



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

The Effect of Lossy Compression on 3D Medical Images Segmentation with Deep Learning

Anvar Kurmukov⁴, Bogdan Zavolovich¹, Aleksandra Dalechina¹, Vladislav Proskurov⁴, and Boris Shirokikh^{2,3}

¹ National Research Nuclear University MEPhI

² Skolkovo Institute of Science and Technology, Moscow, Russia

³ Artificial Intelligence Research Institute (AIRI), Moscow, Russia

⁴ Independent researcher

Abstract. Image compression is a critical tool in decreasing the cost of storage and improving the speed of transmission over the internet. While deep learning applications for natural images widely adopts the usage of lossy compression techniques, it is not widespread for 3D medical images. Using three CT datasets (17 tasks) and one MRI dataset (3 tasks) we demonstrate that lossy compression up to $\times 20$ times have no negative impact on segmentation quality with deep neural networks (DNN). In addition, we demonstrate the ability of DNN models trained on compressed data to predict on uncompressed data and vice versa with no quality deterioration.

Keywords: Lossy compression · Medical imaging

1 Introduction

The advent of digital technology has revolutionized medical imaging, leading to an exponential increase in the volume of clinical data generated and stored globally. As the digital archives of medical images continue to expand, managing the storage requirements of this quantity of data has become a pressing challenge. Conventional approaches to addressing this problem have included the use of lossless compression techniques, such as JPEG and JPEG2000, which have been employed in an attempt to mitigate the issue of disk space scarcity [4].

A more memory-efficient approach is to use lossy compression regimes; however, their application is not without consequences. Particularly in the context of deep learning (DL), where data management and loading efficiency are critical to model performance and GPU utilization, authors have demonstrated models' performance deterioration when trained on heavily compressed images [5,1].

Recent studies, predominantly in non-medical domains, have investigated the impact of lossy image compression on the performance of neural networks, revealing a nuanced relationship between compression and model efficacy. Research by [2,11] on object detection with convolutional neural networks (CNNs) and subsequent inquiries into applications ranging from steel surface detection to

agricultural image segmentation have illustrated both the potential advantages and drawbacks of incorporating lossy compression into data preprocessing workflows. These discoveries emphasize the importance of balancing data efficiency with the preservation of model accuracy and reliability.

Despite the growing body of research in various domains, the exploration of the impact of image compression on medical imaging applications, particularly in the realm of deep learning-based analysis, remains relatively uncharted. Previous studies have primarily focused on the direct effects of compression on image quality rather than its implications for algorithmic performance [13]. However, emerging evidence from the analysis of histopathological images for metastatic cancer [6] detection and mammogram classification [10] suggests that, under certain conditions, models trained on compressed images may not only maintain but potentially enhance their performance on compressed data, challenging preconceived notions about the incompatibility of lossy compression and high-stakes medical image analysis.

In our work, we examine how lossy compression affects DL-based segmentation models on 20 segmentation tasks on 3D computed tomography (CT) and magnetic resonance imaging (MRI) data. Our key contributions are three-fold:

1. Through systematic experiments on four medical image datasets, we demonstrate that storage memory footprint could be reduced at training time up to 20 times via lossy compression with no negative effect on segmentation quality.
2. We illustrate that DNN models based on Unet architecture trained on uncompressed data do not deteriorate while inferenced on compressed data, and vice versa.
3. We show that while trained experts could accurately differentiate between lossy compressed images after a compression rate of 10 times, DL models still preserve their quality for images compressed 20 times.

2 Methods

2.1 JPEG2000

The Joint Photographic Experts Group (JPEG) developed the JPEG 2000 image compression standard, which employs discrete wavelet transformation. This approach ensures the preservation of high image quality during the compression process. A key feature of JPEG 2000 is its support for progressive loading, enabling the stepwise display of image details as more data is received. Furthermore, JPEG 2000 provides the flexibility to either maintain all image details with lossless compression or manage the level of lossy compression according to specific needs. This functionality has been incorporated into the DICOM standard as Supplement 61, substantially enhancing the efficiency of medical image processing and storage. Despite high compression ratio, JPEG 2000 retains critical image details when required, demonstrating an optimal balance between storage efficiency and image quality [4].

