Investigating the Impact of Adversarial Attacks on AI-Based Image Compression Models

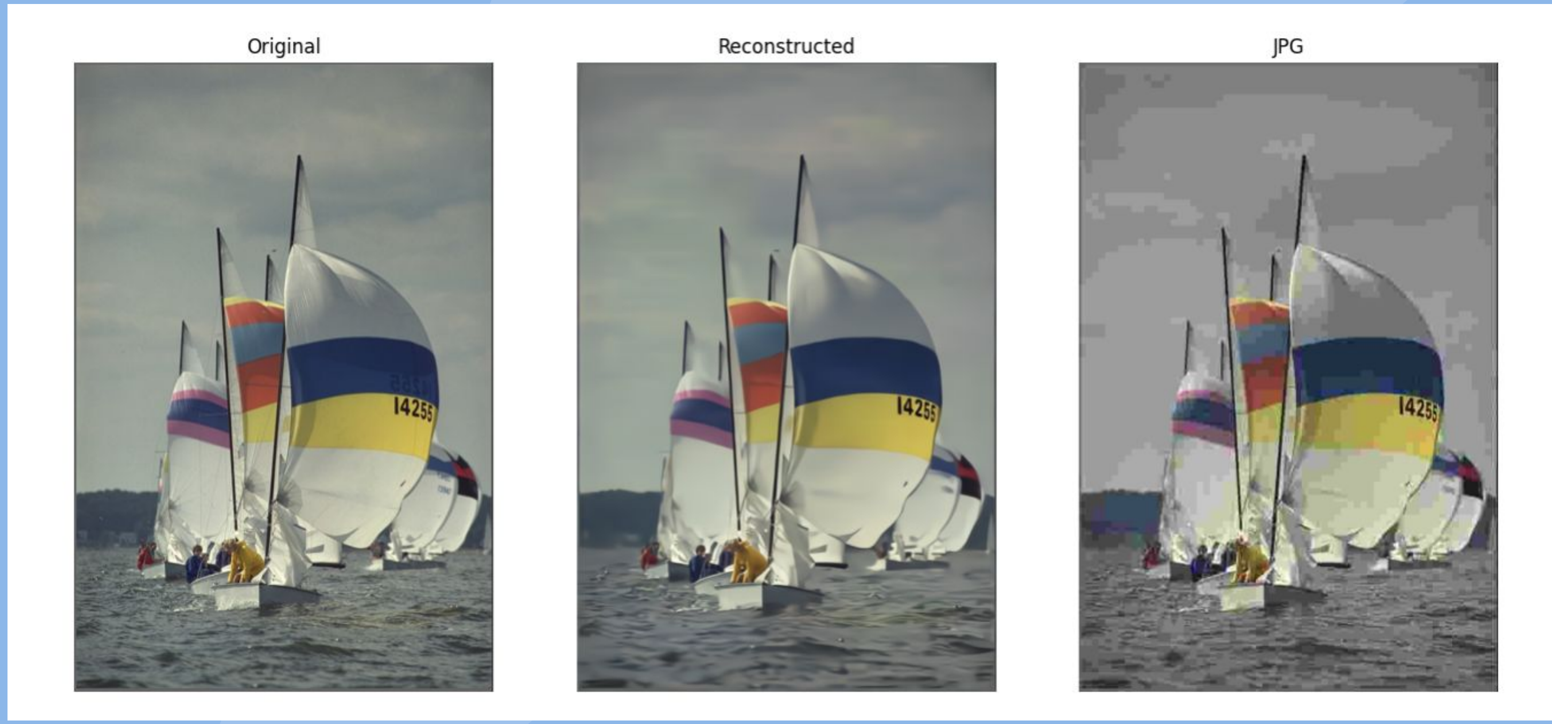
Made by:
Egor Miroshnichenko
Alexey Morozov
Timur Nabiev
Gennady Shutkov

