ETH zürich

Good, Cheap, and Fast: Overfitted Image Compression with Wasserstein Distortion

Jonas Mirlach

15th April 2025

Image compression is already “solved”, isn‘t it?

Why deep learning based image compression?

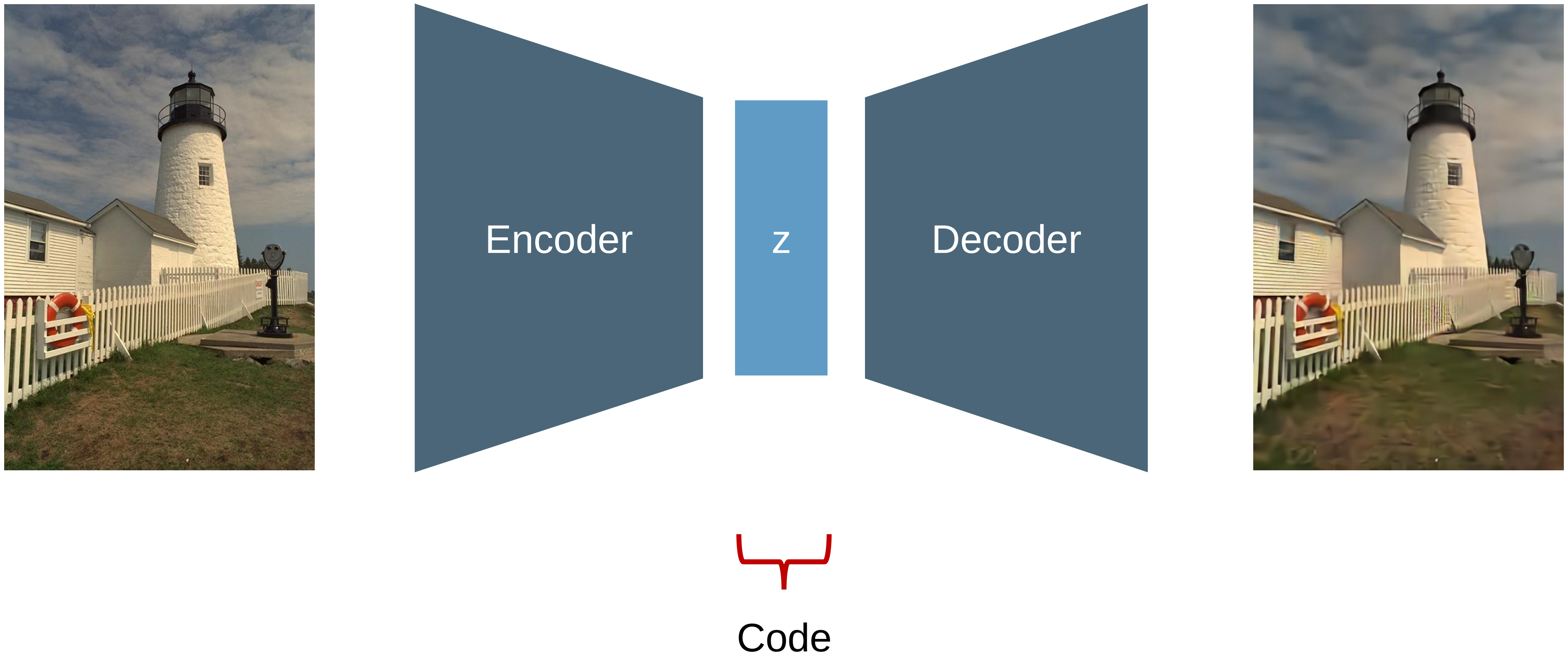
BPP = 0.1

Classical
(HEVC)

JPEG

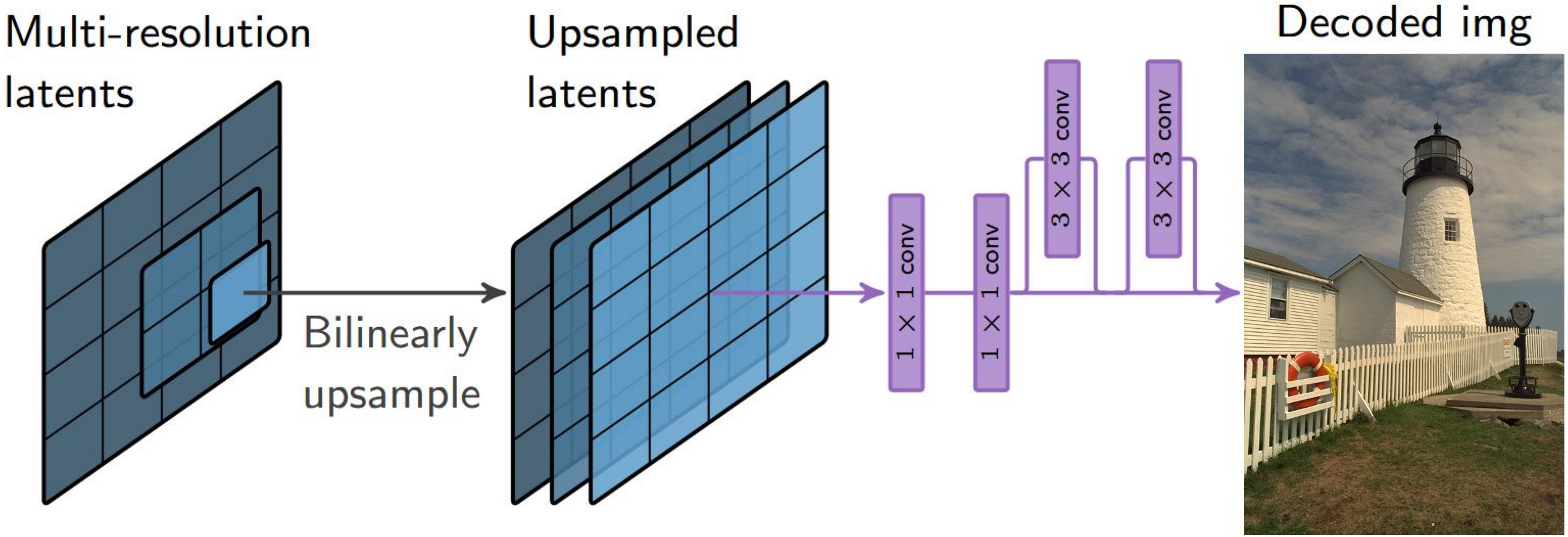
Deep learning
(CompressAI)

How deep learning based image compression?



COOL-CHIC 😎 | Decoding

[1]



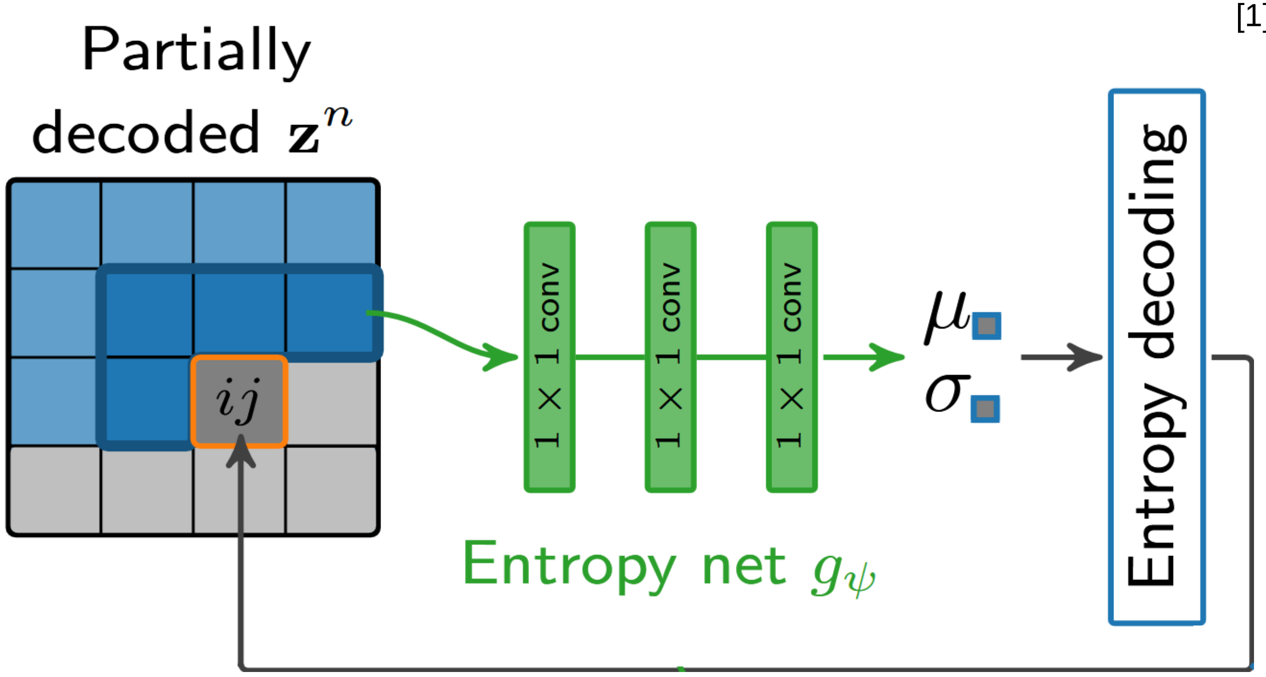
COOL-CHIC 😎 | Entropy coding

Coding rate ≈

$$\underbrace{\mathbb{E}_{z\sim q}\left[-\log_2 p_\psi(z)\right]}_{\text{Cross entropy}} = \underbrace{-\mathbb{E}_{z\sim q}[\log_2 q(z)]}_{\text{Entropy } H(q)} + \underbrace{\mathbb{E}_{z\sim q}\left[\log_2 \frac{q(z)}{p_\psi(z)}\right]}_{\text{KL-Divergence } D_{KL}(q \parallel p_\psi)}$$

