ECRF: Entropy-Constrained Neural Radiance Fields Compression with Frequency Domain Optimization

Soonbin Lee   Fangwen Shu   Yago Sánchez   Thomas Schierl   Cornelius Hellge

Fraunhofer Heinrich-Hertz-Institute (HHI), Germany

{first_name}.{last_name}@hhi.fraunhofer.de

Abstract

Explicit feature-grid based NeRF models have shown promising results in terms of rendering quality and significant speed-up in training. However, these methods often require a significant amount of data to represent a single scene or object. In this work, we present a compression model that aims to minimize the entropy in the frequency domain in order to effectively reduce the data size. First, we propose using the discrete cosine transform (DCT) on the tensorial radiance fields to compress the feature-grid. This feature-grid is transformed into coefficients, which are then quantized and entropy encoded, following a similar approach to the traditional video coding pipeline. Furthermore, to achieve a higher level of sparsity, we propose using an entropy parameterization technique for the frequency domain, specifically for DCT coefficients of the feature-grid. Since the transformed coefficients are optimized during the training phase, the proposed model does not require any fine-tuning or additional information. Our model only requires a lightweight compression pipeline for encoding and decoding, making it easier to apply volumetric radiance field methods for real-world applications. Experimental results demonstrate that our proposed frequency domain entropy model can achieve superior compression performance across various datasets. The source code will be made publicly available.

TensoRF
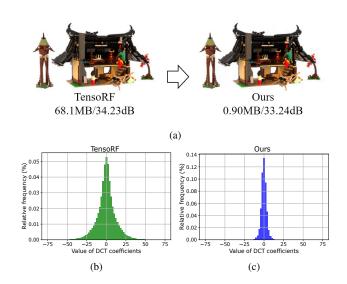68.1MB/34.23dB

Ours
0.90MB/33.24dB

(a)

(b)

(c)

Figure 1. **Reconstructed results** for the *Blacksmith* scene from the Ray-Traced Multi-View (RTMV) dataset [35] are shown in (a). The frequency distribution of transformed coefficients is compared between the (b) TensoRF baseline and (c) the proposed model.

1. Introduction

Recent advancements in implicit neural representations (INRs) have greatly influenced the research of 3D scene modeling and novel-view synthesis [4, 19, 26, 32]. This approach represents a complex volumetric scene as an implicit function that maps positional and directional information of sampled points to corresponding color and density values. This enables the rendering of photo-realistic novel views from desired viewpoints. Further developments have shown the capability to reconstruct various 3D scenes using images and corresponding camera poses, making radiance fields a promising method for representing the real 3D world [5, 34, 43]. However, these methods have a considerable computational overhead during training and inference, which is a significant bottleneck. This often results in several days of convergence time.

On the other hand, representing neural radiance fields using explicit grid-based structures has been shown to significantly improve training and inference efficiency [8, 11, 21, 28, 29]. These volumetric radiance field methods typically use grids to store learnable feature vectors and retrieve the color and density of a 3D point through trilinear interpolation. This can be achieved either without a neural network or with shallow MLPs, resulting in accelerated speed and high synthesis quality [24, 42]. TensoRF is a parameter-efficient yet expressive method for decomposing dense 3D grids into a smaller set of parameters, such as matrices and vectors. This decomposition can significantly re-

duce the memory and computational requirements [9]. Similarly, by incorporating this feature-grid, the training and inference time of grid-based methods has been reduced from days to minutes [21, 29]. While these methods reduce the time complexity for training and inference in representing 3D scenes and objects, they have larger overall sizes compared to MLP-only methods. As a result, the use of grid-based volumetric representations inevitably leads to significant storage costs, which may restrict its practicality in real-world applications.

Therefore, in this work, we propose a new model called entropy-constrained radiance fields (ECRF) for compressing radiance fields. This optimization process leads to a compact feature representation in the frequency domain. As a result, the optimized radiance field can effectively reconstruct complex 3D scenes using a compact representation. Optimization leads to a sparse representation of transform coefficients, which allows the proposed model to outperform other compression models in terms of rate-distortion. As shown in Fig. 1, the volumetric scene can be compressed by using compact transformed coefficients. The proposed model achieves a compression ratio of $75\times$ by reducing the original size, while only resulting in a visual quality loss of 1dB for the given example. This paper presents experimental results to validate the effectiveness of the proposed model, which uses entropy-minimized frequency domain representations. We summarize our contributions as follows:

- We propose a novel entropy-aware training scheme, which aims to compress data by minimizing log-likelihood with an entropy estimation network. During the training process, the optimization objective is to minimize the length of a bit sequence, which corresponds to the size of the entropy-coded bitstream.
- We introduce the frequency-domain entropy parameterization. In order to leverage the sparsity of signals in the frequency domain, the original data is transformed into the frequency domain using block-wise discrete cosine transformation (DCT). To achieve high sparsity and effectively reduce the number of bits, we also apply regularization to the DCT coefficients during training.
- In the TensoRF-based compression models, we achieve state-of-the-art performance with a compression pipeline, which includes 8-bit quantization and entropy coding on the transformed coefficients.