Supervised by:
Razan Dibo

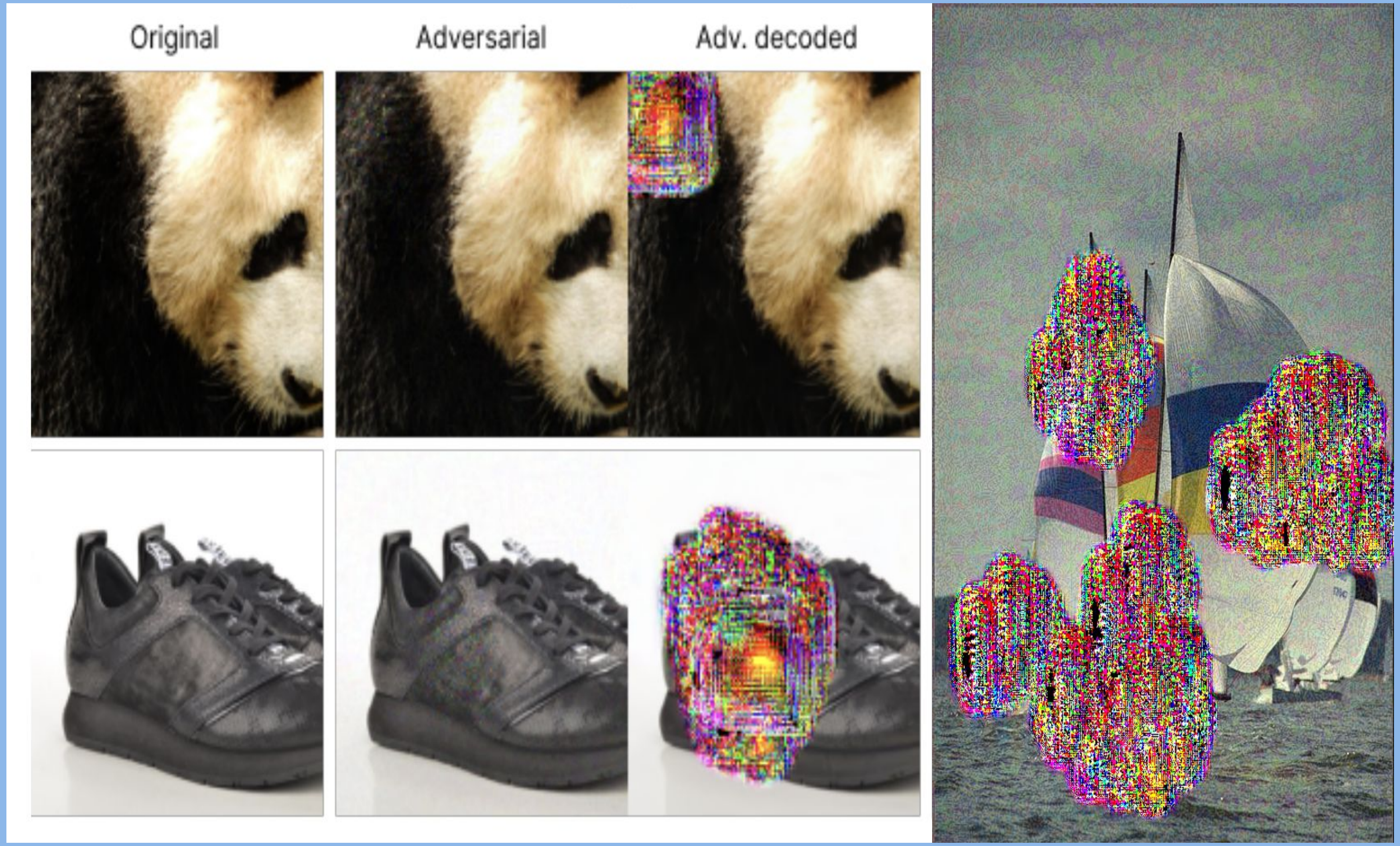
Machine Learning course – 2024
Final project
Team #19

Skoltech

PROBLEM STATEMENT



ML based image compression proved to be more efficient than industry standard JPG.



Focus of Analysis: examines how adversarial attacks impact the compression-decompression process of neural image compression (NIC).

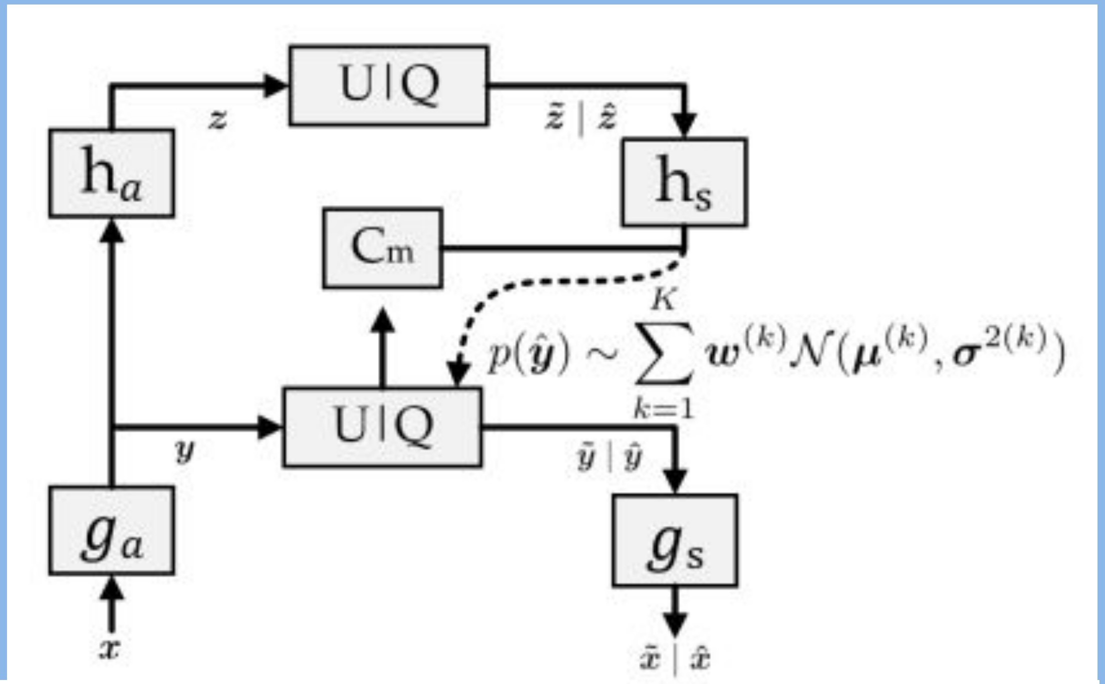
Kovalev, Egor & Bychkov, Georgii & Abud, Khaled & Gushchin, Alexander & Chistyakova, Anna & Lavrushkin, Sergey & Vatolin, Dmitriy & Antsiferova, Anastasia. (2024). Exploring adversarial robustness of JPEG AI: methodology, comparison and new methods. 10.48550/arXiv.2411.11795.

