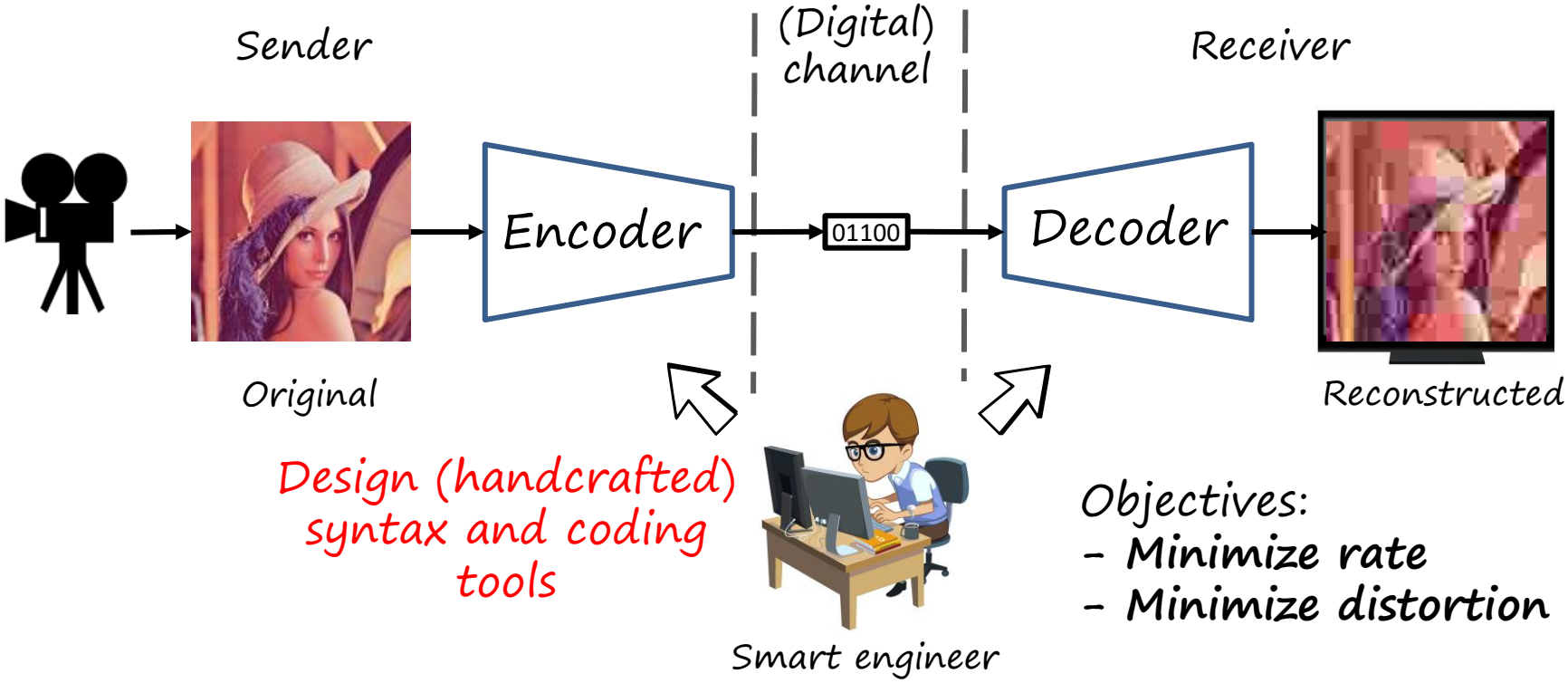
Slimmable Compressive Autoencoders for Practical Neural Image Compression

Fei Yang[1,2,3], Luis Herranz[1,2], Yongmei Cheng[3], Mikhail Mozerov[1,2]

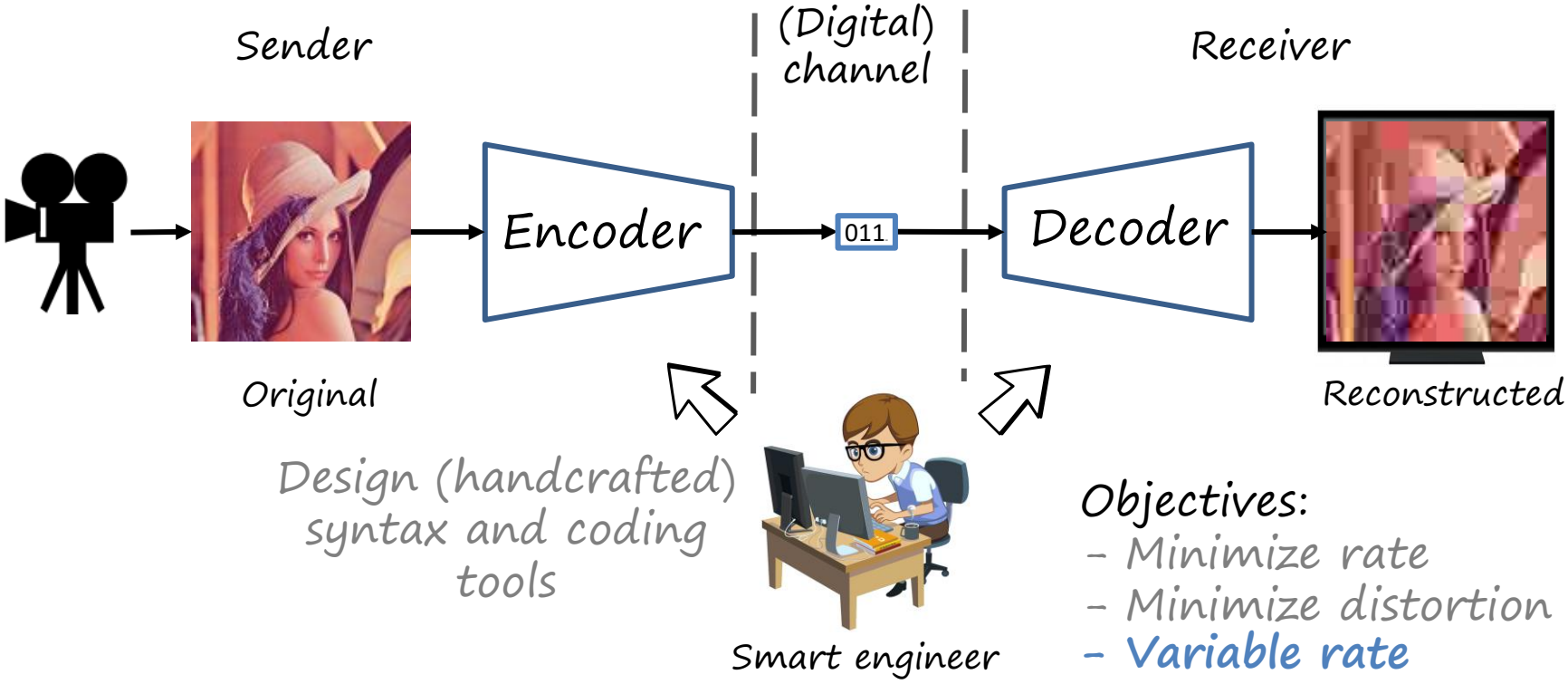
[1]Computer Vision Center
[2]Universitat Autònoma de Barcelona
[3]Northwestern Politechnical University

