Hybrid Representation for 4D Medical Image Compression

Wuyang Zheng[1⋆], Jiarui Meng[1⋆], Jiaqi Zhang[2†], Jian Zhang[1], Siwei Ma[2]
[1]School of Electronic and Computer Engineering, Peking University, China
[2]School of Computer Science, Peking University, China
{shooterhex, mengjiarui}@stu.pku.edu.cn, {jqzhang, zhangjian.sz, swma}@pku.edu.cn

***Abstract*—Due to the substantial storage requirements of the 4D medical images, achieving efficient compression of such images is a crucial topic. Existing traditional image/video coding methods have achieved remarkable results in most compression tasks, but their performance in encoding 4D medical images remain poor. This is because these methods cannot fully exploit the spatio-temporal correlations in 4D images. Recently, implicit neural representation (INR) based image/video compression methods have made significant progress, with coding performance comparable to traditional methods. However, they also suffer from significant performance losses in 4D medical image compression like traditional methods. In this paper, we propose an efficient hybrid representation framework, which includes six learnable feature planes and a tiny MLP decoder. This framework alleviates the issue of previous methods lacking the ability to utilize the spatio-temporal correlations in 4D medical images, enabling it to capture these information more effectively. We also introduce a novel adaptive plane scaling strategy that allocates the numbers of parameter in each plane based on the resolution of the image. This design allows the model to further enhance the reconstruction quality at the same compression ratio. Extensive experiments show that our model achieves better RD performance compared to traditional and INR-based methods, and it also offers faster encoding speeds than INR-based methods.**

***Index Terms*—Medical Data Compression, Hybrid Representation, Feature Planes**

I. Introduction

With advancements in magnetic resonance imaging (MRI) technology, 4D medical images are increasingly being used in clinical applications. These images not only contain the three-dimensional spatial information of organs (width, height, and depth) but also incorporate the time dimension, allowing for the dynamic display of organ and tissue activities. Although these images contain richer information about organ and tissue activities, their higher dimensionality and the usage of high bit-depth formats results in substantially larger storage requirement compared to normal 2D images. For instance, in the Chinese Human Connectome Project [1], storing raw fMRI images of 365 individuals requires 1.85 TB of storage. Such data scales present significant challenges for data storage and transmission, necessitating efficient compression frameworks for these types of medical images.

⋆ means equal contribution and † means corresponding author. This work was supported in part by the National Natural Science Foundation of China no. U21B2012, in part by the China Postdoctoral Science Foundation under Grant no. 2023M740080 and in part by the Postdoctoral Fellowship Program of CPSF under Grant no. GZC20230059.

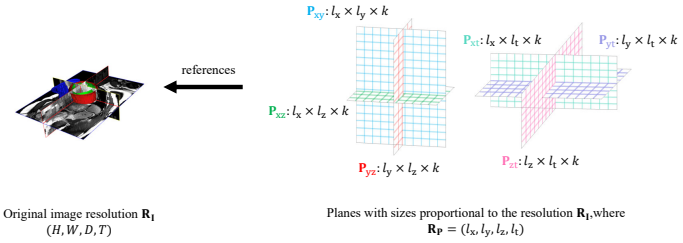

Fig. 1. Adjusting the size of the planes according to the image resolution.

In recent decades, traditional video coding methods have advanced rapidly, with well-known codecs such as H.264 [2] and HEVC [3]. These codecs utilize advanced block partitioning, intra- and inter-frame prediction, and entropy coding to achieve efficient compression of various video types. Since 3D medical images and videos both possess three dimensions, these codecs can also be employed to encode 3D medical images. Besides, some image compression approaches [4]–[6] have made significant progress in 3D medical image compression, which are primarily based on wavelet transform. Among them, JP3D [4] is the most widely used. It extends the 2D wavelet transform [7] into a 3D wavelet transform, achieving efficient compression of 3D data, including 3D medical images. Wavelet transforms can capture more high-frequency signals, such as edges and textures, compared to the DCT transforms used in video coding. However, these image/video coding methods are originally designed for 3D image/video data. When encoding 4D medical images, they must convert the 4D image into 3D slices to encode properly. This approach inevitably loses some inter-dimensional information and cannot fully exploit the spatio-temporal correlations existing among the four dimensions, leading to suboptimal compression performance.

