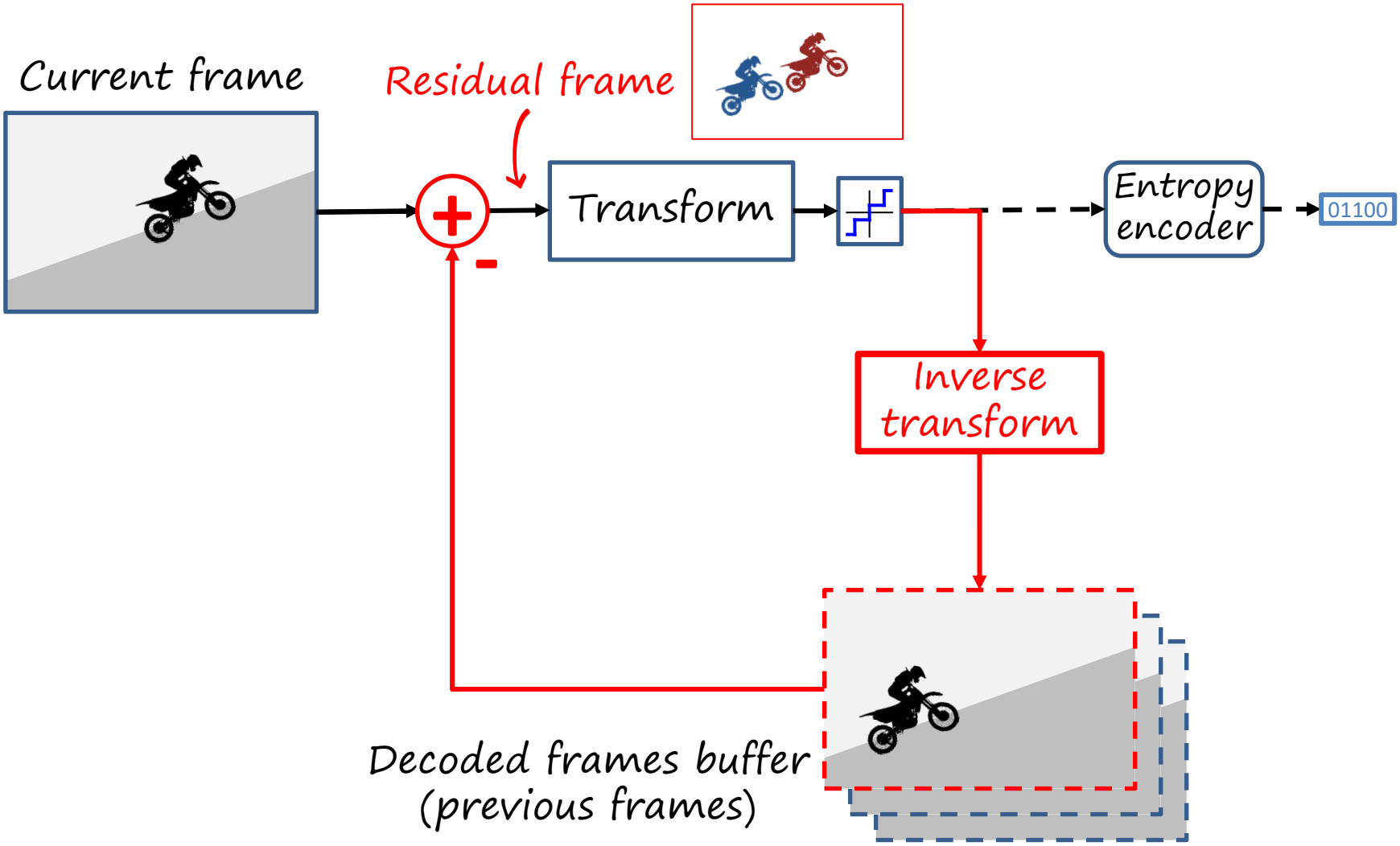
Transform coding pipeline: JPEG



Slide partly adapted from T. Wiegand

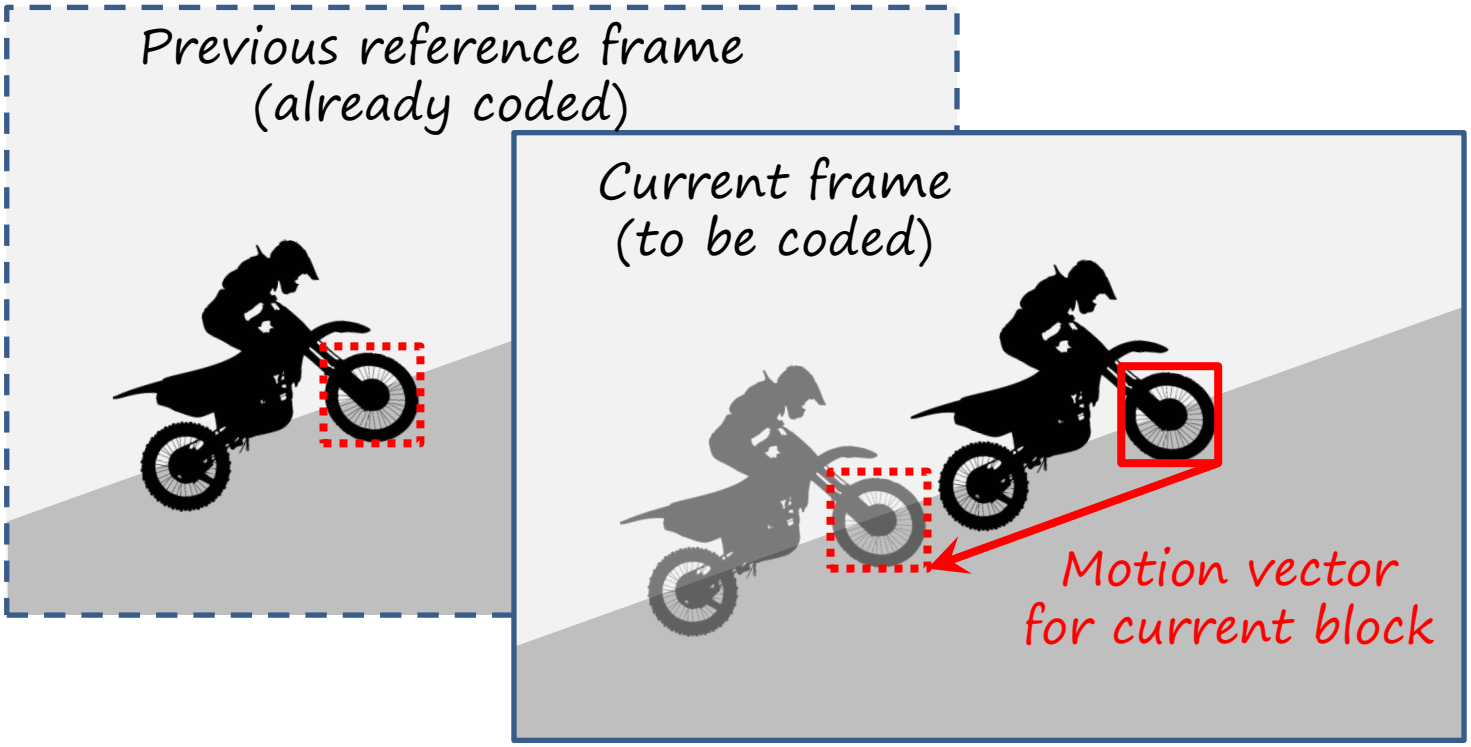
Coding video: temporal redundancy

Estimate current frame from previous coded ones



Motion-compensated prediction

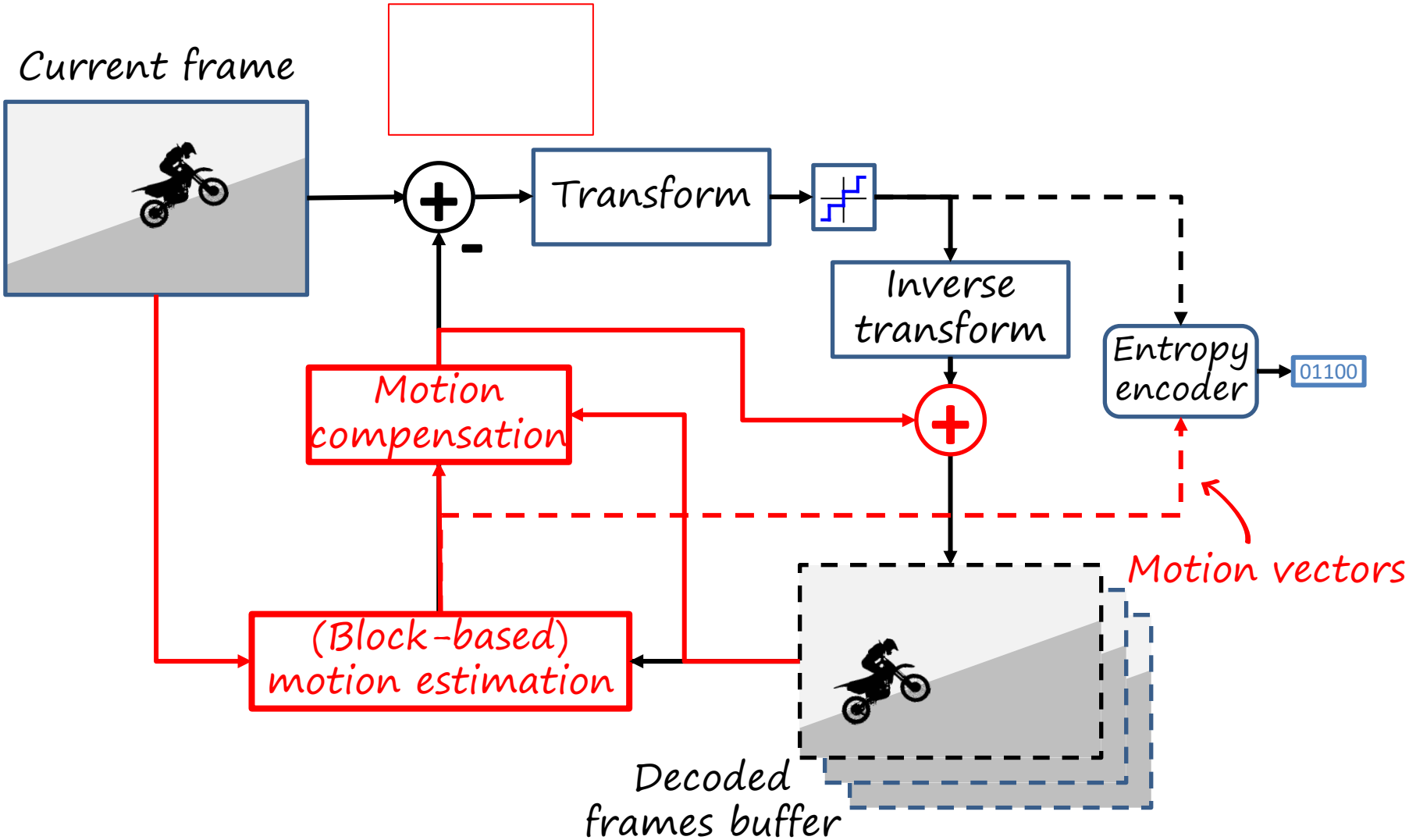
Try to align frames: find most similar blocks in the reference frame



Inspired by T. Wiegand's slide

Motion-compensated video coding

Estimate current frame from previous coded ones
Encode the motion vectors

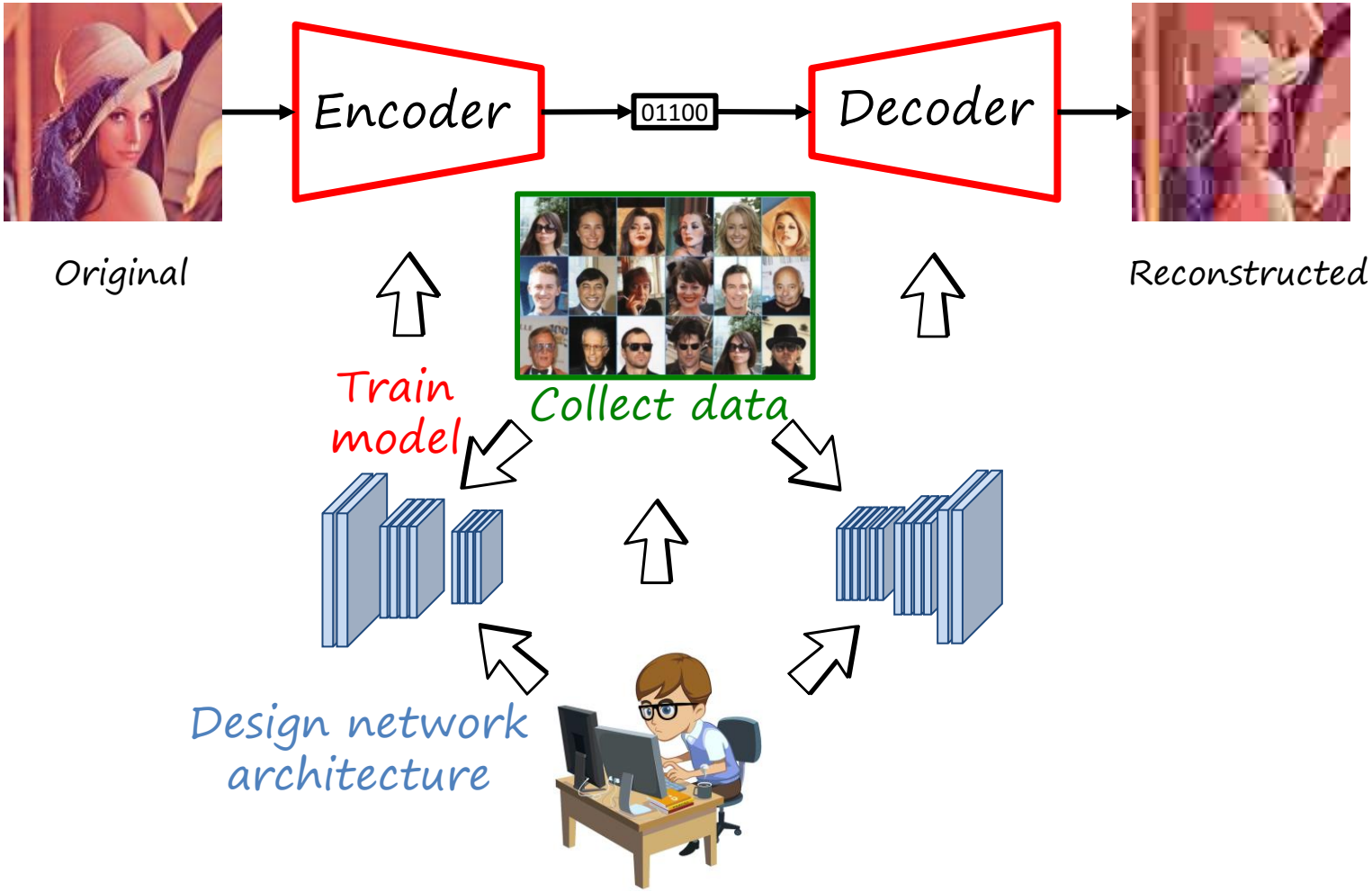


Outline

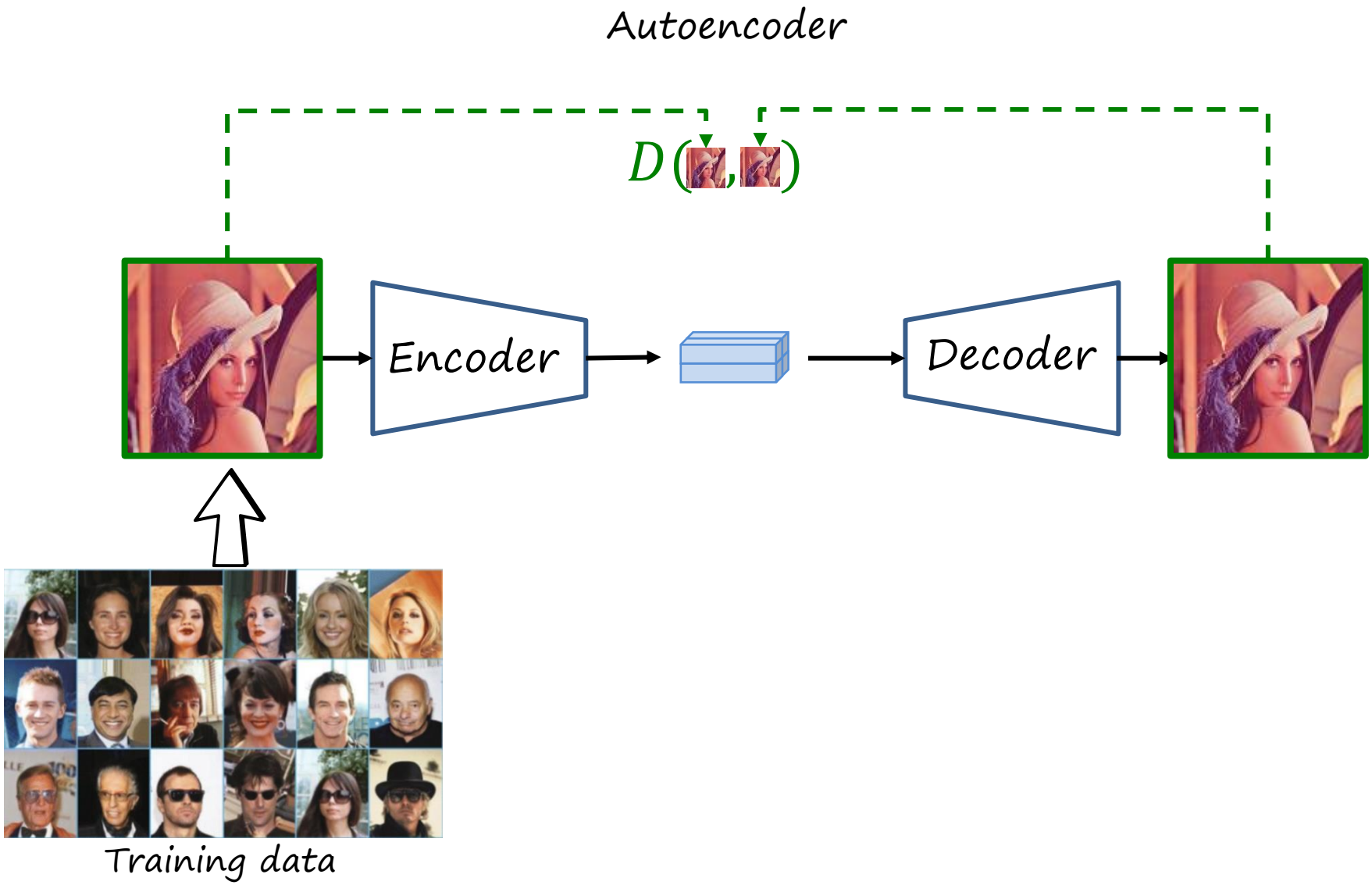
- Introduction: image coding
- Compression with neural networks
- Towards practical image compression