Skoltech

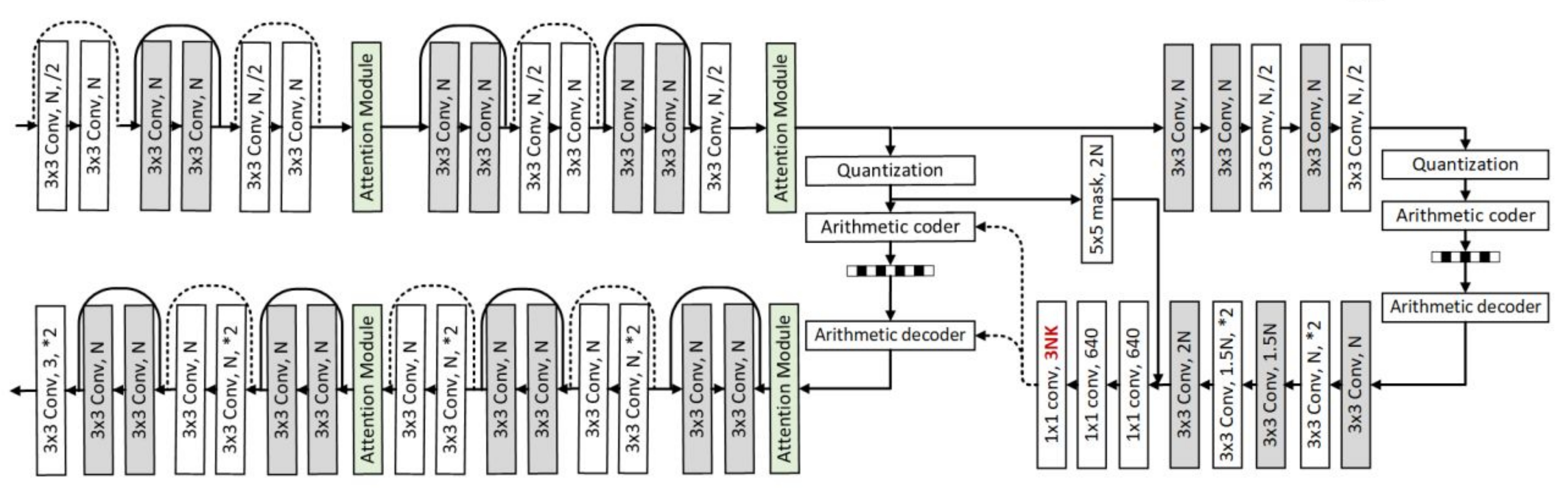
Project Goals

1. Understand how AI compression models work;
2. Test AI compression models on a dataset like Kodak to compress and reconstruct images.
3. Compare performance with traditional compression methods like JPEG. Evaluate results using PSNR, SSIM, bit rate BPP metrics;
4. Apply Adversarial Attacks: Generate adversarial examples (e.g., FGSM, PGD) for all kodak images (24 images);
5. Analyze Robustness: Evaluate attack effectiveness at 3+ compression levels, measuring the difference between the output of AI compression model before and after the attack.