Recently, implicit neural representation (INR) [8], [9] has emerged as a new paradigm for image compression. These methods transform the image compression problem into a model compression problem. [10] further improves the encoding performance of INR-based methods, making them comparable to conventional image codecs [7]. [11], [12] propose an MLP-based framework for compressing 3D medical images. They employ spectral analysis and adaptive block partitioning to allocate the MLP network parameters, allowing for better fitting of high-frequency signals within the images. Despite

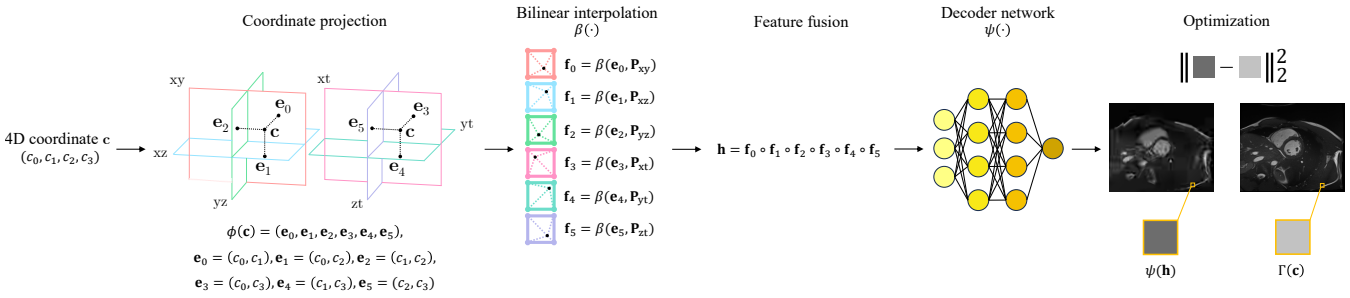

Fig. 2. The overall framework of the proposed hybrid representation for 4D medical image compression.

[11] and [12] demonstrating excellent encoding performance on 3D medical images, they must slice the 4D images like traditional image/video encoding methods, which significantly compromises their encoding performance. Additionally, these INR-based methods require a long encoding time, making them unsuitable for deployment in practical applications.

In the field of radiance field reconstruction, INR-based methods [13], [14] have demonstrated their excellent rendering performance. However, their main bottleneck is the long training time required to achieve high-quality rendering. To address this, some works based on hybrid representations [15]–[17] are proposed. These methods achieve the same rendering quality while offering faster training times, also exhibiting strong compactness and interpretability [16], [17]. Due to the excellent performance of [16] and [17] in dynamic radiance field reconstruction tasks, and considering the similar spatio-temporal properties between 4D medical images and dynamic radiance fields, we believe that such architectures will also perform well in 4D medical image compression. Based on this insight, we propose an hybrid representation framework, which comprises six learnable feature planes and a tiny MLP decoder. Additionally, we introduce a novel adaptive plane scaling strategy, which adjusts the sizes of each plane based on the resolution of the 4D medical image. We empirically confirm that this design further enhances encoding performance. Experimental results demonstrate that our framework achieve better RD performance compared to other methods. Meanwhile, benefiting from the compactness of the feature plane representation, our method offers faster encoding speeds compared to existing INR-based methods. In summary, our main contributions are as follows:

- We propose an efficient hybrid representation framework for 4D medical image compression.
- We introduce an adaptive plane scaling strategy, which can further improve the encoding performance at the same image compression ratio.
- Experiments show our model achieves superior performance over other methods and significantly reduce encoding time compared with INR-based methods.

II. Proposed Method

A. Overview of the proposed model

The framework of the proposed model is shown in Figure 2, which consists of six planes (xy, xz, yz, xt, yt, zt) and a tiny MLP for decoding, where the xy, xz, and yz planes extract the spatial information of the input image, while the xt, yt, and zt planes capture its temporal motion information. After obtaining the feature vectors from these six planes, the MLP decodes the features to produce the final output of the model. Our pipeline is similar to other hybrid representations [16], [17], but the MLP does not take into account the view direction and only outputs the grayscale values of the 4D medical image.