2.2 Data

We used four publicly available datasets, covering two of the most popular 3D medical imaging modalities: computed tomography (CT) and magnetic resonance imaging (MRI), and two types of segmentation tasks: anatomical (15 organs and one anatomical structure) and pathological region segmentation (brain tumors and liver tumors).

LiTS We utilized 131 training CT images with liver tumor segmentation masks from the LiTS dataset [3]. We cropped all images to a given liver mask; the rest of the preprocessing was done in nnunet.

BraTS We employed 1251 training skull-stripped MRI cases from the BraTS21 dataset [14]. Each training case includes native (T1) and post-contrast T1-weighted (T1Gd) images, T2-weighted and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. The BraTS21 dataset contains annotations of the necrotic tumor core, the peritumoral invaded tissue, and GD-enhancing tumor.

AMOS We utilized 300 CT images from the training part of the AMOS dataset [9]. We performed a multi-class segmentation task, predicting 15 abdominal organ masks: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus.

BGPD We employed the Burdenko Glioblastoma Progression Dataset (BGPD) [15], a systematically curated collection of patients diagnosed with primary glioblastoma. For our experiments, we used 176 CT topometric images, predicting segmentation masks of the brain stem.

2.3 Experiments

We selected nnU-Net [8] for our experiments. It is an open-source framework⁵, widely adopted as a strong benchmark for medical image segmentation tasks. Low-resolution (lowres) architecture was used for primary results, fullres (default config) for ablation study. In addition, for ablation study on AMOS dataset, we used a Swin Unetr (base)⁶ model pretrained on BTCV dataset [12].

Lowres architecture has been trained for 300 epochs with an initial learning rate of 10^{-2} reduced to 10^{-4} at the last epoch. Each epoch consists of 250 iterations. Stochastic gradient descent with Nesterov momentum (0.99) is used for training. The loss function is a combination of Dice loss and cross-entropy. Poly learning rate scheduler changes the learning rate during training $LR_{new} = LR_{initial}(1 - epoch/epoch_{max})^{0.9}$. The training of 3 folds takes about 7.5 hours on a 32G NVIDIA Tesla V100.

⁵<https://github.com/MIC-DKFZ/nnUNet>

⁶<https://huggingface.co/darragh/swinunetr-btcv-base>

2.4 Metrics

To assess the quality of the segmentation model, we used the Dice coefficient:

$$\text{Dice}(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (1)$$

where X is the predicted binary segmentation, and Y is the ground truth. The Dice coefficient for multi-class segmentation is then obtained by averaging the Dice scores across all segmentation classes.

We used the Peak Signal-to-Noise Ratio (PSNR) to assess the quality of the compressed images. A higher value indicates a closer match between the original and compressed images. Given the original image I_1 and the compressed image I_2 of size $M \times N$, PSNR is calculated based on the mean squared error (MSE):

$$\text{MSE}(I_1, I_2) = \frac{\sum_{M,N} (I_1(m, n) - I_2(m, n))^2}{M \times N} \quad (2)$$

The PSNR is then defined as:

$$\text{PSNR}(I_1, I_2) = 10 \cdot \log_{10} \left(\frac{r^2}{\text{MSE}(I_1, I_2)} \right), \quad (3)$$

where r is the maximum fluctuation in the input image data type. We set r to 4096 for CT data and to $\max(I) - \min(I)$ for MRI data (measuring the minimum and maximum values over the entire BraTS dataset).

3 Results

First, we assess how lossy compression influences images quality using human evaluation and PSNR. Second, we test the effect on segmentation quality of training a Unet architecture on lossy compressed images. Third, we test how models trained on uncompressed images performs on lossy compressed images. Finally, we demonstrate that in a scenario where storage capacity is limited, compression is a better choice than data subsampling.

3.1 Compressed images quality

To assess how well radiologists differentiate between images of varying compression rates we generated pairs of images with different compression rates: 1 vs 2, 1 vs 5, ..., 20 vs 50. For each pair of compression rates (15 in total), we generated 10 random images for each dataset (600 in total). 2D axial slices with lower compression rates were randomly placed on the left or right side of the image. Each axial slice in a pair was generated from the same uncompressed 2D slice. Two radiologists (5 and 8 years of experience) independently assessed image pairs and provided a decision on which image (left or right) was of better quality or if they could not decide which one was better. We next compare their results with the ground truth using Spearman's Rank correlation (1 if radiologist chosen correct