Neural image codecs

- Coding tools and syntax are **parametric** and **learned**
- Encoders/decoders and probability models **deep neural networks**

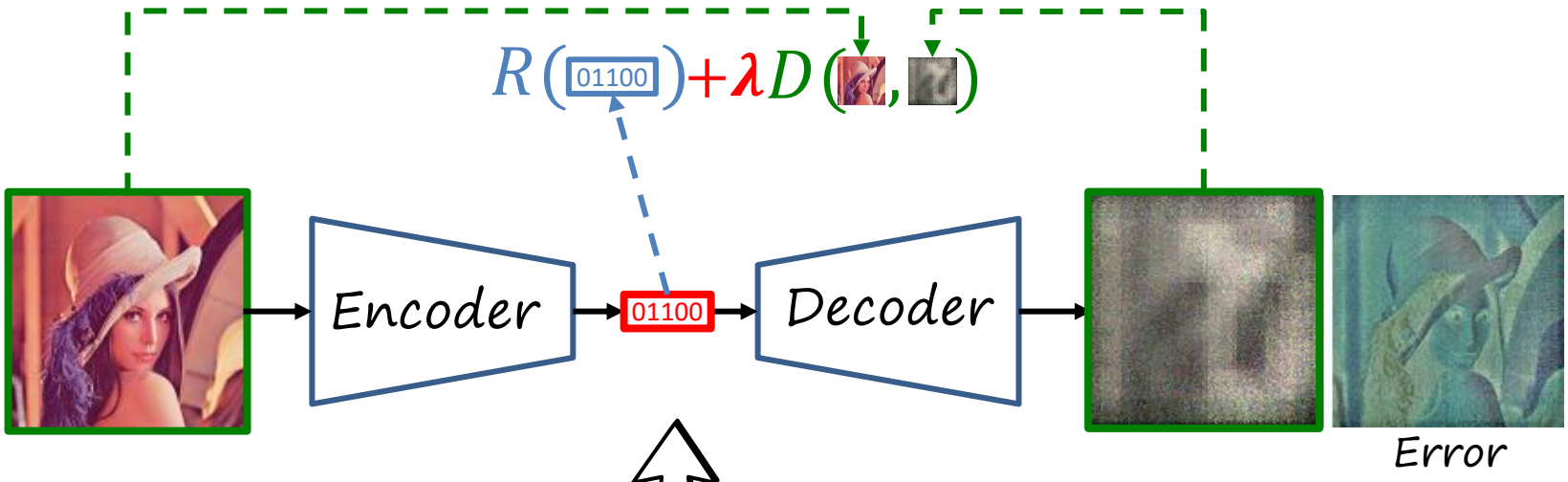


Neural image compression



Neural image compression

Compressive autoencoder (CAE) [Theis2017, Balle2017]
(autoencoder+binary latent representation)



Optimize a weighted rate-distortion loss
(λ controls the **tradeoff**)

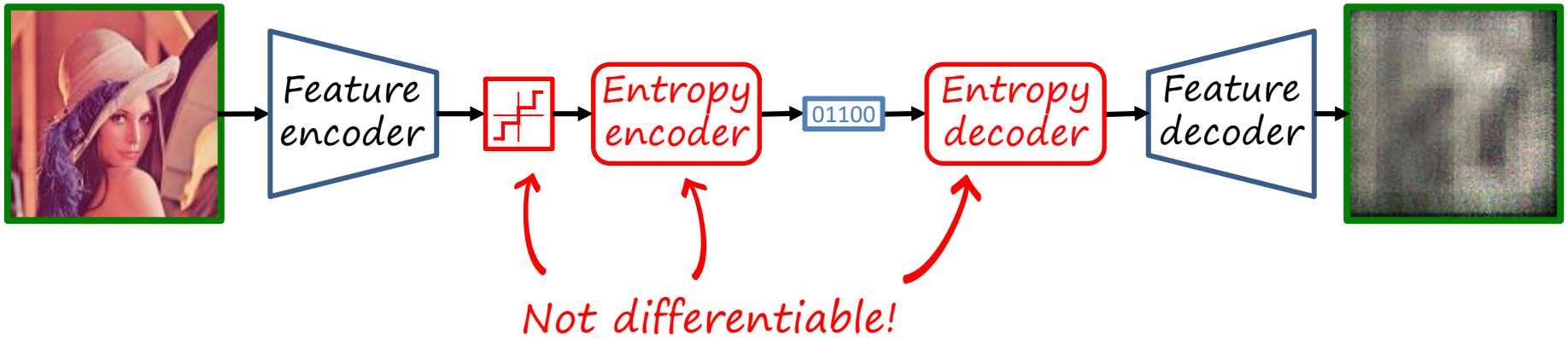
Training data

Distortion is typically mean square error (MSE)

Balle et al. End-to-end Optimized Image Compression, ICLR 2017
Theis et al., Lossy Image Compression with Compressive Autoencoders, ICLR 2017

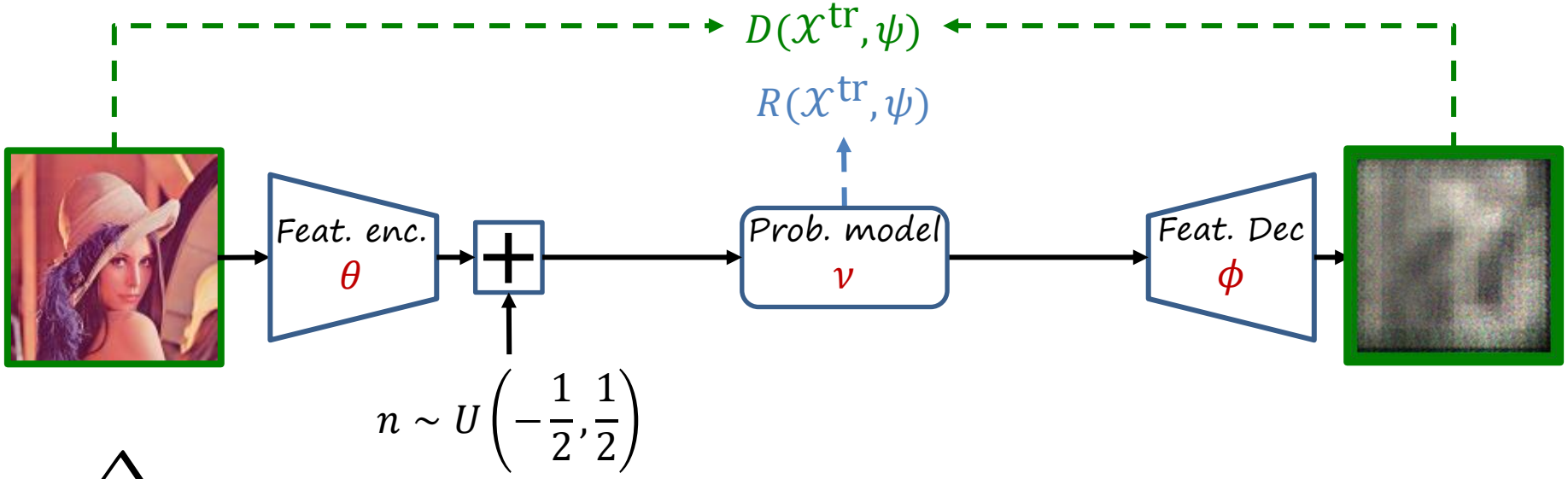
Typical pipeline

Compressive autoencoder (CAE) [Theis2017, Balle2017]
(autoencoder+quantization+entropy coding)



Architecture (training)

Use differentiable proxies for end-to-end training



Training data $\mathcal{X}^{\mathrm{tr}}$

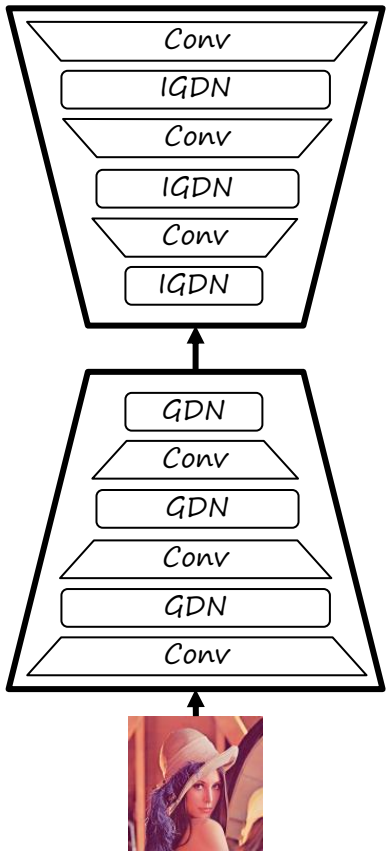
Model parameters $\psi = (\theta, \phi, \nu)$

Loss $J(\mathcal{X}^{\mathrm{tr}}, \psi; \lambda) = R(\mathcal{X}^{\mathrm{tr}}, \psi) + \lambda D(\mathcal{X}^{\mathrm{tr}}, \psi)$

Optimization problem $\psi^* = \min_{\psi} J(\mathcal{X}^{\mathrm{tr}}, \psi; \lambda)$

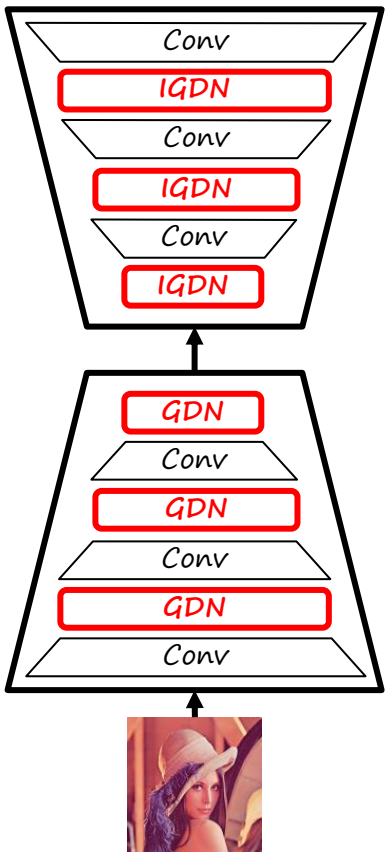
Autoencoder architecture

Balle et al.
[ICLR2017]



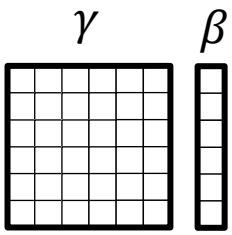
Autoencoder architecture

Balle et al.
[ICLR2017]



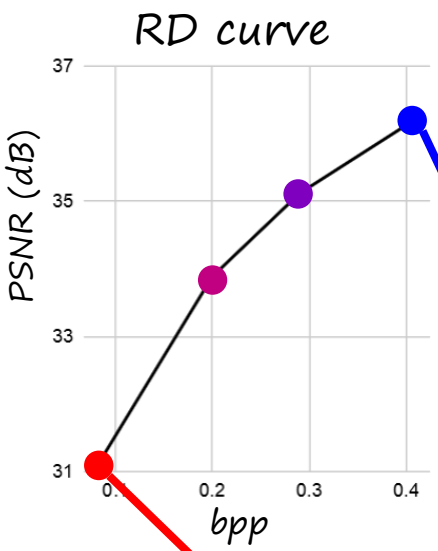
Generalized divisive normalization (GDN) [Balle2016]

$$\hat{y}_i = \frac{y_i}{\left(\beta_i + \sum_j \gamma_{ij}\, y_j^2\right)^{1/2}}$$



