

---

# Learned Image Compression for Earth Observation: Implications for Downstream Segmentation Tasks

---

**Christian Mollière**  
OroraTech GmbH  
Munich

**Iker Cumplido**  
Department of Informatics  
LMU, Munich

**Marco Zeulner**  
Department of Informatics  
LMU, Munich

**Lukas Liesenhoff**  
OroraTech GmbH  
Munich

**Matthias Schubert**  
Department of Informatics  
LMU, Munich

**Julia Gottfriedsen**  
OroraTech GmbH  
Munich

## Abstract

The rapid growth of data from satellite-based Earth observation (EO) systems poses significant challenges in data transmission and storage. We evaluate the potential of task-specific learned compression algorithms in this context to reduce data volumes while retaining crucial information. In detail, we compare traditional compression (JPEG 2000) versus a learned compression approach (Discretized Mixed Gaussian Likelihood) on three EO segmentation tasks: Fire, cloud, and building detection. Learned compression notably outperforms JPEG 2000 for large-scale, multi-channel optical imagery in both reconstruction quality (PSNR) and segmentation accuracy. However, traditional codecs remain competitive on smaller, single-channel thermal infrared datasets due to limited data and architectural constraints. Additionally, joint end-to-end optimization of compression and segmentation models does not improve performance over standalone optimization.

## 1 Introduction

Satellite-based Earth Observation (EO) systems have become essential for monitoring and managing environmental phenomena, particularly wildfires, which pose severe threats to ecosystems and human communities. However, the expanding spatial coverage and increased resolution from EO missions create significant challenges in terms of data transmission and storage. To address these challenges, well-designed image compression algorithms are key. Unlike traditional compression methods, which primarily seek to preserve visual fidelity, task-specific compression aims to retain essential features necessary for accurate downstream analysis, such as image segmentation. Effective task-specific compression would allow significant data volume reductions while maintaining the performance and interpretability of critical tasks like wildfire and cloud detection. The research question that motivates our work is therefore: What is the optimal EO compression regime for various downstream tasks, while maintaining a balanced rate-distortion trade-off? In this paper, we compare the performance of traditional image compression techniques, specifically JPEG 2000, with that of learned compression approaches across three representative EO segmentation tasks: cloud detection, fire detection, and building footprint extraction. By systematically evaluating these methods on diverse datasets differing in spectral modality (thermal infrared versus optical multispectral), dataset size, and spatial resolution, we identify optimal compression strategies tailored to specific EO scenarios.

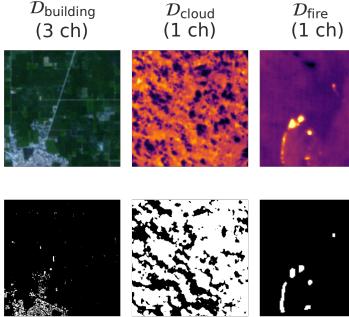


Figure 1a: Samples of the datasets used in this work. The thermal segmentation tasks are evaluated monochromatically, whereas the optical task is providing three channels.

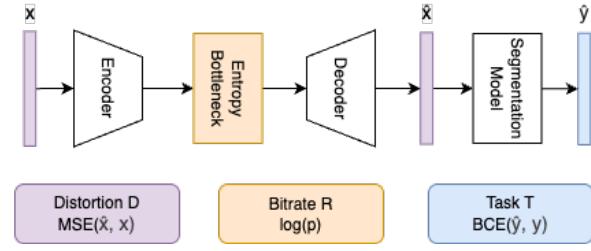


Figure 1b: Overview of experiment setup for task-specific raster data compression. Raw observations  $x$  are compressed onboard the satellite and decompressed at the ground station for segmentation analysis.

## 2 Experimental Setup

**Datasets** We conducted experiments using three diverse EO datasets representing distinct segmentation tasks: Cloud detection, fire detection, and building footprint extraction. Collectively, these datasets allow comprehensive evaluation of compression strategies across various EO scenarios, ensuring our findings are suitable for real-world applications. Samples are shown in Figure 1a.