CVPR 2021

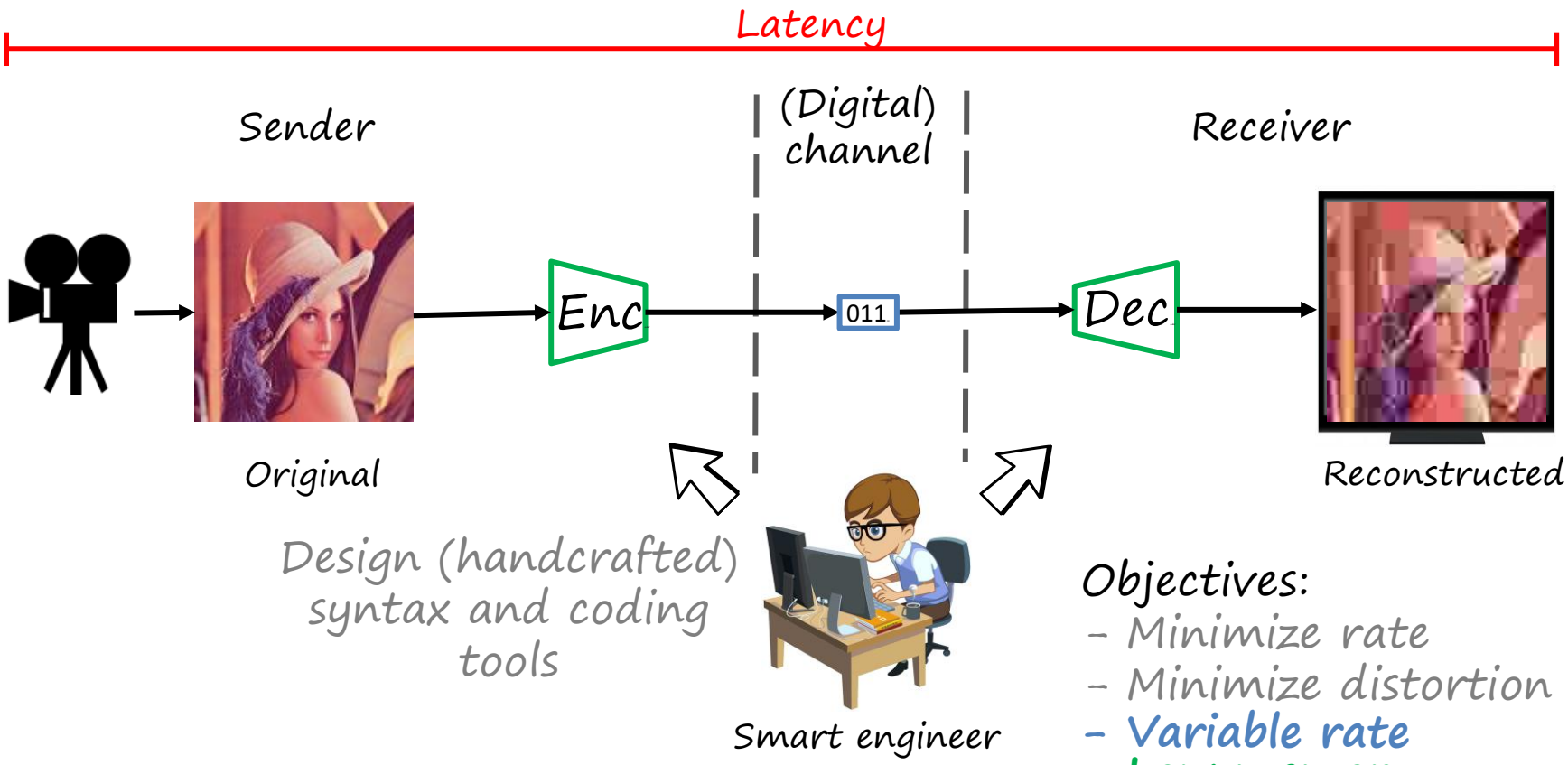
Developing traditional image/video codecs