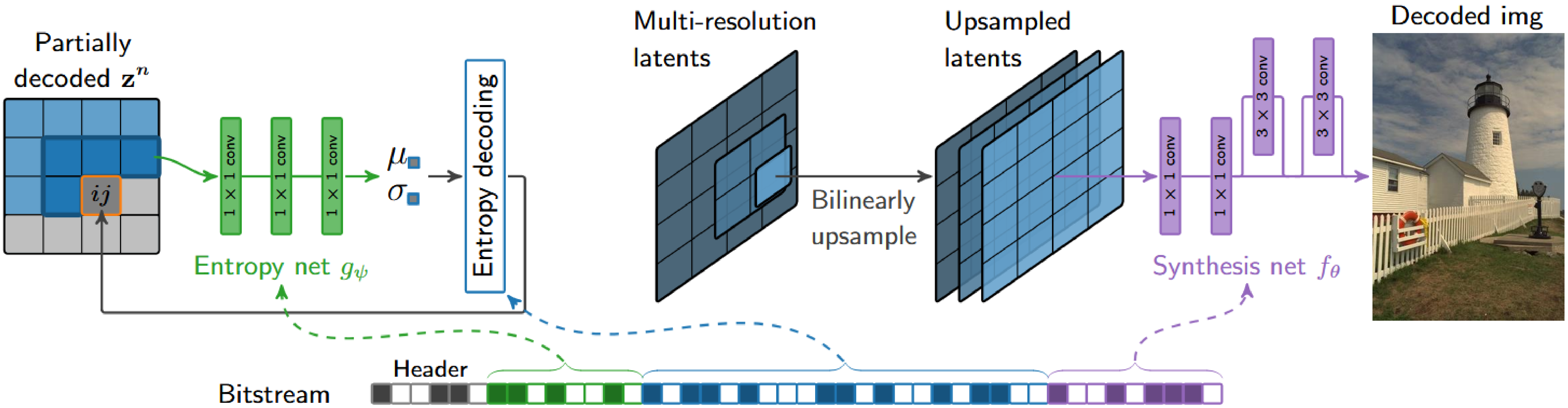
COOL-CHIC 😎 | Entropy coding

[1]



COOL-CHIC 😎

[1]



Partially decoded $\mathbf{z}^n$
ij
1 × 1 conv
1 × 1 conv
1 × 1 conv
Entropy net g_ψ
μ
σ
Entropy decoding
Multi-resolution latents
Bilinearly upsample
Upsampled latents
1 × 1 conv
1 × 1 conv
3 × 3 conv
3 × 3 conv
Synthesis net f_θ
Decoded img
Header
Bitstream

COOL-CHIC 😎 | Objective

$$\min_{\theta,\psi,z} \underbrace{Dist\left(x, f_\theta\big(Upsample(z)\big)\right)}_{\text{Reconstruction quality}} - \lambda \cdot \underbrace{\log_2 p_\psi(z)}_{\text{Estimated latent storage cost}}$$

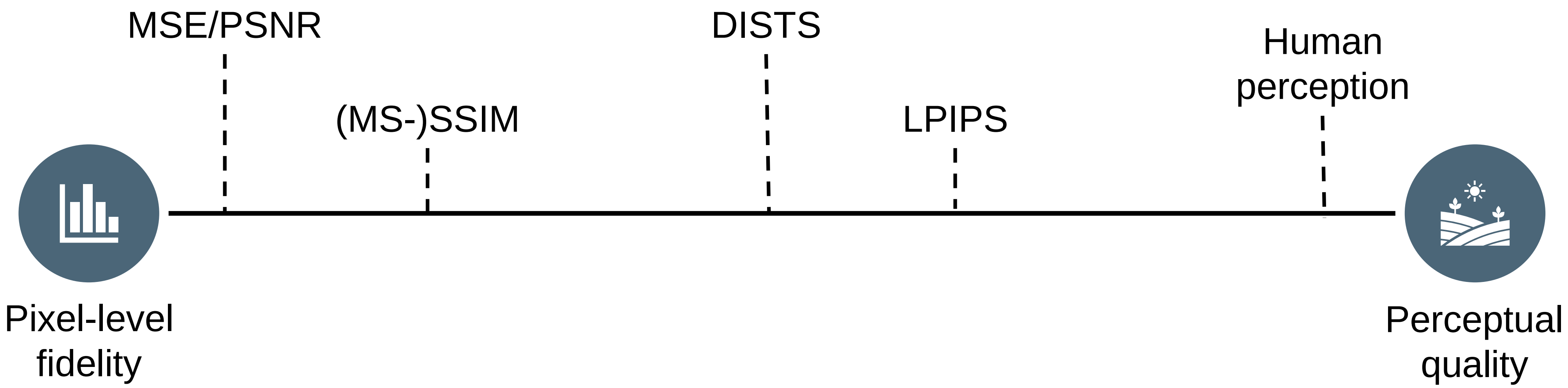
COOL-CHIC 😎 | Quantization awareness

Stage 1 $\nabla_{\theta,\psi,z}\mathcal{L}_{\theta,\psi}(z+u)$ $u \sim Uniform(0,1)$

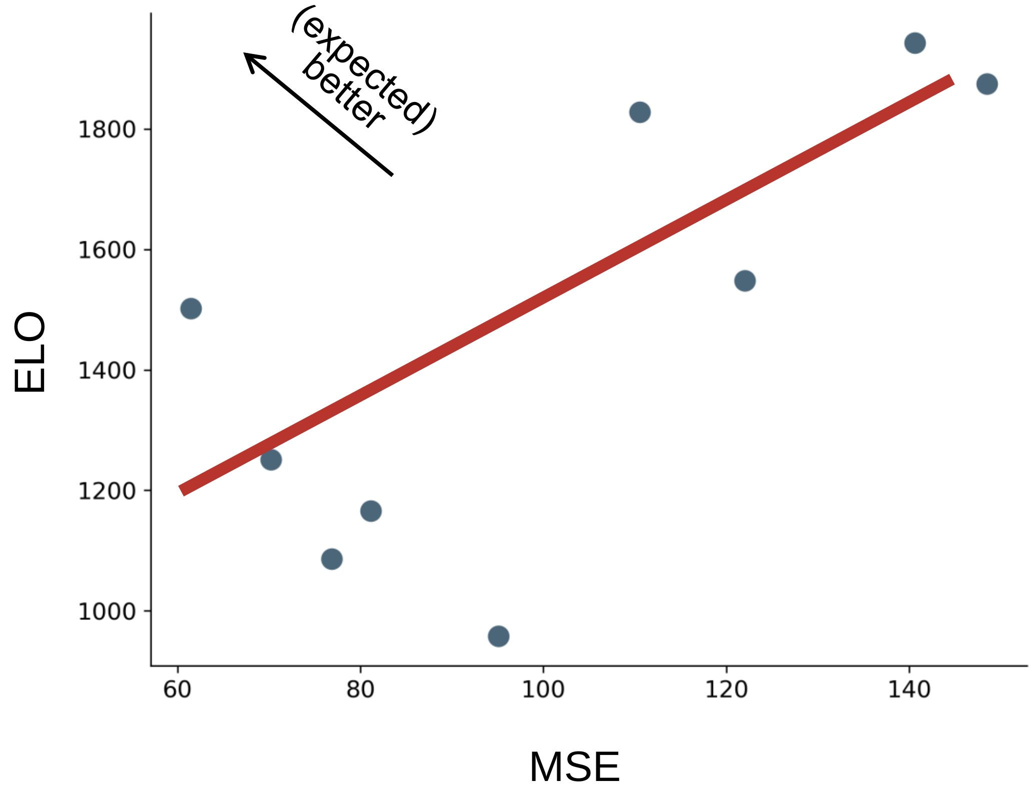
Stage 2 $\nabla_{\theta,\psi}\mathcal{L}_{\theta,\psi}(\lfloor z \rceil)\ and\ \tilde{\nabla}_z\mathcal{L}_{\theta,\psi}(\lfloor z \rceil)$