- **Fire Dataset** (Rötzer et al. [2025]) The fire dataset  $\mathcal{D}_{\text{fire}}$  originates from a commercial CubeSat platform using mid-wave infrared (MWIR) imagery at  $3.8 \mu\text{m}$  wavelength, optimal for detecting active combustion. Initially comprising 40 full scenes, we augment the data by cropping 571 smaller  $32 \times 32$  pixel patches around confirmed fire locations, each accompanied by expert-annotated binary fire masks.
- **Cloud Dataset** (Wölki et al. [2024]) The cloud dataset  $\mathcal{D}_{\text{cloud}}$  comprises thermal infrared imagery also acquired from commercial CubeSat constellation. Each image has a ground sampling distance (GSD) of 200 m, capturing radiance in the long-wave infrared (LWIR) spectrum, centered at  $11.5 \mu\text{m}$ . The dataset consists of 528 manually annotated scenes of cloud presence. Each scene is resized to  $256 \times 256$  pixels.
- **Building Dataset** (Prexl and Schmitt [2023]) The building dataset  $\mathcal{D}_{\text{building}}$  is based on multispectral optical imagery from ESA’s Copernicus Sentinel-2 mission, featuring spatial resolutions of 10–20 m per pixel. We utilized 28,828 10-channel image-mask pairs, each resized to  $128 \times 128$  pixels. The masks represent building footprints extracted from OpenStreetMap and Microsoft global building footprint layers, providing a substantial dataset suitable for training and evaluating the segmentation and compression models.

**Segmentation Baseline Methods** To establish robust baseline performance, we employ a U-Net architecture (Ronneberger et al. [2015]), a widely adopted convolutional neural network designed specifically for segmentation tasks. We use two different encoder backbones, MobileNetV2 (Sandler et al. [2018]) and ResNet34 (He et al. [2016]), selected for their proven effectiveness and computational efficiency. Models are trained with binary cross-entropy (BCE) loss and optimized using the Adam optimizer. Training incorporates data-specific augmentation and early stopping criteria based on validation performance to ensure stability and generalization.

**Compression Algorithms** We compare two distinct image compression methods and systematically analyze the rate-distortion trade-off of each method using standard metrics. This includes the Peak Signal-to-Noise Ratio (PSNR) and bits per pixel (bpp), across different datasets.

- **JPEG 2000:** A classical, wavelet-based compression method widely used for its efficiency, scalability, and support for both lossless and lossy compression. JPEG 2000 (Taubman and Marcellin [2002]) serves as our baseline to benchmark learned compression performance.

- **Discretized Gaussian Mixture Likelihood:** A variational autoencoder-based compression model employing advanced entropy modeling techniques, including Gaussian Mixture Likelihoods and attention mechanisms (Cheng et al. [2020]). It is specifically designed to reduce spatial redundancy in its latent representation, which is of importance for efficient coding when fewer spectral channels are available. It significantly outperforms JPEG 2000 when evaluated on natural RGB images. We selected this architecture from the CompressAI framework (Bégaint et al. [2020]), despite newer SoTA architectures being available, given its accessibility among other implementations and its proven performance over our baseline.

**Optimization Methodology** To investigate potential synergies between compression and segmentation, we implement a fully differentiable pipeline combining both models. This approach allows joint training, optimizing not only for reconstruction quality but also directly for segmentation performance. The combined loss function  $\mathcal{L}$  consists of three terms. The common rate-distortion trade-off balances the distortion term  $\mathcal{D}_{\text{MSE}}$  against the estimated bit-rate  $\mathcal{R}$ . For end-to-end optimization, we add the task-specific BCE segmentation loss  $\mathcal{T}_{\text{BCE}}$  as a third weighted term. The model is trained using separate Adam optimizers (Kingma and Ba [2014]) for each of the three components (auto-encoder, entropy bottleneck and segmentation model), as the entropy bottleneck needs to be optimized using a lower learning rate. The task-specific weight  $\gamma$  is set to zero for standalone compression experiments. An overview of the experiment architecture is shown in Figure 1b. Further details on the used hyperparameters are given in Appendix A.

$$\mathcal{L} = \lambda \cdot \mathcal{D}_{\text{MSE}} + \mathcal{R} + \gamma \cdot \mathcal{T}_{\text{BCE}} \quad (1)$$

### 3 Results

**Segmentation Baseline Performance** The best performing baseline segmentation models are summarized in Table 1. The F1+ score denotes the F1 score of the positive class. It is reported alongside its macro F1 given the class imbalance of fire and building pixels in their respective datasets. Generally, the UNet architecture is able to provide a reasonable performing baseline for all three tasks. However, it is not able to resolve very fine detail contours in the building footprints, hence the comparatively low F1+ result in  $\mathcal{D}_{\text{building}}$ .