… for practical applications



… for practical applications

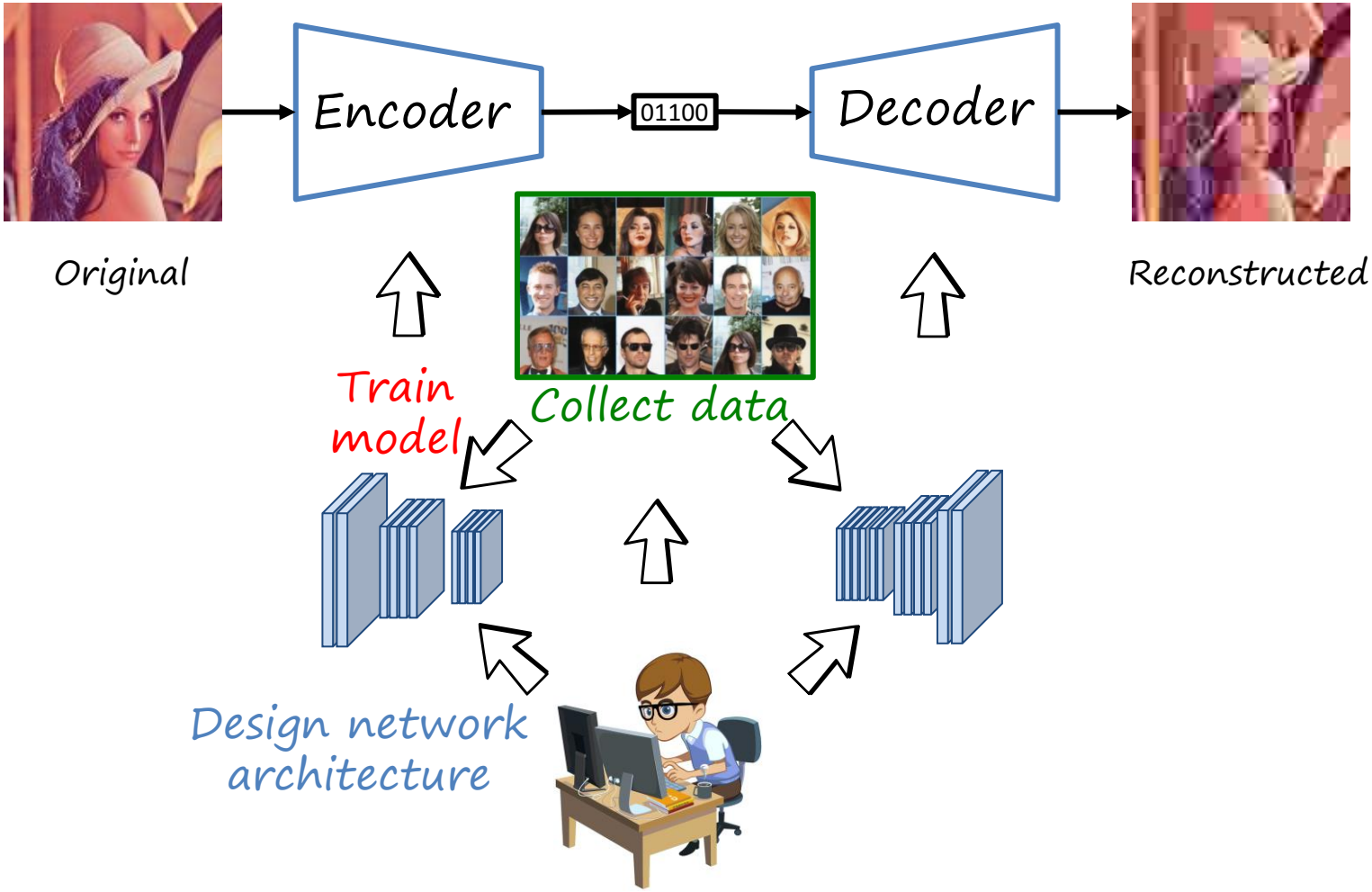


Objectives:

- Minimize rate
- Minimize distortion
- Variable rate
- Low memory
- Low computation
- Low latency

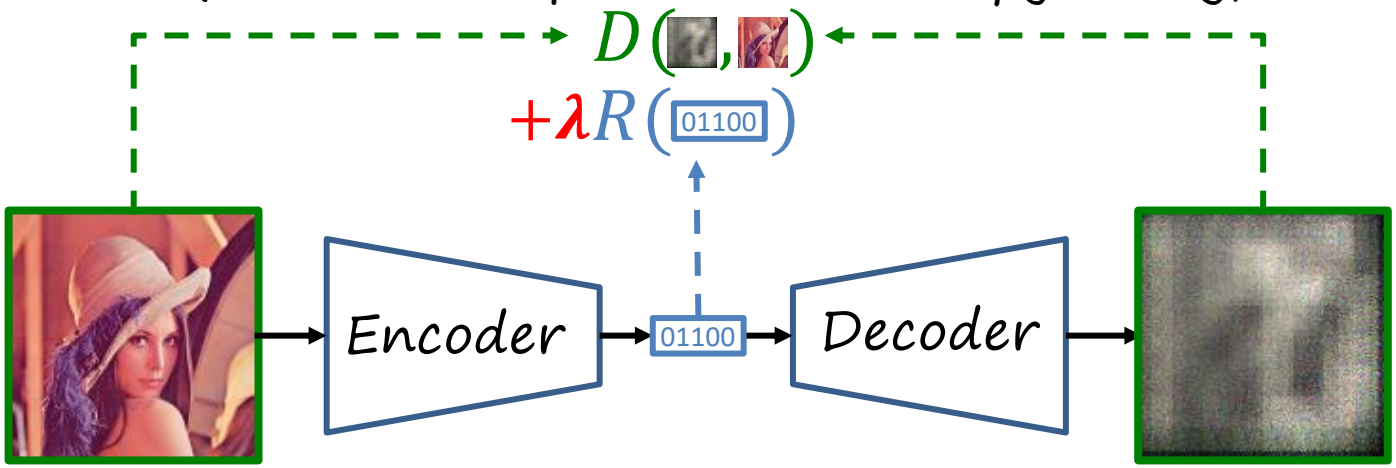
Neural image/video codecs

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



Neural image compression

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



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

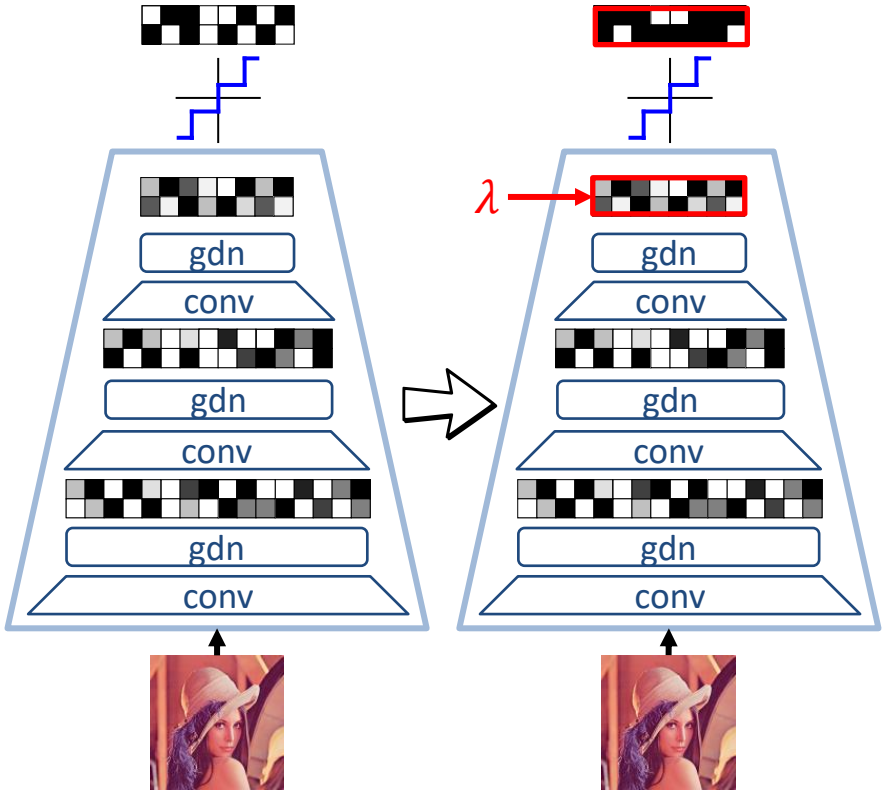
Limitations
- λ is fixed
- Heavy encoders/decoders