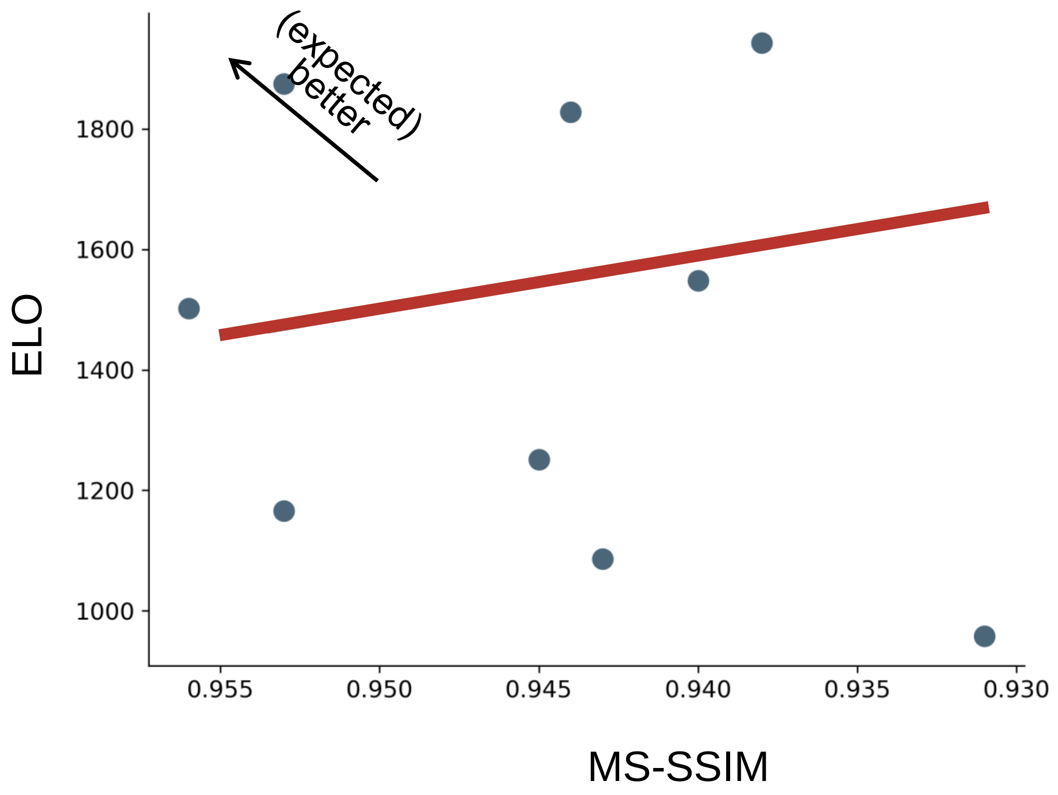
How to measure the quality of reconstructed images?

Distortion metrics

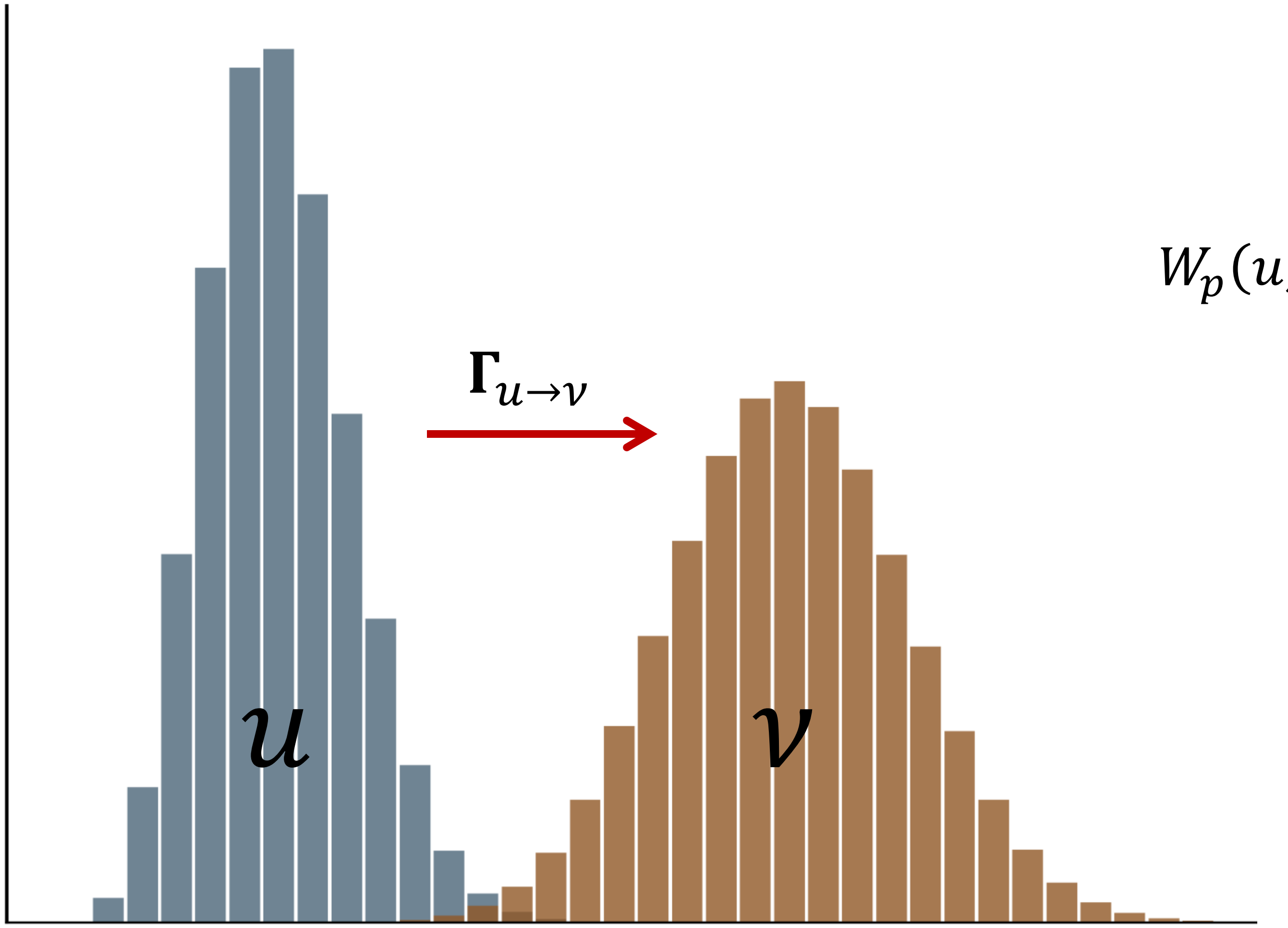


We have everything we need?






Wasserstein distance



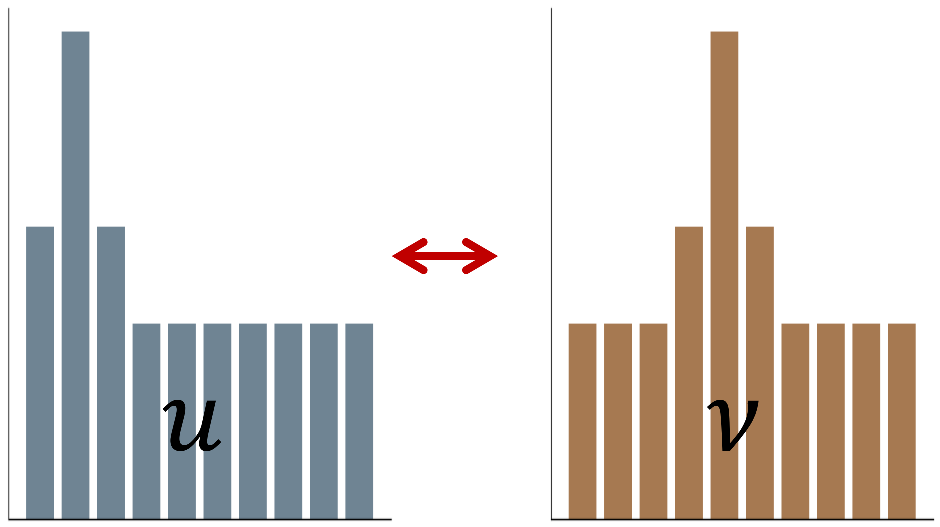
$$W_p(u,v) = \left(\min_{\Gamma \geq 0} \sum_{i,j=1}^{n} \left\| x_i - x_j \right\|^p \cdot \Gamma_{i,j} \right)^{1/p}$$

