Improved RAHT-based Compression of 3D Gaussian Splats

Annalisa Gallina

Department of Information Engineering, University of Padova, Italy

3D Gaussian Splatting (3DGS) has recently emerged as a powerful representation for Novel View Synthesis but it comes with significant storage and bandwidth requirements, which limit its usability in practical applications. In this work, we tackle the problem of compressing Gaussian Splatting by interpreting it as a Point Cloud with rich per-point attributes, and we exploit the Region-Adaptive Hierarchical Transform (RAHT) — a lightweight and well-established tool in Point Cloud compression standards — to decorrelate and compact the attribute information.

While RAHT has been successfully applied in Point Cloud compression, previous attempts to directly integrate it into Gaussian Splatting pipelines have shown limited efficiency due to the redundant and unstructured nature of Gaussian attributes, which reduces the overall benefit of the transform stage. To overcome these limitations, we introduce two key contributions.

First, we propose a quantization-aware fine-tuning strategy that adapts the Gaussian parameters to the transform domain, adjusting them in a way that compensates for quantization noise. We also introduced a sparsity-promoting loss term, based on the ℓ_1 norm of the transform coefficients, that encourages the model to concentrate energy in fewer coefficients. This allows the representation to better align with the compression process, facilitating entropy coding and improving rate-distortion performance.

Second, we design a systematic bit allocation strategy across attribute channels, inspired by quantization matrices in image compression (e.g., JPEG). By allocating bits according to the visual importance and energy distribution of each attribute, our method improves rate–distortion trade-offs without requiring changes to the baseline representation. In addition, the use of a fixed, scene-independent allocation scheme simplifies the encoder design and eliminates the need for transmitting additional side metadata.

Overall, our framework maintains full compatibility with standard Point Cloud coding tools, like Geometry based Point Cloud Compression (GPCC), while reducing storage requirements. Our experimental evaluation on both the Mip-NeRF360, Tanks and Temples and NeRF Synthetic datasets demonstrates that the proposed method, despite its simplicity, achieves state-of-the-art performance among post-training compression approaches, and in several cases, it matches or outperforms more complex methods. While some recent neural methods may yield higher absolute performance, our solution achieves a strong balance between performance, simplicity, and compatibility with existing standards. This shows that efficient compression can be achieved without altering the underlying representation, allowing practical deployment of Gaussian Splatting in real-world applications.