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

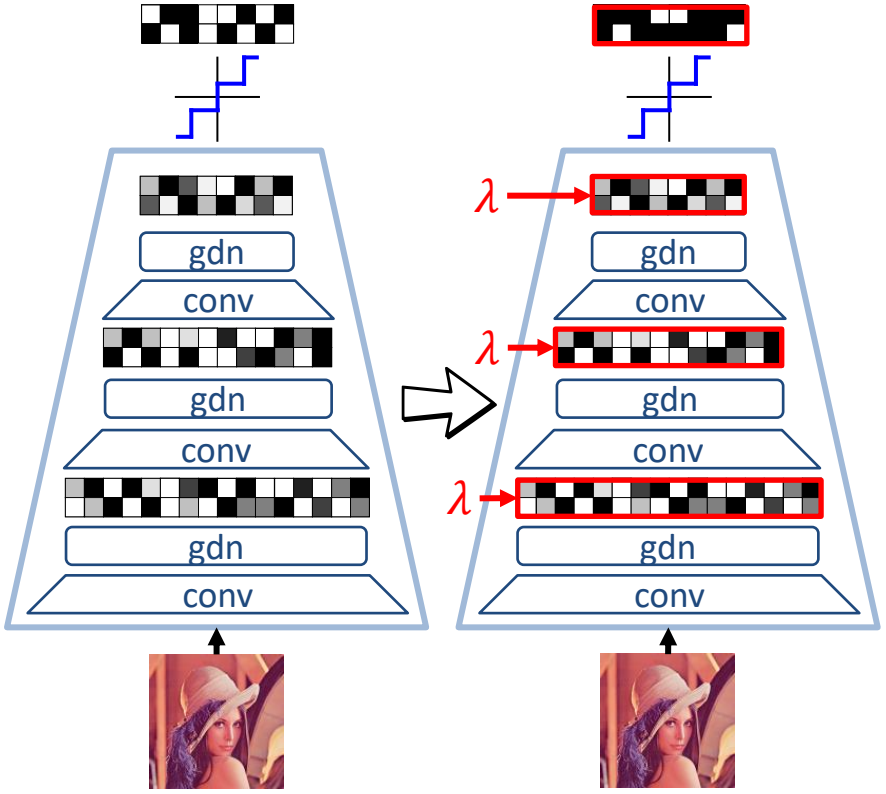
Variable rate neural image compression

Objective: one single model for multiple λ

Bottleneck scaling [Theis2017]



Feature modulation [MAE, cAE]

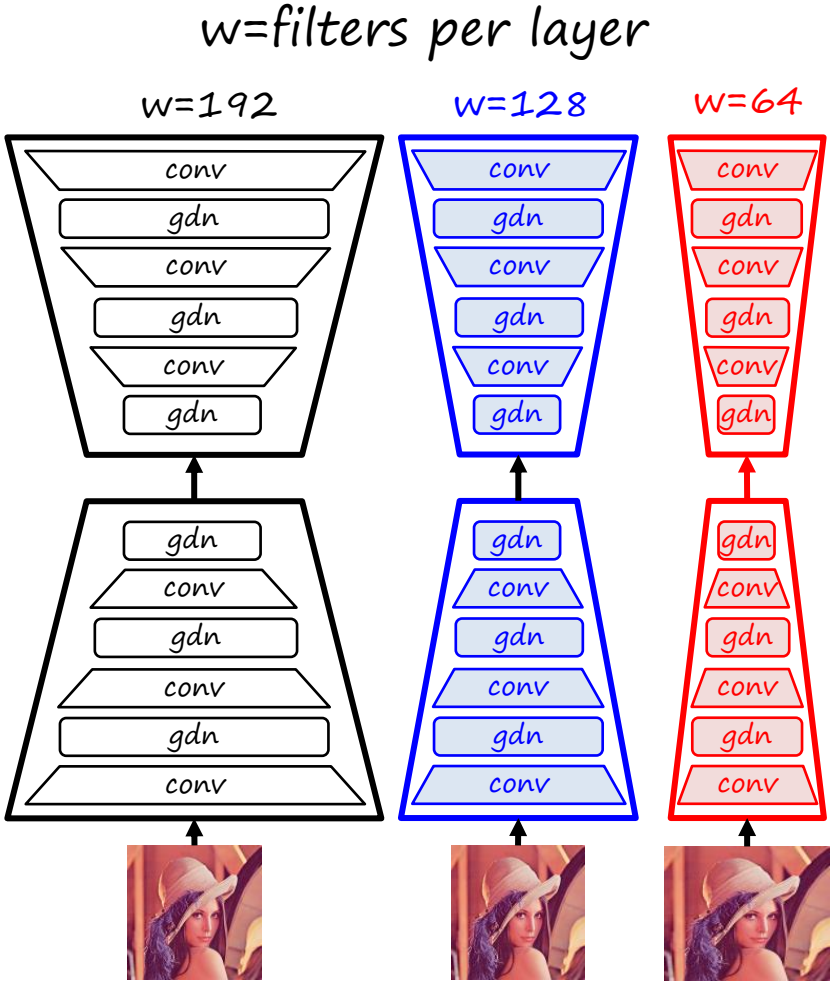


- *Minimize rate* ✓
- *Minimize distortion* ✓
- *Variable rate* ✓

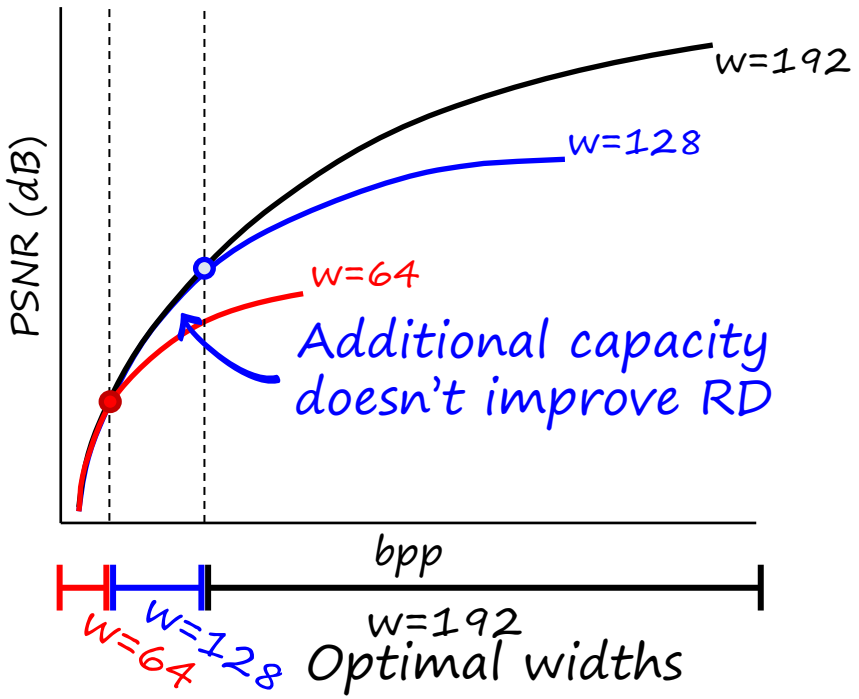
- *Low memory* ✗
- *Low computation* ✗
- *Low latency* ✗