s.t. $\quad \sum_{j=1}^{n} \Gamma_{i,j} = u_i \quad \forall i,$

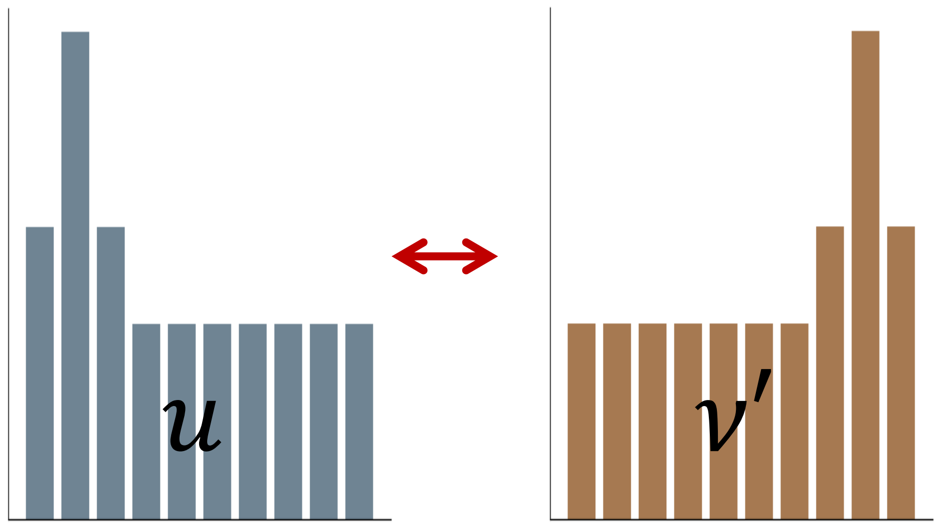
$\quad \sum_{i=1}^{n} \Gamma_{i,j} = v_j \quad \forall j$

Wasserstein distance considers the distributions' geometries

$$W_2(u, v) = 0.96$$
$$D_{KL}(u \parallel v) = 0.14$$

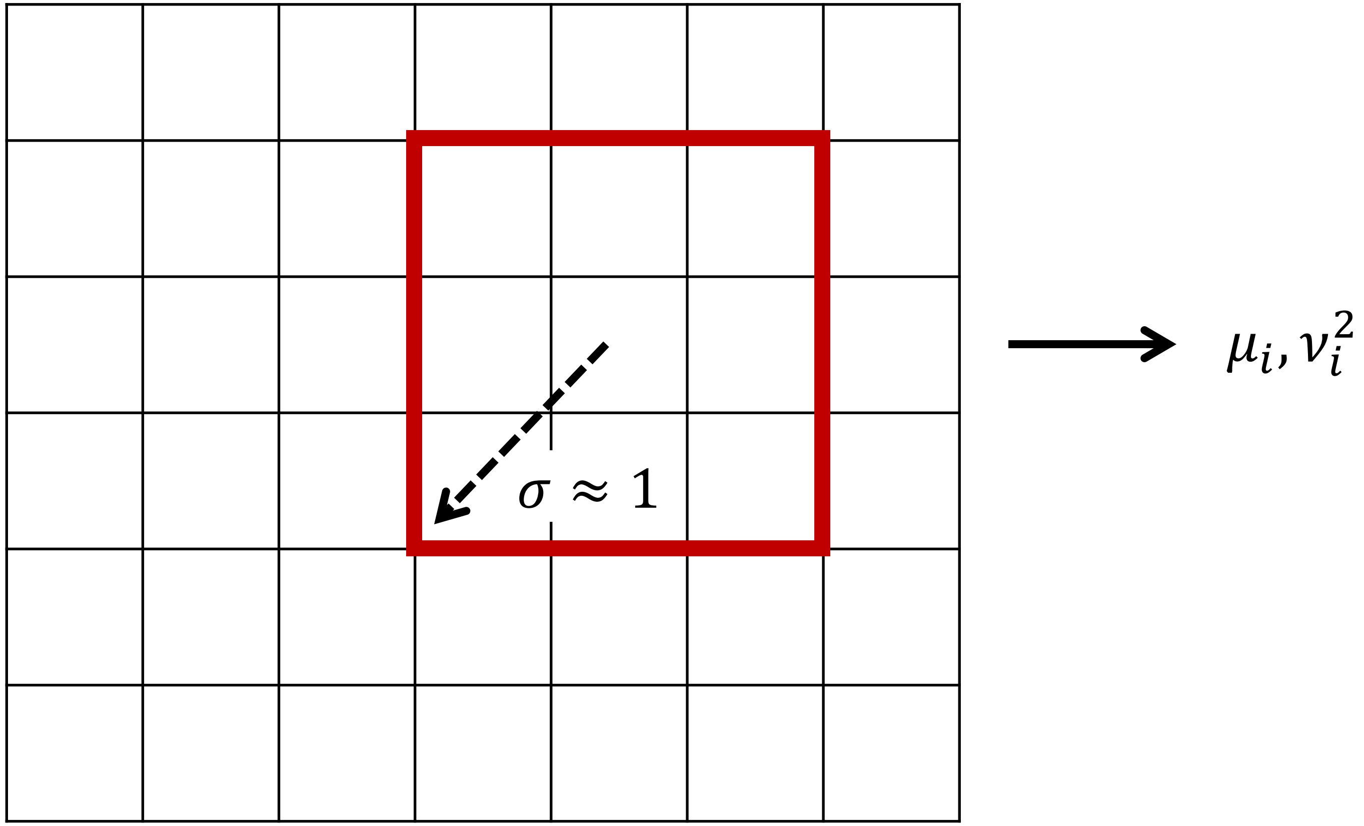


$$W_2(u, v') = 1.71$$
$$D_{KL}(u \parallel v') = 0.14$$

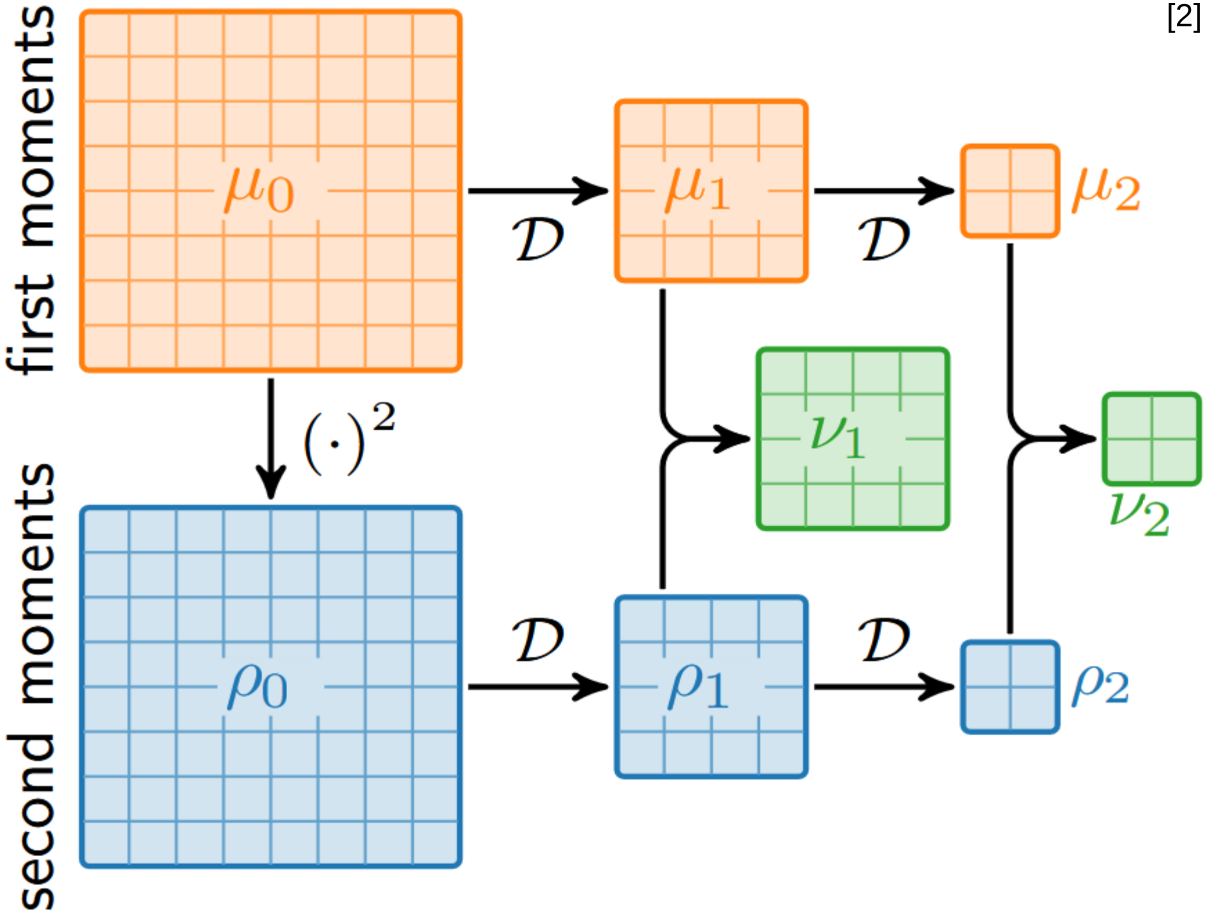


Approximating Wasserstein distance via Gaussians

$$Q_{0/1} \sim \mathcal{N}\left(\mu_{0/1}, \nu_{0/1}^2\right): \qquad W_2(Q_0, Q_1) = \sqrt{(\mu_0 - \mu_1)^2 + (\nu_0 - \nu_1)^2}$$