Learnable parameters

Rate-distortion tradeoff λ

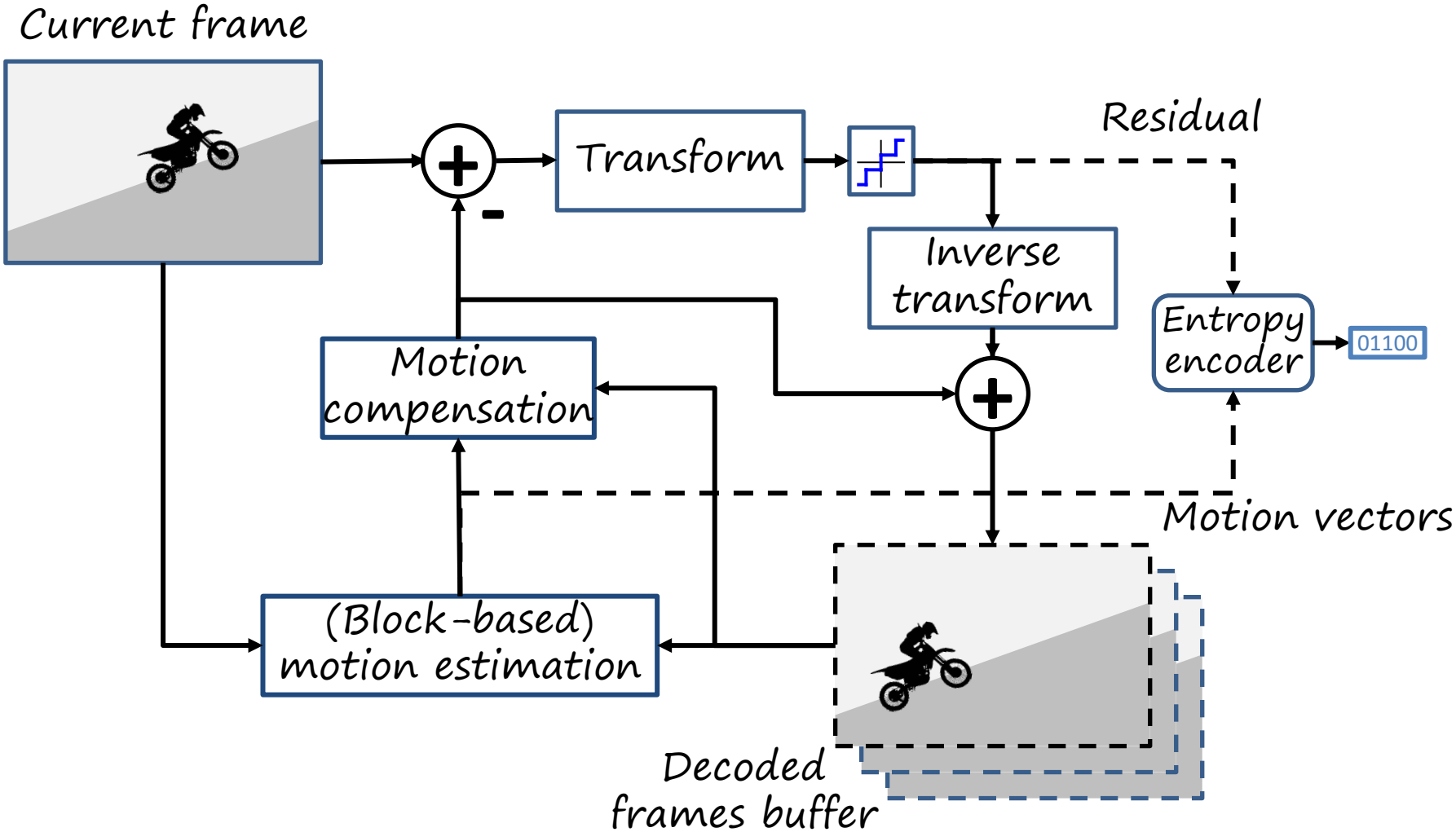


High rate (λ=0.032) PSNR= 36.2 dB Rate= 0.41 bpp

Low rate (λ=0.002) PSNR= 31.1 dB Rate= 0.08 bpp

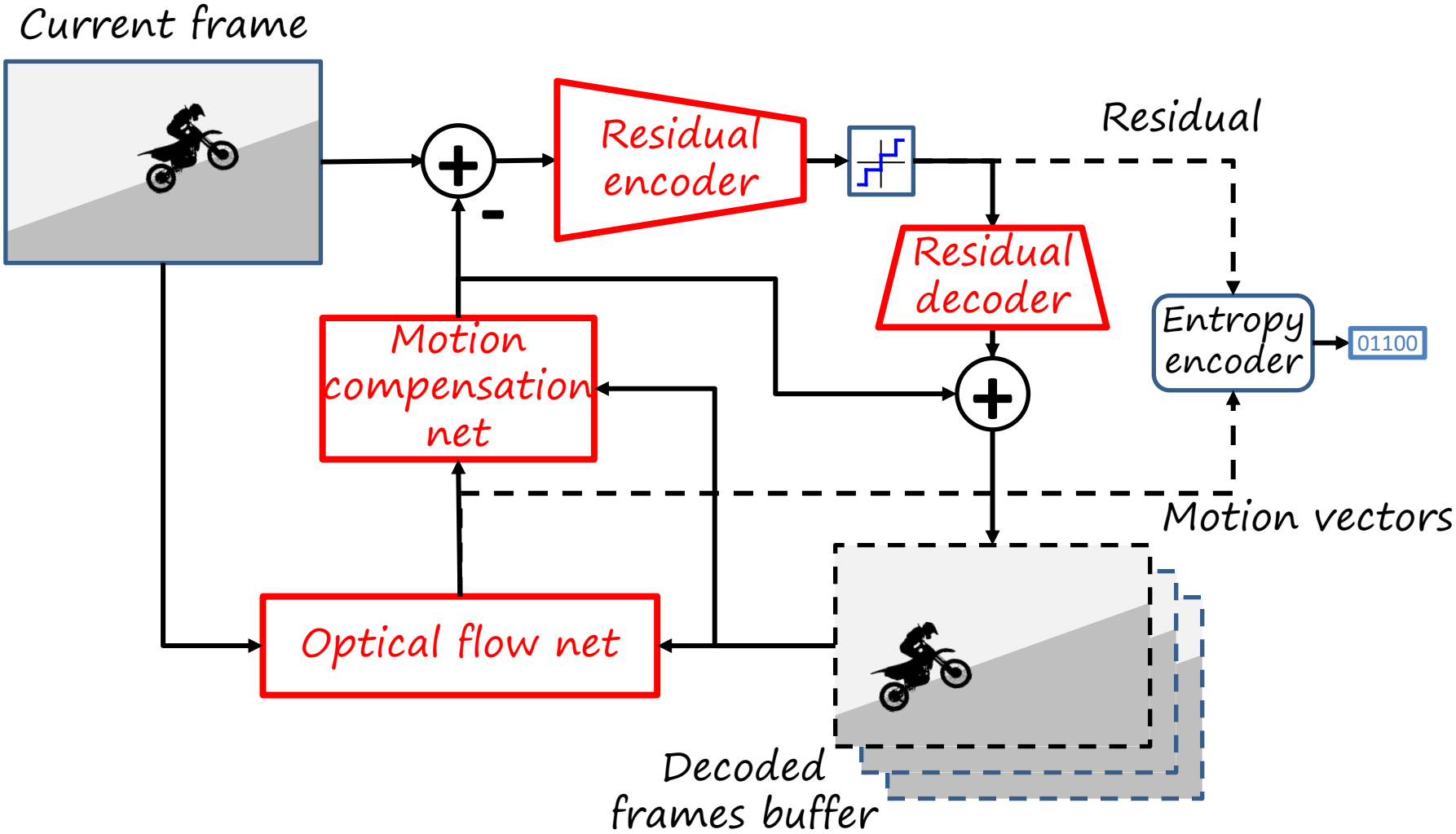
Traditional video compression

Replace modules by trainable neural networks



Neural video compression

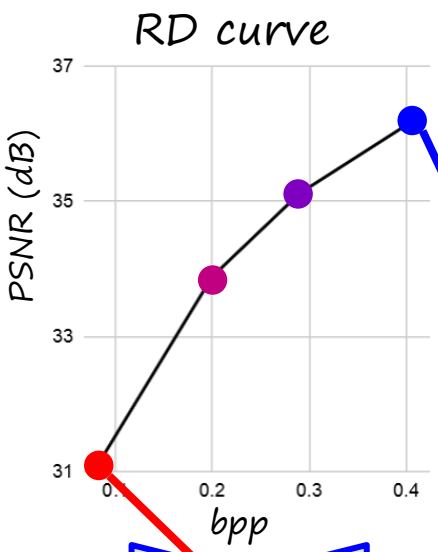
Replace modules by trainable neural networks



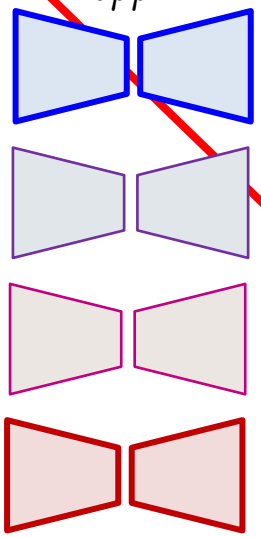
Outline

- Introduction: image coding
- Compression with neural networks
- Towards practical image compression

Rate-distortion tradeoff λ



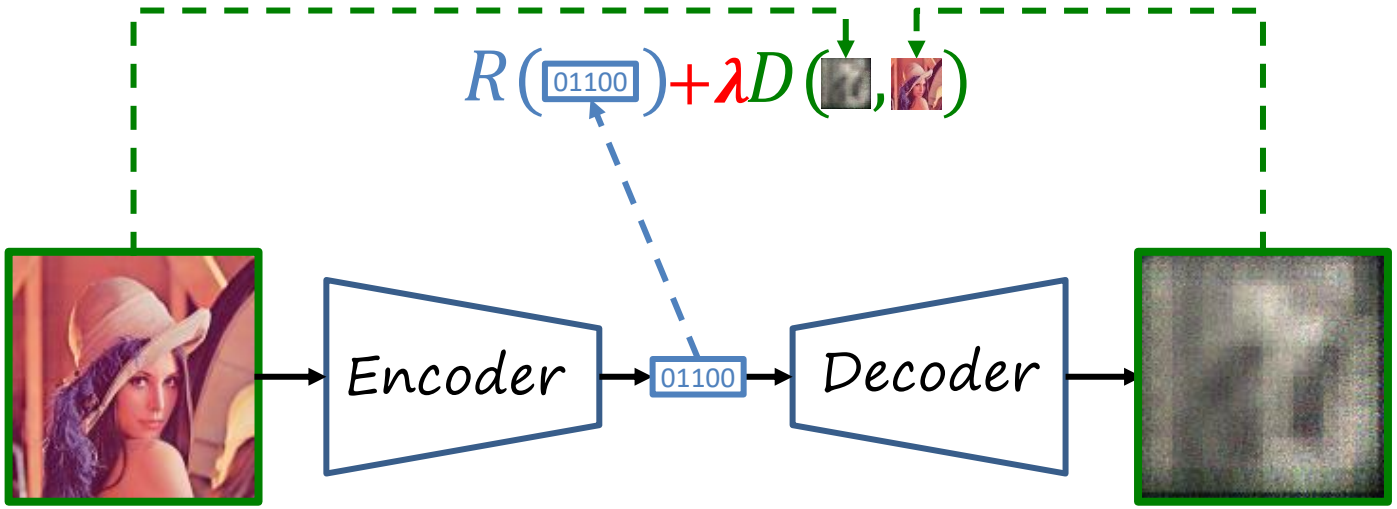
High rate (λ=0.032) PSNR= 36.2 dB Rate= 0.41 bpp



Low rate (λ=0.002) PSNR= 31.1 dB Rate= 0.08 bpp

Problems: total memory, total training time

Is neural image compression practical?



Practical neural image compression?
- Minimize rate ✓
- Minimize distortion ✓
- Variable rate ✗
- Low memory ✗
- Low computation ✗
- Low latency ✗

MAE [SPL2020]

SlimCAE [CVPR2021]

Limitations
- λ is fixed
- Heavy encoders/decoders

Other practical considerations
- **Domain-specific codecs** (e.g. videoconference, screencast)
- **Back./forw. compatibility** (with legacy encoders/decoders)

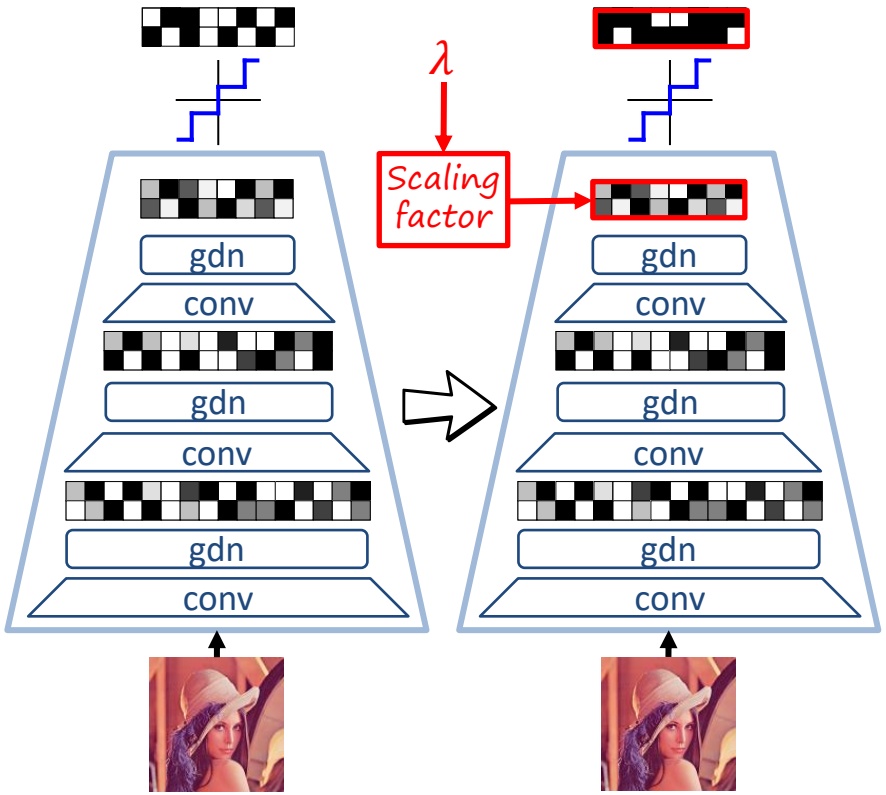
DANICE [CLIC2021]

[SPL2020] Variable Rate Deep Image Compression with Modulated Autoencoder, Signal Processing Letters 2020
[CVPR2021] Slimmable compressive autoencoders for practical imaga compression, CVPR 2021
[CLIC2021] DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021

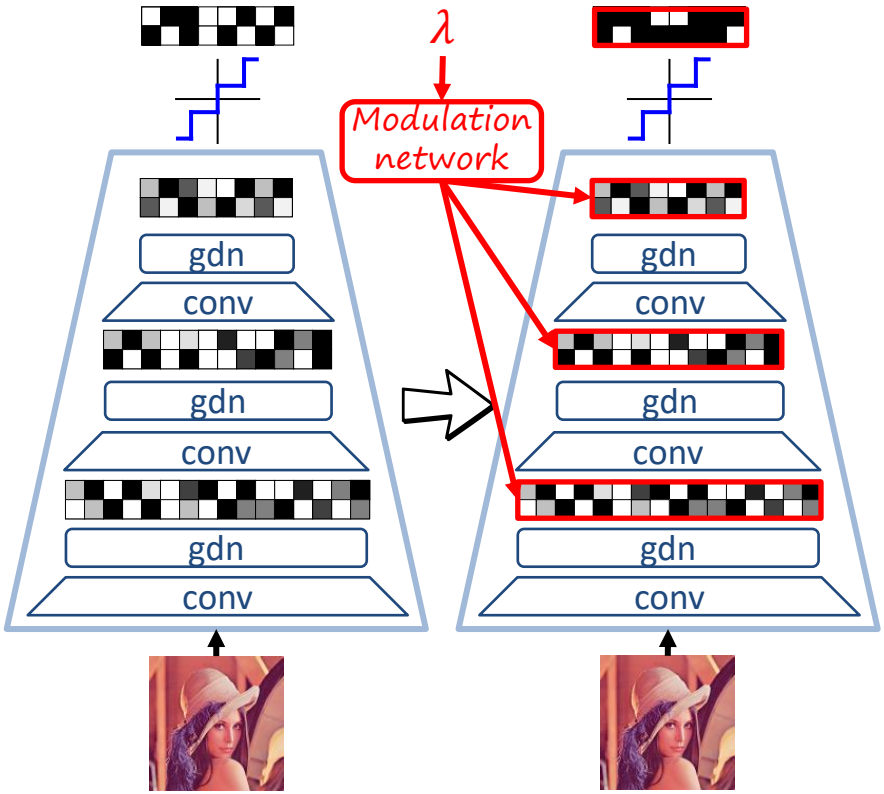
Variable rate with modulated autoencoders

Objective: one single model for multiple λ

Bottleneck scaling [Theis2017]



Feature modulation [MAE, cAE]



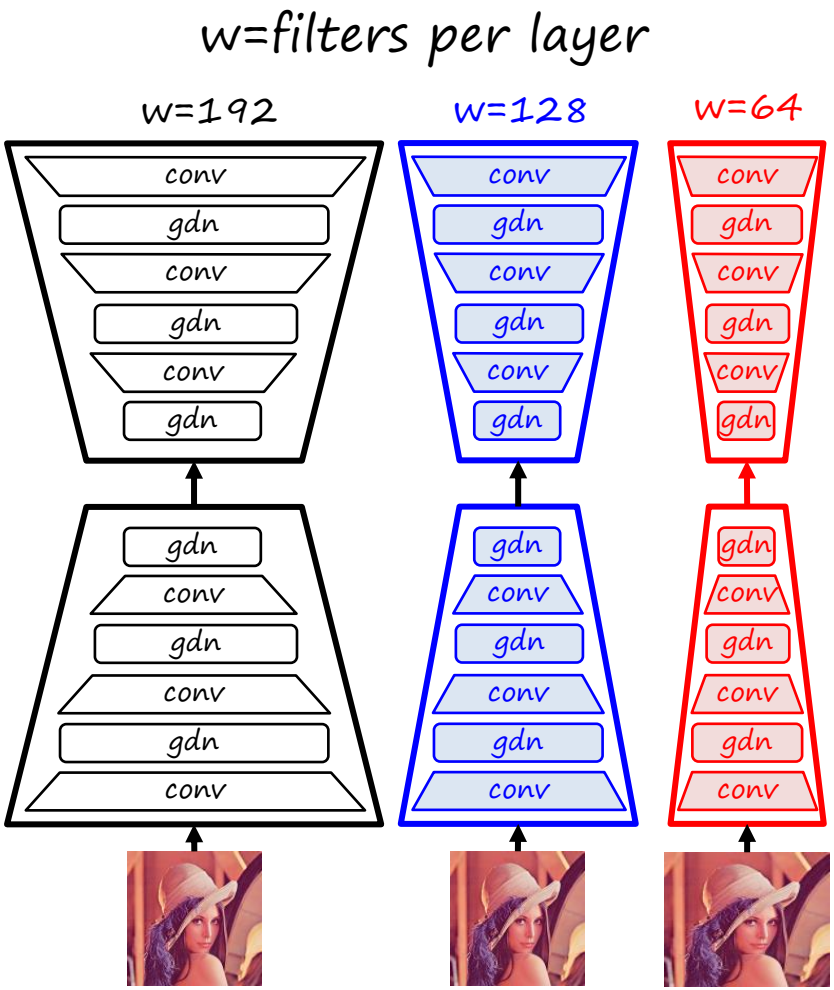
- *Minimize rate* ✓
- *Minimize distortion* ✓
- *Variable rate* ✓

- *Low memory* ✗
- *Low computation* ✗
- *Low latency* ✗

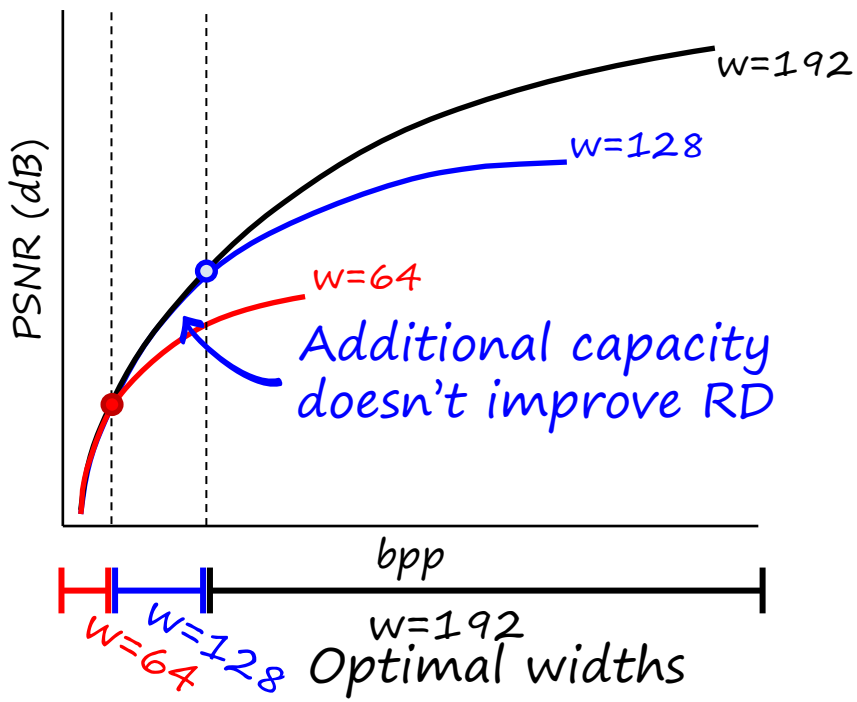
[Yang et al., Variable Rate Deep Image Compression with Modulated Autoencoder](), Signal Processing Letters 2020

cAE: conditional autoencoder [Choi2019]
MAE: modulated autoencoder [Yang2020]

Model capacity and rate-distortion



There is a **minimal capacity** for every RD tradeoff



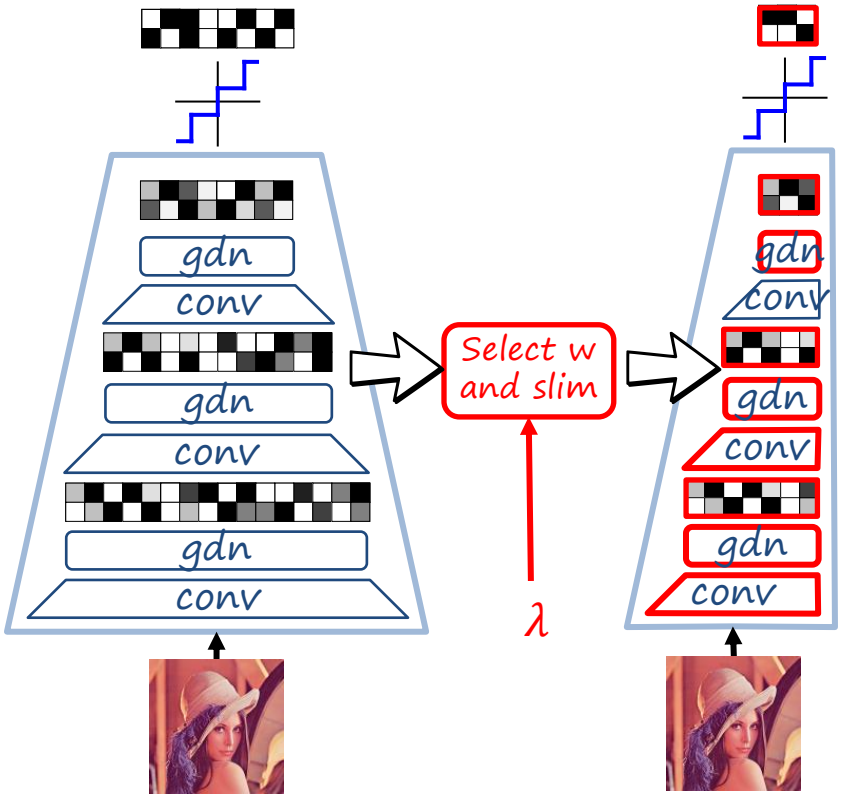
Lower w results in less memory and computation!!

Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Slimmable compressive autoencoder

Approach: slim the network to the minimal capacity for a given λ

Slimming [SlimCAE]

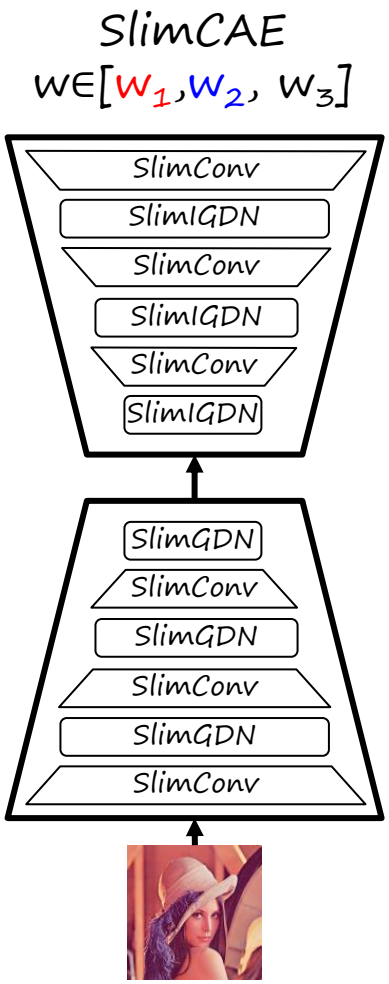


- Minimize rate ✓
- Minimize distortion ✓
- Variable rate ✓
- Lower memory ✓
- Lower computation ✓
- Lower latency ✓

(for low-mid rates)

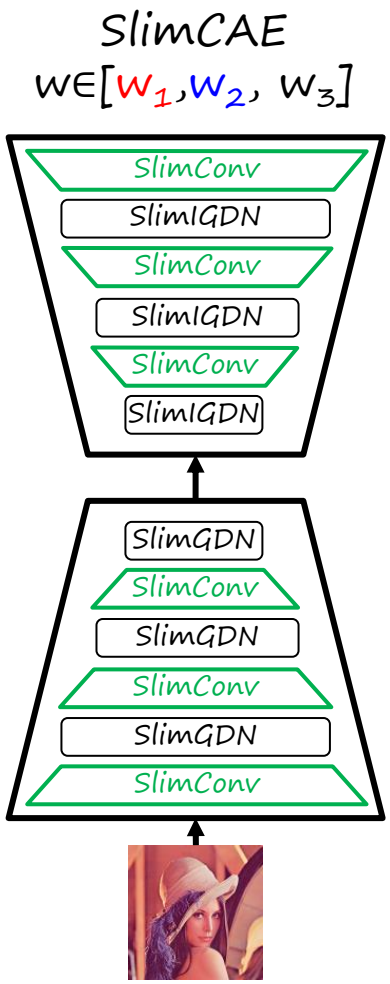
Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

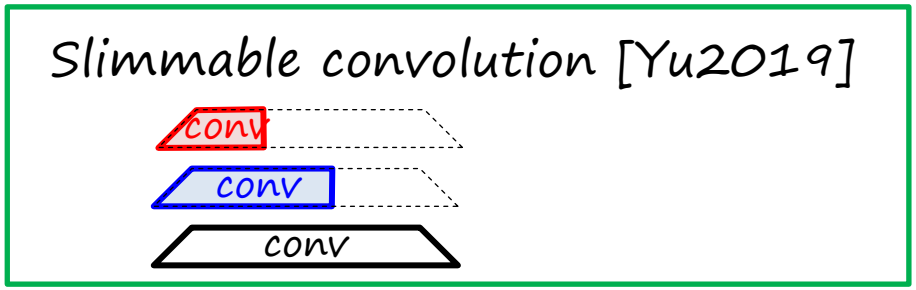
Slimmable layers in SlimCAE



[Slimmable compressive autoencoders for practical imaga compression](), CVPR 2021

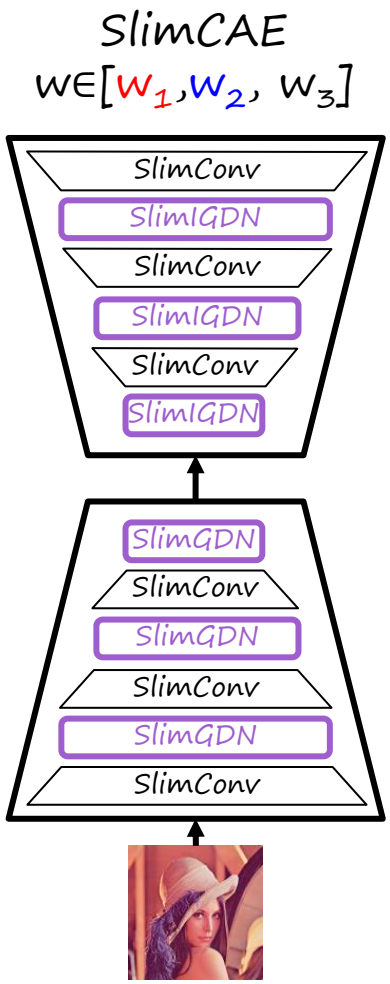
Slimmable layers in SlimCAE

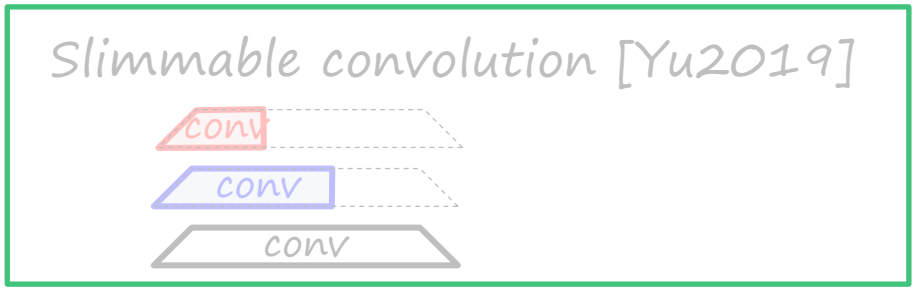


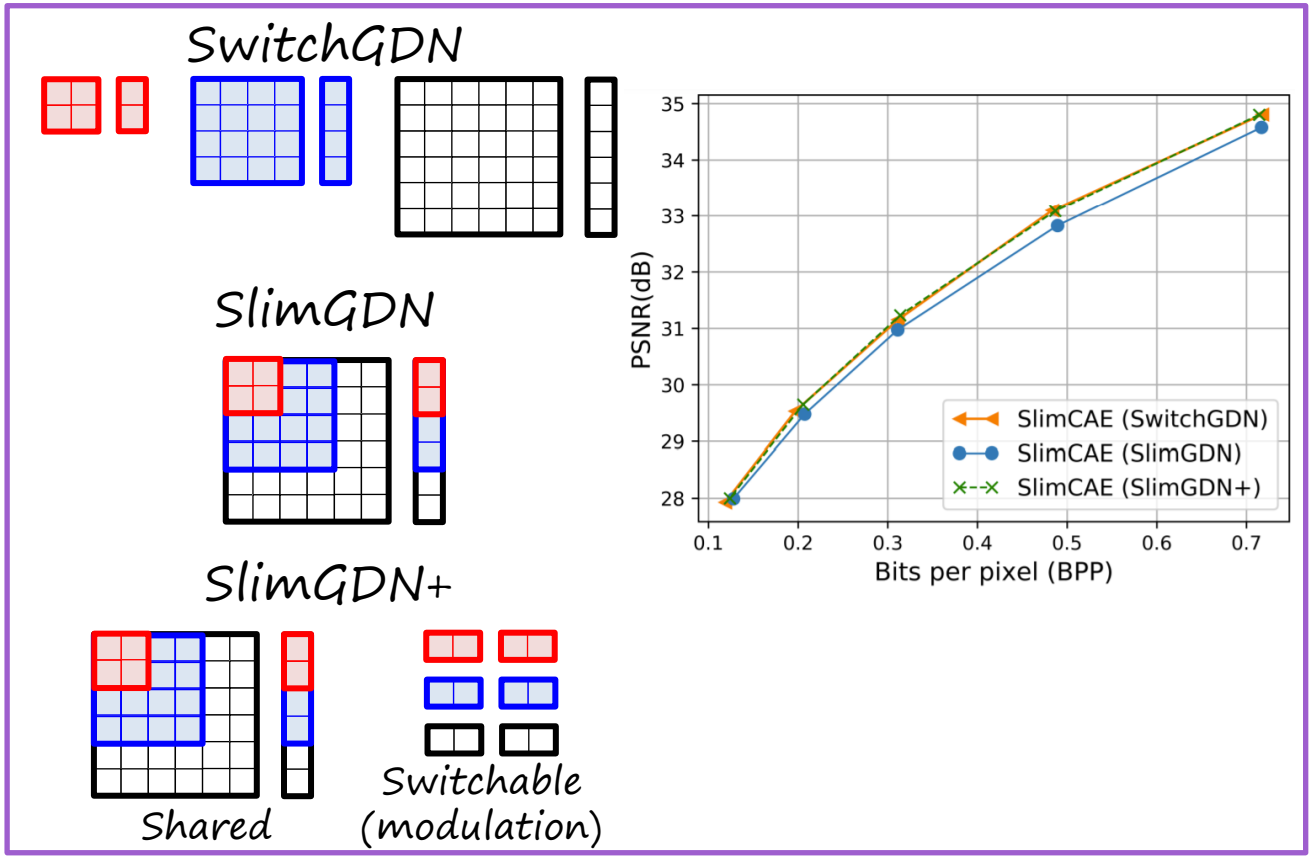


Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Slimmable layers in SlimCAE





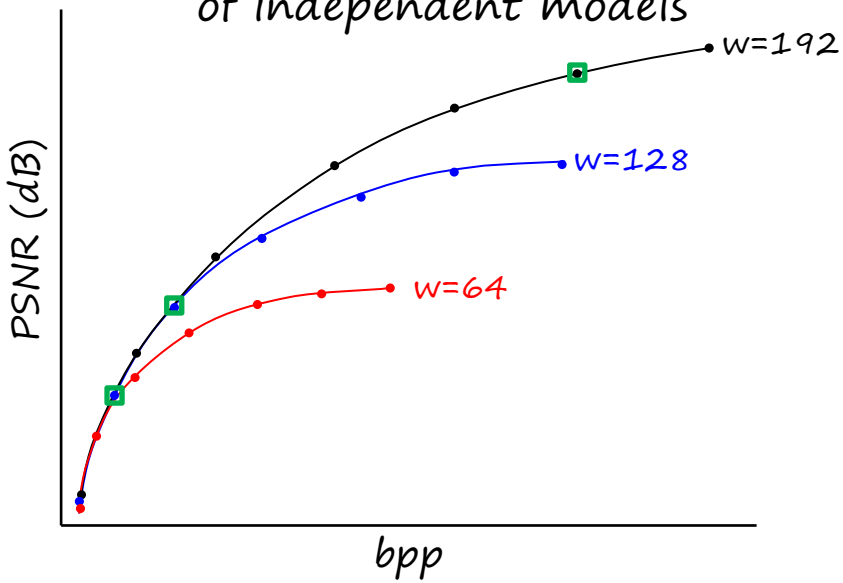


Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Training SlimCAE

Problem: we need the **optimal λs** to train the SlimCAE

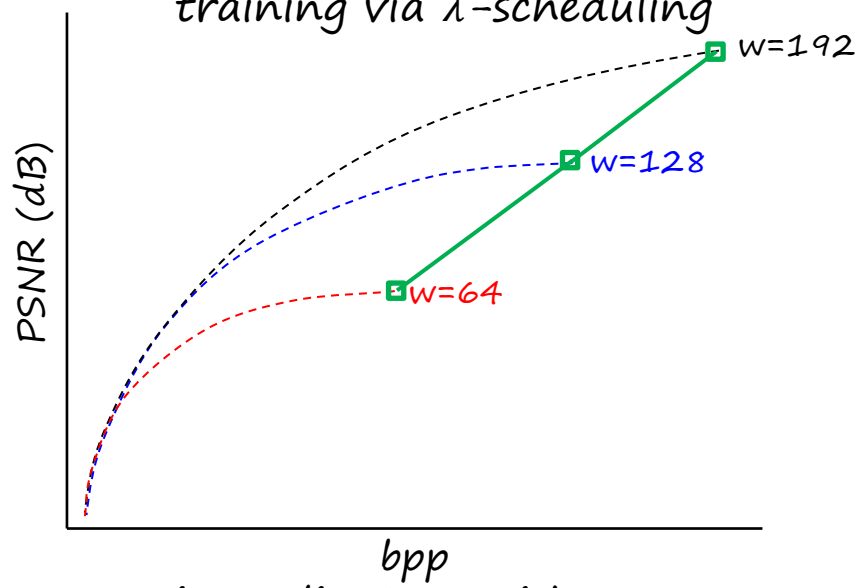
Estimate from RD curves of independent models



1. Train several independent models for different w
2. Plot RD curves and find critical points
3. Estimate optimal λs from trained models

Problem: extremely expensive!

Automatically estimate during training via λ-scheduling



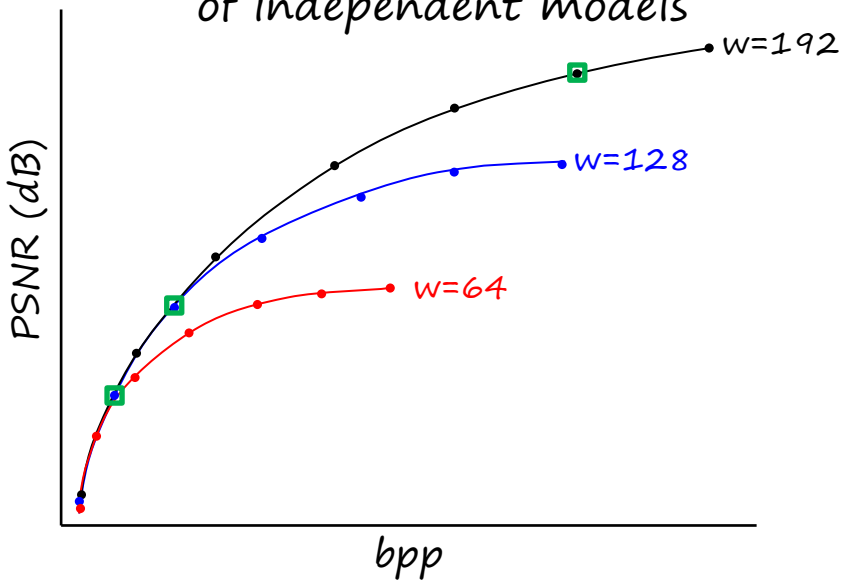
1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Training SlimCAE

Problem: we need the optimal λs to train the SlimCAE

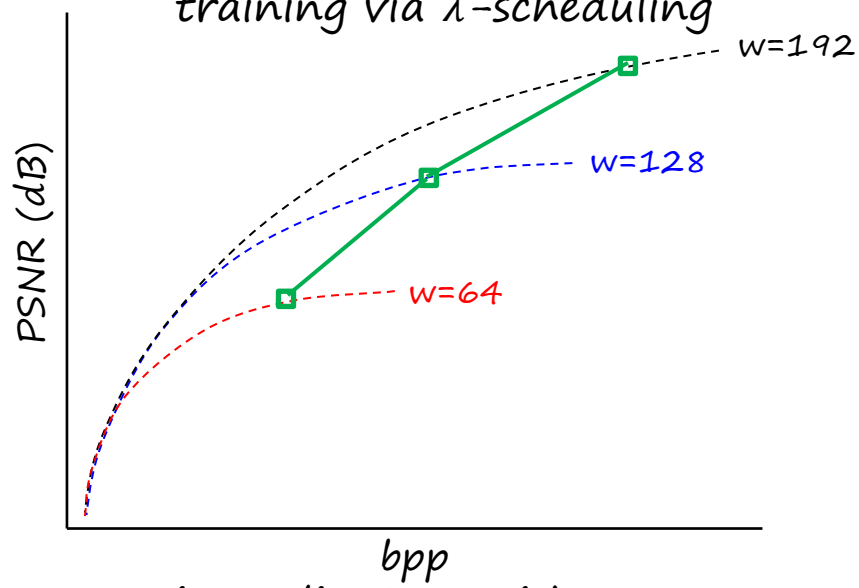
Estimate from RD curves of independent models



1. Train several independent models for different w
2. Plot RD curves and find critical points
3. Estimate optimal λs from trained models

Problem: extremely expensive!