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

Model capacity and rate-distortion



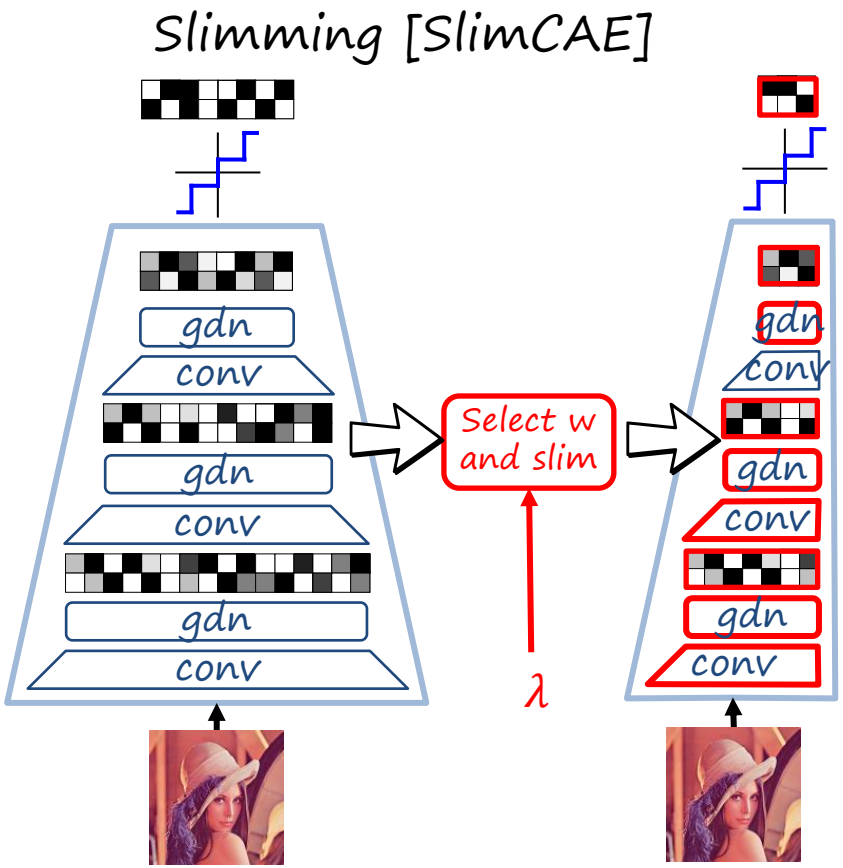
There is a **minimal capacity** for every RD tradeoff



Lower w results in less memory and computation!!

Slimmable compressive autoencoder

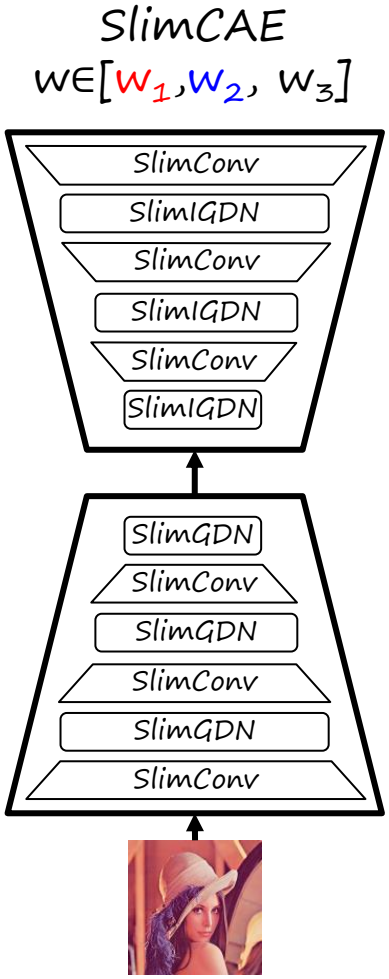
Approach: slim the network to the minimal capacity for a given λ



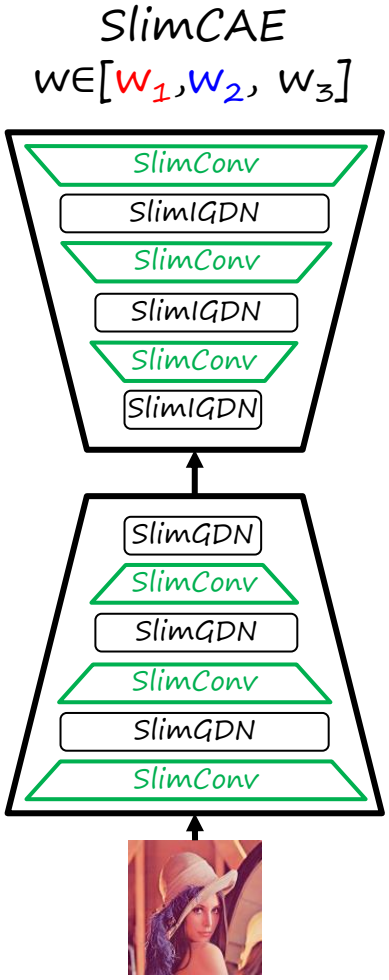
- Minimize rate ✓
- Minimize distortion ✓
- Variable rate ✓
- Lower memory ✓
- Lower computation ✓
- Lower latency ✓

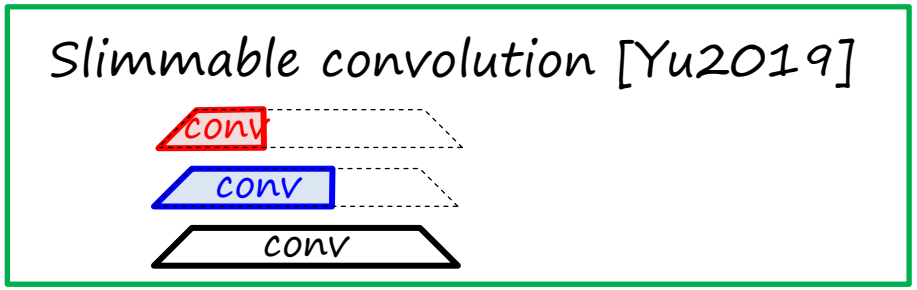
(for low-mid rates)