Given a 4D coordinate $\mathbf{c} = (c_0, c_1, c_2, c_3)$ of an input image which is normalized to [-1, 1], we first perform a projection operation ϕ on $\mathbf{c}$. This operation splits the coordinate $\mathbf{c}$ into six 2D coordinates, projecting $\mathbf{c}$ onto six different planes. We denote the six resulting coordinates as $\mathbf{e}_0$ to $\mathbf{e}_5$. These coordinates are then sent to their respective feature planes $\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{M \times N \times k}$, with M, N and k representing the number of rows, columns, and channels of the planes, respectively. Since $\mathbf{e}_0$ to $\mathbf{e}_5$ are normalized coordinates, they can extract local features and combine information from their four nearest neighboring points in these learnable feature planes through bilinear interpolation β. The results of β are denoted as $\mathbf{f}_0$ to $\mathbf{f}_5$, which are the features output by each plane after projecting coordinate $\mathbf{c}$. Here, $\mathbf{f}_0, \mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_5 \in \mathbb{R}^k$. To effectively utilize the features from each plane, we need to perform an aggregation operation. There are various aggregation methods, including summation, product, and networks. Empirically, we find that element-wise multiplication maximizes the representational capacity of the model. Therefore, we aggregate the features $\mathbf{f}_0$ to $\mathbf{f}_5$ using element-wise multiplication $\circ$. The result is denoted as $\mathbf{h}$, which can be expressed as

$$\mathbf{h} = \mathbf{f}_0 \circ \mathbf{f}_1 \circ \mathbf{f}_2 \circ \mathbf{f}_3 \circ \mathbf{f}_4 \circ \mathbf{f}_5 \tag{1}$$

B. Adaptive scaling planes

We observe that different 4D medical images exhibit variations in resolution across the four dimensions (width, height, depth, and time). Some images may have a larger resolution in temporal dimension while some may have an unusual aspect ratio. Based on this observation, we design a novel plane parameter allocation strategy that adjusts the sizes of each plane according to the resolution of the original image and a given compression ratio.

We define the resolution $\mathbf{R_P} = (l_x, l_y, l_z, l_t)$, which represents the height, width, depth, and time dimension of the planes, respectively. We can determine the resolution for

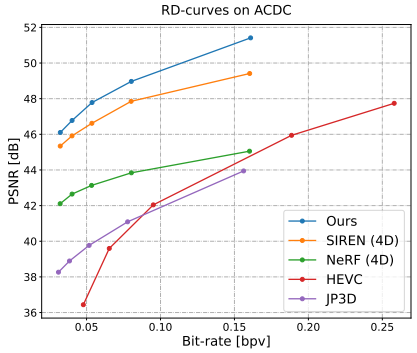

Fig. 3. The RD performance of different methods on the ACDC dataset

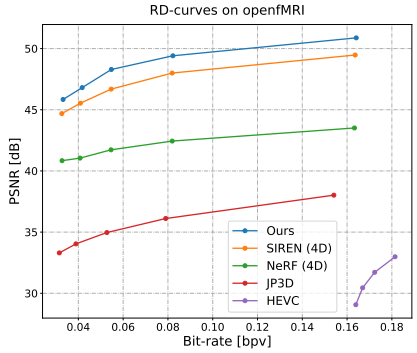

Fig. 4. The RD performance of different methods on the openfMRI dataset

each of the six planes through $\mathbf{R_P}$. For example, $\mathbf{P}_{xy}$ has a resolution of $l_\mathrm{x} \times l_\mathrm{y}$, and $\mathbf{P}_{xz}$ has a resolution of $l_\mathrm{x} \times l_\mathrm{z}$ and so on. Assuming the original image occupies a storage space of size S, the current compression ratio is r, and the output feature dimension of a plane is k. The proportional relationship between the size of the original image and the total size of the six planes can be derived by

$$(l_\mathrm{x}l_\mathrm{y} + l_\mathrm{x}l_\mathrm{z} + l_\mathrm{x}l_\mathrm{t} + l_\mathrm{y}l_\mathrm{z} + l_\mathrm{y}l_\mathrm{t} + l_\mathrm{z}l_\mathrm{t}) \times k = \frac{S}{r}. \tag{2}$$