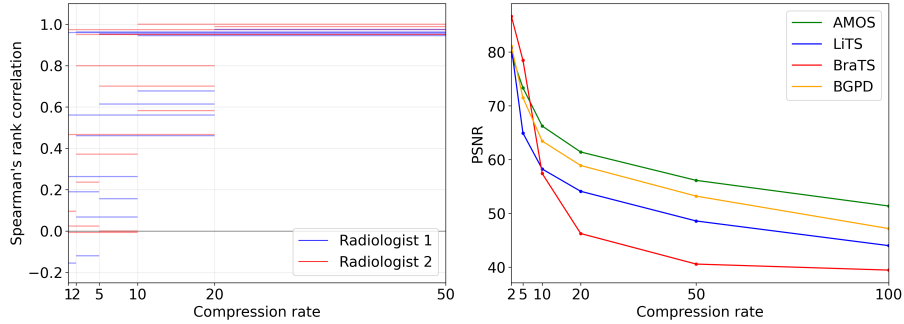


Fig. 1: Left: results of human assessment of image pairs with different compression ratios. Each line represents the average Spearman’s rank correlation (x-axis) between two compression rates (two ends of each line, y-axis). Right: peak signal-to-noise ratio of images compressed at different rates.

image of higher quality, 0 if she can not decide, -1 if image of lower quality is selected as image of higher quality). See example images in the Supplementary materials.

First, we observe that both radiologists are unable to distinguish between uncompressed images and images with compression rates up to 5. Their performance steadily improves at compression rates 10 and 20, achieving almost perfect quality for compression rate 50 Figure 1 (left). Figure 1 (right) shows the PSNRs achieved via JPEG2000 lossy compression at different compression rates. We observe a significant drop in PSNR at rate 10, which then steadily declines. This result intuitively aligns with the outcomes from human evaluation.

3.2 Effect of training on lossy compressed images

To investigate the influence of image compression on the downstream segmentation task, we trained the Unet segmentation architecture on images compressed at different rates and on uncompressed original images (separate models) Figure 2 (left). For all 4 datasets, we observe stable segmentation quality up to compression rates 20, with rapid deterioration of the quality at higher compression rates. This finding provides strong evidence that the Unet architecture with a standard training pipeline is robust to artifacts introduced by the JPEG2000 compression algorithm. Table 1 presents the by-class segmentation performance on the AMOS dataset. We observe no segmentation drop, even for smaller organs like adrenal glands and the gallbladder.

3.3 Inference models trained on original data

Next, we examine how models trained on images of higher compression rates perform on images of lower compression rates, as shown in Figure 3. On the

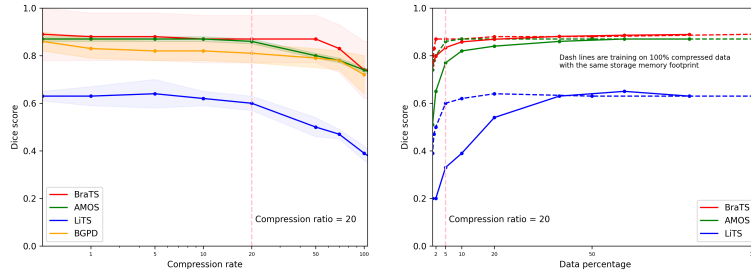


Fig. 2: Left: segmentation performance of models trained on data compressed at different rates. Right: storage memory footprint tradeoff, few-shot (uncompressed) vs full-shot (compressed).

AMOS and LiTS datasets, we observe no performance drop for compression rates up to 20. The selection of compression rate 20 is due to the fact that beyond this rate, models' performance decreases compared to models trained on uncompressed data, as discussed in Section 3.2. This scenario is important if one wants to use models already trained on compressed data on uncompressed data, potentially saving time on compression itself.

Furthermore, we investigated how models trained on uncompressed data perform on data with various compression rates. For this experiment, we utilized the Unet (trained via the full resolution nn-Unet pipeline) and publicly available weights for the SWIN Unetr model (trained on the BTCV dataset). The results are presented in Table 2. Once again, we observe no deterioration in segmentation quality for either model. The significance of this result lies in the fact that researchers already have a plethora of trained models (both public and private), typically on uncompressed data. Our findings suggest that adopting an "inference on compressed data strategy" does not introduce any additional costs. Despite relatively modest (by an absolute value) segmentation performance of pretrained SWIN Unetr⁷, the purpose of this experiment was to demonstrate that JPEG compression has almost no effect on the performance of publicly available network trained on an uncompressed data.