Slimmable layers in SlimCAE

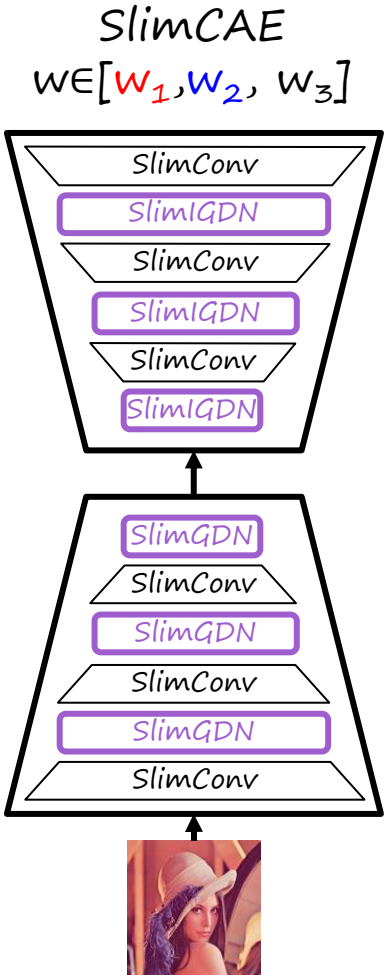


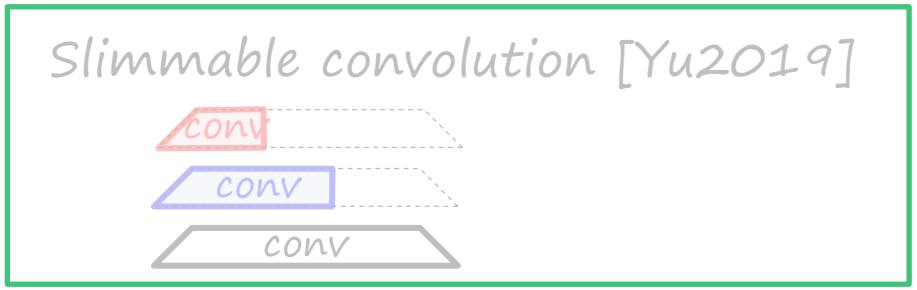
Slimmable layers in SlimCAE

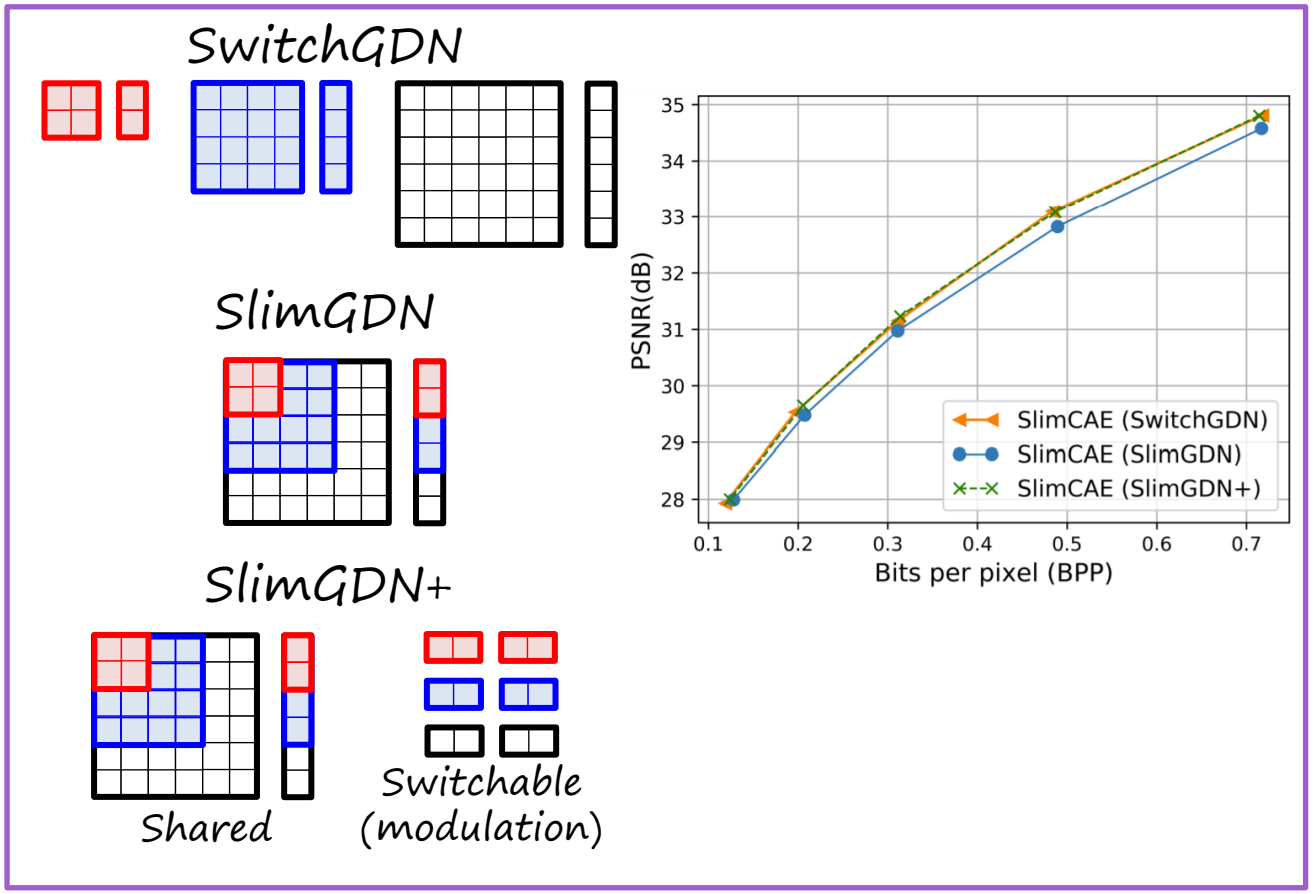




Slimmable layers in SlimCAE



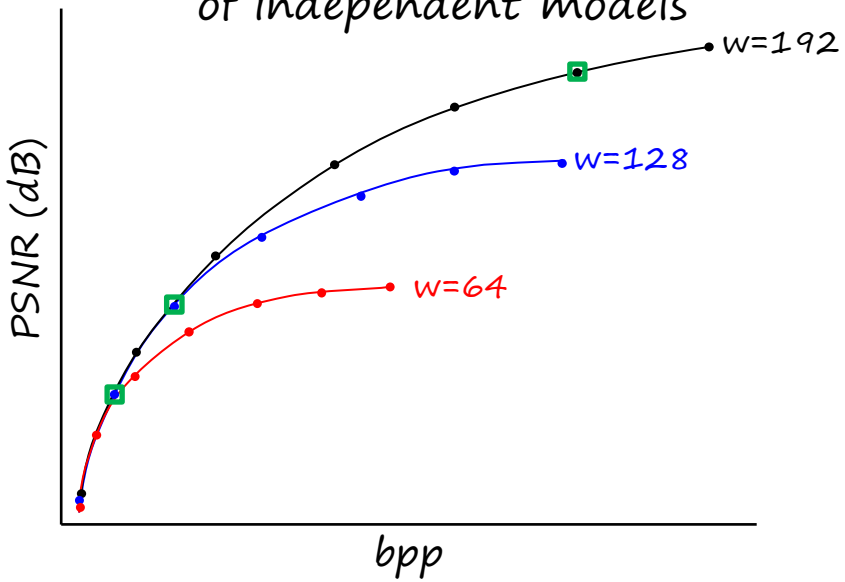




Training SlimCAE

Problem: we need the optimal λs to train the SlimCAE

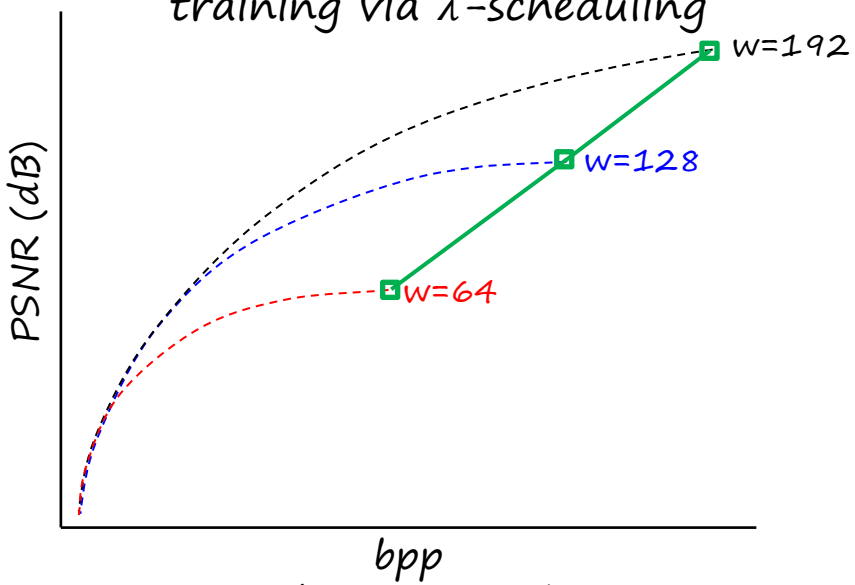