**Segmentation under Bit Rate Reduction** Figure 2 shows the impact of bit rate reduction on segmentation performance (F1), when training on the decompressed image data. This is done to directly quantify the effect of compression on our selection of downstream tasks. For all task, we observe that a certain level of image quality is needed before reaching a plateau in segmentation performance. This is in agreement with existing work, as many segmentation tasks in EO do not seem to require the full fidelity of the used observation data Garcia-Sobrino et al. [2020].

Table 1: Results of segmentation baseline models.

Task	Backbone	F1+ ↑	F1 ↑
$\mathcal{D}_{\text{fire}}$	MobileNetV2	0.665	0.830
$\mathcal{D}_{\text{cloud}}$	ResNet34	0.792	0.857
$\mathcal{D}_{\text{building}}$	ResNet34	0.448	0.720

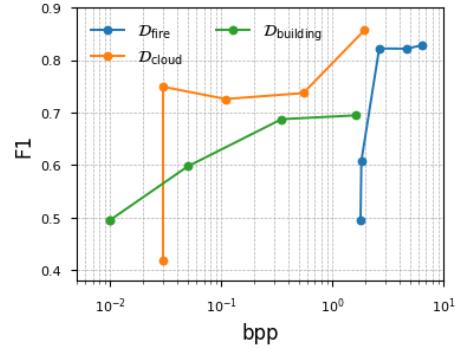


Figure 2: Segmentation performance under bit rate reduction using JPEG 2000.

**Standalone Compression Performance** Prior to end-to-end optimization, we evaluate the standalone performance of the different lossy compression methodologies on the given set of tasks by comparing the rate-distortion tradeoff. This is commonly measured by the reconstruction (PSNR) and its bitrate per pixel (bpp).

- **Single-Channel Tasks** Figure 3a shows the results on the two single-channel tasks  $\mathcal{D}_{\text{fire}}$  and  $\mathcal{D}_{\text{cloud}}$ . JPEG2000 remains superior in the monochromatic tasks, despite testing multiple learning architectures including FullyFactorizedPrior, ScaleHyperPrior (Ballé et al. [2018]) and the Chenge2020Anchor (Cheng et al. [2020]). We accredit this result due to the lack of redundant spectral information that is usually present in multi-spectral datasets. Therefore, the algorithms can only leverage the spatial context to find meaningful representations to reduce the effective entropy of the data.
- **Multi-channel Task** Figure 3b summarizes the performance on the multi-channel task  $\mathcal{D}_{\text{building}}$ . Here, the learned compression algorithm clearly outperforms the JPEG baseline both, when trained from scratch or fine-tuned. Additionally, the pretrained, not fine-tuned model variants from the CompressAI model zoo are shown in comparison (Bégaint et al. [2020]).

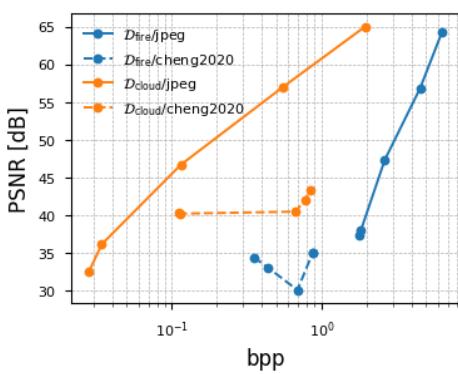


Figure 3a: Results on single-channel tasks comparing JPEG2000 against the learned compression algorithm. All models were trained from scratch.

