SoftMax-NeRF Attribute Compression with Vector-Quantized Color Tables for Point Clouds

Zhu Li*, Ning Xu†
*University of Missouri–Kansas City (UMKC)
Email: lizhu@umkc.edu
†Adeia Inc.
Email: ning.xu@adeia.com

***Abstract*—Volumetric point clouds enable immersive communication and telepresence but remain challenging to store and transmit due to the massive size of geometry and color attributes. This paper presents a learning-based attribute codec that combines a vector-quantized (VQ) color table with a compact neural radiance field (NeRF) whose output layer is a K-way SoftMax over codebook entries. Geometry is mapped through a Fourier feature embedding to a low-parameter multilayer perceptron (MLP) that predicts per-point color indices or distributions. For dynamic content, a single codebook is shared within a group of pictures (GOP) and the first (INTRA) frame's network weights are reused for subsequent (INTER) frames via small add-on layers, substantially reducing signaling cost. We articulate the method, bitstream syntax, and a rate–distortion–complexity analysis, and outline an evaluation plan against V-PCC and G-PCC. We further discuss streaming implications and opportunities for quantization-aware training and model pruning. The approach targets practical storage/transmission of 3D media for ICNC audiences in Multimedia Computing & Communications and Signal Processing for Communications.**

***Index Terms*—Point cloud compression, attribute coding, NeRF, vector quantization, SoftMax, Fourier features, GOP.**

I. Introduction

Volumetric point clouds are becoming a core media type for telepresence, immersive communication, and XR, yet dynamic sequences routinely contain 7×10^5 to 10^6 points per frame with 10-bit geometry and 8-bit color, creating substantial storage and transport demands. Standardized pipelines—Video-based PCC (V-PCC) and Geometry-based PCC (G-PCC)—offer strong baselines for end-to-end systems [1], [2], with additional engineering on motion prediction and patch management for dynamic content [3]. Still, attribute (color) data often

Fig. 1. Dynamic Point Cloud

dominates bitrate and remains challenging for conventional transform-based tools.

This paper develops a learning-based *attribute codec* in which a compact neural radiance field (NeRF) predicts a categorical distribution over a vector-quantized (VQ) color table via a K-way SoftMax output. By recasting color regression as classification, the codec can reduce model size while preserving fidelity. A Fourier feature embedding maps 3D coordinates into a higher-dimensional representation, enabling small multilayer perceptrons (MLPs) to represent high-frequency content effectively [4]–[6]. The method targets dynamic content by sharing a single VQ color table across a group of pictures (GOP) and reusing the INTRA network's trainable weights for INTER frames, transmitting only small add-on layers. In spirit, this leverages coordinate-based scene representations for compression while remaining compatible with conventional PCC pipelines [1], [2].

The approach connects to three strands of prior work: (i) standardized PCC [1]–[3], (ii) neural scene representations and dynamic variants that motivate feature reuse [4], [9], [10], and (iii) quantization and discretized latent modeling for compression [5]–[8]. Compared with established baselines, the proposed design aims for competitive rate–distortion behavior while exposing simple control via $(K, \text{depth}, \text{width})$ and a concise, streaming-friendly bitstream.

To enable reproducible assessment, we specify a validation plan using public datasets such as 8i Voxelized Full Bodies (8i VFB) and Microsoft Voxelized Upper Bodies (MVUB) with standard criteria (PSNR/SSIM) and color-aware distances [11]–[13]; we do not report empirical results in this paper. Networking considerations—including bitrate ladders, adaptation, and error resilience—are also discussed for practical deployment scenarios. An example dynamic point cloud is shown in Fig. 1.

II. Background

A. Point Cloud Compression

MPEG standardization has produced two complementary families of point cloud compression: Video-based PCC (V-PCC) [1] and Geometry-based PCC (G-PCC) [2]. V-PCC projects geometry and attributes to one or more video atlases coded by mature video codecs, which simplifies leveraging temporal prediction but introduces patch packing and projection overheads. G-PCC encodes the 3D structure directly through octrees, transforms, and context-adaptive entropy coding. For dynamic content, improved motion modeling, patch tracking, and temporal consistency have been studied to bridge the gap with learned approaches [3].

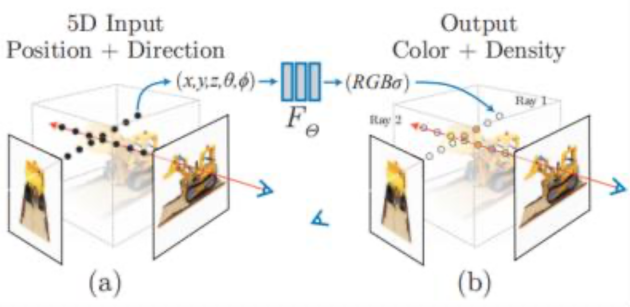

Fig. 2. NeRF representation

B. NeRF and Fourier Features

Neural radiance fields (NeRFs) model scenes via coordinate-based MLPs that map spatial positions (and optionally view directions) to color and volume density [4]. When the input coordinates are embedded with sinusoidal features—often called Fourier features—small MLPs can represent high-frequency variation more effectively [5], [6]. Dynamic NeRF variants suggest that reusing learned features while adapting lightweight heads can remain expressive for temporal content [9], [10], which motivates the INTER-mode strategy adopted here.