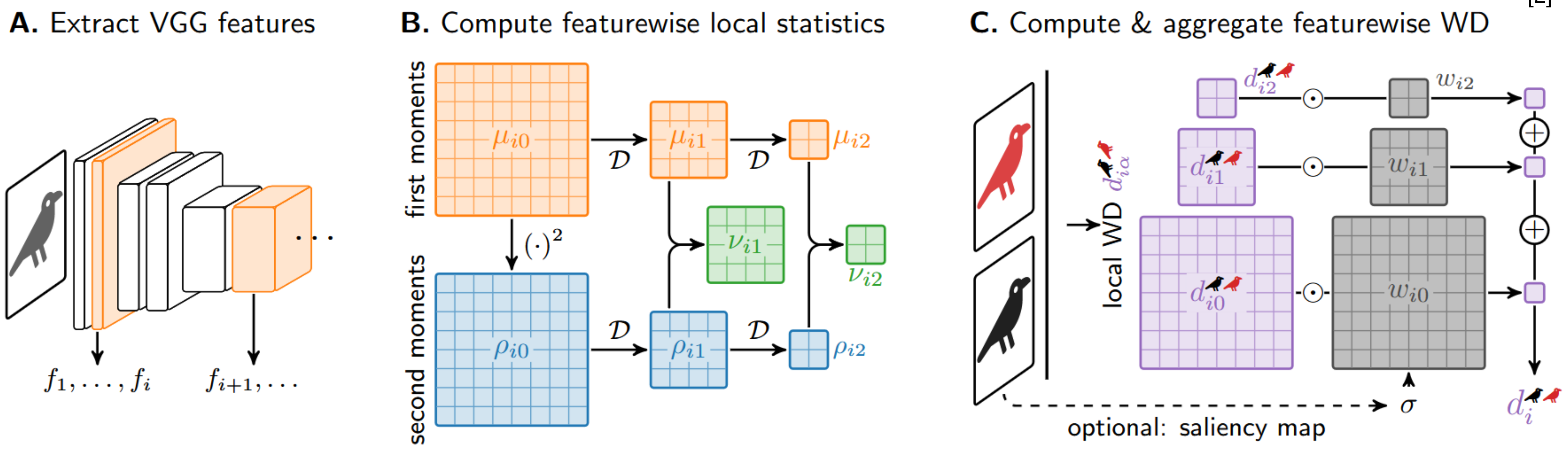
Approximating the Gaussian approximation



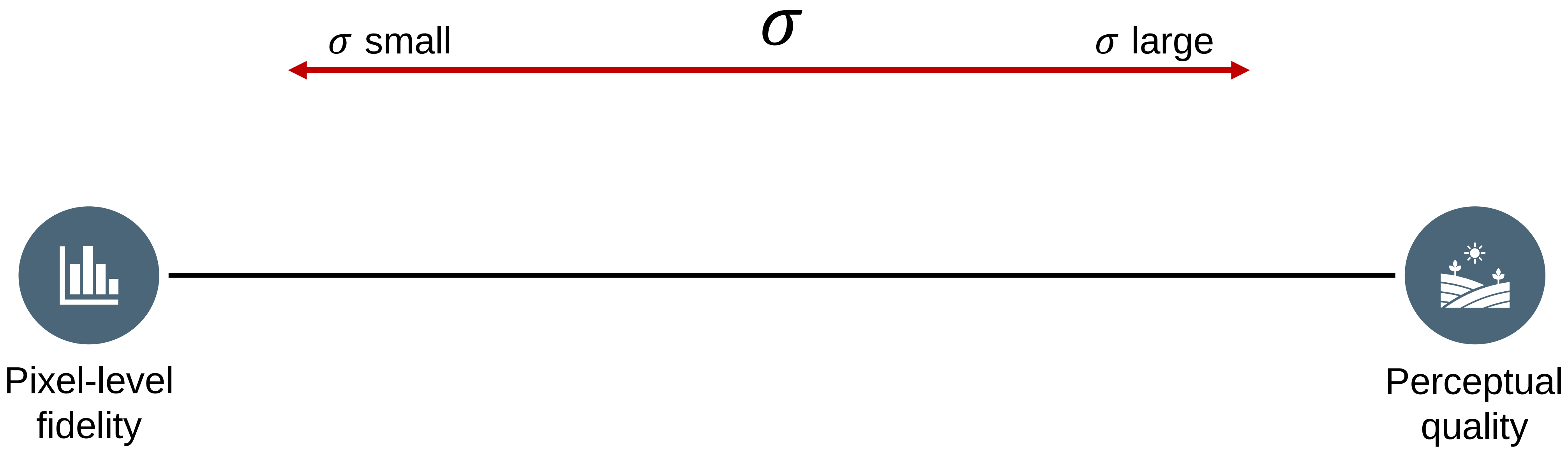
[2]

Wasserstein distortion (WD)

[2]



$\boldsymbol{\sigma}$ enables controlling the “Pixel-level fidelity”-“Perceptual quality” trade-off