Estimate from RD curves of independent models



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

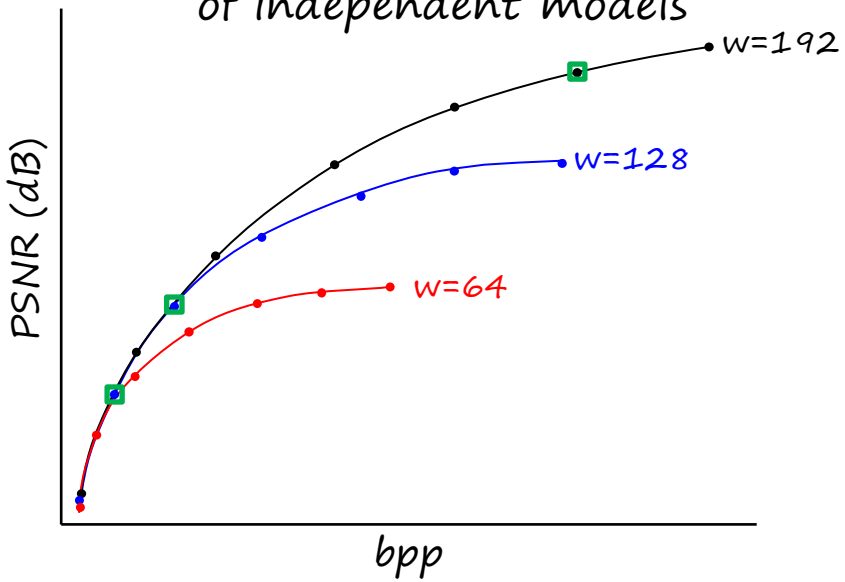


1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

Training SlimCAE

Problem: we need the optimal λs to train the SlimCAE

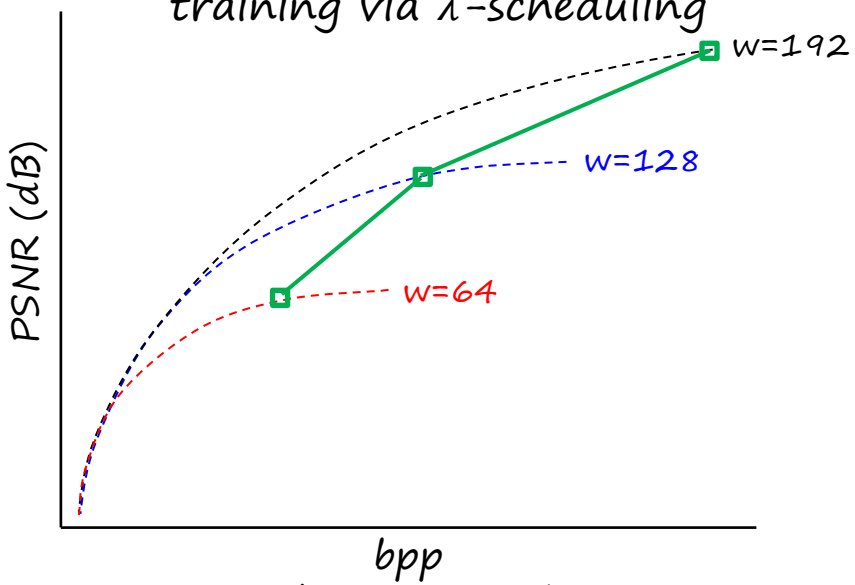
Estimate from RD curves of independent models