Models architecture



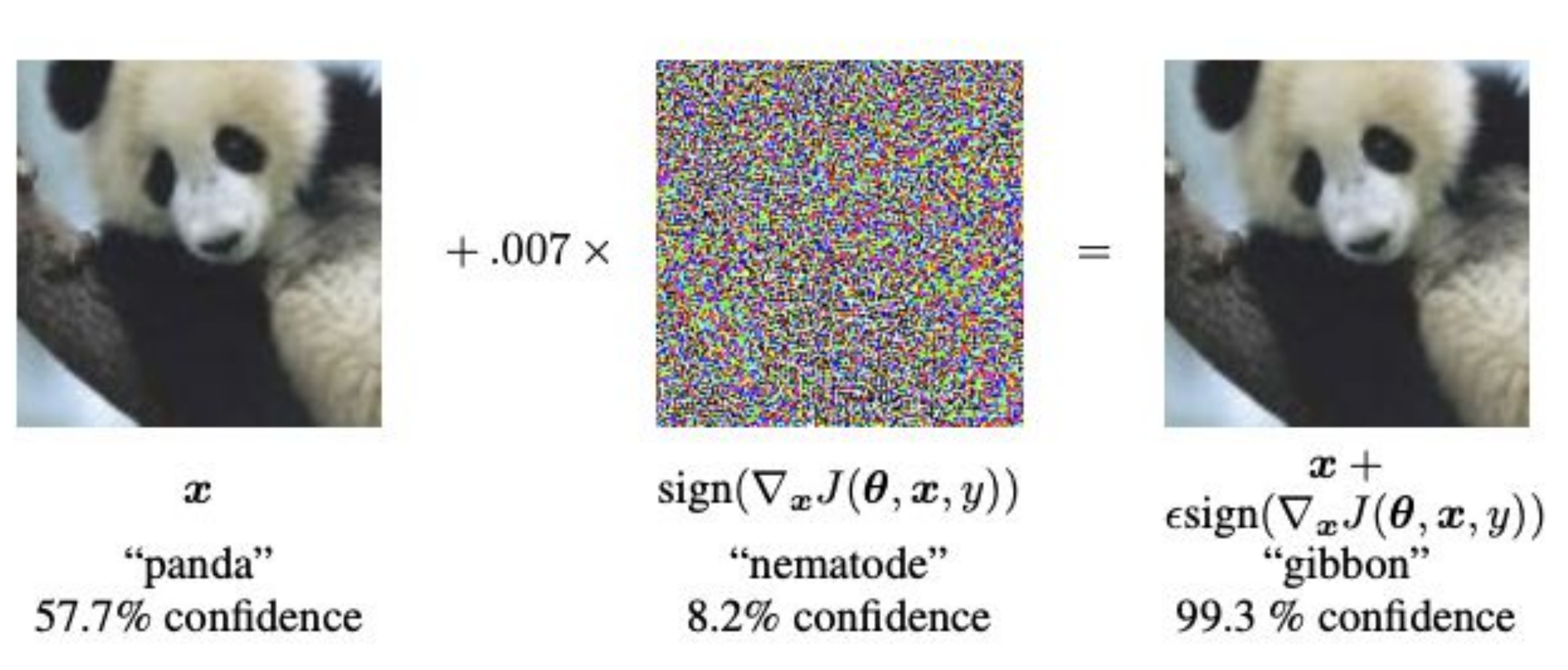
Cheng 2020 Anchor architecture



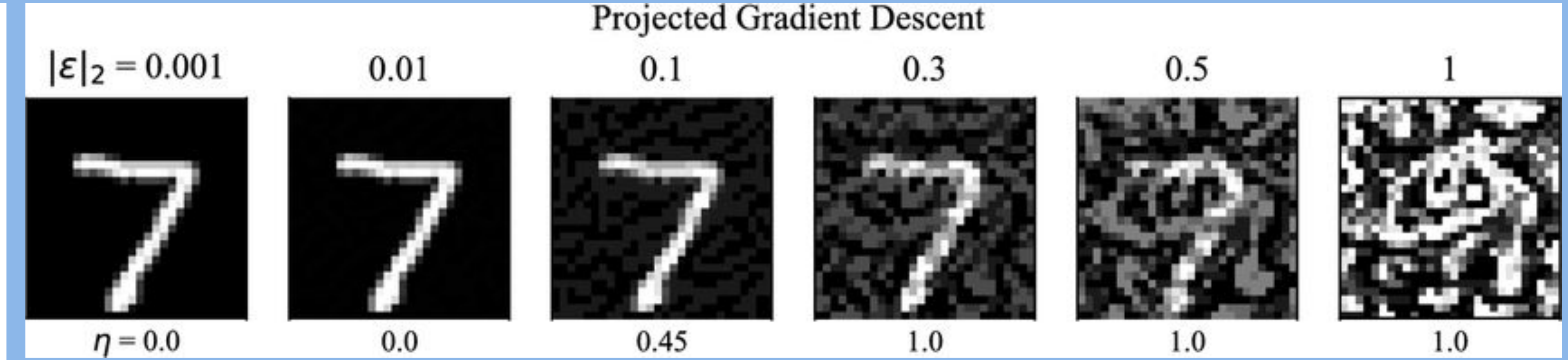
Cheng, Zhengxue & Sun, Heming & Takeuchi, Masaru & Katto, Jiro. (2020). Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules. 10.48550/arXiv.2001.01568. https://doi.org/10.48550/arXiv.2001.01568

Adversarial attack methods

Fast Gradient Sign Method (FGSM)



Projected Gradient Descent (PGD)



Single-step perturbation in gradient direction:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, \theta))$$

Iterative refinement with projection:

$$x^{t+1} = \text{Proj}_\epsilon \left(x^t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(x^t, \theta)) \right)$$

Skoltech

Goodfellow, Ian & Shlens, Jonathon & Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.

Adversarial attack methods

Fast Gradient Sign Method (FGSM)

Formula:

$$x' = x + \epsilon \cdot \mathrm{sign}(\nabla_x J(\theta, x, y))$$

Key Parameters:

- Original Image: x
- Perturbation Size: ϵ
- Loss Function: $J(\theta, x, y)$

Iterative Fast Gradient Sign Method (I-FGSM)

Formula:

$$x^{t+1} = x^t + \alpha \cdot \mathrm{sign}(\nabla_x J(\theta, x^t, y))$$

Key Parameters:

- Step Size: α
- Iteration Number: t
- Total Perturbation: ϵ

Momentum Iterative Fast Gradient Sign Method (M-FGSM)

Formula:

$$g^{t+1} = \mu \cdot g^t + \nabla_x J(\theta, x^t, y)$$

$$x^{t+1} = x^t + \alpha \cdot \mathrm{sign}(g^{t+1})$$

Key Parameters:

- Momentum Factor: μ
- Step Size: α

Projected Gradient Descent (PGD)

Formula:

$$x^{t+1} = \Pi_{\mathcal{B}(x,\epsilon)}(x^t + \alpha \cdot \mathrm{sign}(\nabla_x J(\theta, x^t, y)))$$

Key Parameters:

- Step Size: α
- Total Perturbation: ϵ

https://doi.org/10.48550/arXiv.1710.06081

Skoltech

Metrics

Cheng-2020 Anchor

(Anchor Variant)

- Input Image → Encoder → Entropy Model → Decoder → Reconstructed Image

Cheng-2020 Attn

(Attention Variant)