We believe that if one dimension of a 4D medical image is significantly larger than the others, the size of its corresponding plane should reflect this as well. For instance, in an image with a resolution of $512 \times 512 \times 10 \times 30$, the width and height dimensions are clearly larger than the other two dimensions. Consequently, the resolution of the planes related to the width and height should also be larger. Based on this principle, we set the following constraints for the plane resolution $\mathbf{R_P}$. As shown in Fig. 1, to ensure that $\mathbf{R_P}$ is proportional to the resolution of the original image, we assume that the resolution of the image is $\mathbf{R_I} = (H, W, D, T)$. We then define $U = l_\mathrm{x} + l_\mathrm{y} + l_\mathrm{z} + l_\mathrm{t}$ and $V = H + W + D + T$, where U and V represent the sum of all dimensions in the planes and the original image, respectively. Combining this with Eq. (2), we can obtain the following:

$$\begin{cases} (l_\mathrm{x}l_\mathrm{y} + l_\mathrm{x}l_\mathrm{z} + l_\mathrm{x}l_\mathrm{t} + l_\mathrm{y}l_\mathrm{z} + l_\mathrm{y}l_\mathrm{t} + l_\mathrm{z}l_\mathrm{t}) \times k = \dfrac{S}{r}, \\ \dfrac{l_\mathrm{x}}{U} = \dfrac{H}{V}, \quad \dfrac{l_\mathrm{y}}{U} = \dfrac{W}{V}, \quad \dfrac{l_\mathrm{z}}{U} = \dfrac{D}{V}, \quad \dfrac{l_\mathrm{t}}{U} = \dfrac{T}{V}. \end{cases} \tag{3}$$

Through Eq. (3), we can obtain the specific sizes for each dimension of the planes. We demonstrate empirically that this design exhibits better expressive power compared to the version where the sizes of each plane are equally divided.

C. MLP decoder and optimization

After obtaining the feature vectors from the plane outputs, we decode them using a tiny MLP decoder ψ. The output of ψ is the grayscale value fitted by the model, which is optimized by calculating the photometric loss with respect to the original 4D medical image. The optimization objective is defined as

$$\mathcal{L} = \sum_{\mathbf{c}} \|\Gamma(\mathbf{c}) - \psi(\mathbf{h})\|_2^2, \tag{4}$$

where $\Gamma(\mathbf{c})$ is the ground-truth grayscale value at coordinate $\mathbf{c}$ and $\psi(\mathbf{h})$ denotes the predicted grayscale value at the same location in the original image, respectively. The MLP decoder consists of two layers. The activation function of the first layer is an exponential function with gradient clipping, and the second layer uses ReLU [18]. Empirically, we find this design effectively prevents the gradient vanishing and explosion issues during training.

D. Differences with video coding methods

Our model and other INR-based methods can be collectively referred to as model-based methods. Compared to video coding methods, model-based methods differ in two key aspects: (i) Unlike video coding methods that output a bitstream after encoding, model-based methods directly save their parameters in binary format to a bitstream after training. Once the model parameters are reloaded locally on the decoding side, all coordinates of the 4D medical image can be sampled, and a single inference is enough to reconstruct the image; (ii) While video coding methods use λ to control the compression level in rate-distortion optimization, model-based methods control the number of network parameters using the compression rate r mentioned in II-B. Additionally, model-based methods require network re-initialization and re-training for different r values, as changes in r lead to variations in the number of network parameters.

III. Experiments

A. Datasets

We evaluate our method on the ACDC [19] and openfMRI [20] datasets. The ACDC dataset contains a total of 150 cardiac MRI images. We only use images with high bit depth (bit depth $>$ 8 bits) from the training set for our experiments, totaling 35 images. The openfMRI is another dataset that focuses on brain imaging. We choose the ds000007 and ds000101 from the openfMRI dataset for our experiments, with a total of 41 images, all stored in high bit depth format.