2. Related Work

Radiance Field Compression. To achieve real-time rendering with reasonable memory requirements, Instant-NGP proposed the use of multi-resolution voxel grids augmented with hash tables [21]. Another approach is to decompose 3D grids into lower-dimensional representations, such as vector-matrix and voxel grids [9, 28]. To reduce the

size of these data structures, Variable Bitrate Neural Fields (VQAD) proposed using codebooks with vector quantization (VQ) to achieve variable bitrate and level of detail [31]. Similarly, VQRF proposes a compression framework pipeline that utilizes VQ with importance-based voxel pruning on the feature-grid [15]. VQRF compresion framework combined an 8-bit weight quantization with entropy coding to reduce the model size to 1MB, while still maintaining visual quality. CCNeRF suggests a low-rank approximation method for tensor decomposition [33], and Re:RF proposes simple pruning techniques to reduce the size of the feature-grid [10].

The work most closely related to ours is the masked wavelet radiance field, which achieves competitive compression performance by using wavelet transformation and binarized masks for wavelet coefficients [23]. ReRF introduces DCT to compress the coefficients of the dynamic voxel grid using Huffman coding [39]. Similarly, TinyNeRF also incorporates DCT into the voxel grid and utilizes pruning and quantization-aware training to achieve high signal sparsity [44]. Another approach, Binary Radiance Fields (BiRF), utilizes binary encoding for parameter quantization, restricting values to either -1 or +1 to minimize feature vector storage requirements [25]. In this work, we demonstrate experimentally that using entropy parameterization can lead to more efficient representations, resulting in a smaller model size.

Frequency Domain Transformation. Several studies have examined the use of frequency-based parameterization to generate efficient neural field representations [32, 37, 40]. Fourier Plenoctree has demonstrated that applying the Fourier transform can enhance both the parameterization efficiency and training speed of the octree structure [38]. However, these studies have focused on enhancing the expressiveness of 3D scenes through frequency representation, rather than fundamentally reducing model size. Meanwhile, conventional image and video codecs commonly utilize handcrafted techniques to reduce spatial redundancy. For example, the Joint Photographic Experts Group (JPEG) uses the wavelet transform to convert images from the pixel domain to the frequency domain [17, 36]. In addition, High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC) employ DCT with varying block sizes [7, 27]. Inspired by this consideration, our work is to incorporate these video coding pipeline into the radiance field compression.

Entropy Model for Learned Compression. Several works for learned compression have been proposed to utilize the advanced capabilities of neural networks [3, 18, 41]. The main idea behind neural compression is to use entropy models that are parameterized by neural networks [1, 2, 20]. This is achieved by establishing an entropy model that optimizes the size of its bitstream for latent features [6, 12, 22].
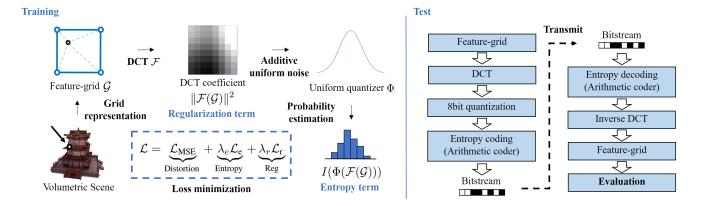
Figure 2. **Overall architecture of the proposed model.** After training with DCT representation as the optimization target, the feature-grid is entropy-coded in the frequency domain using the compression pipeline.

3. Proposed Method

In this section, we introduce the entropy-constrained radiance field, a novel approach for efficient storage that employs entropy parameterization of transformed coefficients. Fig. 2 illustrates the overall scheme based on TensoRF [9] of our radiance field reconstruction. We will begin by explaining the parameterization of entropy using an entropy estimation network. Then, we will introduce a training approach that includes frequency domain parameterization and coefficient regularization. This approach aims to enhance data compression by leveraging the advantages of frequency domain representation. It is worth noting that our proposed model does not require any fine-tuning or additional storage cost (e.g., mask, index mapping). Once the feature-grid is trained with our loss function, the DCT is applied to transform the feature-grid into coefficients. These transformed coefficients are then quantized to 8-bit, following the same approach as other compression models [15, 23]. Finally, the quantized coefficients are encoded using an arithmetic coder to calculate the overall storage cost [14].