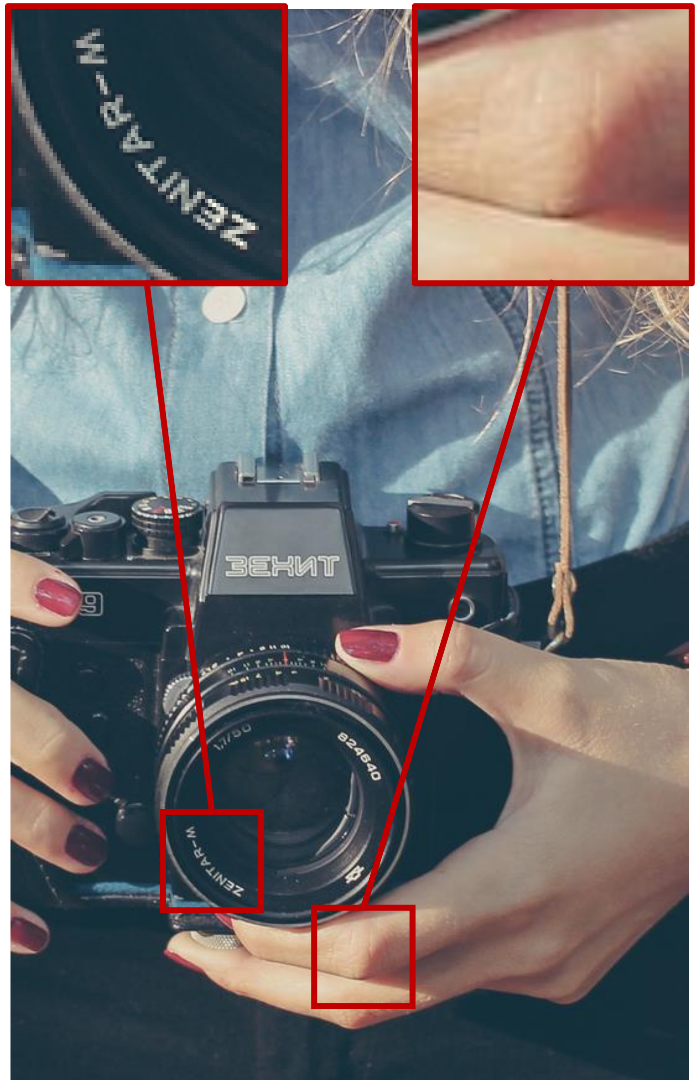


Putting this all together …

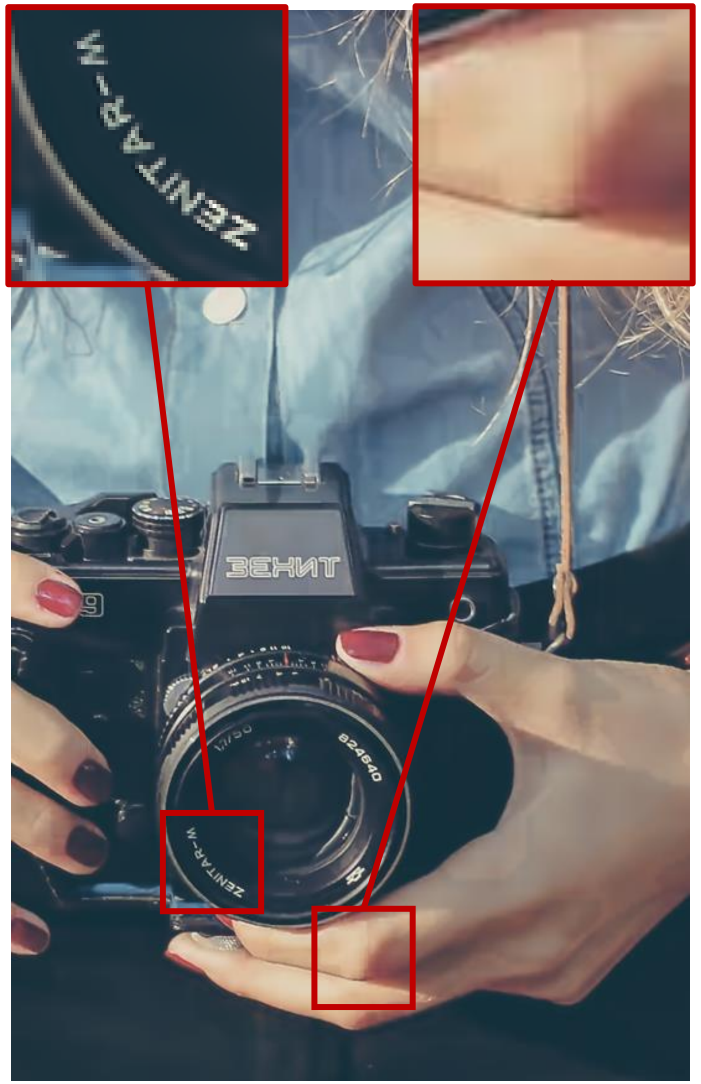
BPP = 0.075

C3/MSE

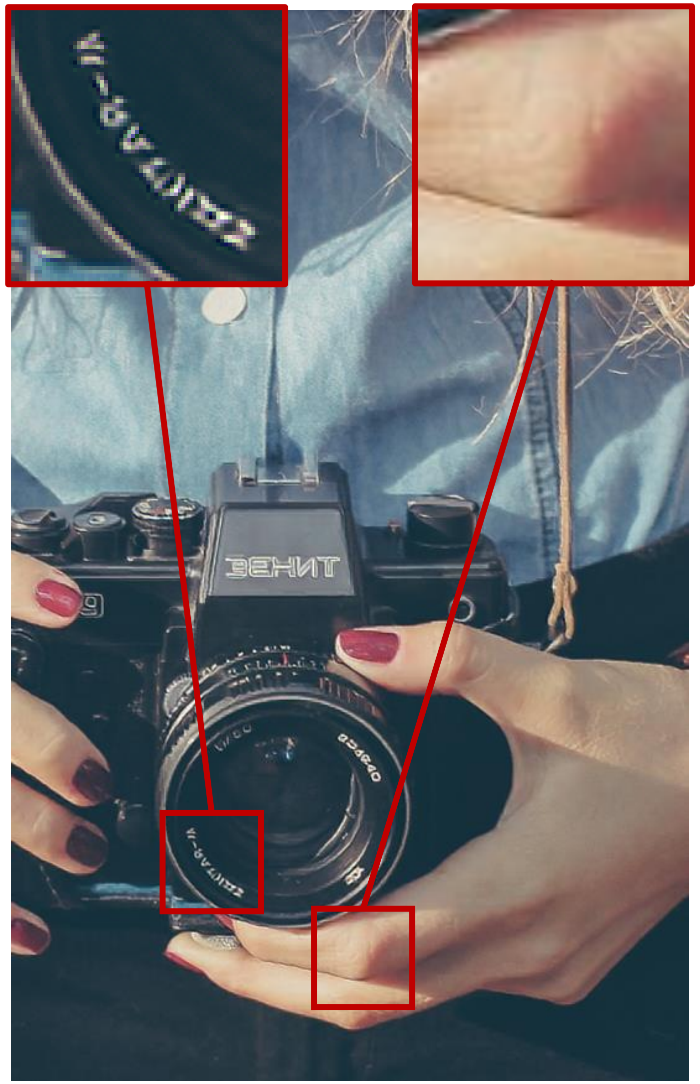
C3/WD



Original

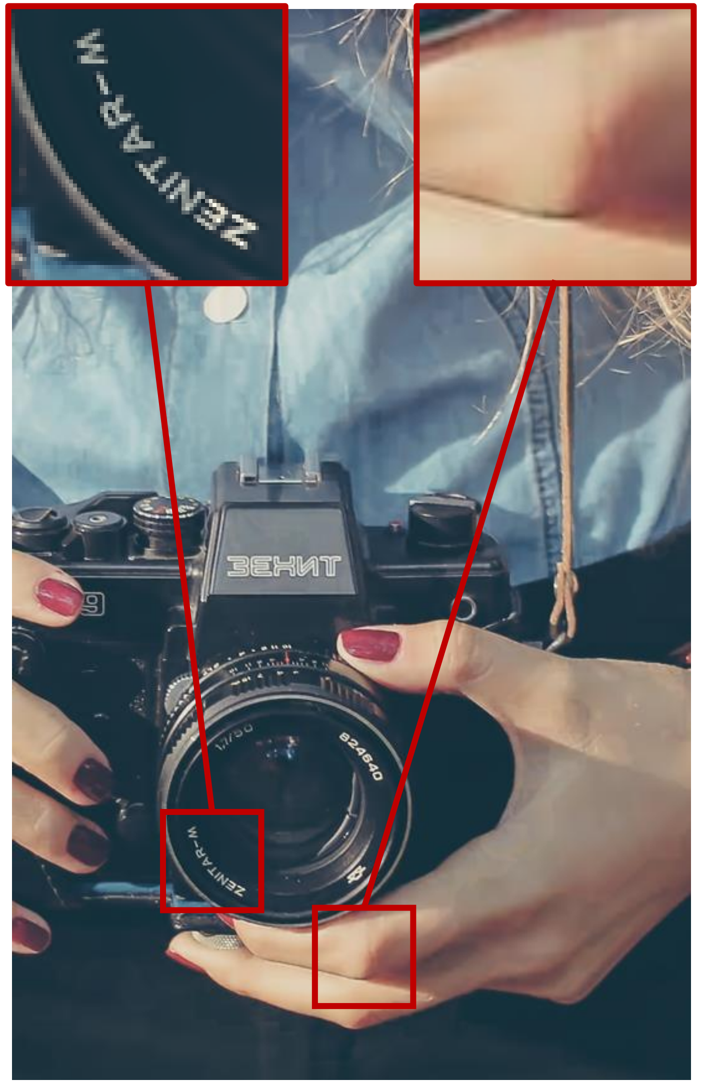


C3/MSE