3.4 Few-shot vs lossy compressed full training

Finally, we examine the memory footprint tradeoff of training on compressed images. We trained nn-Unet in a few-shot regime, using 1%, 5%, 10%, 20%, 50%, and 100% of the data. We then compared the segmentation quality of the resulting algorithms using the same holdout set. Figure 2 (right) illustrates the Dice score of networks trained in a few-shot regime. Dashed lines visualize the segmentation quality of the networks trained on 100% of the data at different

⁷Likely due to the fact it was trained on a BTCV dataset, using only 24 training samples.

Table 1: Per-organ segmentation quality for different compression rates on AMOS. Numbers are mean Dice score averaged over 3-fold cross-validation.

compression rate	1	2	5	10	20	50	70	100
SPL	0.96	0.96	0.96	0.96	0.96	0.94	0.92	0.90
RKI	0.84	0.84	0.84	0.84	0.83	0.76	0.72	0.67
LKI	0.76	0.76	0.76	0.76	0.73	0.63	0.59	0.54
GBL	0.76	0.76	0.76	0.76	0.75	0.62	0.57	0.51
ESO	0.80	0.80	0.80	0.80	0.78	0.70	0.67	0.62
LIV	0.87	0.87	0.88	0.87	0.85	0.79	0.76	0.73
STO	0.82	0.81	0.81	0.82	0.80	0.75	0.73	0.70
AOR	0.96	0.96	0.96	0.96	0.95	0.93	0.91	0.89
IVC	0.95	0.95	0.95	0.95	0.95	0.92	0.91	0.89
PAN	0.82	0.81	0.82	0.82	0.79	0.66	0.61	0.52
RAD	0.84	0.84	0.84	0.84	0.83	0.78	0.75	0.72
LAD	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.94
DUO	0.90	0.90	0.90	0.90	0.89	0.85	0.83	0.80
BLA	0.95	0.95	0.95	0.95	0.95	0.93	0.92	0.90
PRO/UTE	0.90	0.90	0.90	0.90	0.89	0.85	0.83	0.80
average	0.87	0.87	0.87	0.87	0.86	0.80	0.78	0.74

Table 2: Models trained on uncompressed data tested on compressed data. Results on AMOS dataset. nn-Unet results are averaged over three cross validation folds. Swin Unetr results are averaged over all AMOS dataset.

compression rate	2	5	10	20	50
nnUnet (high resolution)	0.971	0.971	0.970	0.969	0.946
Swin Unetr [7]	0.614	0.614	0.614	0.614	0.603

compression rates with the same memory footprint. For example, training a network on all training data compressed 20 times is equivalent (in terms of storage memory) to training on 5% of uncompressed data.

The results on all 4 datasets suggest that at a compression rate of 20, training on 100% compressed data is superior compared to training the network on a fraction of uncompressed data with the same storage memory footprint.

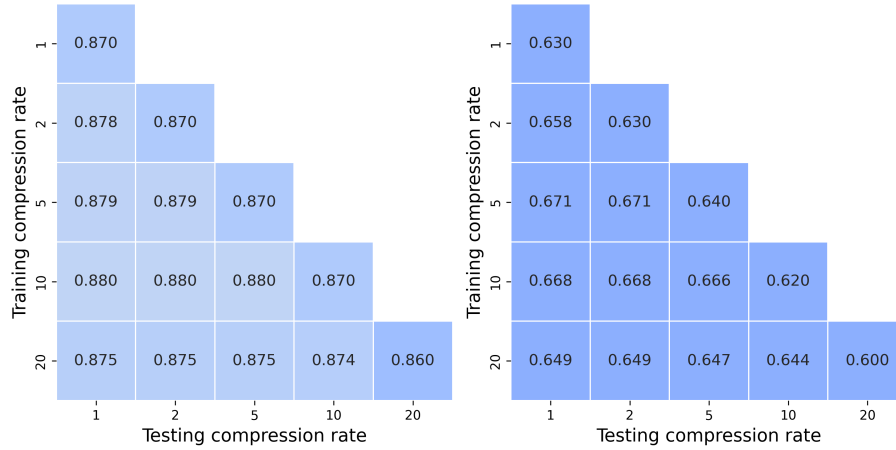


Fig. 3: Testing model trained on images at different compression rates on images of other compression rates. Results for AMOS dataset (left), LiTS dataset (right).