3.1. Feature-Grid Neural Radiance Fields

Neural Radiance Fields (NeRF) use an implicit function to represent scenes. This function maps spatial point $\mathbf{x} = (x, y, z)$ and view direction $\mathbf{d} = (\theta, \phi)$ to density σ and color $\mathbf{c}$. By integrating the color $\mathbf{c}_i$ and density σ_i of the spatial points $\mathbf{x}_i = \mathbf{o} + t_i \mathbf{d}$ sampled along a ray $\mathbf{r}$, we can estimate the RGB value $\hat{C}(\boldsymbol{r})$ of the corresponding pixel:

$$\hat{C}(\boldsymbol{r}) = \sum_i^N T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \mathbf{c}_i \quad (1)$$

In the above equation, $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_i \delta_i\right)$, and δ_i represents the distance between adjacent samples. For effi-

cient representation of 3D objects and scenes, recent studies have proposed the use of lower-dimensional grids, such as 2D planes or 1D lines [8, 9]. TensoRF introduces a VM (vector-matrix) decomposition to reduce the memory usage caused by the large tensor $\mathcal{T} \in \mathbb{R}^{I \times J \times K}$ into low-rank matrices $\mathbf{M}$ and vectors $\mathbf{v}$ as follows:

$$\mathcal{T} = \sum_{r=1}^R \mathbf{v}_r^X \circ \mathbf{M}_r^{YZ} + \mathbf{v}_r^Y \circ \mathbf{M}_r^{XZ} + \mathbf{v}_r^Z \circ \mathbf{M}_r^{XY} \quad (2)$$

Here, $\mathbf{v}_r^X \in \mathbb{R}^I$ denote the vector for X axis, $\mathbf{M}_r^{YZ} \in \mathbb{R}^{J \times K}$ is the matrix for Y and Z axes, R is the rank, and the symbol $\circ$ is outer product. Hence, $\mathbf{v}_r^X \circ \mathbf{M}_r^{YZ} \in \mathbb{R}^{I \times J \times K}$. TensoRF has shown its capability to decrease memory usage and training complexity by utilizing VM decomposition. However, this method results in total VM parameters that consume over 60MB of storage. In this work, we propose a compression model that reduces the size of the VM parameters by referring to them as a feature-grid. We will refer to the compression targets for all matrices $\mathbf{M}$ and vectors $\mathbf{v}$ as the feature-grid $\mathcal{G}$, and denote them as follows:

$$\mathcal{G} = \{\mathbf{M}_\sigma, \mathbf{M}_c, \mathbf{v}_\sigma, \mathbf{v}_c\} \quad (3)$$

A feature vector for a 3D point is calculated using trilinear interpolation from the feature-grid $\mathcal{G}$ and decoded by a small MLP to determine color and density values.

3.2. Entropy Parameterization of Feature-Grid

We propose an entropy parameterization technique to the feature-grid $\mathcal{G}$ as an optimization target. By using this technique, the trained model minimizes its entropy I, resulting in a much shorter bit length:

$$I(\mathcal{G}) = -\sum_{i=1}^N \log_2 P_i(\mathcal{G}) \quad (4)$$

3

To calculate the total entropy $I(\mathcal{G})$, we need to introduce a probability density function (PDF) model P_i for N samples $i \in \{1, \ldots, N\}$. However, estimating the PDF presents a significant challenge due to its non-differentiability. Because the derivatives of the naive rounding operation are almost zero everywhere, gradient descent becomes ineffective. To address this issue, the learned compression model proposed the use of additive uniform noise as an approximation for quantized functions [1]. Similarly, we introduce a uniform quantization function Φ to obtain proxy gradients of the PDF:

$$\Phi(\theta) = \theta + u, u \sim \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right) \qquad (5)$$

where $\mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)$ is independent and identically distributed and uniformly distributed noise. This uniform quantization has been shown to be a good approximation for entropy when using the negative log-likelihood based on noisy discretization. Then, the discrete range is convolved with P to obtain the discretized PDF:

$$\begin{aligned} P(\Phi(\theta_i))) &= \int_{\Phi(\theta_i)-\frac{1}{2}}^{\Phi(\theta_i)+\frac{1}{2}} P(\theta_i)\mathrm{d}\theta_i \\ &= P_c\left(\Phi(\theta_i) + \frac{1}{2}\right) - P_c\left(\Phi(\theta_i) - \frac{1}{2}\right) \end{aligned} \qquad (6)$$

To simplify the calculation of PDF, [2] constructed a cumulative distribution function (CDF) P_c with small MLPs. This CDF P_c maps $\mathbb{R} \to [0, 1]$ and it is a monotone increasing function that represents the CDF of the underlying PDF P. In this way, the calculation of P_c simplifies the estimation of the PDF. Our proposed model only requires small MLPs to calculate P_c. After training is completed, these MLPs can be discarded. The detailed architecture of this entropy estimation network is described in the supplementary materials.