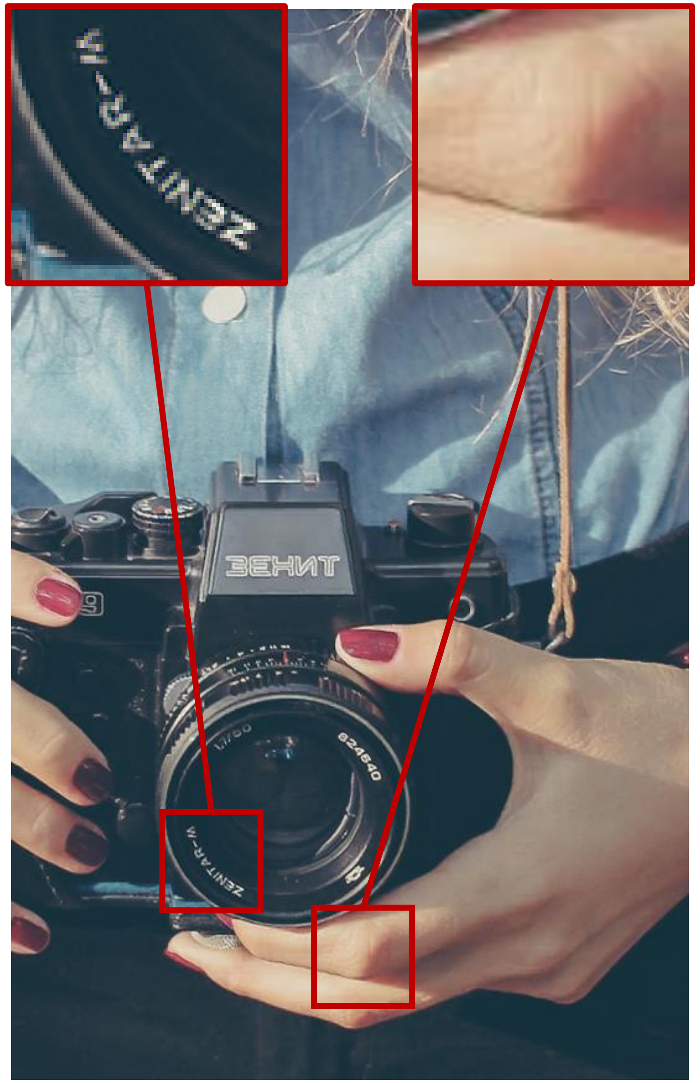


C3/WD

Saliency



C3/MSE + saliency



C3/WD + saliency

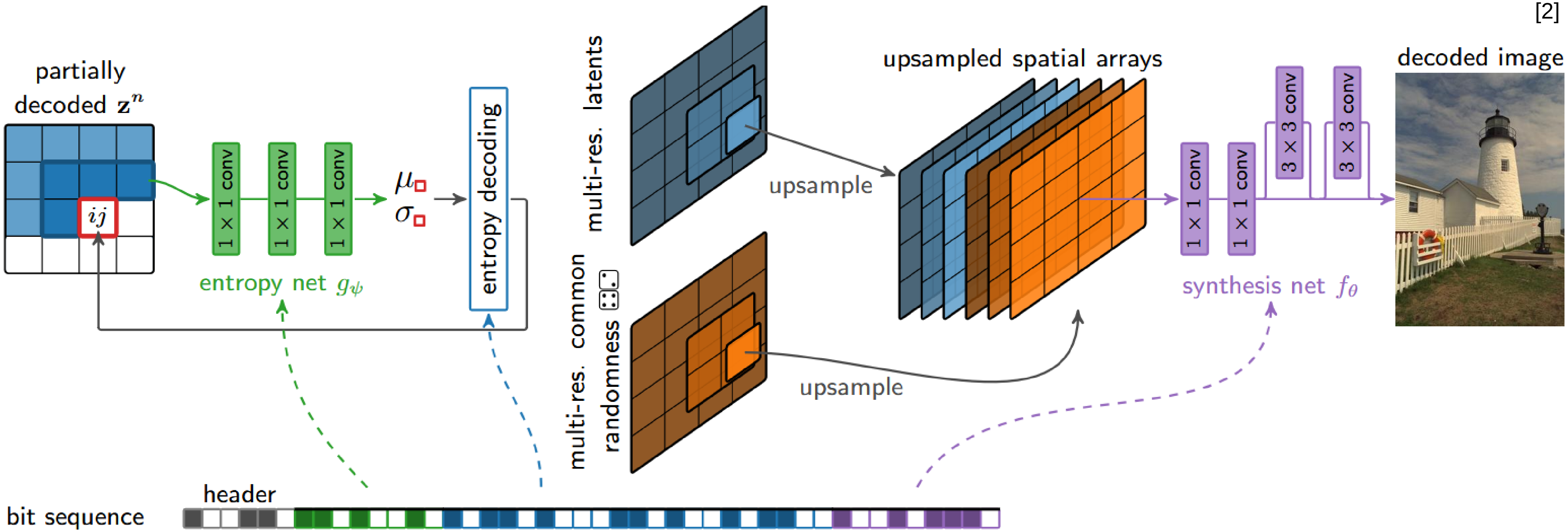
Integrating common randomness (CR)

C3 without CR

C3 with CR

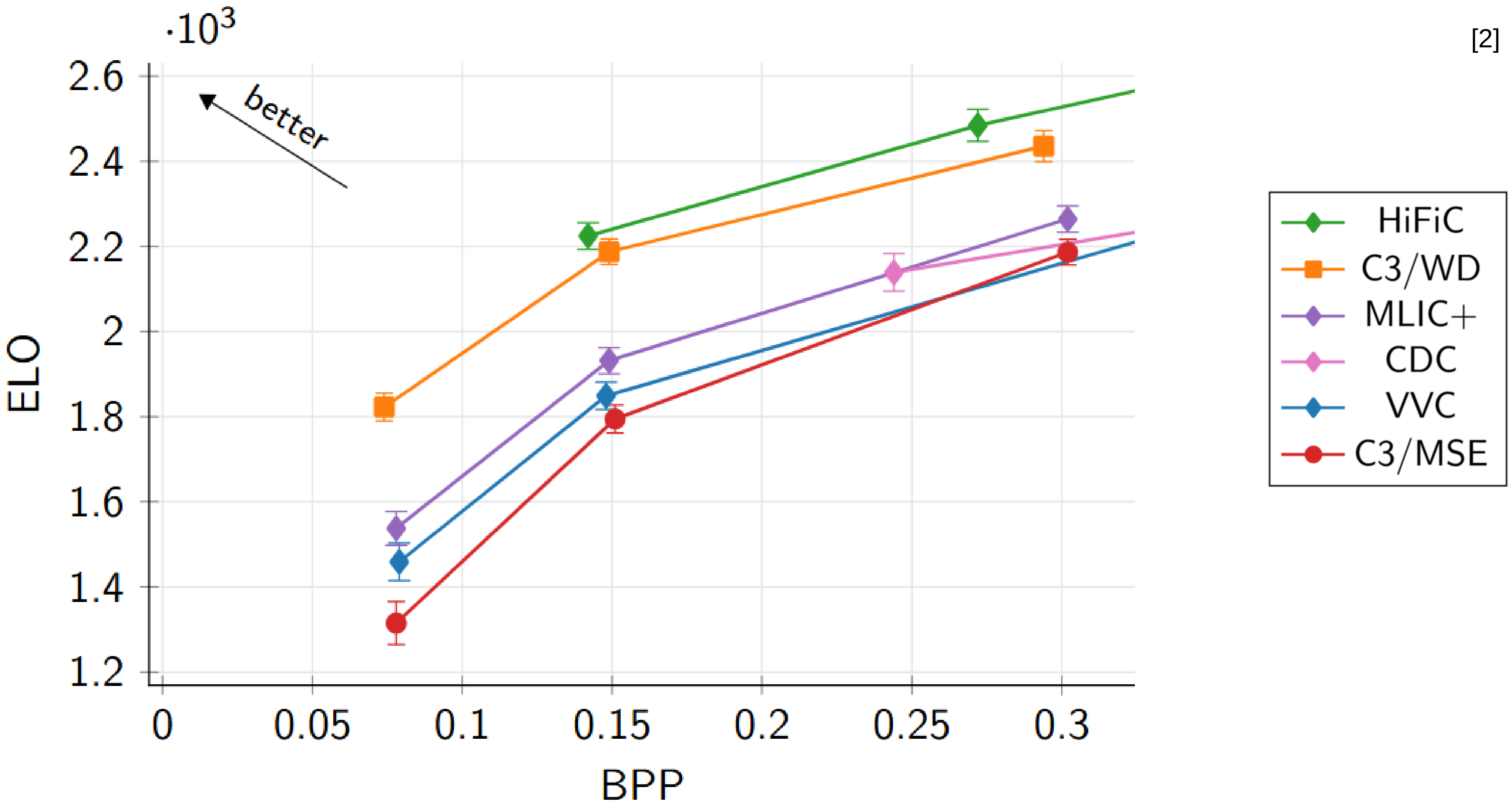
The (now truly) final pipeline

[2]