4 Conclusion

In this study, we conducted a comprehensive investigation into the impact of lossy image compression on the performance of deep learning-based segmentation models in the context of 3D medical imaging. Our findings demonstrate that significant reductions in storage memory footprint can be achieved through the use of lossy compression without compromising the segmentation quality of the DL models. We hope that our work will encourage medical imaging researchers, working with 3D images to start integrating lossy compression in their pipelines, without the unreasonable fear of (segmentation) quality loss.

4.1 Study limitations

First, we solely tested the effect of lossy compression using JPEG2000. While we believe this choice is justified by the fact that this is the only lossy compression algorithm currently supported by DICOM, investigating the effect of other algorithms might be interesting. Second, we only focus our experiments on a single

neural network architecture Unet, with a limited analysis of a publicly available pretrained Swin Unetr. Finally, human evaluation experiment only tested the ability of radiologists to differentiate between images of varying compression, but not their performance at segmenting images after compression.

References

1. Behme, L., Thirumuruganathan, S., Mahdiraji, A.R., Quiané-Ruiz, J.A., Markl, V.: The art of losing to win: Using lossy image compression to improve data loading in deep learning pipelines. In: 2023 IEEE 39th International Conference on Data Engineering (ICDE). pp. 936–949 (2023). <https://doi.org/10.1109/ICDE55515.2023.00077>
2. Benbarrad, T., Eloutouate, L., Arioua, M., Elouaai, F., Laanaoui, M.D.: Impact of image compression on the performance of steel surface defect classification with a cnn. *Journal of Sensor and Actuator Networks* **10**(4), 73 (2021)
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). *Medical Image Analysis* **84**, 102680 (2023)
4. Foos, D.H., Muka, E., Slone, R.M., Erickson, B.J., Flynn, M.J., Clunie, D.A., Hildebrand, L., Kohm, K.S., Young, S.S.: Jpeg 2000 compression of medical imagery. In: *Medical Imaging 2000: PACS Design and Evaluation: Engineering and Clinical Issues*. vol. 3980, pp. 85–96. SPIE (2000)
5. Gandor, T., Nalepa, J.: First gradually, then suddenly: understanding the impact of image compression on object detection using deep learning. *Sensors* **22**(3), 1104 (2022)
6. Ghazvinian Zanjani, F., Zinger, S., Piepers, B., Mahmoudpour, S., Schelkens, P., de With, P.H.: Impact of jpeg 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images. *Journal of Medical Imaging* **6**(2), 027501–027501 (2019)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266* (2022)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
9. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems* **35**, 36722–36732 (2022)
10. Jo, Y.Y., Choi, Y.S., Park, H.W., Lee, J.H., Jung, H., Kim, H.E., Ko, K., Lee, C.W., Cha, H.S., Hwangbo, Y.: Impact of image compression on deep learning-based mammogram classification. *Scientific Reports* **11**(1), 7924 (2021)
11. Júnior, J.D.D., Ribeiro, J.B., Backes, A.R.: Assessing the impact of jpeg compression on the semantic segmentation of agricultural images. *Signal, Image and Video Processing* pp. 1–7 (2023)
12. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In: *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. vol. 5, p. 12 (2015)

13. Liu, F., Hernandez-Cabronero, M., Sanchez, V., Marcellin, M.W., Bilgin, A.: The current role of image compression standards in medical imaging. *Information* **8**(4), 131 (2017)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**(10), 1993–2024 (2014)
15. Zolotova, S.V., Golanov, A.V., Pronin, I.N., Dalechina, A.V., Nikolaeva, A.A., Belyashova, A.S., Usachev, D.Y., Kondrateva, E.A., Druzhinina, P.V., Shirokikh, B.N., Saparov, T.N., Belyaev, M.G., Kurmukov, A.I.: Burdenko’s glioblastoma progression dataset (burdenko-gbm-progression) (2023). <https://doi.org/10.7937/E1QP-D183>, <https://www.cancerimagingarchive.net/collection/burdenko-gbm-progression/>