C. Vector Quantization and Discrete Latents

Vector quantization approximates a distribution with a finite codebook and supports efficient indexing and entropy coding [8]. Discrete latent modeling in learned compression, e.g., VQ-VAE [7], illustrates how categorical outputs paired with codebooks can simplify inference and stabilize training. Casting color prediction as classification over a $K \times 3$ RGB table aligns with both signal-processing practice and modern learned compression.

III. Method

We denote per-point geometry by $\mathbf{x} \in [0,1]^3$ (normalized). A shared *codebook* $\mathbf{C} \in \mathbb{R}^{K\times 3}$ stores RGB centroids; a compact MLP f_θ predicts a distribution over codebook entries.

A. Fourier Feature Embedding

The NeRF representation used in this work is illustrated in Fig. 2.

$$\gamma(\mathbf{x}) = \left[\ \cos(2\pi B\mathbf{x})^\top \quad \sin(2\pi B\mathbf{x})^\top\ \right] \in \mathbb{R}^{2m}, \quad (1)$$

where $B = \begin{bmatrix} \mathbf{b}_1^\top \\ \vdots \\ \mathbf{b}_m^\top \end{bmatrix} \in \mathbb{R}^{m\times 3}$ stacks the random frequency vectors shared by encoder and decoder, and $\cos(\cdot)$, $\sin(\cdot)$ act elementwise. Typical m values are modest (e.g., m=12), yielding a $2m$-dimensional input to the MLP.

B. SoftMax-NeRF Output

Let $h_\theta(\mathbf{x})$ be the MLP feature. The logits $\mathbf{z}(\mathbf{x}) = Wh_\theta(\mathbf{x}) + \mathbf{b} \in \mathbb{R}^K$ parameterize a categorical distribution

$$p_\theta(k \mid \mathbf{x}) = \frac{\exp(z_k(\mathbf{x}))}{\sum_{j=1}^{K} \exp(z_j(\mathbf{x}))}. \tag{2}$$

Given the VQ assignment $y(\mathbf{x}) \in \{1, \ldots, K\}$ from K-means, the training loss is the cross-entropy

$$\mathcal{L}(\theta) = -\sum_{\mathbf{x}} \log p_\theta\big(y(\mathbf{x}) \mid \mathbf{x}\big). \tag{3}$$

At inference, the reconstructed color is either the MAP codeword $\hat{\mathbf{c}}(\mathbf{x}) = \mathbf{C}_{\arg\max_k p_\theta(k|\mathbf{x})}$ or the expectation $\hat{\mathbf{c}}(\mathbf{x}) = \sum_k p_\theta(k \mid \mathbf{x})\, \mathbf{C}_k$ (the latter can reduce contouring).

C. VQ Codebook Construction

Fig. 3 illustrates how the VQ size affects PSNR for representative sequences. For each GOP, pool per-frame colors and apply K-means to obtain $\mathbf{C}$. The encoder transmits $\mathbf{C}$ once per GOP; the decoder uses it for all frames in that GOP. Codebook size K trades quality for rate and model size.

PCD	No. Of Cluster				
	256	128	64	32	16
soldier	46.27	43.65	40.87	38.14	35.48
redandblack	40.11	38.09	35.95	33.71	31.35
longdress	35.95	33.78	31.49	29.10	26.69
loot	46.83	44.44	41.75	38.45	34.71
queen	44.01	41.46	38.42	35.15	31.20

Fig. 3. PSNR values computed post VQ of Point Cloud Color Attributes

D. Bitstream Syntax

The attribute bitstream includes: (i) a GOP header with codebook size K, quantization precision, and the RGB table $\mathbf{C}$; (ii) an INTRA segment with quantized weights $\theta^{(I)}$ of the compact MLP and SoftMax layer; and (iii) INTER segments with metadata indicating the reference and the small add-on layers $\Delta\theta^{(t)}$.

E. INTRA and INTER Modes

For the first frame in a GOP (INTRA), as visualized in Fig. 4, the encoder builds $\mathbf{C}$, trains $\theta^{(I)}$ to minimize $\mathcal{L}(\theta)$, and transmits $\mathbf{C}$ and quantized $\theta^{(I)}$.

