BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Compression of Large Volumetric Datasets Using Deep Learning for Streaming and Interactive Rendering

Jan Svoboda

Abstract

This article focuses on design and implementation of a solution for lossy compression of large volumetric datasets using deep learning. The solution uses an implicit neural representation with input encoding through multi-resolution hash grids to compress volumetric data divided into blocks. Moreover, the proposed data structure allows storing compressed blocks across multiple compression levels simultaneously. The block representation reduces computational hardware requirements and, together with multiple compression levels, also enables the solution to be used for streaming volumetric data. The solution also contains automatic selection of compression parameters based on the user's desired output quality, using a proposed metric that describes the complexity of the input data. Furthermore, this work focuses on integrating the developed compression solution into an existing system for streaming and displaying volumetric data, designed by the author in his previous work, and extending it with realistic GPU rendering. The resulting solution demonstrated high efficiency of the proposed compression while maintaining high fidelity of representation. The high potential of the proposed compression for streaming volumetric datasets was also demonstrated, as the developed solution enables realistic rendering of large datasets stored on a server with significantly lower network load compared to streaming without compression.

jan.svoboda16@seznam.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

With ongoing technological progress, the amount of volumetric data generated by various acquisition methods is steadily increasing. Higher sensor resolutions, faster acquisition speeds, greater memory capacities, and more powerful computing resources all contribute to this trend. As a result, very large datasets, often hundreds of gigabytes in size, are now common. However, performing computationally intensive tasks on these datasets, such as real-time 3D rendering, often exceeds the capabilities of even high-end desktop computers. Managing and processing such larg datasets also remains a major challenge.

In earlier work [1], we presented a system for streaming large volumetric datasets from server storage in blocks of multiple levels of detail, allowing real-time visualization using CPU Ray Casting. This made it possible to display very large datasets even on relatively modest hardware. However, the main limitation of this solution is the high network bandwidth required, which makes broader deployment difficult. While lossy compression offers a possible way to reduce the network load, it must be applied carefully to avoid significantly degrading the quality of important details.

Advances in machine learning have led to the development of models for implicit neural representation — a technique that learns continuous approximations of data, such as images or volumetric datasets, using neural networks. Müller et al. [2] introduced the idea of encoding neural network inputs using a multi-resolution hash grid, achieving very good results. Wu et al. [3] later extended this concept to interactive volumetric data visualization using direct network inference and volumetric data compression. However, their approach treats the data as a single entity, making it unsuitable for streaming, and impractical for very large datasets, where the resulting model can easily exceed the available accelerator memory.

In this work, we propose a method for volumetric data compression based on implicit neural representations and multi-resolution hash grids, designed specifically

with streaming in mind. Our approach compresses the data block-by-block and applying several levels of compression simultaneously. This allows the use of smaller models, reduces training time, and supports efficient streaming. We also present an automatic mechanism for selecting model parameters based on user-defined requirements. The solution has been integrated into an existing system for streaming volumetric data, which we have extended with realistic GPU rendering.

Testing shows that our method can achieve high compression ratios (up to 30:1 depending on the dataset) while maintaining very good image quality. The results also demonstrate that the solution enables realistic rendering of large server-stored datasets with significantly lower network load compared to uncompressed streaming.

2. Compression

A library for volumetric data compression was developed in *C++*, based on the *Tiny Cuda NN* library and several supporting libraries. First, the dataset is partitioned into blocks, and for each block, an implicit neural representation model is trained, using a multi-resolution hash grid to encode the input parameters. The principle of this model is illustrated in **Fig. 2**, taken from [2]. The compressed blocks (the trained models) are then stored in an output directory structure together with metadata describing the entire dataset. The resulting structure supports multiple compression levels per block, which is particularly advantageous for streaming applications.

Our goal was also to design an automatic system for predicting model parameters based on the input data and the desired compression quality. The solution allows users to specify the compression quality through four *SSIM* values at different levels of detail. This composite metric helps reduce the impact of noise on the compression ratio. To assess the complexity of the input data, we propose a new metric based on analyzing changes in the gradient direction of the image. The final prediction of model parameters is then performed using a neural network.

An example of the Stamina dataset detail compressed to 31:1 compression ratio is visible in the **Fig. 5**.

3. Streaming

The developed compression library was integrated into an existing system we had previously designed for streaming and displaying volumetric data. This system consists of a server repository built using .NET technology and a client application developed in C++.

Data are streamed from the server in blocks and decompressed as needed for the current display. The priority of the retrieved blocks is determined by the proposed heuristic. The overall architecture of the system is illustrated in **Fig. 1**.

The effectiveness of compression during streaming is well shown in the graphs **Fig. 3** and **Fig. 4**.

4. Visualization

The original solution for streaming volumetric data uses only CPU rendering, which is insufficient for more advanced imaging techniques. Since image quality was crucial for comparing volumetric data across different compression levels, a realistic renderer was designed and implemented using *CUDA*. This renderer enables real-time display of streamed objects, including advanced material properties and local illumination calculations and CUDA provides code flexibility.

The quality of the display can be seen in **Fig. 5**, **Fig. 6**, **Fig. 7** and **Fig. 8**. The *CTLiver* dataset was obtained from the public repository of the 3D Slicer tool [4]. The *Chameleon* dataset was obtained from [5]. The *Stamina* dataset was provided by *Tescan 3DIM*, and the *Chaetoblast* dataset was obtained from the public repository of *Empiar* [6].

Acknowledgements

I would like to thank doc. Ing. Michal Španiel, Ph.D. for excellent guidance of this work and for providing valuable expert information. I would also like to thank all the employees of Tescan 3DIM who provided me with valuable information.

References

[1] Jan Svoboda. Zobrazení rozsáhlých volumetrických dat na cpu. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, 2023.

[2] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics*, 41(4):1–15, 2022.

[3] Qi Wu, David Bauer, Michael J. Doyle, and Kwan-Liu Ma. Interactive volume visualization via multi-resolution hash encoding based neural representation, 2023.

[4] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe

Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V. Miller, Steve Pieper, and Ron Kikinis. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012. Quantitative Imaging in Cancer.

[5] Pavol Klacansky. Open-scivis-datasets, 2025.

[6] Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 1 2023.