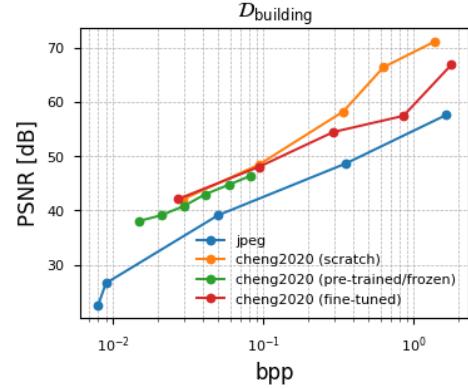


Figure 3b: Results on the multi-channel task including both training from scratch and different transfer learning strategies.

**Task-specific Compression Performance** During our experiments, we were unable to demonstrate the benefits of joint optimization of the compression and segmentation components compared to individual optimization of the components. Details of the experiments are given in Appendix B.

## 4 Conclusion

**Channel-dependent Efficacy** We observe that learned compression models substantially outperform classical codecs, such as JPEG 2000, particularly when dealing with abundant, multi-channel optical imagery datasets. The building footprint extraction task exemplifies this scenario, demonstrating clear advantages in both reconstruction quality (PSNR) and downstream segmentation accuracy. However, classical codecs remain competitive and effective for tasks involving smaller datasets or single-channel thermal imagery, such as cloud and fire detection, due to limitations in available training data and the architectural constraints of current learned compression models. This result also stood when compression and segmentation were optimized end-to-end.

**Implications for Downstream Tasks of EO Missions** Primarily, learned image compression methodologies work best for multi-spectral datasets with a high number of spectral features. When designing compression methodologies for datasets with low spectral count (less than three channels) we recommend starting with traditional compression codecs first. Secondly, many tasks in EO do not need the full fidelity of raw observations. Depending on the intended set of tasks lossy compression is a viable option to drastically reduce downlink overhead, while maintaining acceptable performance. Still there is a potential decrease in segmentation quality which must be weighed against the lower transfer volumes.

**Future Directions** Future work will aim to extend learned compression architectures specifically for single-channel EO imagery. Current methods exhibit significant limitations when applied to datasets such as thermal infrared imagery, which hinders their practical deployment. To address these challenges, it will be essential to collect larger, well-annotated datasets and to explore compression techniques that more effectively capture the spatial correlations inherent in the data.

## 5 Impact Statement

This paper addresses learned compression algorithms and their impact on downstream applications. This is of interest to any edge-based imaging system in EO but also other domains. Probing new emerging algorithmic solutions have to potential to significantly ease access to larger remote sensing constellations in space as they can operate under stricter mass, power and budget constraints. However, many of these technologies are of dual use. Therefore, potential advances might also be used by the defense & intelligence community, which lie outside of the author's governance.

## References

- Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior, 2018. URL <https://arxiv.org/abs/1802.01436>.
- Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.
- Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules, 2020. URL <https://arxiv.org/abs/2001.01568>.
- Joaquin Garcia-Sobrino, Armando J. Pinho, and Joan Serra-Sagrista. Competitive segmentation performance on near-lossless and lossy compressed remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 17(5):834–838, May 2020. ISSN 1558-0571. doi: 10.1109/LGRS.2019.2934997. URL <http://dx.doi.org/10.1109/LGRS.2019.2934997>.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Jonathan Prexl and Michael Schmitt. The potential of sentinel-2 data for global building footprint mapping with high temporal resolution. In *2023 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4, 2023. doi: 10.1109/JURSE57346.2023.10144166. URL <https://ieeexplore.ieee.org/document/10144166>. ISSN: 2642-9535.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, pages 234–241. Springer, 2015.
- Matthias Rötzer, Lukas Liesenhoff, Max Bereczky, Martin Ickerott, Jayendra Chorapalli, and Julia Gottfriesen. Self-supervised learning for fire segmentation in forest-2 images. In *ESA-NASA International Workshop on AI Foundation Model for EO*, Frascati, Italy, May 2025. ESA-ESRIN.
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- D.S. Taubman and M.W. Marcellin. Jpeg2000: standard for interactive imaging. *Proceedings of the IEEE*, 90(8):1336–1357, 2002. doi: 10.1109/JPROC.2002.800725.
- Niklas Wölki, Lukas Kondmann, Christian Mollière, Martin Langer, Martin Werner, and Julia Gottfriesen. Exploring machine learning for cloud segmentation in thermal satellite images of the forest-2 mission. In *Proceedings of SPAICE2024: The First Joint European Space Agency/IAA Conference on AI in and for Space*, pages 362–366, 2024.