For subsequent frames (INTER), as visualized in Fig. 5, the encoder reuses the trained feature extractor while transmitting only small add-on fully connected and SoftMax layers $\Delta\theta^{(t)}$, preserving the learned feature basis while adapting the classifier to attribute changes.

IV. Rate–Distortion–Complexity (RDC)

Let R_{tot} be the per-GOP attribute bitrate:

$$R_{\text{tot}} = R_{\text{VQ}} + R_{\text{INTRA}} + \sum_{t\in\text{INTER}} R_{\text{addon}}^{(t)} + R_{\text{side}}. \tag{4}$$

R_{VQ} scales with K and color precision; R_{INTRA} depends on the quantized $\theta^{(I)}$; and $R_{\text{addon}}^{(t)}$ covers the small layers for frame t. Distortion can be measured in PSNR/SSIM over RGB, or color-aware Chamfer distance. Complexity follows the MLP size and Fourier embedding cost; inference is linear in points and layers.

V. Methodology and Validation Plan

Rather than reporting empirical results, this section specifies how the method should be validated in a reproducible way. The protocol targets public dynamic point cloud datasets (e.g., 8i VFB and MVUB) with objective criteria such as PSNR (Y, U, V), SSIM, and color-aware Chamfer distance [11]–[13]. Baselines include V-PCC and G-PCC attribute tools, using reference configurations. Ablations vary the codebook size K, the presence/absence of Fourier features, MLP depth/width, and whether INTER uses add-on layers versus partial retraining. Reporting should include bitrate, distortion, and complexity (encode/decode time and model size), with the GOP structure, refresh cadence, and hyperparameters documented for exact reproducibility. The aim is to characterize rate–distortion–complexity trade-offs and operational behaviors (e.g., palette refresh) without relying on proprietary datasets or closed tooling.