B. Experiment settings

The proposed model is compared with two types of compression methods: (i) traditional compression methods, including HEVC [3] and JP3D [4]; and (ii) INR-based compression

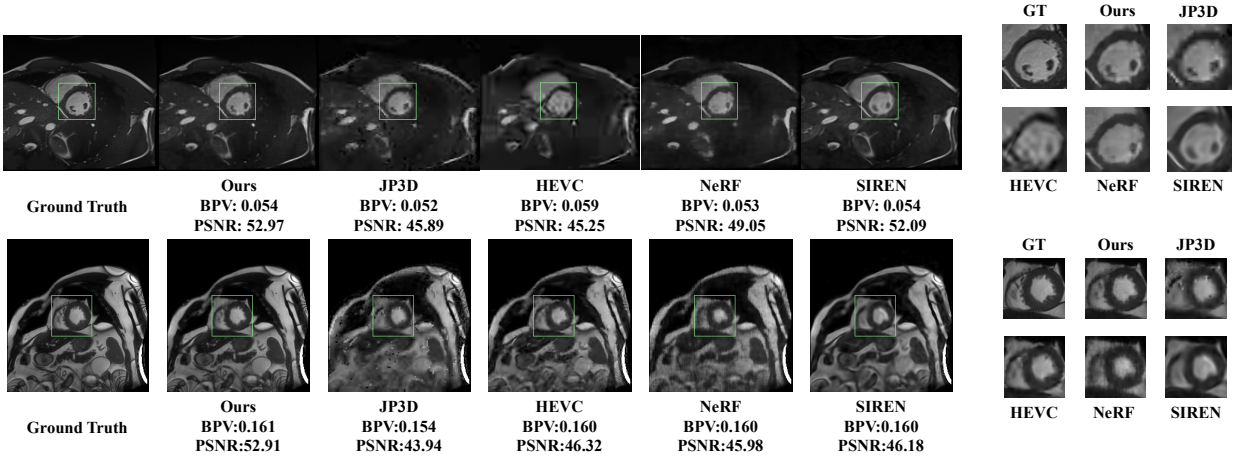

Fig. 5. The subjective quality results from baseline codecs. The images above belong to patient062, while the ones below belong to patient021.

methods, including NeRF [13] and SIREN [14]. We use peak signal-to-noise ratio (PSNR) and bits per voxel (bpv) to measure the encoding performance. The bpv is calculated by dividing the size of encoded bitstream by the resolution of the 4D medical image.

For traditional methods, they cannot compress 4D medical images directly. Therefore, we slice all 4D images along the time dimension, extracting the three spatial dimensions (xyz). Then we treat the z-dimension as the time axis and compress it using HEVC and JP3D. For INR-based methods and our proposed model, we first normalize the grayscale range of the 4D images to [0, 100], and the image coordinates to [-1, 1]. We modify the input layer of NeRF and SIREN, in order to support 4D coordinates as input. For NeRF, we set the frequency hyper parameter of positional encoding to 10. For SIREN, we set its hyper parameter w_0 to 20. Both models utilize a 5-layer MLP network structure. For our approach, we employ the adaptive plane scaling strategy. The training process is conducted on a single NVIDIA RTX 4090 GPU using the Adam [21] optimizer.

C. Quantitative results

Figure 3 and 4 illustrate the RD performance of the baseline codecs on the ACDC and openfMRI datasets, respectively. We observe that HEVC exhibits relatively poor encoding performance on both datasets. We attribute this to the fact that the xyz slices are essentially traversals of slices in human organ tissues, resulting in significant movement of the screen contents between adjacent slices. This massive screen contents movement hinders the ability of HEVC to exploit inter-frame similarities for further compression, leading to suboptimal compression performance. Our method not only addresses this issue but also outperforms other INR-based methods in terms of encoding performance.

Table I presents the average encoding and decoding times for INR-based methods and our approach on the ACDC and openfMRI datasets, with HEVC used as the anchor for BD-rate calculation. Due to the compact nature of the multi-plane representation, our method converges within fewer training epochs compared to INR-based methods, resulting in faster encoding times. Moreover, our method surpasses other INR-based methods in terms of image reconstruction quality.