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

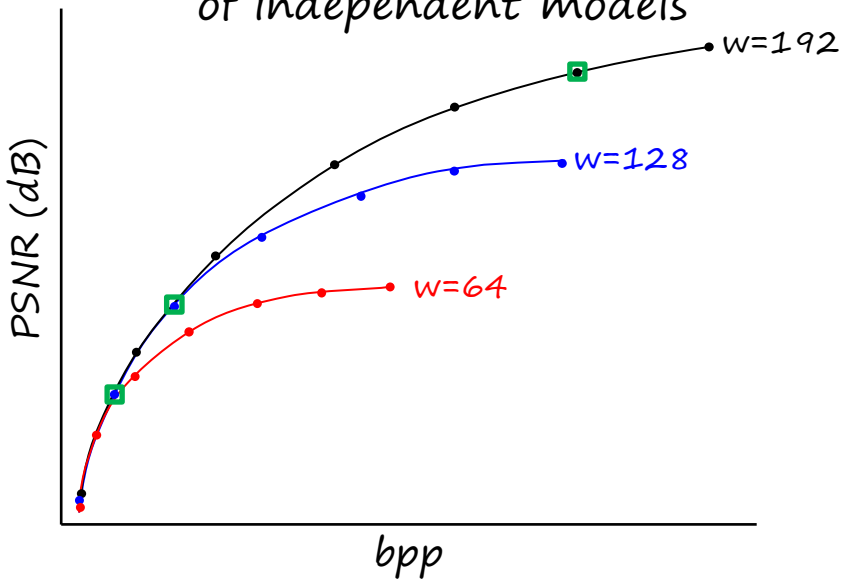


1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

Training SlimCAE

Problem: we need the optimal λs to train the SlimCAE

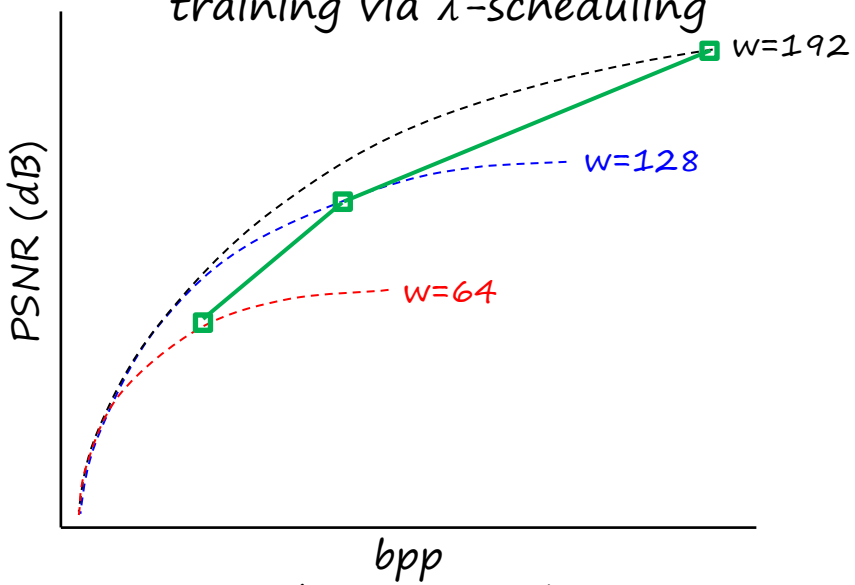
Estimate from RD curves of independent models



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

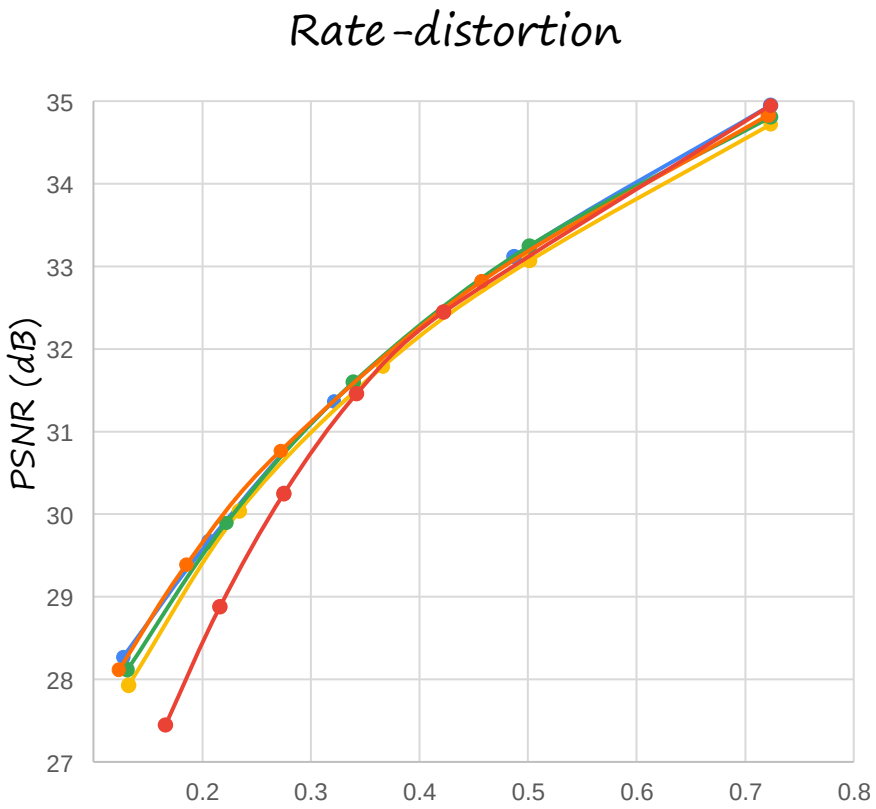


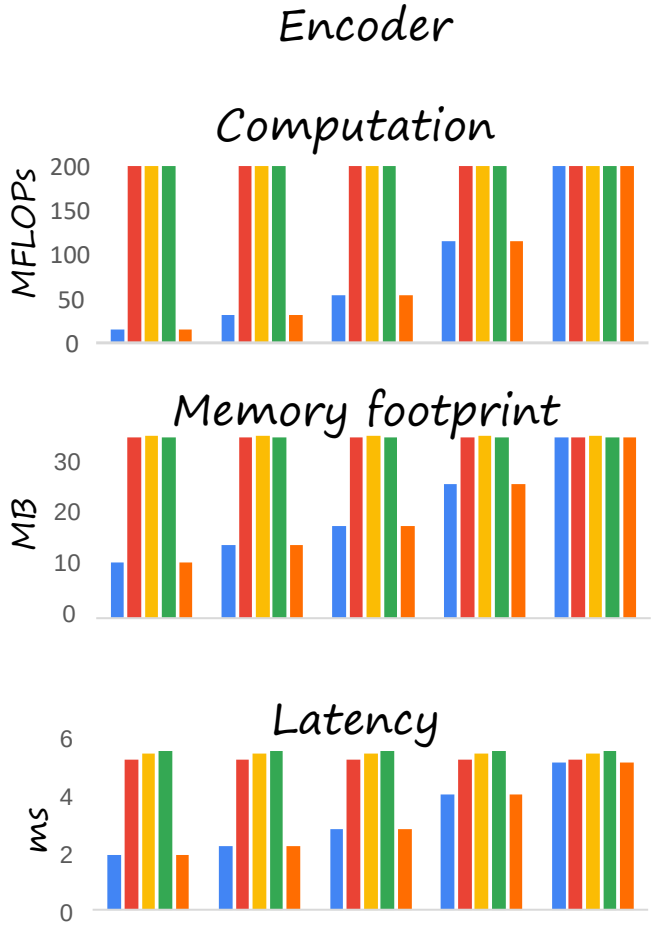
1. Train a SlimCAE with $\lambda_1 = \lambda_2 = \lambda_3$
2. While not converged do
   - Update λs according to schedule
   - Optimize CAE

Directly train one model!

Performance comparison

Independent CAEs Scaling [Theis2017] MAE [Yang2020] cAE [Choi2019] SlimCAE (ours)





Thanks!

https://arxiv.org/abs/2103.15726

https://github.com/FireFYF/SlimCAE

Fei Yang

Luis Herranz

Yongmei Cheng

Mikhail Mozerov

Universitat Autònoma de Barcelona