3.3. Frequency Domain Parameterization

The DCT is a commonly used method for Fourier-related transformations in image and video compression and it is known for its efficient energy compaction of spectral information. Hence, our proposed method focuses on signal compression by optimizing it in the frequency domain, rather than directly estimating entropy in the feature-grid. We estimate entropy in the frequency domain, which consists of transformed coefficients of the feature-grid. Specifically, the matrices $\mathbf{M}$ and vectors $\mathbf{v}$ are transformed into the frequency domain using the DCT. Given a feature-grid $\mathcal{G} \in \mathbb{R}^{N \times C \times H \times W}$, where N is the number of matrices and vectors, C is the dimension of the feature channel, and $H \times W$ the shape of the feature-grid. The transformed coefficients $\mathcal{F}(\mathcal{G})$ after applying 3D DCT with $K_1 \times K_2 \times K_3$
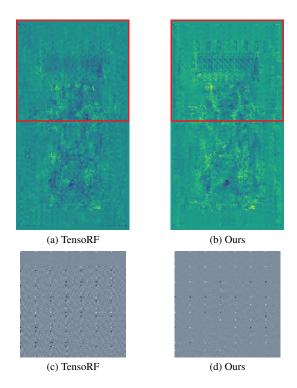

(a) TensoRF   (b) Ours

(c) TensoRF   (d) Ours

Figure 3. **Visualization of the first channel in the XY plane** $\mathbf{M}_c^{XY}$ is shown for (a) TensoRF baseline and (b) our proposed model. The coefficients are transformed using a 2D DCT 32×32 block $\mathcal{F}(\mathbf{M}_c^{XY})$ shown for (c) TensoRF baseline and (d) the proposed model. All results are based on 30k training iterations.

block size can be defined as:

$$\mathcal{F}(\mathcal{G})_{n,u,v,w} = \sum_{x=0}^{K_1-1} \sum_{y=0}^{K_2-1} \sum_{z=0}^{K_3-1} \mathcal{G}_{n,x,y,z}$$
$$\cos\left[\frac{\pi}{K_1}\left(x+\frac{1}{2}\right)u\right] \cos\left[\frac{\pi}{K_2}\left(y+\frac{1}{2}\right)v\right] \cos\left[\frac{\pi}{K_3}\left(z+\frac{1}{2}\right)w\right] \qquad (7)$$

where $\mathcal{F}(\mathcal{G}) \in \mathbb{R}^{N \times C \times H \times W}$ is the transformed coefficients, $[x, y, z]$ and $[u, v, w]$ are block indices for $\{0, 1, 2, \ldots, K_i{-}1\}$. Fig. 3 provides a visualization of the XY plane tensor $\mathbf{M}_c^{XY}$, using a 2D DCT with block size 32×32 for *Lego* scene. The signal on the tensor plane projected onto the XY-axis can be seen in the figure. During training, our proposed method can use frequency domain transformation to generate optimized representations for all transformed coefficients. After training with the entropy minimization, it is shown that the transformed coefficients $\mathcal{F}(\mathbf{M}_c^{XY})$ can restore the original tensor signal with a more sparse representation. For each volumetric scene, our proposed model jointly optimizes the entropy of the transformed coefficients, resulting in a sparse representation. The DCT operation is differentiable, allowing for backpropagation with entropy parameterization.

Method	Synthetic NeRF			Synthetic NSVF			Tanks and Temples		
	Size (MB)↓	PSNR↑	SSIM↑	Size (MB)↓	PSNR↑	SSIM↑	Size (MB)↓	PSNR↑	SSIM↑
TensoRF-VM [9]	71.3	33.15	0.962	71.8	36.70	0.984	72.0	28.58	0.922
Instant-NGP [21]	39.5	33.08	0.960	39.5	36.04	0.981	39.5	28.85	0.924
TensoRF-CP [9]	3.9	31.66	0.950	3.9	34.48	0.971	4.0	27.59	0.897
Re:TensoRF High [10]	7.9	32.81	0.956	8.5	36.14	0.978	6.7	28.24	0.907
TinyNeRF (4MB) [44]	4	31.90	0.956	4	34.88	0.975	4	28.32	0.910
TinyNeRF (2MB) [44]	2	31.72	0.954	2	34.62	0.968	2	28.19	0.908
VQ-TensoRF [15]	3.5	32.88	0.960	4.2	36.04	0.979	3.4	28.25	0.913
VQ-DVGO [15]	1.4	31.77	0.954	1.3	34.72	0.974	1.4	28.26	0.909
BiRF (Rate-3) [25]	2.8	**33.26**	0.961	2.9	36.17	0.983	2.9	**28.62**	0.924
Masked (Rate-3) [23]	2.4	32.38	0.958	2.7	35.57	0.976	2.4	28.27	0.915
Ours (Rate-3)	2.5	33.05	0.961	2.6	**36.22**	0.983	2.5	28.41	0.921
BiRF (Rate-2) [25]	1.4	32.64	0.959	1.5	35.40	0.976	1.5	**28.44**	0.916
Masked (Rate-2) [23]	1.5	32.23	0.958	1.6	35.33	0.976	1.7	28.01	0.904
Ours (Rate-2)	1.3	**32.72**	0.960	1.4	**35.73**	0.978	1.4	28.31	0.915
BiRF (Rate-1) [25]	0.7	31.53	0.949	0.8	34.26	0.971	0.8	28.02	0.906
Masked (Rate-1) [23]	0.8	31.95	0.956	0.9	34.67	0.974	0.9	27.77	0.901
Ours (Rate-1)	0.8	**32.20**	0.958	0.8	**35.00**	0.975	0.8	**28.14**	0.908

Table 1. **Quantitative results** of model size and reconstruction quality are presented for the Synthetic NeRF, NSVF dataset, and Tanks and Temples dataset. To provide a more detailed comparison, we report the performance at three rate levels: {0.8, 1.4, 2.5}MB. Note that these models use different experimental configurations for each rate level, as described in Sec 4.2. The results of BiRF, TinyNeRF and Re:TensoRF are directly adopted from the original paper.