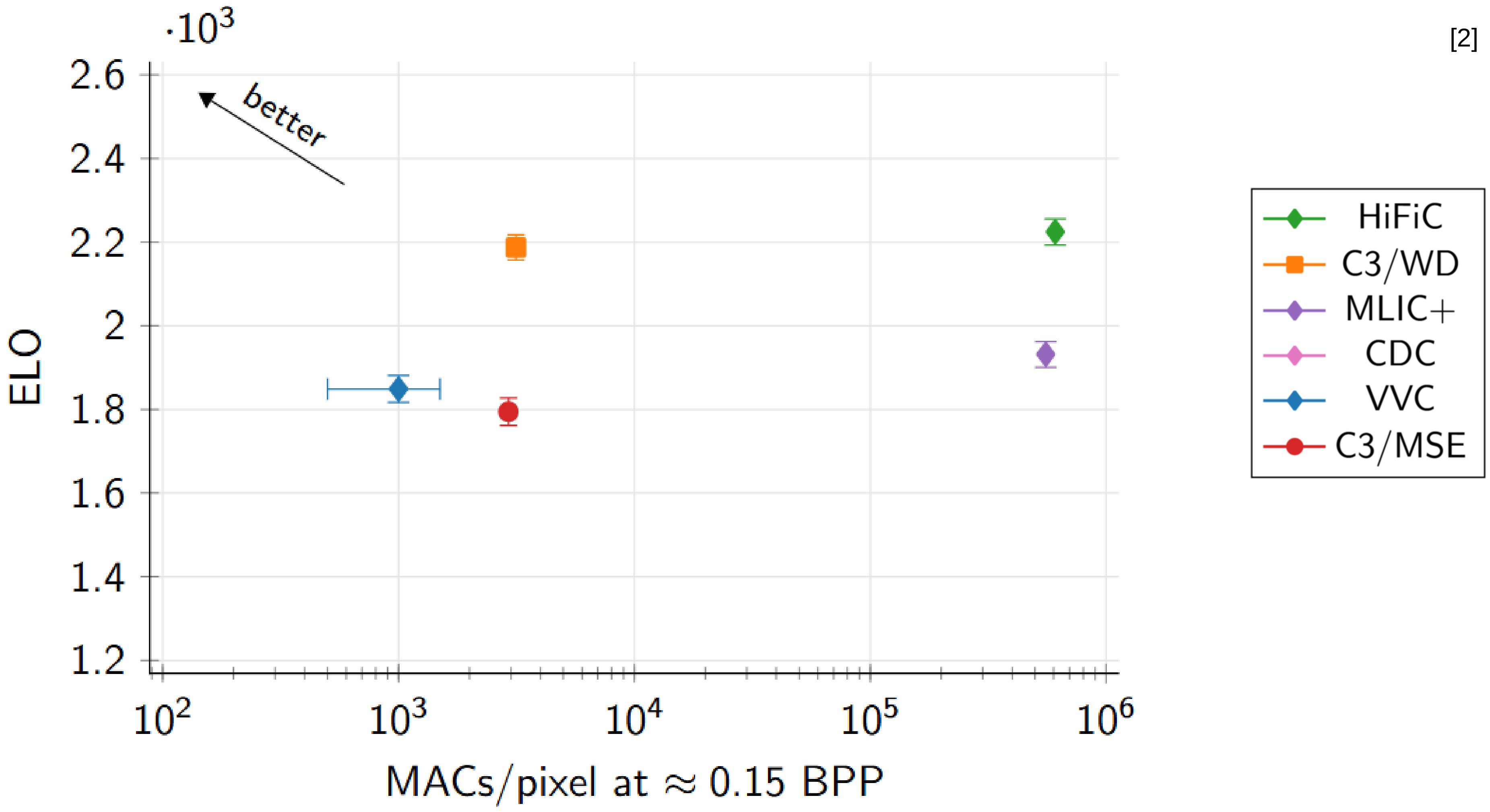
C3/WD exhibits competitive reconstruction quality …

[2]



... with very high decoding speed

[2]



Wasserstein distortion as a predictor for human evaluation

[2]

Metric	% correct	Correlation (PCC)
PSNR	61%	.36
MS-SSIM	65%	.54
NLPD	64%	.54
LPIPS	70%	.71
DISTS	67%	.73
PIM-5	70%	.76
WD fixed σ	**73%**	**.94**
WD saliency map	**73%**	**.94**

So, a perfect image compression technique?

Encoding speed …

“Good, Cheap, and Fast”?

Conclusion

Approach

Good visual reconstruction quality,

Cheap in terms of bit rate,

Fast decoding, at the cost of

Slow encoding

Methodological critique

? Somewhat “chaotic” comparisons

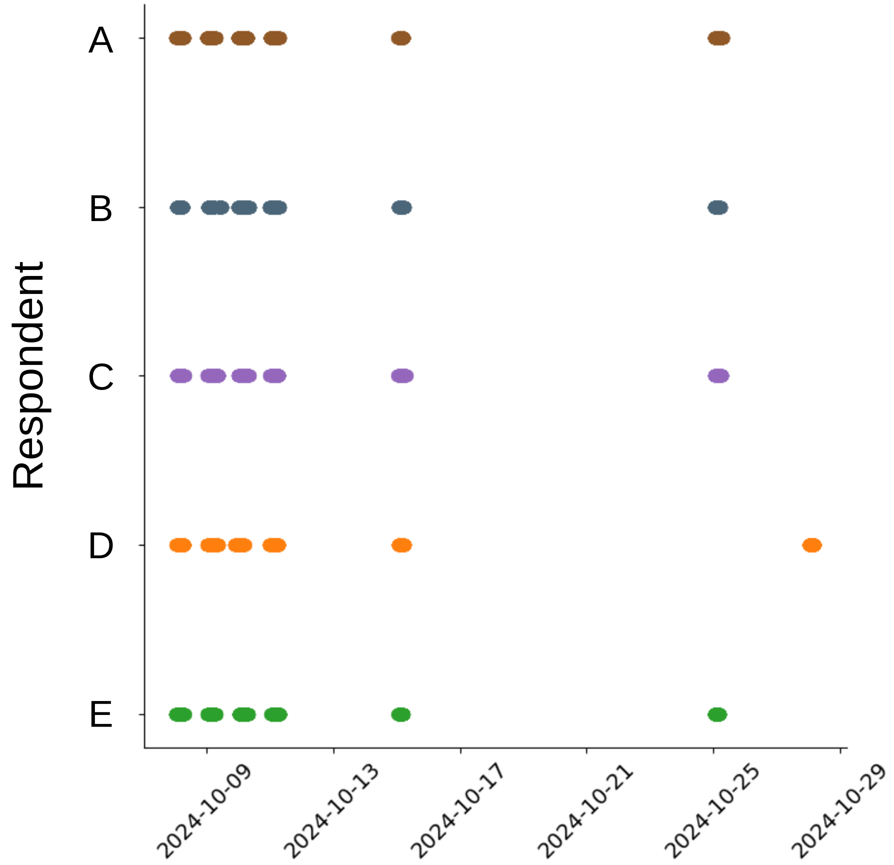
Limited human study (5 participants)

\+ Transparency: Provided images and evaluation results

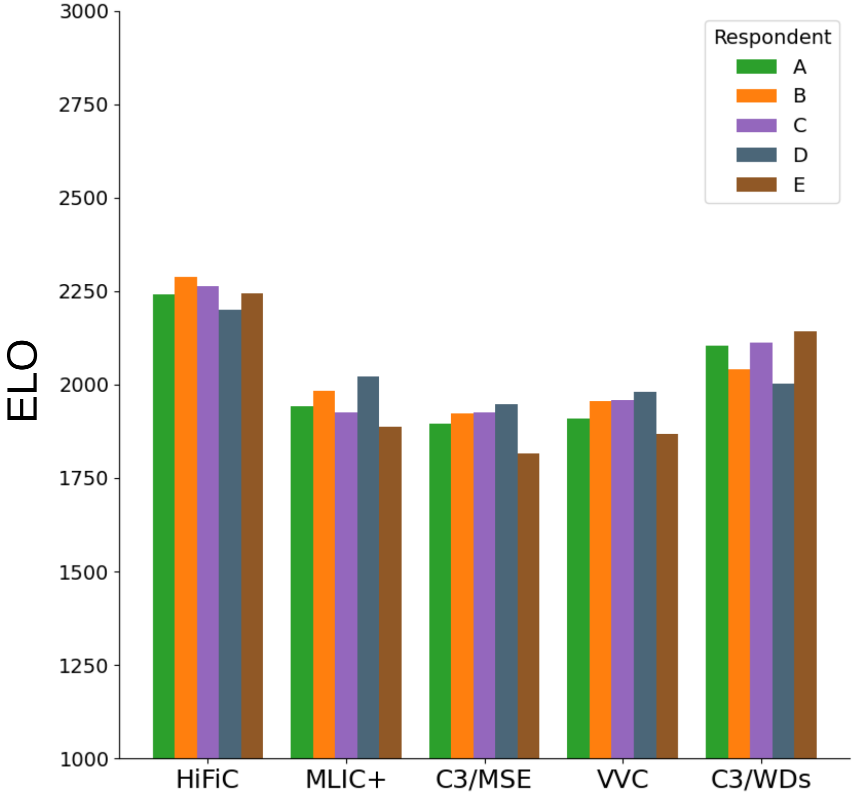
❯ **Not entirely novel approach, but a successful integration of existing techniques**

Looking at the raw survey results

When did the respondents take the survey?



How is the deviation between the respondents?



Discussion

References

[1] Kim, H., Bauer, M., Theis, L., Schwarz, J. R., & Dupont, E. (2024). *C3: High-performance and low-complexity neural compression from a single image or video*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Ballé, J., Versari, L., Dupont, E., Kim, H., & Bauer, M. (2024). Good, cheap, and fast: Overfitted image compression with Wasserstein distortion. arXiv preprint arXiv:2412.00505

[3] Qiu, Y., Wagner, A. B., Ballé, J., & Theis, L. (2024). *Wasserstein distortion: Unifying fidelity and realism*. In *Proceedings of the 58th Annual Conference on Information Sciences and Systems (CISS).*

[-] Ladune, T., Philippe, P., Henry, F., Clare, G., & Leguay, T. (2023). *COOL-CHIC: Coordinate-based low complexity hierarchical image codec*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

Further sources

- Images: Kodak lossless true color image suite (link) and CLIC 2020 validation dataset (download)
- Memes: Copied from [a] link and [b] link
- Data on slide 13: CLIC 2024 challenge task “image@0.075bpp” (link)