- Input Image → Encoder (with Self-Attention) → Entropy Model → Decoder (with Self-Attention) → Reconstructed Image

Metric	Loss function
MSE	$\mathcal{L} = \lambda * 255^2 * \mathcal{D} + \mathcal{R}$
MS-SSIM	$\mathcal{L} = \lambda * (1 - \mathcal{D}) + \mathcal{R}$

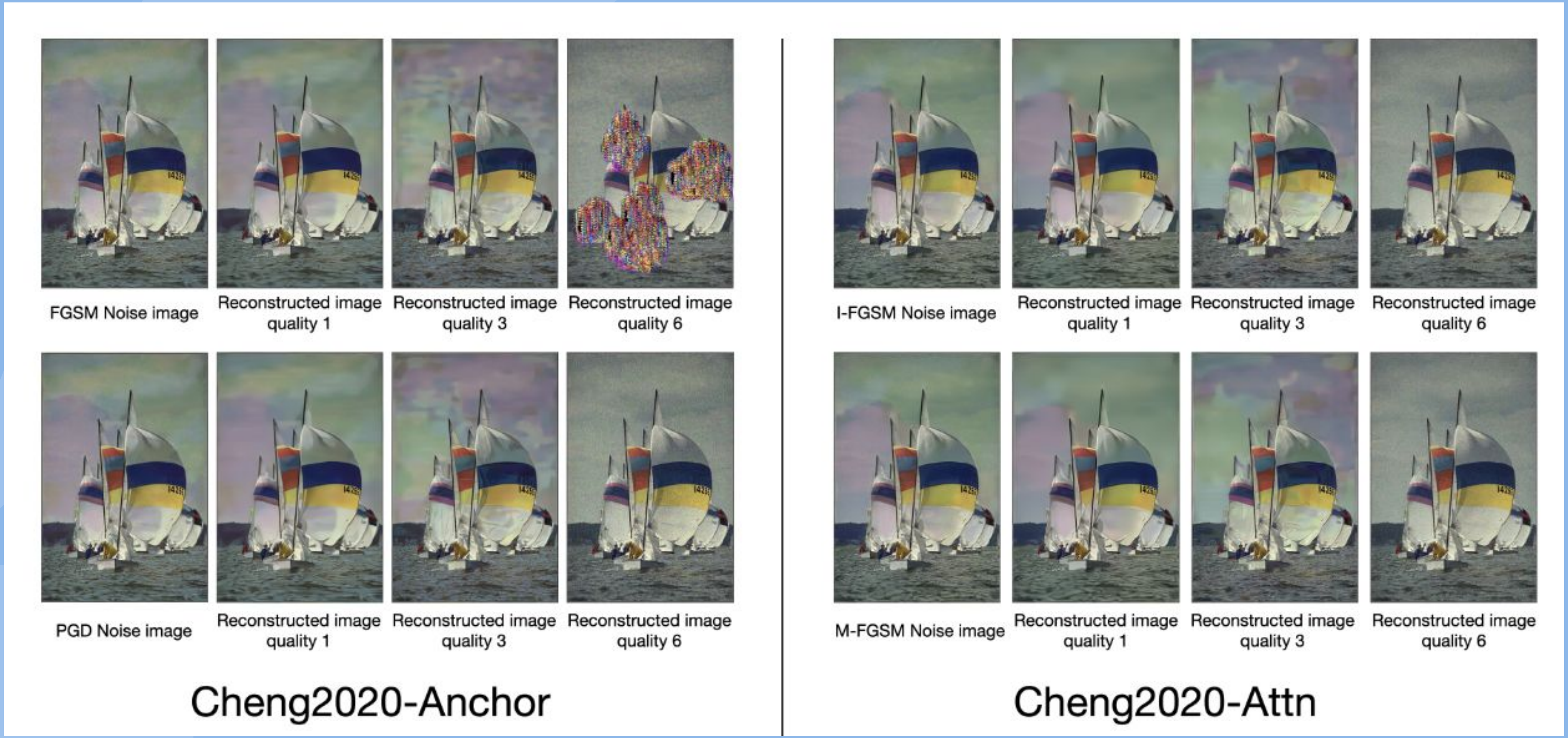
with ***D*** and ***R*** respectively the mean distortion and the mean estimated bit-rate.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