Finally, our proposed model uses a differentiable entropy of transformed coefficients $\mathcal{F}(\mathcal{G})$ as a loss function $\mathcal{L}_e$:

$$\mathcal{L}_e = I(\Phi(\mathcal{F}(\mathcal{G}))) \tag{8}$$

Additionally, we apply a coefficient regularization term. This regularization term is expected to have an impact on the quantization and rounding of the DCT coefficients, similar to the compression pipeline used in traditional video codecs. Experimental results have shown that this term effectively improves sparsity, leading to a smaller bitstream size. Specifically, we apply $\mathcal{L}_2$ norm regularization by squaring all values of the transformed coefficients for the following optimization objective:

$$\mathcal{L}_r = \|\mathcal{F}(\mathcal{G})\|^2 \tag{9}$$

To establish a rate-distortion trade-off, we consider the mean squared error (MSE) of the RGB value as the measure of distortion:

$$\mathcal{L}_{MSE} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \|\hat{C}(r) - C(r)\|^2 \tag{10}$$

where $\mathcal{R}$ is the set of camera rays sampled in a mini-batch, and $C(r)$ is the ground-truth pixel value corresponding to the camera ray r. Thus, the overall loss function is a combi-nation of the distortion loss and entropy parameterizations of the transformed coefficients with the regularization term:

$$\mathcal{L} = \underbrace{\mathcal{L}_{MSE}}_{\text{Distortion}} + \underbrace{\lambda_e \mathcal{L}_e}_{\text{Entropy}} + \underbrace{\lambda_r \mathcal{L}_r}_{\text{Reg}} \tag{11}$$

Our proposed model incorporates this loss function that minimizes entropy in the transformed representation. It results in a relatively sparse discrete representation of the feature-grid, significantly reducing the size of the compressed model when entropy coding is performed. By adjusting the parameter λ_e in different experiments, this model can explore the trade-off between compressed size and reconstructed visual quality.

To render the original scene, the compression pipeline described above only needs one inverse DCT and entropy decoding for the transformed coefficients. Therefore, the additional complexity costs to the original feature-grid model are negligible. For TensoRF, a shallow decoder MLP is used to predict color and density from the trilinear interpolated feature-grid. The total storage required for a decoder MLP is approximately 0.14 MB, according to the default experimental configurations of TensoRF. In our quantitative results presented in Tab. 1, we report the model size including the decoder MLP, and do not apply any compression techniques to it.
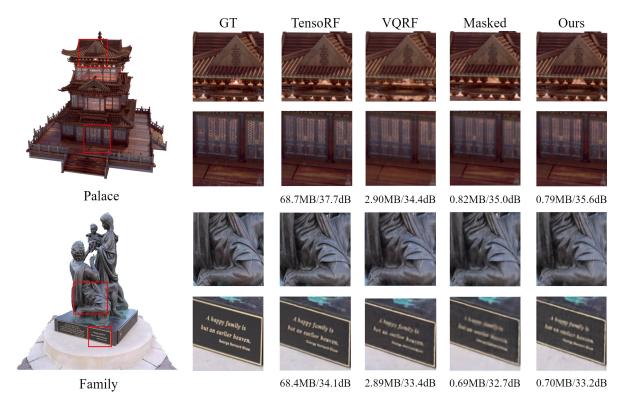
Figure 4. **Qualitative results** for visual comparison are provided for each scene of the Synthetic NSVF and Tanks and Temple datasets using TensoRF-based compression models. Each subfigure displays the storage size and PSNR.

4. Experiments and Results

We present a comprehensive evaluation of the proposed compression model for tensorial radiance fields. First, we describe our configurations and the datasets used. Next, we compare our approach with previous and concurrent works. Finally, we present ablation studies for each component and parameter.