Automatically estimate during training via λ-scheduling



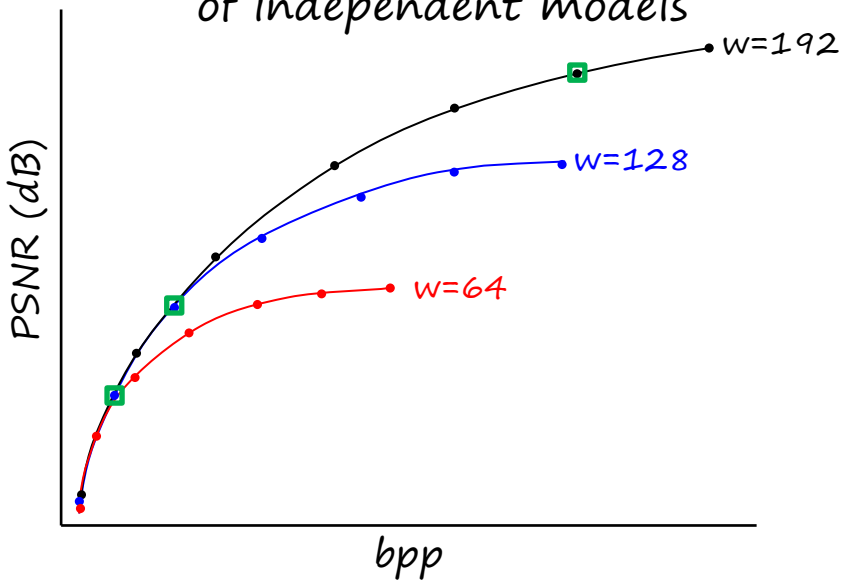
1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Training SlimCAE

Problem: we need the optimal λs to train the SlimCAE

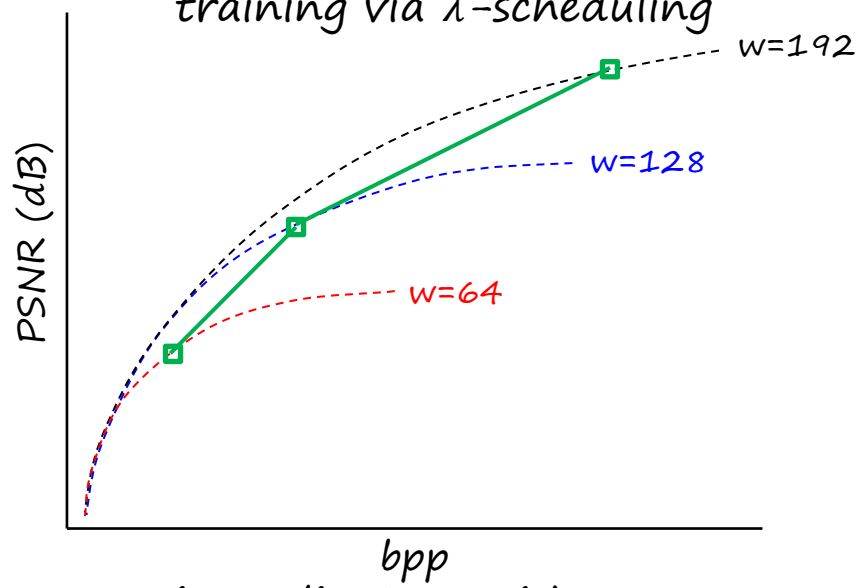
Estimate from RD curves of independent models



1. Train several independent models for different w
2. Plot RD curves and find critical points
3. Estimate optimal λs from trained models

Problem: extremely expensive!

Automatically estimate during training via λ-scheduling



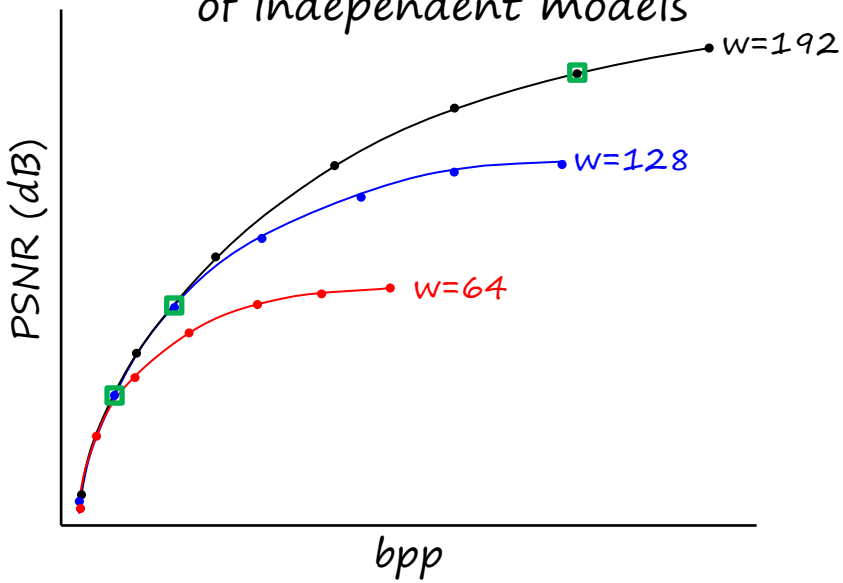
1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Training SlimCAE

Problem: we need the optimal λs to train the SlimCAE

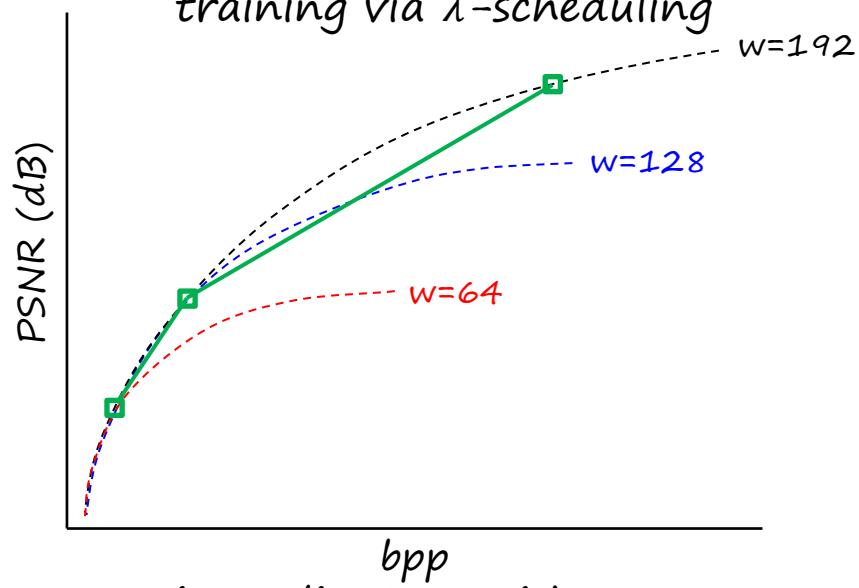
Estimate from RD curves of independent models



1. Train several independent models for different w
2. Plot RD curves and find critical points
3. Estimate optimal λs from trained models

Problem: extremely expensive!

Automatically estimate during training via λ-scheduling



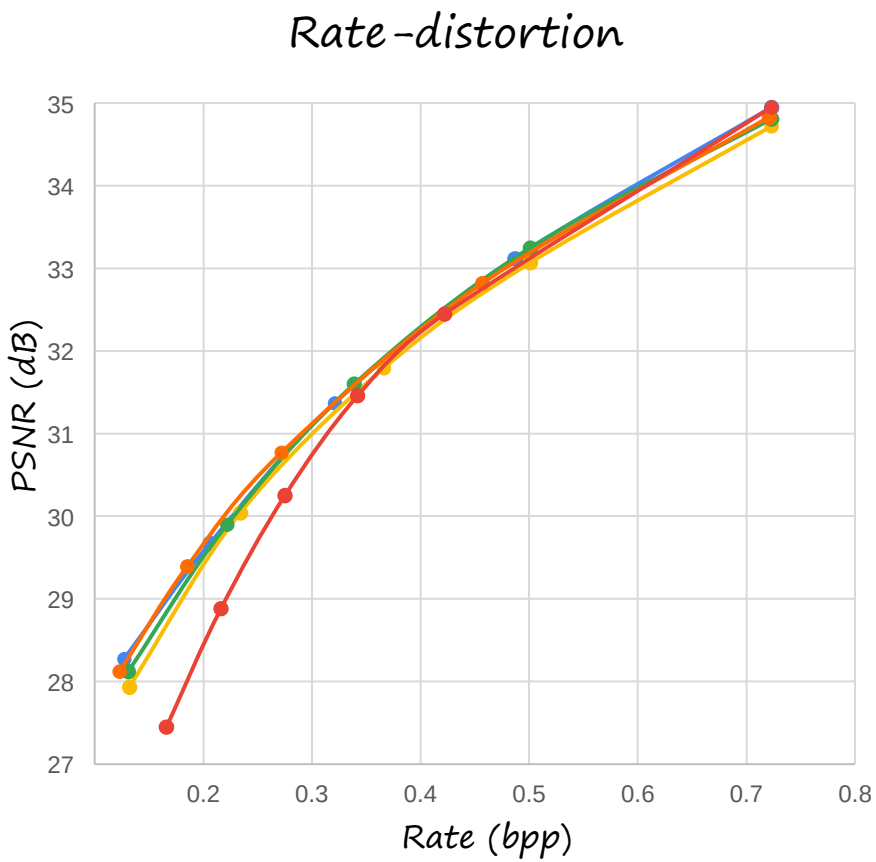
1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

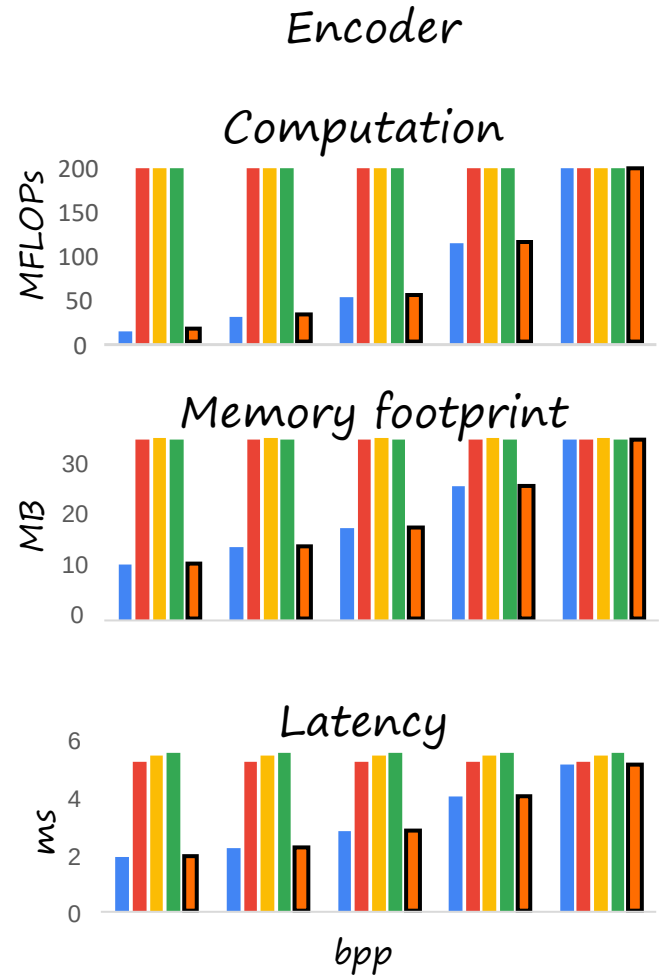
Directly train one model!

Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Performance comparison

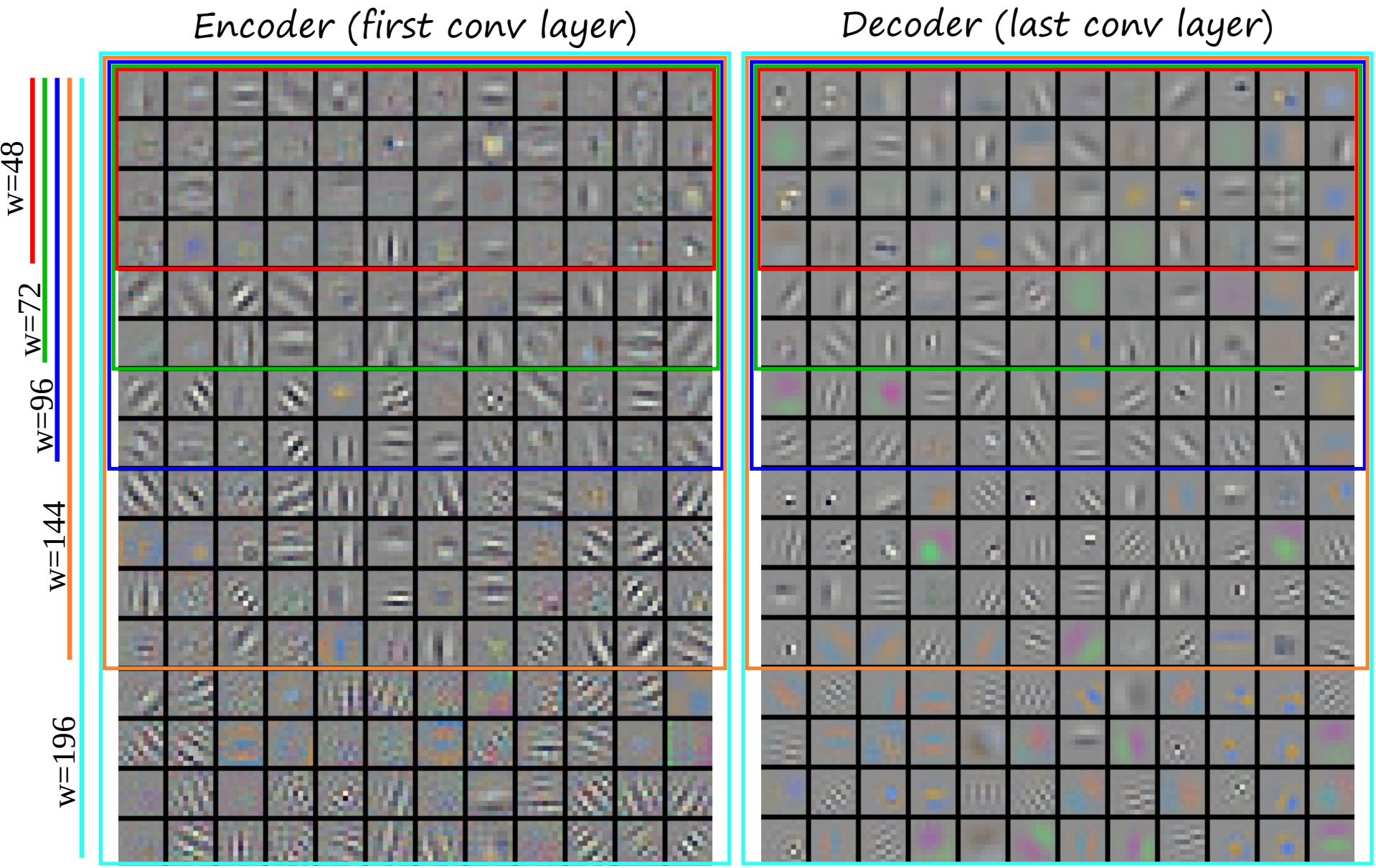
Independent CAEs (each with minimal capacity) | Scaling [Theis2017] | MAE [Yang2020] | cAE [Choi2019] | SlimCAE (ours)





Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

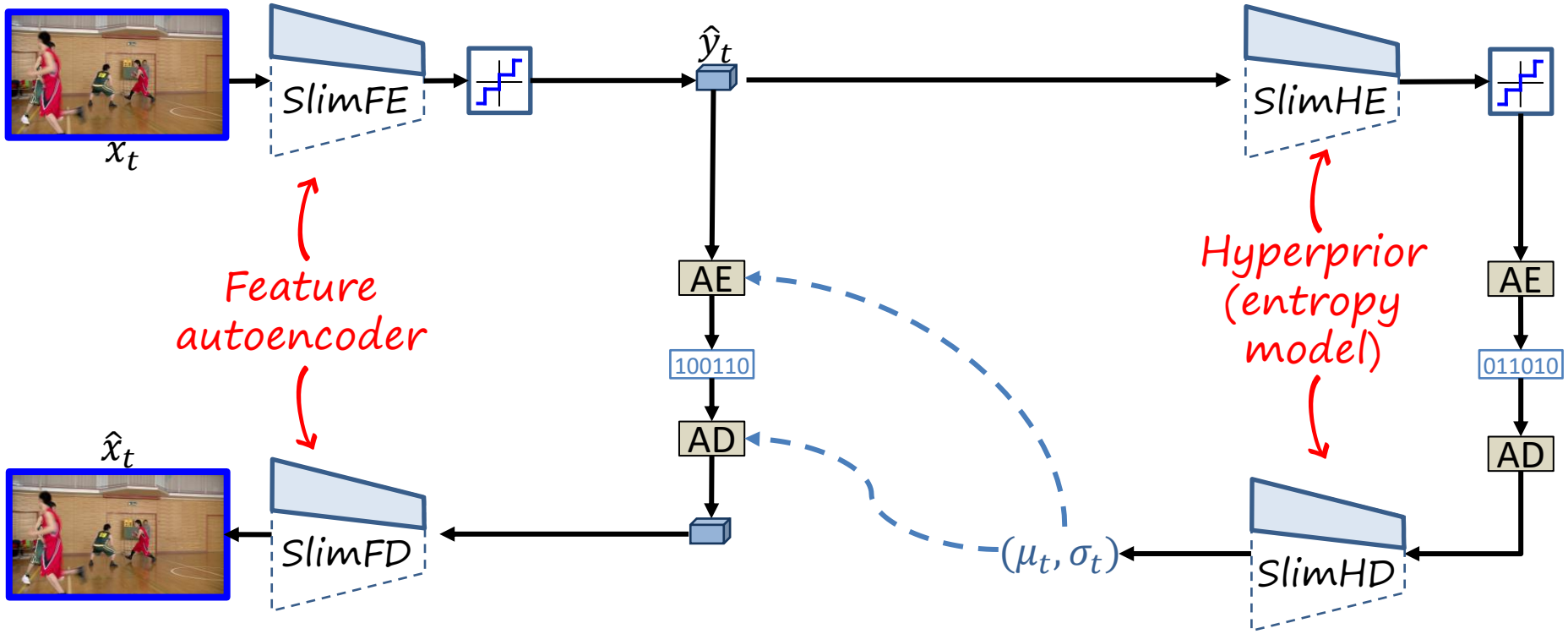
Visualizing some parameters



Slimmable compressive autoencoders for practical imaga compression, CVPR 2021

Slimmable video codec (SlimVC)