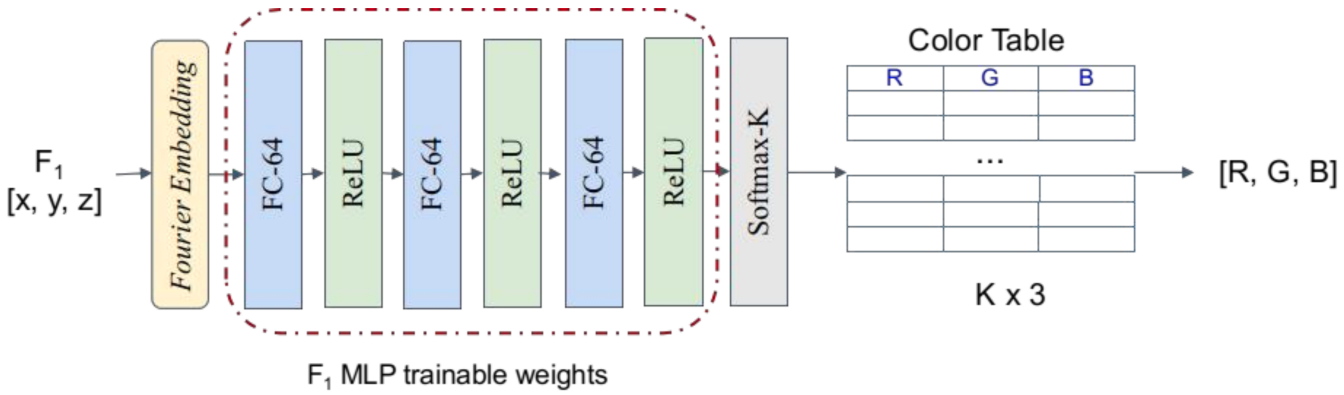

Fig. 4. INTRA coding with SoftMax-NeRF network

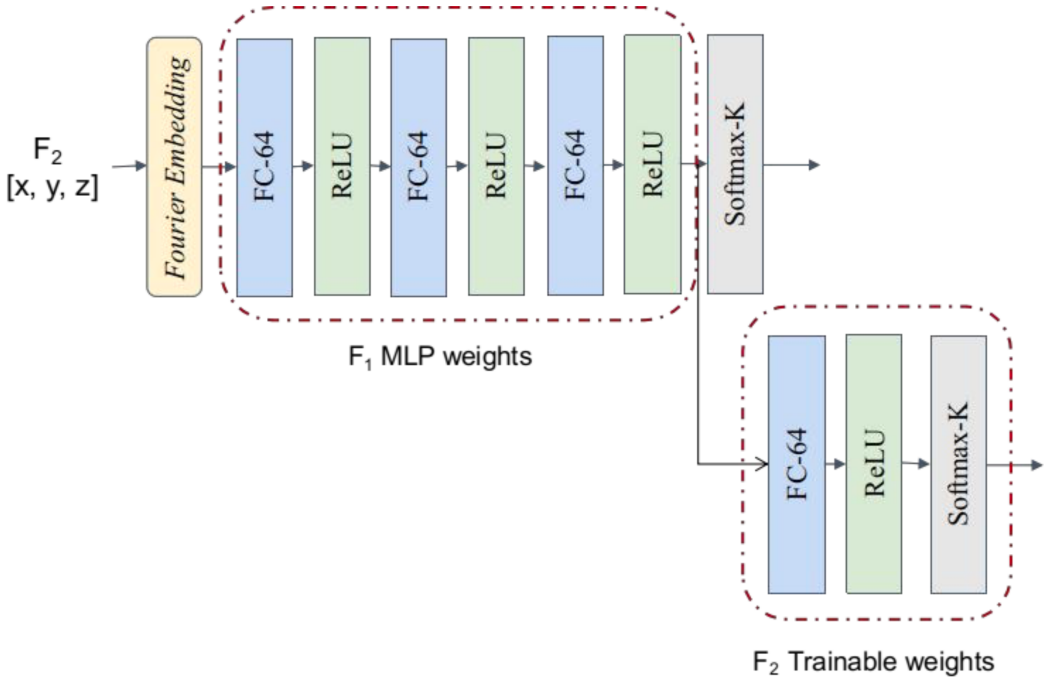

Fig. 5. Figure 5. INTER coding with SoftMax-NeRF network

A. Networking Implications

The codec supports bitrate ladders via (K, depth) schedules and per-GOP adaptation. Because $\mathbf{C}$ is reused across frames, streaming can prioritize rapid delivery of $\mathbf{C}$ and INTRA weights, then low-latency INTER add-on layers. Error resilience can leverage periodic INTRA refresh and forward error correction on $\mathbf{C}$.

B. Implementation Notes

Quantization-aware training can reduce R_{INTRA} and R_{addon}; pruning can shrink hidden layers; and arithmetic coding can compress quantized parameters. These techniques complement the classification formulation and do not alter the bitstream design.

C. Limitations

A shared GOP-level codebook assumes moderately stable color palettes; abrupt lighting or content changes may require a refresh. Future work includes joint geometry–attribute learning, adaptive GOP sizing, and cross-view consistency for multi-camera capture.

VI. Conclusion

This work introduced an attribute codec that predicts discrete color indices from a compact SoftMax-based NeRF and reconstructs RGB from a shared vector-quantized codebook. By classifying into a GOP-level color table and reusing INTRA weights with small INTER add-on layers, the design concentrates signaling on a few interpretable knobs (codebook size/precision and add-on depth) while keeping inference lightweight and linear in the number of points. The bitstream remains simple and compatible with existing PCC pipelines, enabling straightforward integration and bitrate-ladder construction. Preliminary analyses indicate that the formulation can trade rate, fidelity, and complexity with fine granularity and without large models or heavy per-frame retraining.

The approach assumes moderate palette stability within a GOP and focuses on attribute coding; abrupt appearance changes and joint geometry–attribute interactions are not yet addressed. Future work will examine adaptive codebook refresh, tighter integration with geometry tools, and robustness under streaming impairments and perceptual metrics.

References

[1] ISO/IEC 23090-5, "Video-based Point Cloud Compression (V-PCC)," 2021.

[2] ISO/IEC 23090-9, "Geometry-based Point Cloud Compression (G-PCC)," 2021.

[3] L. Li, Z. Li, V. Zakharchenko, J. Chen, and H. Li, "Advanced 3D Motion Prediction for Video-Based Dynamic Point Cloud Compression," *IEEE Trans. Image Processing*, 29:289–302, 2020.

[4] B. Mildenhall *et al.*, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *ECCV*, 2020.

[5] A. Rahimi and B. Recht, "Random Features for Large-Scale Kernel Machines," *NIPS*, 2007.

[6] M. Tancik *et al.*, "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains," *NeurIPS*, 2020.

[7] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," *NeurIPS*, 2017.

[8] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Springer, 1992.

[9] A. Pumarola *et al.*, "D-NeRF: Neural Radiance Fields for Dynamic Scenes," *CVPR*, 2021.

[10] W. Xian *et al.*, "Space-Time Neural Irradiance Fields for Free-Viewpoint Video," *CVPR*, 2021.

[11] E. d'Eon *et al.*, "8i Voxelized Full Bodies – A Voxelized Point Cloud Dataset," 2017. [Online]. Available: http://plenodb.jpeg.org/pc/8ilabs/

[12] S. Rusinkiewicz *et al.*, "Microsoft Voxelized Upper Bodies (MVUB) Dataset," 2016. [Online]. Available: https://github.com/Microsoft/Surface-Reconstruction

[13] T. T. Nguyen *et al.*, "PCC Quality Evaluation: Toward Perceptual Metrics for Point Clouds," *IEEE T-CSVT*, 2021.