4.1. Experimental Configurations

We follow the experimental configurations of TensoRF, which includes 30k iterations of training and a minibatch size of 4096 rays, along with other configurations [9]. TensoRF uses a coarse-to-fine training strategy that dynamically adjusts the tensor size and alpha mask (occupancy) to achieve accurate rendering. During training, the process of tensor upsampling and alpha mask updating has the potential to disrupt the accurate estimation of entropy. To update the alpha mask, we use a slightly different version from the masked wavelet model [23], which is at the $\{2000, 4000, 6000, 11000, 16000\}$ iterations. The masked wavelet model utilizes a wider range of intervals, but this configuration does not have a significant impact on our model. Therefore, the loss function of our proposed model is applied starting from the 16k iteration, when the size of

Method	Size (MB)↓	PSNR (dB)↑	SSIM↑
TensoRF-VM [9]	75.3MB	33.67	0.971
VQRF [15]	2.18MB	30.60	0.949
NGLOD-NeRF [30]	20MB	32.72	0.970
VQAD (6bw) [31]	0.49MB	30.76	0.956
Ours ($\lambda_e = 2e^{-10}$)	0.41MB	**31.40**	**0.961**
VQAD (4bw) [31]	0.33MB	30.09	0.948
Ours ($\lambda_e = 5e^{-10}$)	0.29MB	**30.33**	**0.949**

Table 2. **Quantitative results** for 10 brick scenes from RTMV dataset at a 400×400 resolution. The results of VQAD and NGLOD-NeRF are directly adopted from the original paper.

the feature-grid is completely fixed. It has been observed that the proposed model converges sufficiently under this condition.

We then compare our model with other compression models using various datasets, including Synthetic NeRF and Neural Sparse Voxel Fields (NSVF) for 800×800 resolution and Tanks And Temples (T&T) for 1920×1080 resolution [13, 16, 19]. In addition, 10 brick scenes from the Ray-Traced Multi-View (RTMV) dataset are used for comparison with VQAD at a 400×400, which matches the resolution used in the VQAD [31, 35]. The proposed model is based on TensoRF VM-192 and adjusts the parameter λ_e
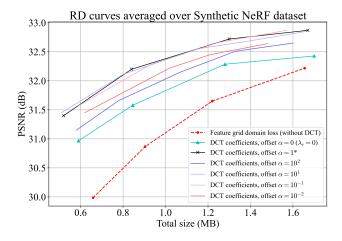
6

Figure 5. **Ablation study: the impact of the regularization term.** The rate-distortion (RD) curves demonstrate the effect of varying the regularization factor λ_r with offset α, where * indicates the default configuration.

to control the bitrate. Regarding the DCT block size, this paper chooses a 3D block size of $16\times16\times16$ for matrices **M**, and a 2D block size of 8×8 for vectors **v**. More details are presented in section 4.3.

4.2. Quantitative Result

For a fair comparison, different models are categorized into grid-based models and compression models in Tab. 1. In this table, 'BiRF' indicates binary radiance fields that is based on Instant-NGP with binary encoding [25], and 'Masked' represents the masked wavelet model that is based on the TensoRF with masking for wavelet coefficients [23]. To compare the performance of compression models, we assess the visual quality at three different rate levels: about 0.8MB (Rate-1), 1.4MB (Rate-2), and 2.5MB (Rate-3). In the low bitrate range (Rate-1), our proposed model achieved the highest performance across all datasets compared to the other compression models. We also present the performance of the masked wavelet model based on the configuration of VM-192 for Rate-1. It is important to note that the proposed model is based on the VM-192 configuration for all bitrate levels. This ensures a fair comparison with other TensoRF-based compression models.

In the middle bitrate range (Rate-2), BiRF achieves 0.13dB better visual quality than ours on the T&T dataset, with a slightly larger size. It is worth noting that BiRF is based on Instant-NGP, which demonstrates superior quality on the T&T dataset. Additionally, we report the performance of the masked wavelet model on VM-384 to achieve the better results at Rate-2 and -3. In VM-384, the tensor channel size is doubled from the default configuration to achieve the better results. Increasing the number of channels can increase the model capacity, but this configuration

DCT block	Total block size	Size (MB)↓	PSNR (dB)↑
2D DCT ($1\times8\times8$)	$T=64$	0.60/1.05	29.86/32.22
2D DCT ($1\times16\times16$)	$T=256$	0.63/1.08	31.81/33.04
2D DCT ($1\times32\times32$)	$T=1024$	0.69/1.08	33.21/33.55
3D DCT ($4\times4\times4$)	$T=64$	0.61/1.11	31.56/33.21
3D DCT ($8\times8\times8$)	$T=512$	0.68/1.05	33.14/33.53
3D DCT ($16\times16\times16$)*	$T=4096$	0.65/1.09	33.17/33.51

Table 3. **Ablation study: the impact of varying DCT block size** on the *Ficus* scene in the Synthetic NeRF dataset (low/high bitrates), where * indicates the default configuration. These experimental configurations are only for matrices **M**.

also leads to higher computational costs for both training and testing.

In the high bitrate range (Rate-3), our proposed model demonstrates better performance on the NSVF dataset, while BiRF shows better quality in the other two datasets. The base model clearly affects the compression performance results. Moreover, BiRF increases the number of feature dimensions F from 2 to 4 at Rate-3, which is twice the default configuration. As a result, BiRF is reported to have high performance, even surpassing the TensoRF and Instant-NGP baseline in terms of visual quality.