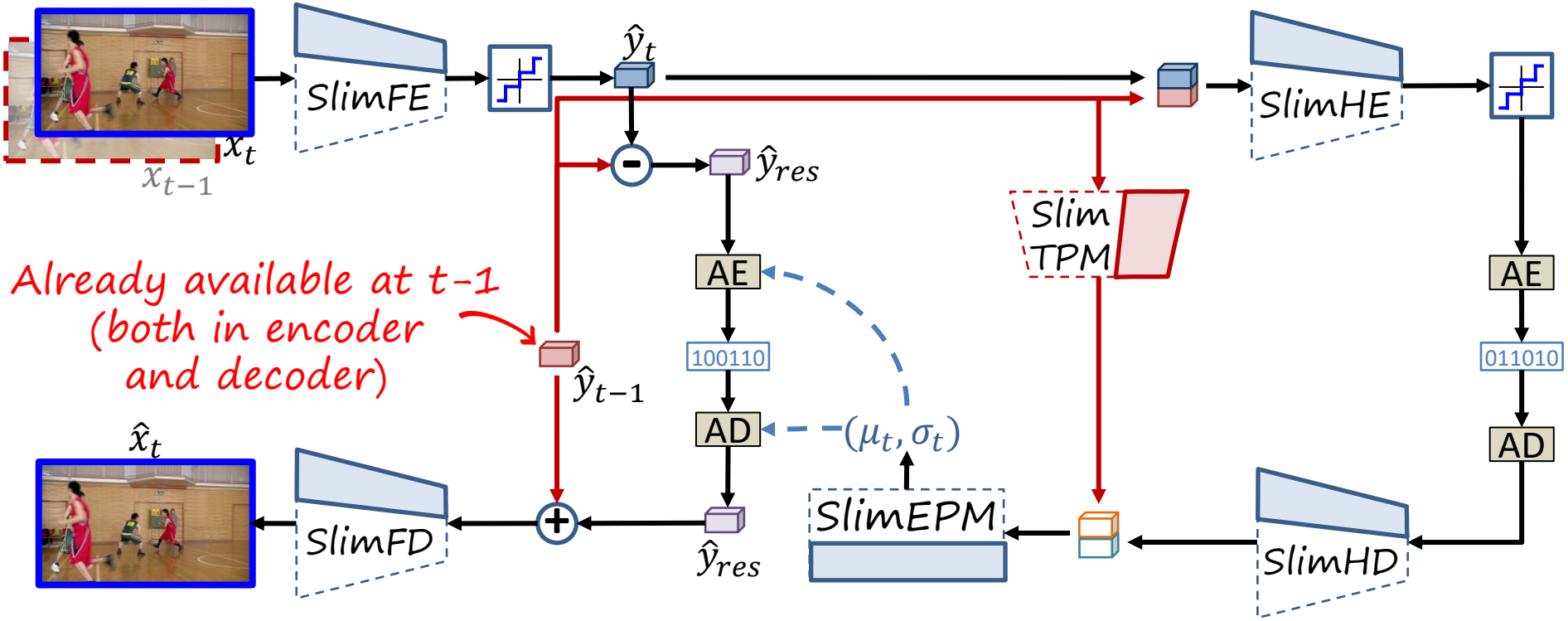
Extending SlimCAE to video



Slimmable video codec, 2022 (under review)

Slimmable video codec (SlimVC)

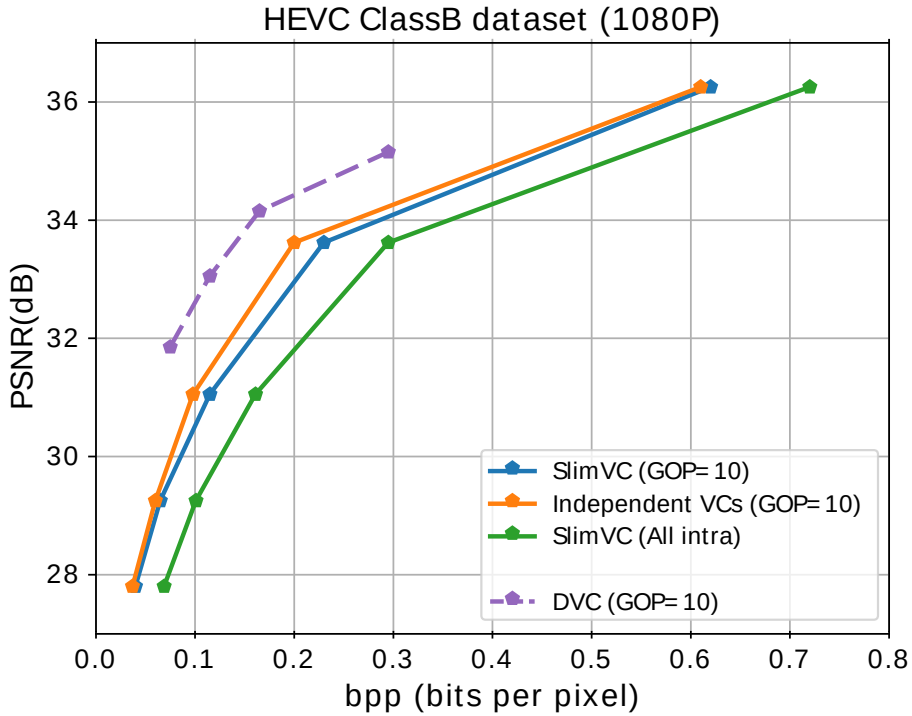
Extending SlimCAE to video



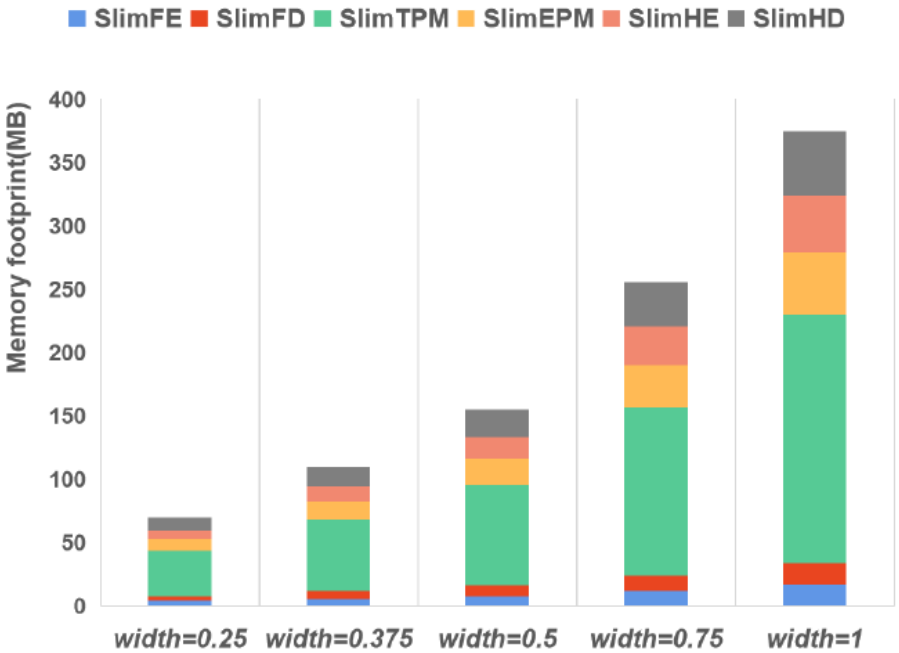
Slimmable video codec, 2022 (under review)

Slimmable video codec (SlimVC)

RD performance



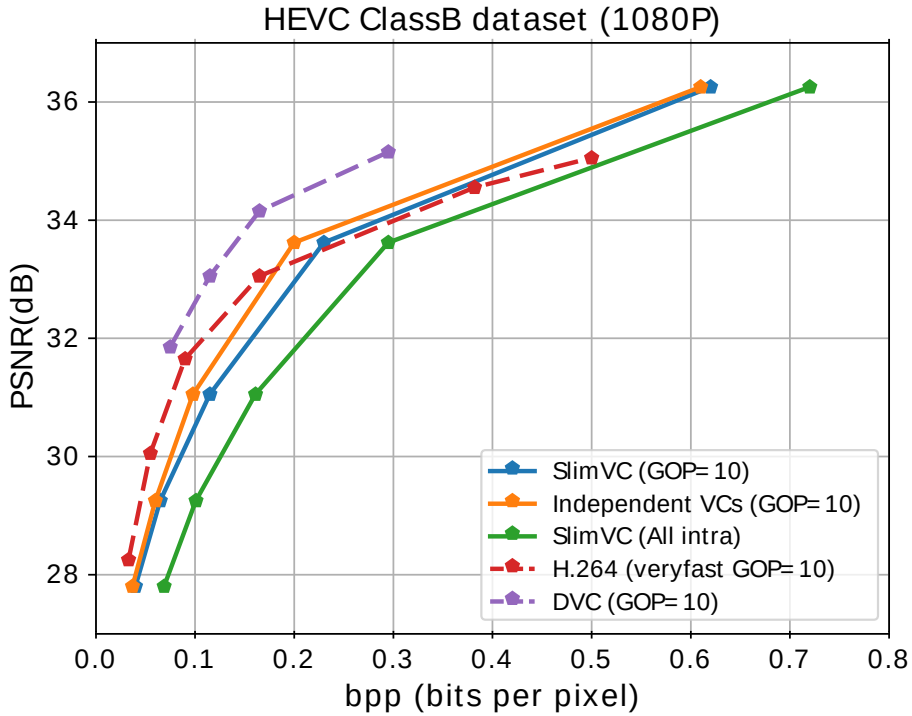
Memory footprint



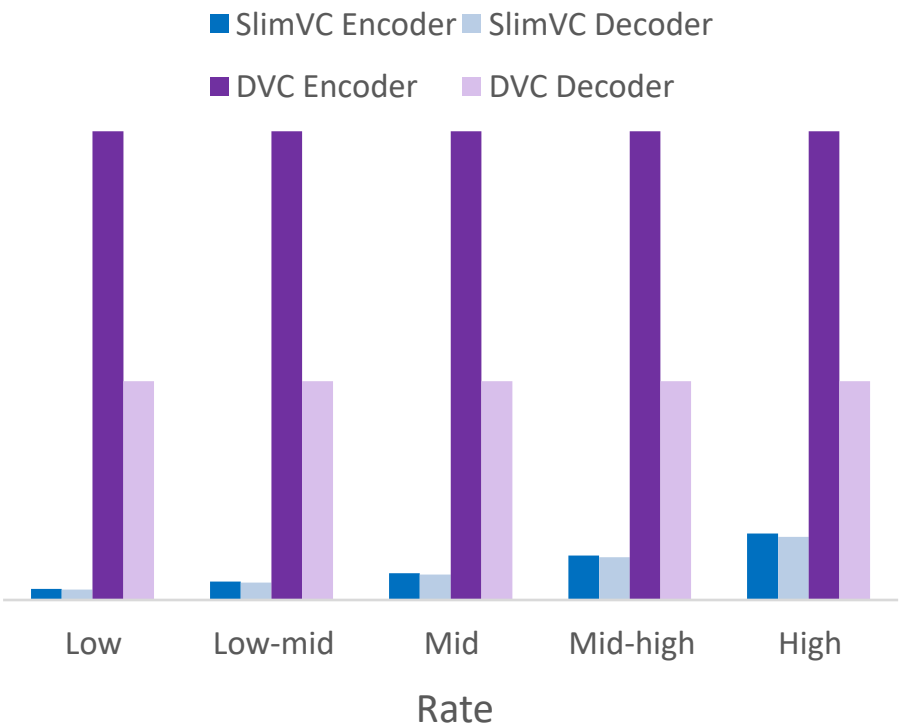
Slimmable video codec, 2022 (under review)

Slimmable video codec (SlimVC)

RD performance

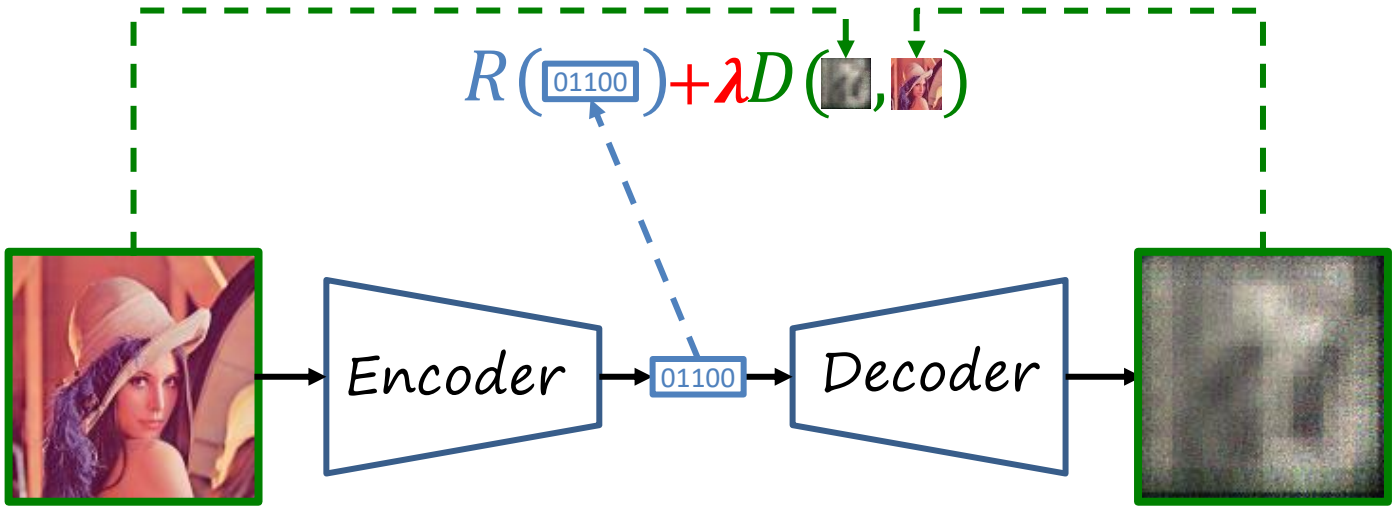


Computational cost (GFLOPs)



Slimmable video codec, 2022 (under review)

Is neural image compression practical?



Practical neural image compression?
- Minimize rate ✓
- Minimize distortion ✓
- Variable rate ✗
- Low memory ✗
- Low computation ✗
- Low latency ✗

MAE [SPL2020]

SlimCAE [CVPR2021]

Limitations
- λ is fixed
- Heavy encoders/decoders

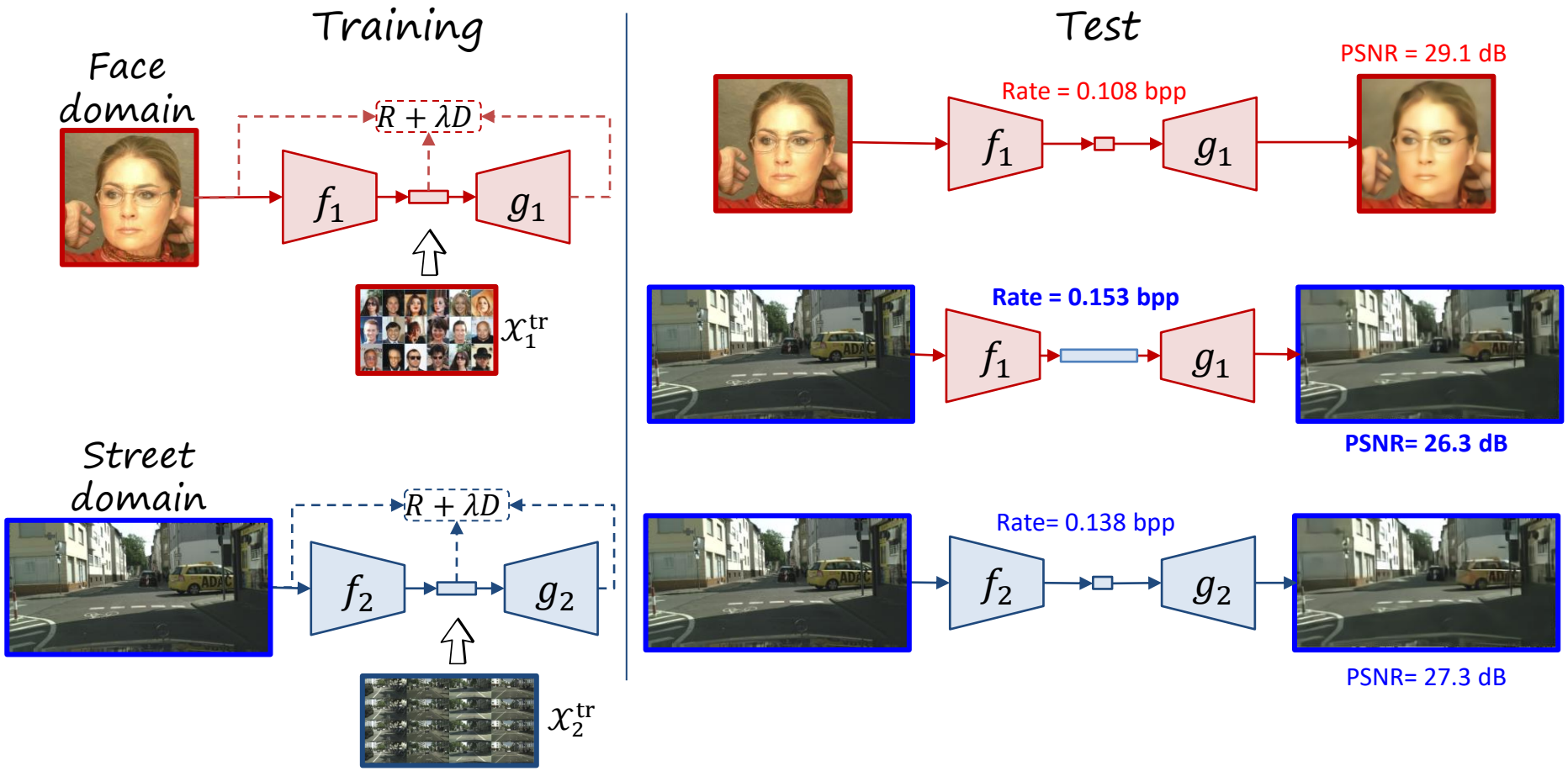
Other practical considerations
- **Domain-specific codecs** (e.g. videoconference, screencast)
- **Back./forw. compatibility** (with legacy encoders/decoders)