where $MAX_i^{\ 2}$ is a maximum pixel value

Skoltech

Experiments and Results

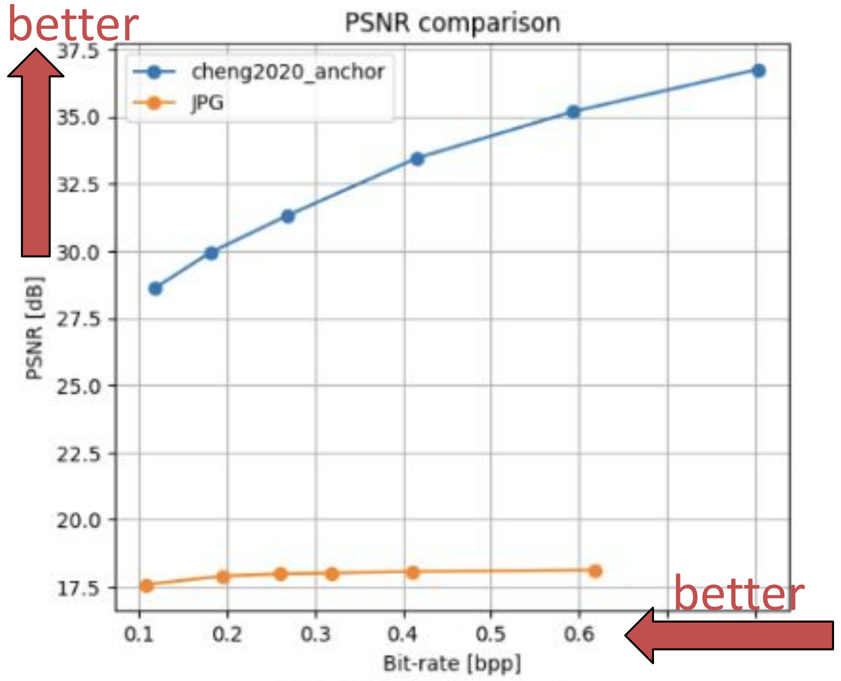


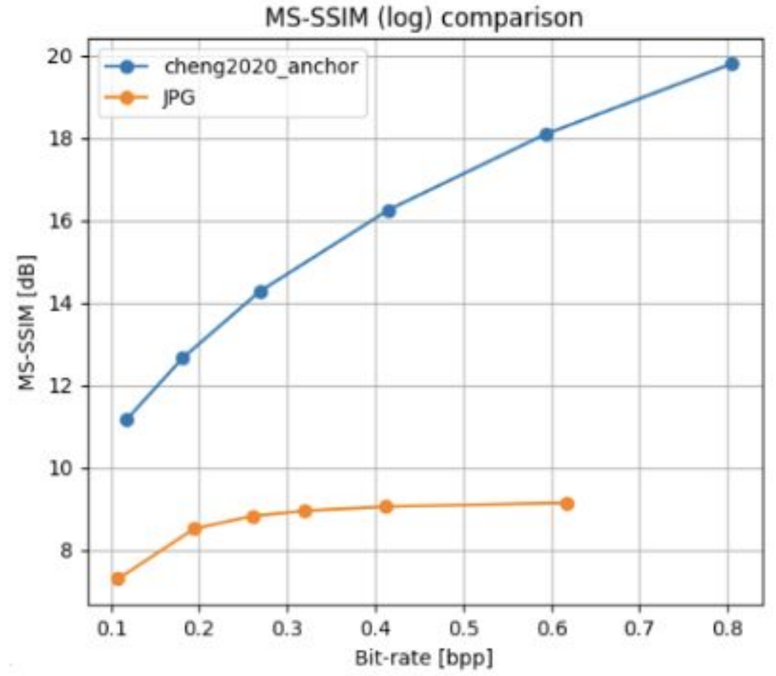
FGSM has the greatest impact on compression quality