The rendering results for visual comparison of each compression model are shown in Fig. 4, where we illustrate that our proposed model can maintain better visual quality even in smaller sizes. Tab. 2 presents the comparison results between our proposed model and VQAD. VQAD employs vector quantization of the feature-grid by learning an integer codebook. Ours demonstrates better compression performance on the RTMV dataset.

4.3. Ablation Study

Effect of the Regularization Term. This section analyzes how the $\mathcal{L}_r$ term affects compression performance. This term encourages the transformed coefficient values to become zero, which leads to a significant reduction in size when combined with entropy minimization. However, the bitrate is highly sensitive to the weighting parameters λ_r and λ_e, making it difficult to analyze their effect. To avoid exhaustive search, we introduce a offset α, which automatically determines λ_r as $\lambda_r = \alpha\lambda_e$. As shown in Fig. 5, it can be observed that adding the $\mathcal{L}_r$ term improves compression performance. When the offset α is close to 1 ($\lambda_r = \lambda_e$), it performs more effectively. However, when the value of α is too large or too small, it can negatively impact compression performance. In addition, in Fig. 5, we illustrate the performance of entropy minimization in the spatial feature-grid domain instead of the frequency domain. It clearly demonstrates the advantages of entropy minimization in the frequency domain.

Impact of DCT Block Size. The impact of DCT size is another important aspect of the proposed model. The en-

VQRF
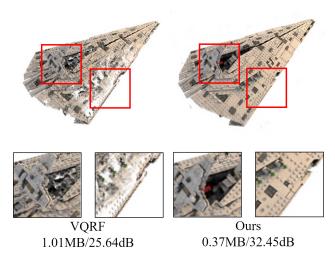1.01MB/25.64dB

Ours
0.37MB/32.45dB

Figure 6. **Compression artifacts** of VQRF [15] and our proposed model ($\lambda_e = 2e^{-10}$) for *Destroyer* scene. VQRF is a pruning-based compression model.

Method	Size (MB)↓	PSNR (dB)↑	SSIM↑
TensoRF-VM baseline, 30k	71.3MB	33.15	0.962
VQRF (fine-tuned) [15]	21.9MB	32.98	0.960
+8bit quantization	5.47MB	32.88	0.959
+Entropy coding	**3.52MB**	32.88	0.959
Ours, 30k ($\lambda_e = 1e^{-11}$)	71.6MB	33.12	0.961
+DCT, 8bit quantization	18.1MB	33.05	0.960
+Entropy coding	**2.51MB**	33.05	0.960
Ours, 30k ($\lambda_e = 2.5e^{-9}$)	71.6MB	32.45	0.957
+DCT, 8bit quantization	18.1MB	32.20	0.956
+Entropy coding	**0.80MB**	32.20	0.956

Table 4. **Comparison of VQRF [15] and our proposed model with the compression pipeline** discussed in 4.3. The experimental results are averaged across all scenes from the Synthetic NeRF dataset.

tropy and regularization term involves applying the block-wise DCT to the feature-grid. Therefore, it is conceivable that the size of the DCT block will impact the performance. This ablation study focuses on the *Ficus* scene, which shows clear trends for DCT block size. Tab. 3 shows the performance according to the DCT block size with matrices $\mathbf{M}$. It can be observed that a smaller total block size ($T < 512$) has a significantly negative impact on performance. However, if the total block size T is bigger than 512, there is no meaningful improvement regardless of the dimension. For vectors $\mathbf{v}$, the size of the DCT block does not have a significant impact. In addition, the size of the vectors is limited to applying a larger block, so we fixed 2D DCT 8×8 for vectors $\mathbf{v}$.

Compression Artifacts. Fig. 6 illustrates the reconstruction errors that occur at extremely low bitrates. For the proposed model, the optimized tensor signal only contains low-frequency components of the DCT, which leads to block artifacts in the feature-grid. The reconstructed scene suffers from a loss of high-frequency information. However, using a pruning model like VQRF for compression may result in the loss of spatial information. Our proposed model has the advantage of reducing size while preserving visual quality even at extremely low bitrates.

Compression Pipeline. While the feature-grid is optimized after training, sparse representations themselves do not decrease the total size. Our design follows the compression pipeline of VQRF, but the compression target of our model is the transformed coefficients. The experimental results with the compression pipeline are listed in Tab. 4. The VQRF model uses pruning and joint fine-tuning techniques to eliminate unnecessary weights without significant quality loss. Then, they apply 8-bit quantization to the feature-

grid, resulting in a size of approximately 5.4MB. Finally, entropy coding is applied to reduce the size of the data and pack it into a binary bitstream for transmission. In contrast, our proposed model directly minimizes the entropy with our loss function. After performing DCT and 8-bit quantization, applying entropy coding to the transformed coefficients leads to a significant reduction in size. Compared to the uncompressed baseline, our proposed model achieves a compression ratio of 28× with a negligible drop in PSNR of only 0.1 dB. The size of our proposed model is 0.8MB for higher λ_e, with a loss of 0.95dB.