## A Hyperparameter Setup of Experiments

**Input Normalization.** For each experiment we normalized the dataset by dividing by the global maximum of the input data to yield a distribution between [0,1].

**Hyperparameter Tuning.** We conducted a systematic hyperparameter search to find the optimal configuration for the Mixed Gaussian Likelihood model. The search space explored the following key hyperparameters using grid search.

- **Quality Level ( $q$ ):** The CompressAI library offers discrete quality levels from 1 to 6. These levels primarily control the network’s capacity by setting the dimensionality of the latent space. We observed that levels  $q \in \{1, 2, 3\}$  corresponded to a latent dimension of 128, while levels  $q \in \{4, 5, 6\}$  used a dimension of 192. To represent these two main configurations, we focused our final experiments on levels 3 and 6.
- **Batch Size:** We explored batch sizes in the set {4, 8, 16, 32}.
- **Learning Rates ( $\eta$ ):** Separate learning rates were tuned for the main network ( $\eta$ ) and the auxiliary hyperprior network ( $\eta_{aux}$ ), both within the range of  $[10^{-6}, 10^{-1}]$ .
- **Rate-Distortion Weight ( $\lambda$ ):** This crucial hyperparameter balances the trade-off between the rate and distortion terms in the loss function. We explored values in the range  $[10^{-4}, 10^3]$ .

**Loss Function and Optimizer.** The training objective for the Mixed Gaussian Likelihood model is the rate-distortion loss function described in Section 2. This function combines a rate term, encouraging compact representations, with a distortion term, enforcing reconstruction accuracy. We employed the Adam optimizer to train the compression and segmentation models for all experiments (Kingma and Ba [2014]).

**Computing Resources** All training has been done using a single NVIDIA RTX 4000. Training times were usually in the range from hours to days.

## B Results of Joint Optimization

We focus on  $\mathcal{D}_{building}$  for the evaluation of end to end optimization, given the apparent unsuitability of monochromatic tasks for learned compression algorithms. The results prior and post optimization are summarized in Table 2. In summary, end-to-end optimization is able to converge both the compression and the segmentation components when starting from low quality models. However, the final performance does not exceed standalone optimization of the individual components.

The used loss weights for these results are ( $\lambda = 10.0, \gamma = 1e - 3$ ) for the high-quality start and ( $\lambda = 10.0, \gamma = 1e - 2$ ) for the low-quality start respectively. The start scenarios are initialized using pretrained components for both, compression and segmentation. In the high-quality start scenario, the compression model has been optimized for quality, whereas the low-quality scenario uses a very lossy compression.

Table 2: Performance of experiments prior and post to end-to-end optimization.

Scenario	bpp <sub>prior</sub>	PSNR <sub>prior</sub> [dB] $\uparrow$	F1 <sub>prior</sub> $\uparrow$	bpp	PSNR [dB] $\uparrow$	F1 $\uparrow$
high-quality start	0.6121	66.29	0.7281	0.6106	66.64	0.7238
low-quality start	8e-5	25.76	0.4925	0.4407	<b>49.80</b>	<b>0.7090</b>

Generally, we explored a range of different hyperparameters for the joint optimisation.

- **Rate-Distortion Weight ( $\lambda$ ):** {0.1, 1, 10, 100}.
- **Task-specific Weight ( $\gamma$ ):** {0.001, 0.01, 0.05, 0.1, 1}.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract only states facts that are either referenced in the paper or directly demonstrated by results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss limitations and future directions in Section 4.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not contain any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The proprietary datasets are currently not public. Similar results could be reproduced using data from publicly available missions carrying thermal bands.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: As stated in previous answer. However, we are working on open sourcing a dataset in a coming publication. The model components are publicly available through the compressAI framework.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The used hyperparameters are disclosed in Section 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The resulting performance of our experiments was largely deterministic given by the used hyperparameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computing resources are stated in A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We could not identify harm through our research and explicitly stated the expected impact in Section 5.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, all used frameworks, models and datasets are cited accordingly.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This publication does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.