DANICE [CLIC2021]

[SPL2020] Variable Rate Deep Image Compression with Modulated Autoencoder, Signal Processing Letters 2020
[CVPR2021] Slimmable compressive autoencoders for practical imaga compression, CVPR 2021
[CLIC2021] DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021

Rate-distortion optimality of learned codecs

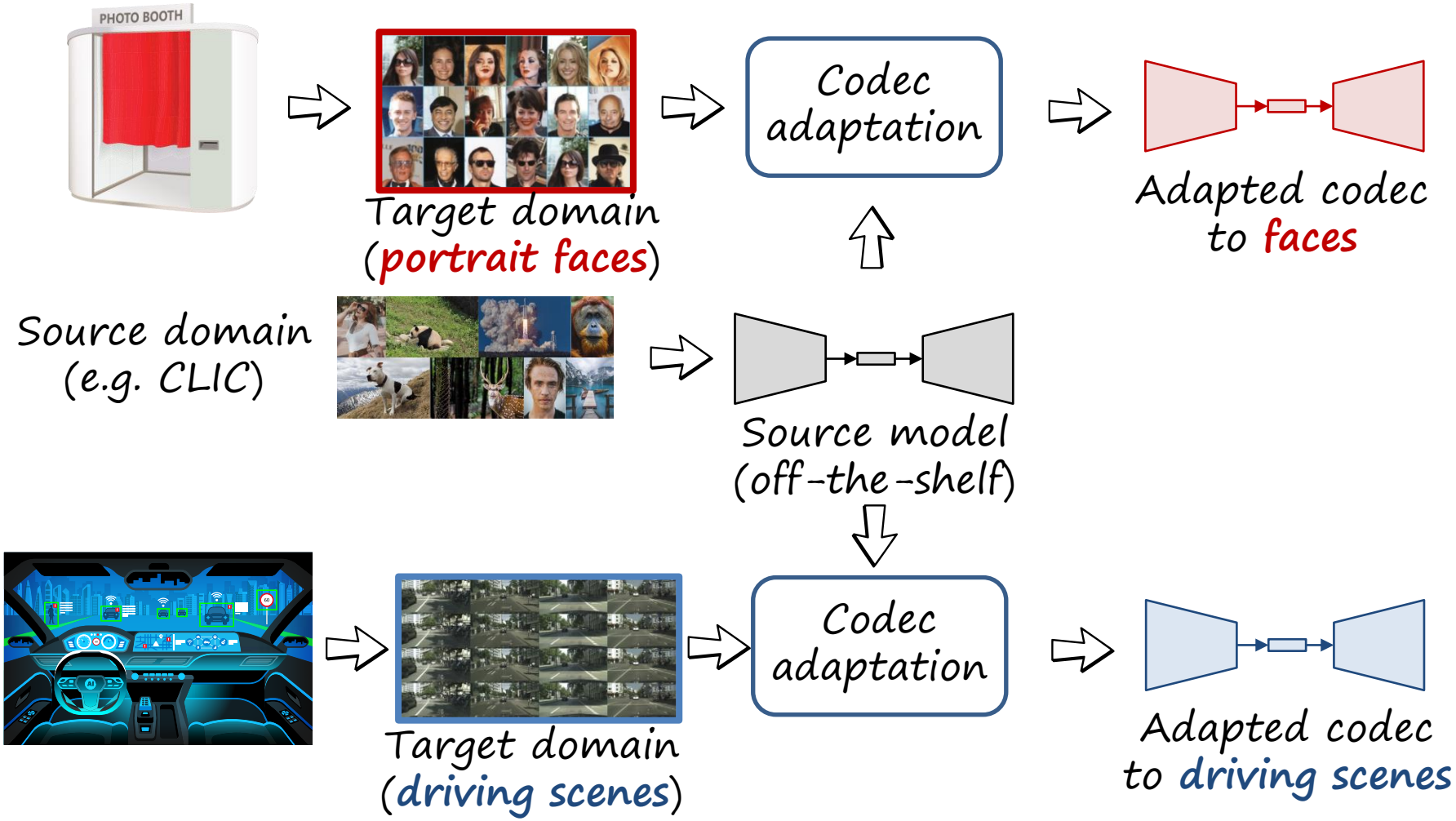
Learned codecs are only optimal in the domain of the training data



DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021

Domain Adaptation in Neural Image ComprEssion (DANICE)

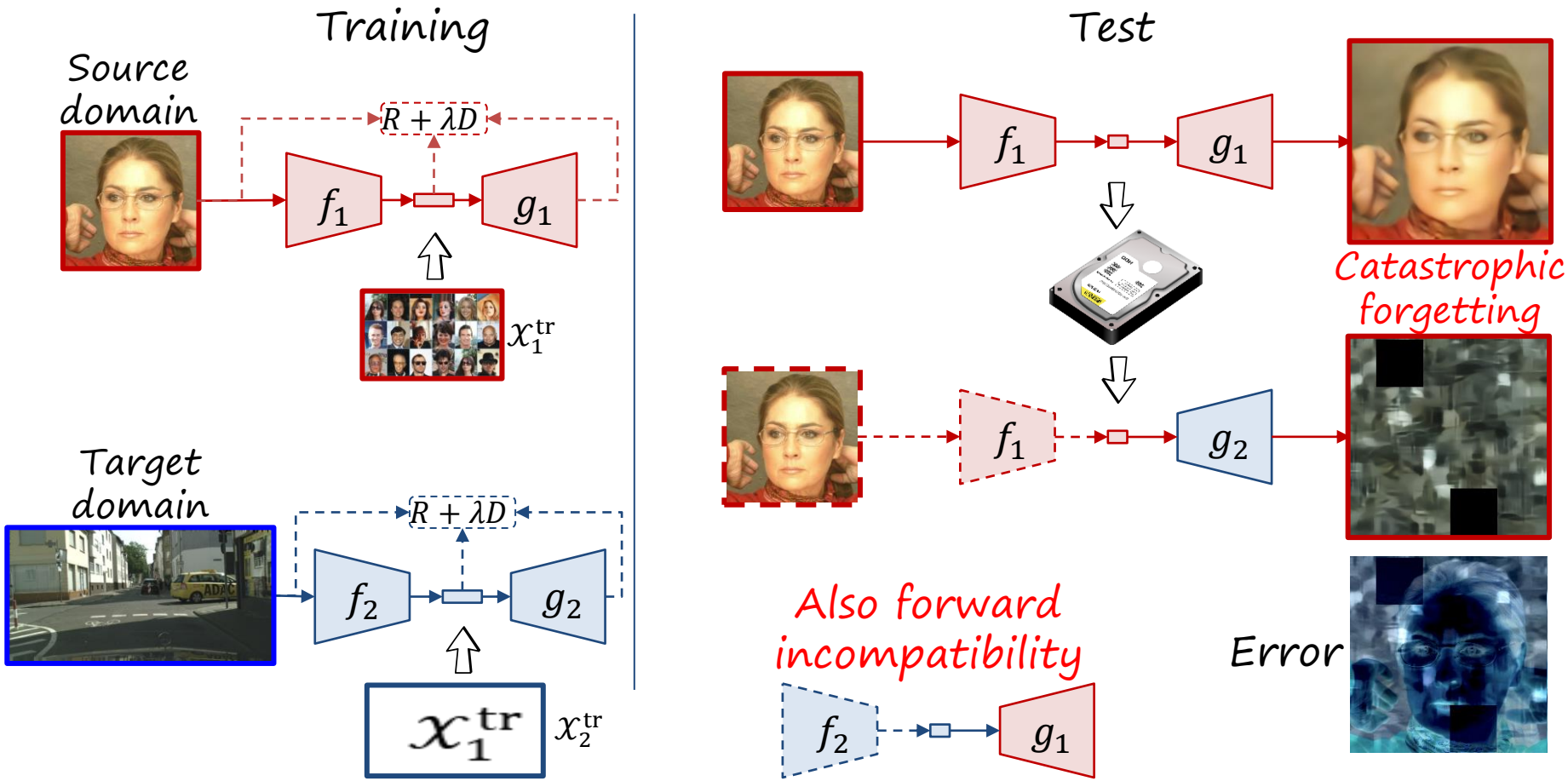
Learned codecs can be **customized with user content** to specific domains
Problem: usually **not enough custom data**; training is **expensive**
Solution: **transfer pre-trained codecs**



DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021

Backward incompatibility with legacy bitstreams: catastrophic forgetting

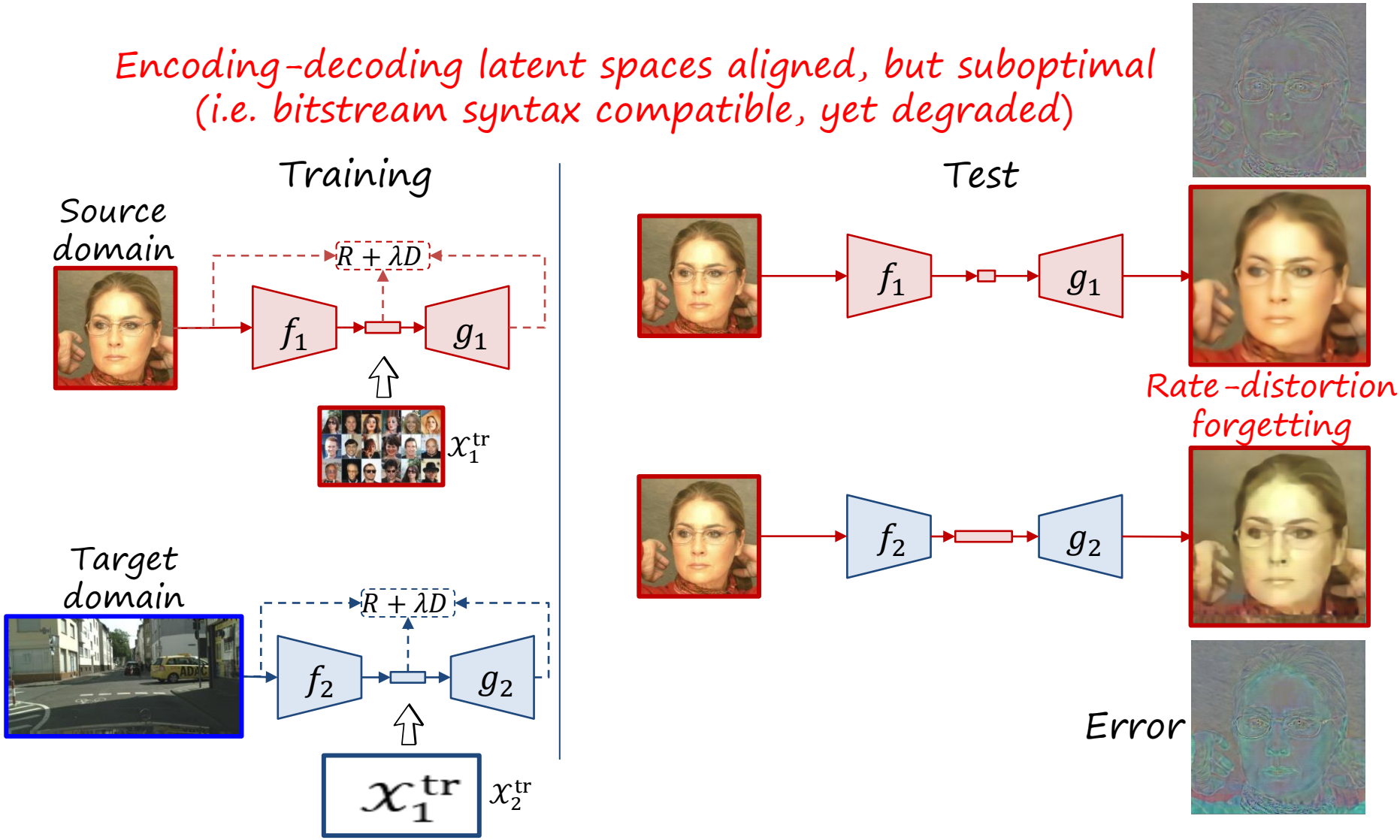
Misalignment between encoding-decoding latent spaces (i.e. bitstream syntax incompatible)



DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021

Rate-distortion forgetting

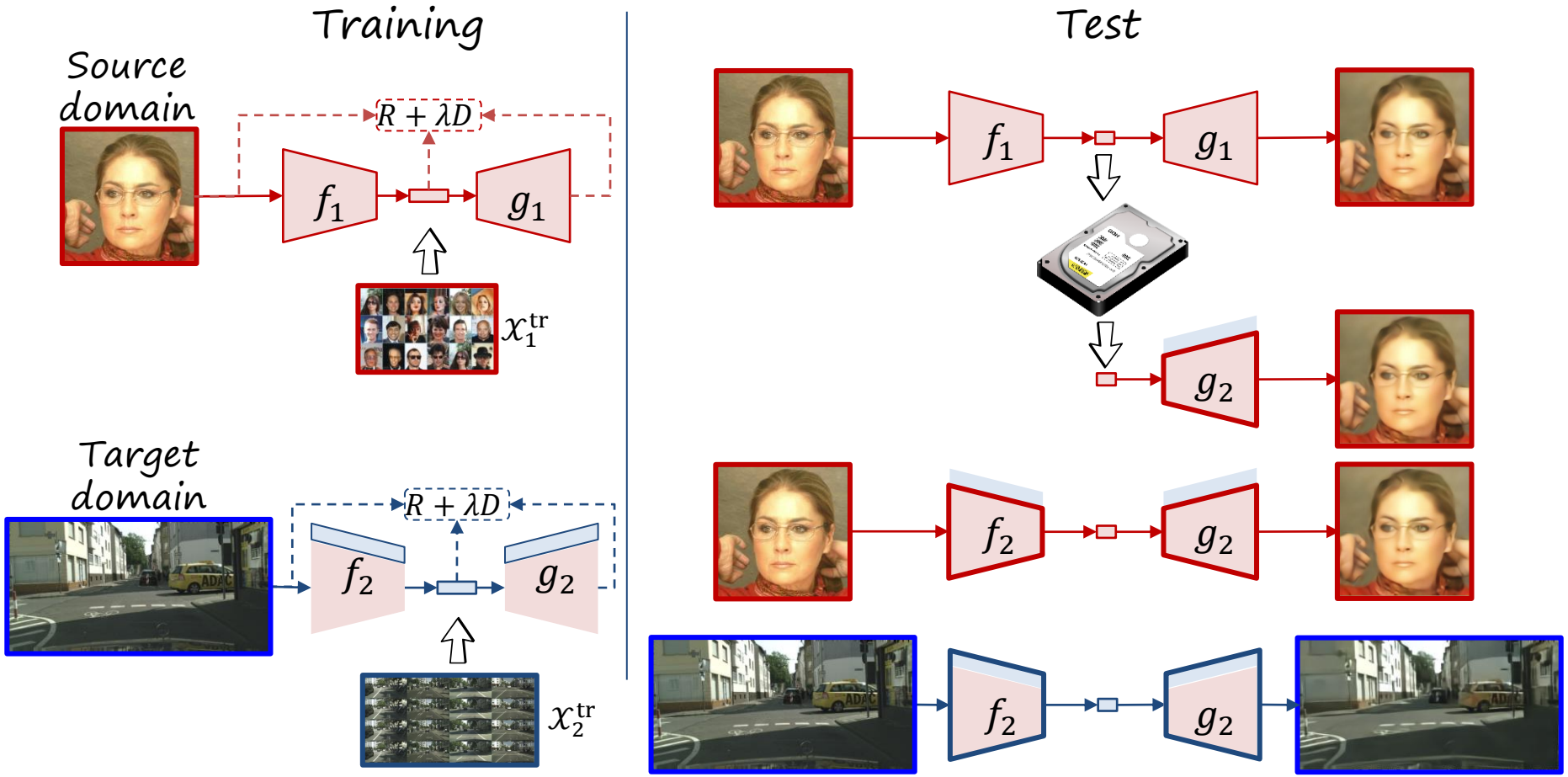
Encoding-decoding latent spaces aligned, but suboptimal
(i.e. bitstream syntax compatible, yet degraded)



DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021

Codec adaptation without forgetting (CAwF)

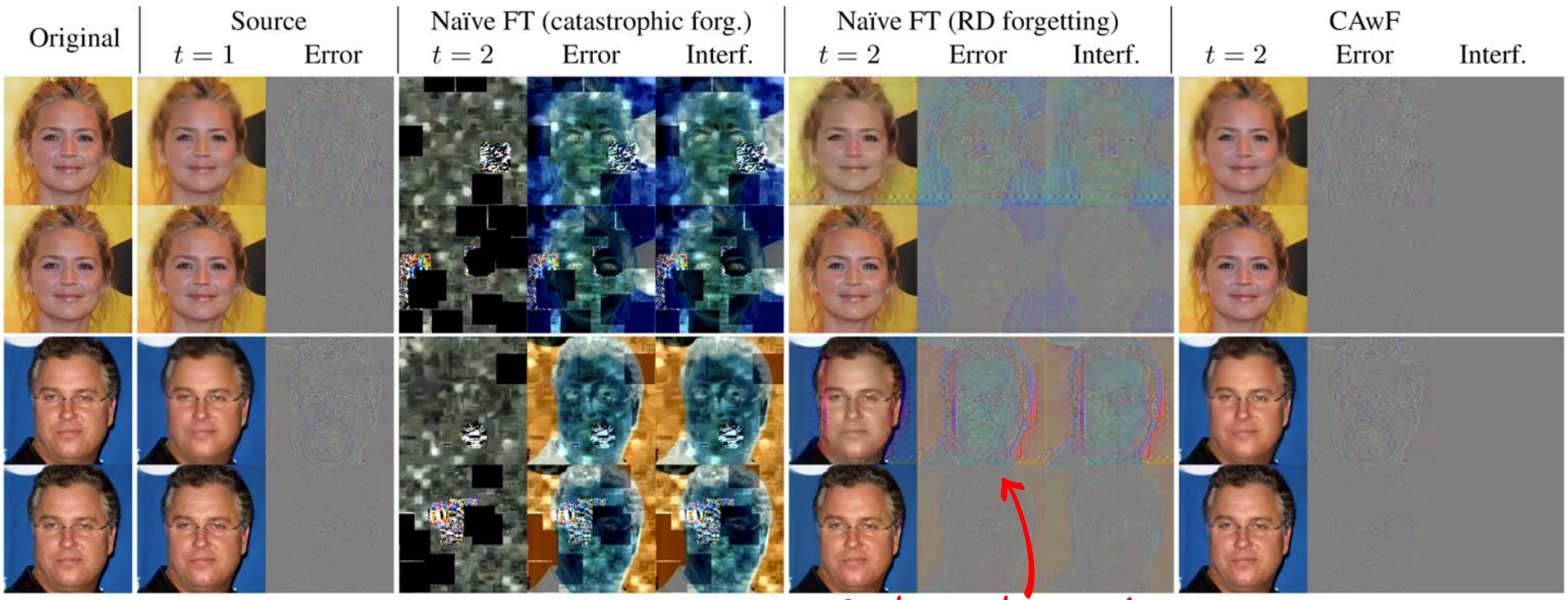
Freeze source codec, and learn target codec as an enhancement layer
Drawback: adds additional parameters