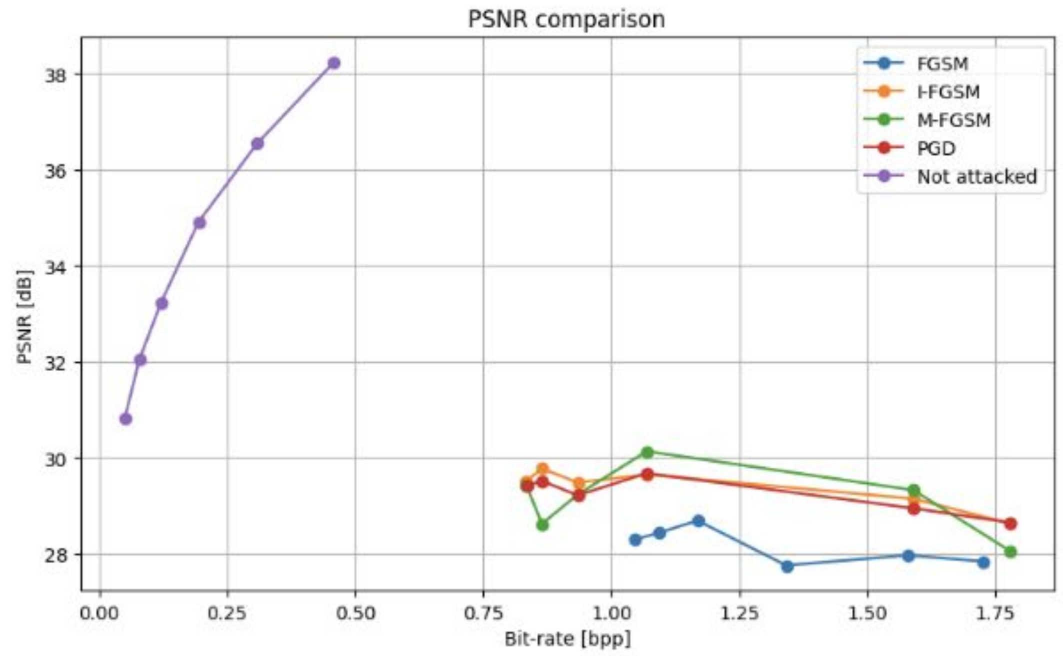
Skoltech

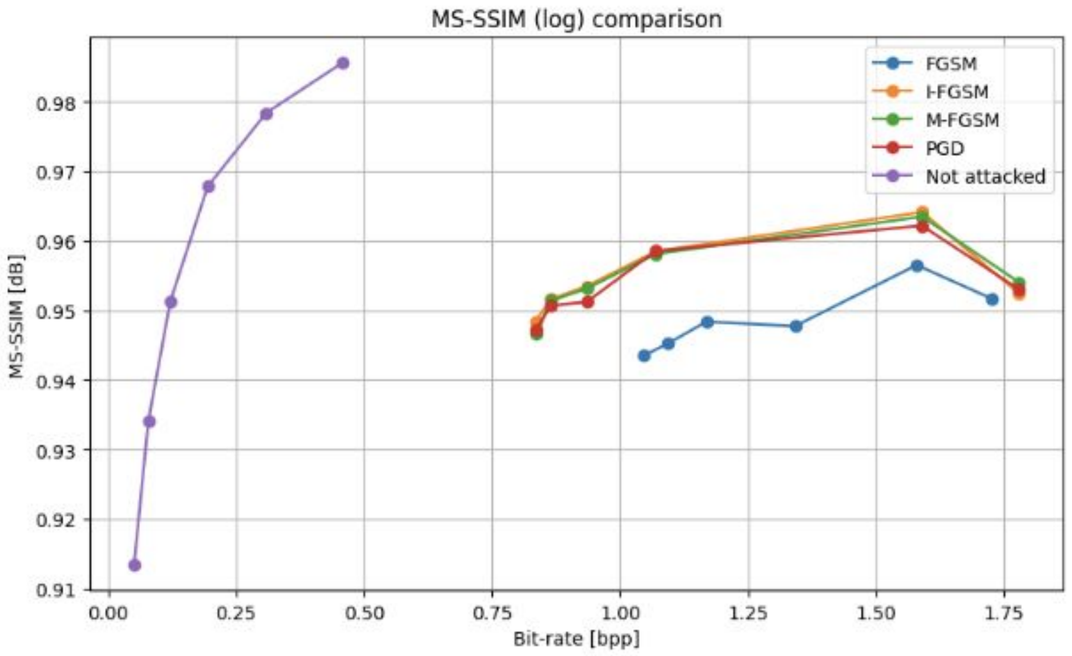
Experiments and Results

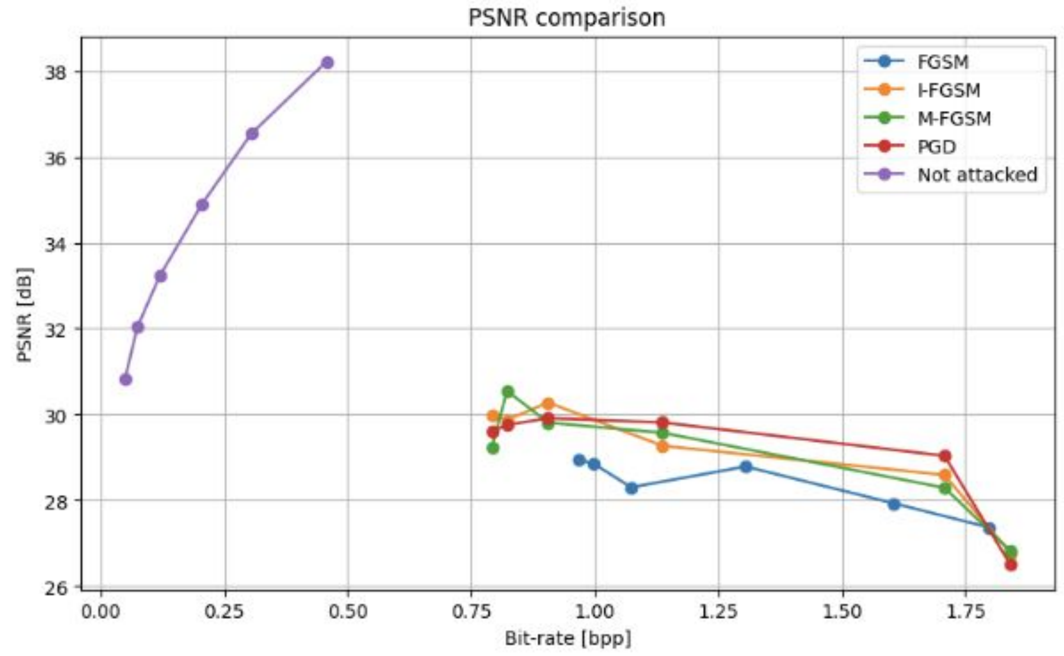
Metric	JPEG	Ch-Anchor	Ch-Attn
PSNR	27.13	30.86	30.74
MS-SSIM	0.89	0.96	0.96
Bit-Rate	0.075	0.071	0.070