Complexity Overhead. The training time for the proposed model (VM-192) is calculated using an NVIDIA A100 with 40 GB of memory. The DCT operation slightly increases the training time, while the entropy calculation significantly increases it. Since the entropy loss is applied after the feature-grid is fixed in coarse-to-fine training, our proposed model typically takes about 4 to 5 minutes longer than the baseline. The rendering time of the compressed model is comparable to the original one, as it only requires entropy decoding and inverse DCT.

5. Conclusion

We introduce ECRF, a novel compression framework specifically designed for tensorial radiance fields, but it can be applied to any grid-based neural fields. Unlike traditional compression methods, ECRF's primary goal is to directly minimize the entropy of the compression target. This optimization process takes place in the DCT coefficient domain, leading to more sparse representations for efficient compression. This is achieved through the utilization of a frequency-domain entropy parameterization, which is integrated into a compression pipeline that includes 8-bit quantization and entropy coding. Our proposed model effectively reduces the model size without compromising rendering quality, as demonstrated by experimental results showing superior performance, especially at low bitrates.

References

[1] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 4

[2] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 4

[3] Johannes Ballé, Philip A Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2020. 2

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 1

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 1

[6] Thomas Bird, Johannes Ballé, Saurabh Singh, and Philip A Chou. 3d scene compression through entropy penalized neural representation functions. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. 2

[7] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video technology (TCSVT)*, 31(10):3736–3764, 2021. 2

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 1, 3

[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 2, 3, 5, 6

[10] Chenxi Lola Deng and Enzo Tartaglione. Compressing explicit voxel grid representations: fast nerfs become also small. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1236–1245, 2023. 2, 5

[11] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 1

[12] Sharath Girish, Kamal Gupta, Saurabh Singh, and Abhinav Shrivastava. Lilnetx: Lightweight networks with EXtreme model compression and structured sparsification. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[13] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 6

[14] Glen G. Langdon. An introduction to arithmetic coding. *IBM Journal of Research and Development*, 28(2):135–149, 1984. 3

[15] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Liefeng Bo. Compressing volumetric radiance fields to 1 mb. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4222–4231, 2023. 2, 3, 5, 6, 8

[16] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in neural information processing systems (NIPS)*, 33:15651–15663, 2020. 6

[17] Michael W Marcellin, Michael J Gormish, Ali Bilgin, and Martin P Boliek. An overview of jpeg-2000. In *Proceedings DCC 2000. Data Compression Conference*, pages 523–541. IEEE, 2000. 2

[18] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems (NIPS)*, 33, 2020. 2

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 6

[20] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in Neural Information Processing Systems (NIPS)*, 31, 2018. 2

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 2, 5

[22] Deniz Oktay, Johannes Ballé, Saurabh Singh, and Abhinav Shrivastava. Scalable model compression by entropy penalized reparameterization. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[23] Daniel Rho, Byeonghyeon Lee, Seungtae Nam, Joo Chan Lee, Jong Hwan Ko, and Eunbyung Park. Masked wavelet representation for compact neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20680–20690, 2023. 2, 3, 5, 6, 7

[24] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022. 1

[25] Seungjoo Shin and Jaesik Park. Binary radiance fields. *arXiv preprint arXiv:2306.07581*, 2023. 2, 5, 7

[26] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems (NIPS)*, 33:7462–7473, 2020. 1

[27] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology (TCSVT)*, 22(12):1649–1668, 2012. 2

[28] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, 2022. 1, 2

[29] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Improved direct voxel grid optimization for radiance fields reconstruction. *arXiv preprint arXiv:2206.05085*, 2022. 1, 2

[30] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11358–11367, 2021. 6

[31] Towaki Takikawa, Alex Evans, Jonathan Tremblay, Thomas Müller, Morgan McGuire, Alec Jacobson, and Sanja Fidler. Variable bitrate neural fields. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 6

[32] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems (NIPS)*, 33:7537–7547, 2020. 1, 2

[33] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *Advances in neural information processing systems (NIPS)*, 35:14798–14809, 2022. 2

[34] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 1

[35] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, and Stan Birchfield. Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *IEEE/CVF European Conference on Computer Vision Workshop (Learn3DG ECCVW), 2022*, 2022. 1, 6

[36] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992. 2

[37] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 2

[38] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13524–13534, 2022. 2

[39] Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. Neural residual radiance fields for streamably free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 76–87, 2023. 2

[40] Zhijie Wu, Yuhe Jin, and Kwang Moo Yi. Neural fourier filter bank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14153–14163, 2023. 2

[41] Yibo Yang, Stephan Mandt, and Lucas Theis. An introduction to neural data compression. *Foundations and Trends in Computer Graphics and Vision*, 15(2):113–200, 2023. 2

[42] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5752–5761, 2021. 1

[43] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1

[44] Tianli Zhao, Jiayuan Chen, Cong Leng, and Jian Cheng. Tinynerf: Towards 100 x compression of voxel radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3588–3596, 2023. 2, 5