[DANICE: Domain adaptation without forgetting in neural image compression](), CLIC 2021 at CVPR 2021

Codec adaptation without forgetting (CAwF)

CelebA→Cityscapes
(source domain)



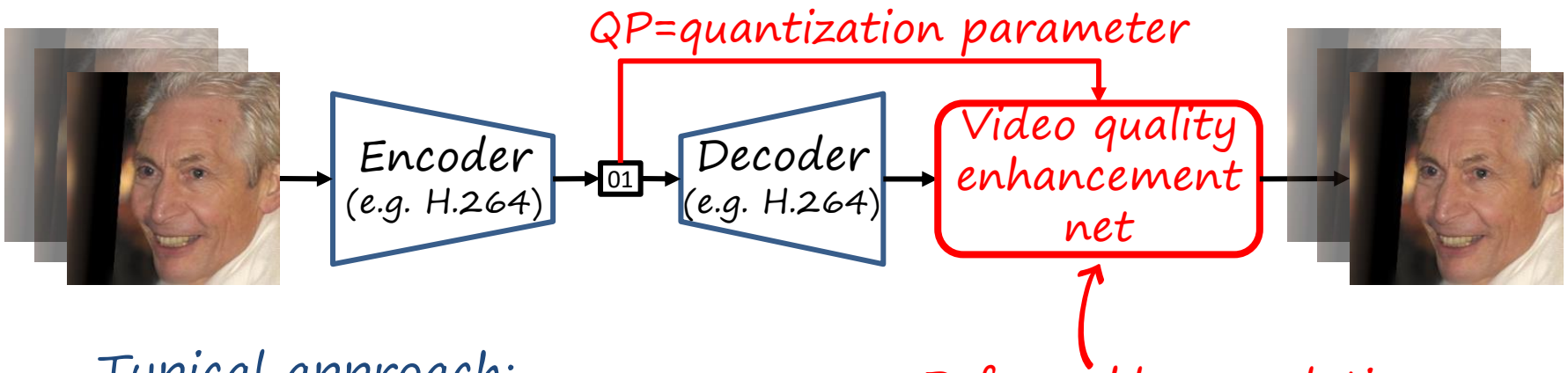
Codec adaptation
artifacts

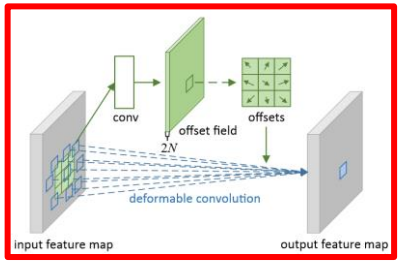
[DANICE: Domain adaptation without forgetting in neural image compression](), CLIC 2021 at CVPR 2021

Outline

- Introduction: image coding
- Compression with neural networks
- Towards practical image compression
- Briefly: other works

Video quality enhancement and artifact removal



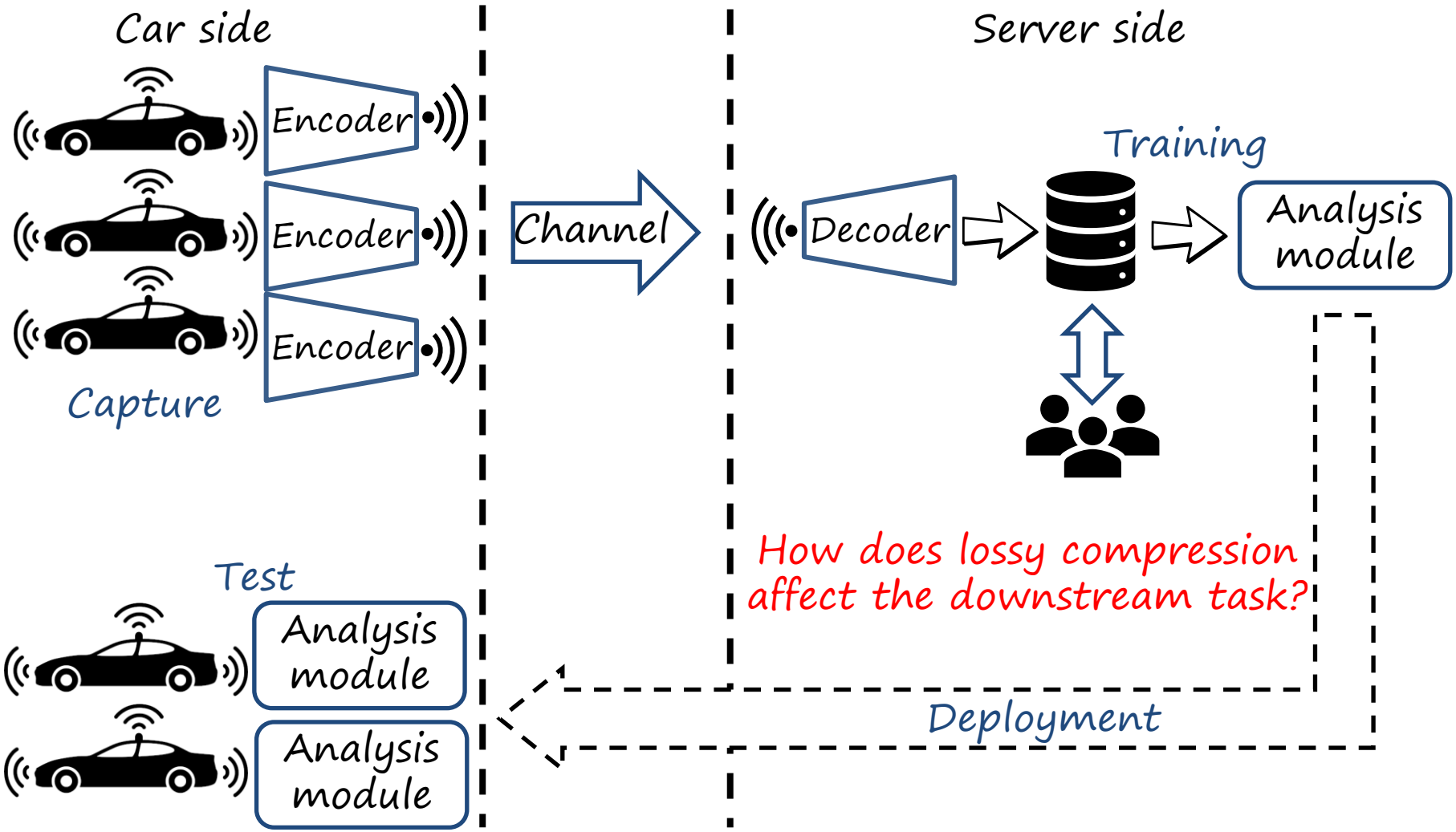


Typical approach:
- Align several frames
- Aggregate the aligned information to alleviate noise/artifacts

Our specific **contribution**:
- Use **deformable convolutions** for multiframe alignment
- **QP-conditional** quality enhancement network

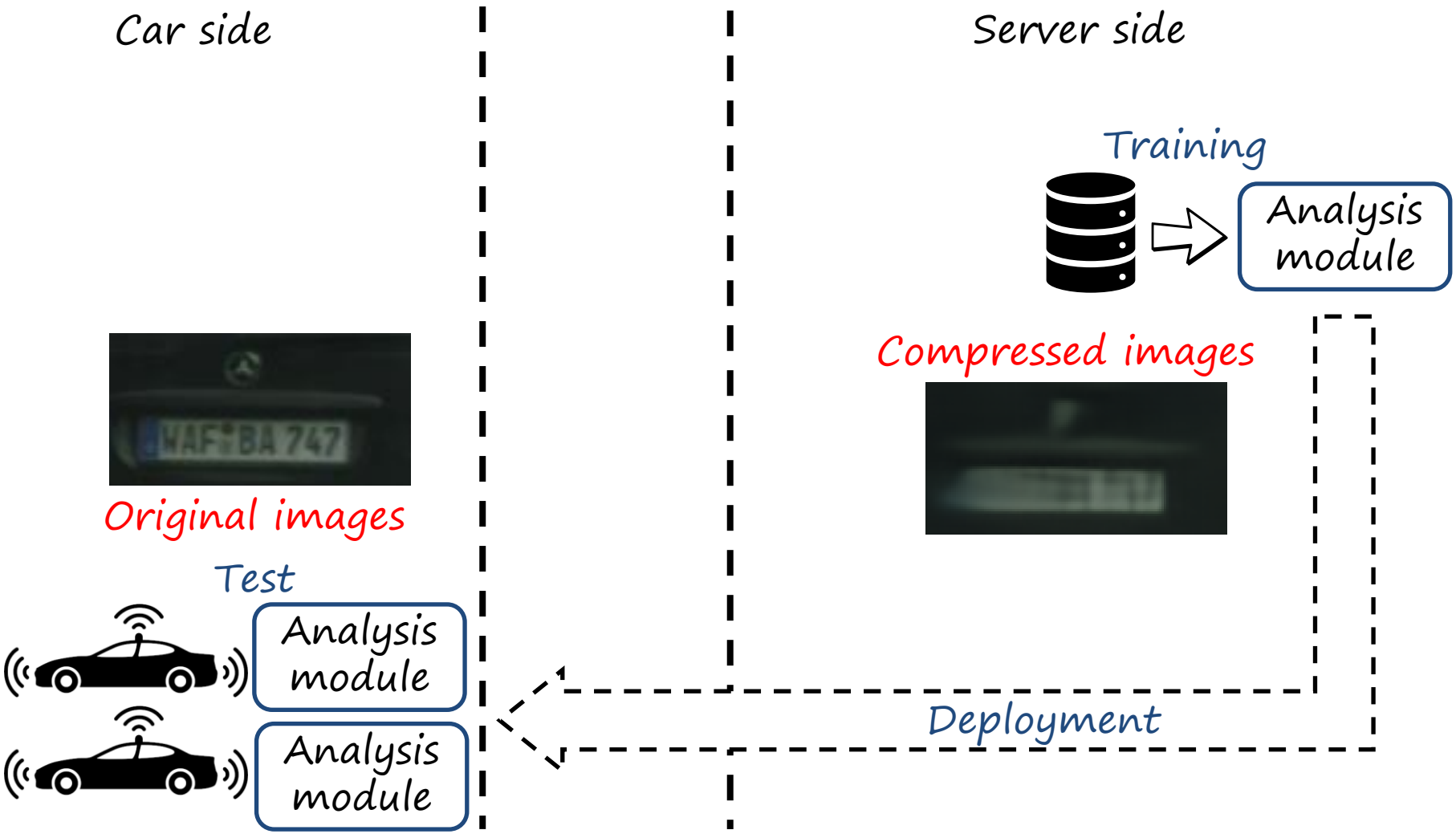
Dai et al., Deformable Convolutional Networks, ICCV 2017
DCNGAN: A deformable convolution-based GAN with QP adaptation for perceptual quality enhancement of compressed video, ICASSP 2022

Distributed data collection for autonomous driving



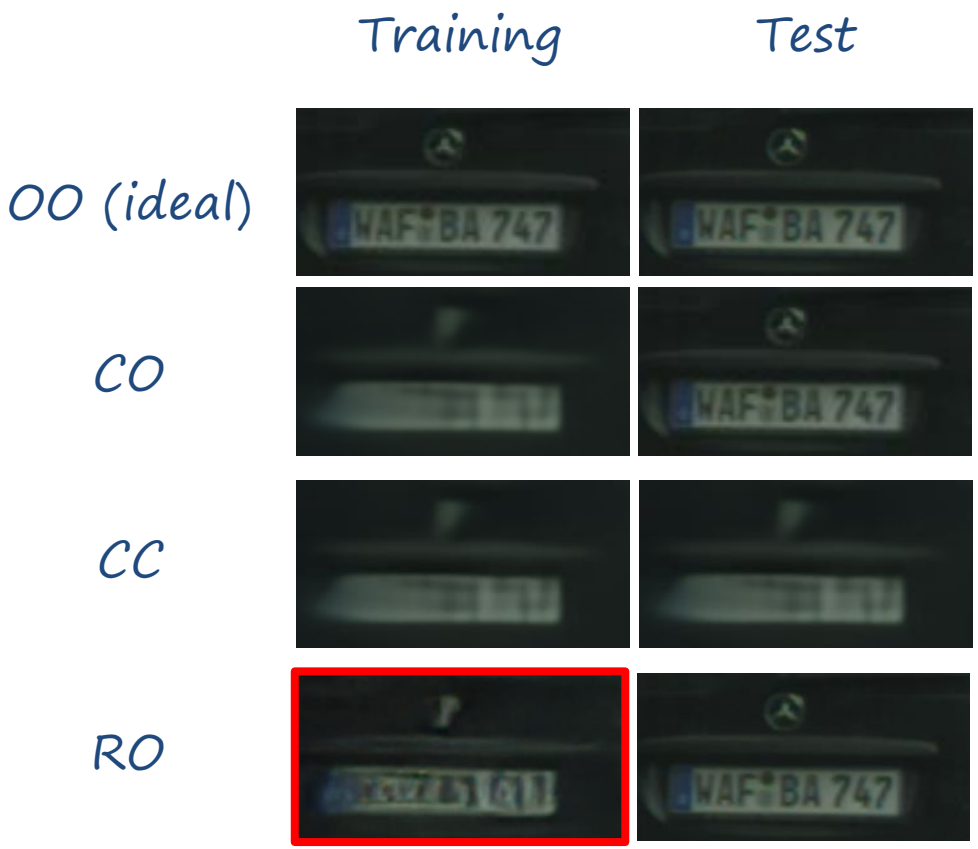
Distributed Learning and Inference with Compressed Images, IEEE Trans. Image Processing 2021

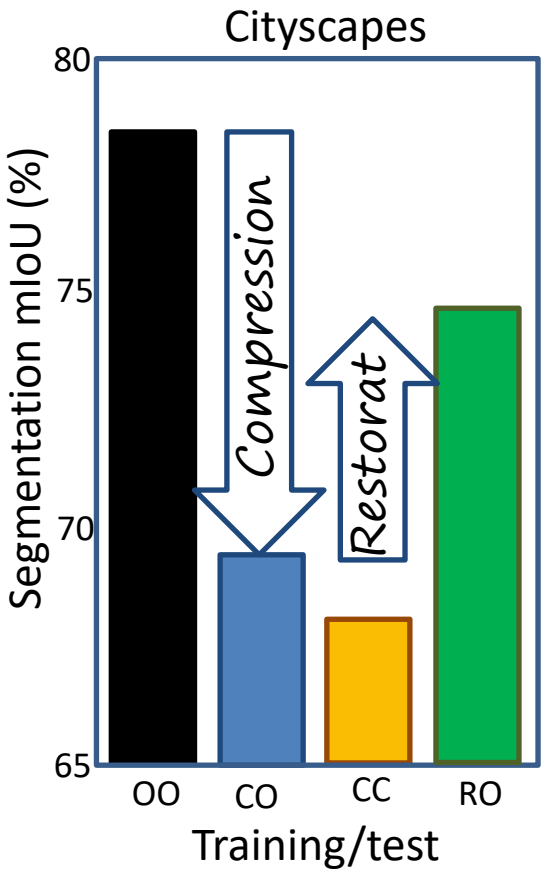
Distributed data collection



Distributed Learning and Inference with Compressed Images, IEEE Trans. Image Processing 2021

Effect on downstream task





Proposed approach: dataset restoration

Distributed Learning and Inference with Compressed Images, IEEE Trans. Image Processing 2021

Training images vs test images

Distributed Learning and Inference with Compressed Images, IEEE Trans. Image Processing 2021

References

General references

- Balle et al. End-to-end Optimized Image Compression, ICLR 2017
- Theis et al., Lossy Image Compression with Compressive Autoencoders, ICLR 2017
- Dai et al., Deformable Convolutional Networks, ICCV 2017
- Sun et al., Spatiotemporal Entropy Model is All You Need for Learned Video Compression, arxiv 2021

Works by our group and collaborators (with Marta Mrak's group at BBC R&D, London, UK and Shuai Wan's group at Nortwestern Politechnic University/Xidian Uni., Xi'an, China)

- Yang et al., Variable Rate Deep Image Compression with Modulated Autoencoder, Signal Processing Letters 2020
- Yang et al., Slimmable compressive autoencoders for practical imaga compression, CVPR 2021
- Katakol et al., DANICE: Domain adaptation without forgetting in neural image compression, CLIC 2021 at CVPR 2021
- Zhang et al., DCNGAN: A deformable convolution-based GAN with QP adaptation for perceptual quality enhancement of compressed video, ICASSP 2022
- Katakol et al., Distributed Learning and Inference with Compressed Images, IEEE Trans. Image Processing 2021
- Liu et al., Slimmable Video Codec, 2022 (under review)

THANK YOU!

lherranz@cvc.uab.es
www.lherranz.org

Universitat Autònoma de Barcelona
MINISTERIO DE CIENCIA E INNOVACIÓN
UNIÓN EUROPEA
FONDO SOCIAL EUROPEO
El FSE invierte en tu futuro
AGENCIA ESTATAL DE INVESTIGACIÓN