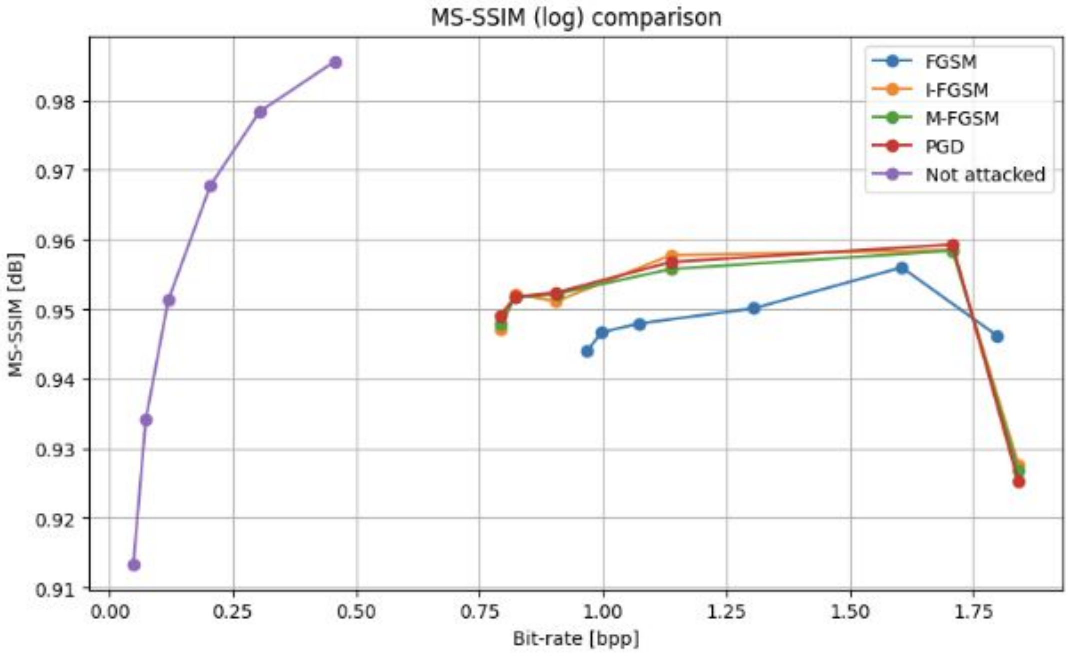




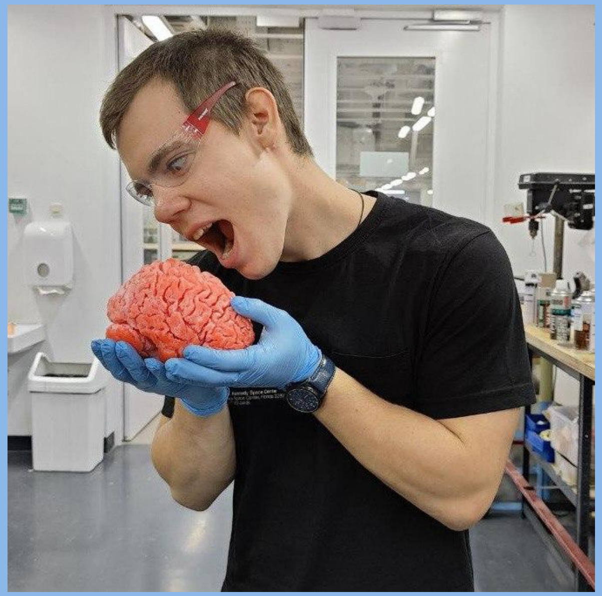








Our team #19

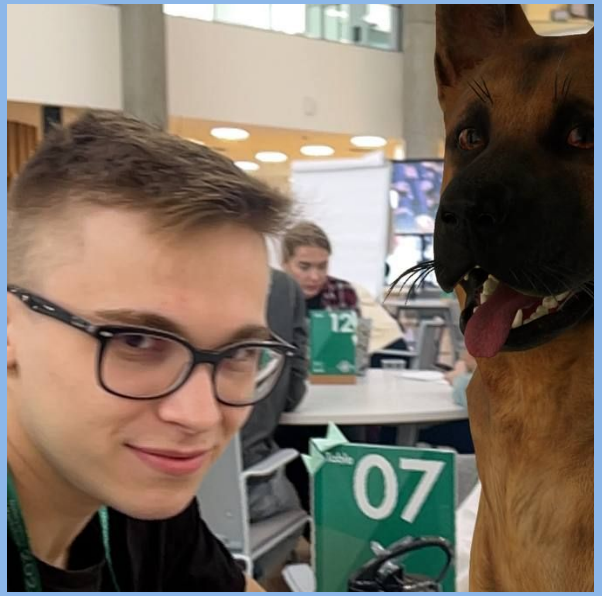


Gennady Shutkov
MS-2 IoT

- Coding main algorithm
- Post Processing

Egor Miroshnichenko
MS-1 DS

- Literature review
- Reporting



Timur Nabiev
MS-1 DS

- Literature review
- Data collection

Alexey Morozov
MS-1 LS

- Preparing the GitHub Repo
- Coding main algorithm

Thanks our lecturers and TA’s for teaching, guidance and support!

Reference

- Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436, 2018.
- Bégaint, J., Racapé, F., Feltman, S., and Pushparaja, A. Compressai: a pytorch library and evaluation platform for end-to-end compression research. arXiv preprint arXiv:2011.03029, 2020.
- Cheng, Z., Sun, H., Takeuchi, M., and Katto, J.Learned image compression with discretized gaussian mixture likelihoods and attention modules. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7939–7948, 2020.
- Cihang Xie, Zhishuai Zhang, Y. Z. S. B. J. W. Z. R. A. Y. Improving transferability of adversarial examples with input diversity. https://doi.org/10.48550/arXiv.1803.06978, 2019.
- Kamisli, F., Racapé, F., and Choi, H. Variable-rate learned image compression with multi-objective optimization and quantization-reconstruction offsets. 2024.

Skoltech

Thx!

Skoltech