TABLE I
ENCODING/DECODING TIME AND BD-RATE USING HEVC AS ANCHOR

Dataset	Method	Encoding Time (s)	Decoding Time (s)	BD-Rate
ACDC [19]	NeRF	771.99	0.36	-41.98%
	SIREN	508.58	**0.01**	-74.80%
	Ours	**228.84**	0.61	**-80.67%**
openfMRI [20]	NeRF	903.29	0.67	-87.82%
	SIREN	614.10	**0.02**	-95.55%
	Ours	**288.51**	0.97	**-96.42%**

D. Qualitative results

Figure 5 presents the subjective quality comparison of images encoded by the baseline codecs from the ACDC dataset. We select patient062 and patient021 from the ACDC dataset as ground truth, with pixel intensity ranges of [128, 1128] for patient062 and [128, 4183] for patient021. The bit rates used for these images are approximately 0.054 and 0.160 bpv, respectively. It can be observed that our method maintains consistent encoding performance across images with different intensity ranges, without significant performance degradation.

IV. CONCLUSION

In this work, we propose a novel six-plane hybrid representation framework, where each feature plane learns the relationships between specific dimensions in 4D medical images. We also employed a tiny MLP decoder to transform the features into the model output. Additionally, we introduced an adaptive plane scaling strategy, which further enhances the representational capacity of the model while maintaining the same model size. Experimental results demonstrated that our model not only outperformed traditional and INR-based methods in terms of 4D medical image encoding performance but also significantly reduced the encoding time compared to other INR-based methods.

REFERENCES

[1] J. Ge, G. Yang, M. Han, S. Zhou, W. Men, L. Qin, B. Lyu, H. Li, H. Wang, H. Rao *et al.*, "Increasing diversity in connectomics with the Chinese Human Connectome Project," *Nature Neuroscience*, vol. 26, no. 1, pp. 163–172, Jan. 2023.

[2] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[4] P. Schelkens, A. Munteanu, A. Tzannes, and C. Brislawn, "JPEG2000. Part 10. Volumetric data encoding," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, 2006.

[5] D. Xue, H. Ma, L. Li, D. Liu, and Z. Xiong, "iWave3D: End-to-end Brain Image Compression with Trainable 3-D Wavelet Transform," in *Proceedings of International Conference on Visual Communications and Image Processing (VCIP)*, 2021.

[6] D. Xue, H. Ma, L. Li, D. Liu, and Z. Xiong, "aiWave: Volumetric Image Compression With 3-D Trained Affine Wavelet-Like Transform," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 606–618, Mar. 2023.

[7] M. Marcellin, M. Gormish, A. Bilgin, and M. Boliek, "An overview of JPEG-2000," in *Proceedings of Data Compression Conference (DCC)*, 2000.

[8] E. Dupont, A. Goliński, M. Alizadeh, Y. W. Teh, and A. Doucet, "COIN: COmpression with Implicit Neural representations," *arXiv preprint arXiv:2103.03123*, 2021.

[9] E. Dupont, H. Loya, M. Alizadeh, A. Golinski, Y. W. Teh, and A. Doucet, "COIN++: Neural compression across modalities," *Transactions on Machine Learning Research*, 2022.

[10] Y. Strümpler, J. Postels, R. Yang, L. V. Gool, and F. Tombari, "Implicit Neural Representations for Image Compression," in *Proceedings of Computer Vision – ECCV*, 2022.

[11] R. Yang, T. Xiao, Y. Cheng, Q. Cao, J. Qu, J. Suo, and Q. Dai, "SCI: A Spectrum Concentrated Implicit Neural Compression for Biomedical Data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[12] R. Yang, "TINC: Tree-Structured Implicit Neural Compression," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Proceedings of Computer Vision – ECCV*, 2020.

[14] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit Neural Representations with Periodic Activation Functions," in *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[15] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "TensoRF: Tensorial Radiance Fields," in *Proceedings of Computer Vision – ECCV*, 2022.

[16] A. Cao and J. Johnson, "HexPlane: A Fast Representation for Dynamic Scenes," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-Planes: Explicit Radiance Fields in Space, Time, and Appearance," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.

[19] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[20] R. A. Poldrack and K. J. Gorgolewski, "OpenfMRI: Open sharing of task fMRI data," *NeuroImage*, vol. 144, no. Pt B, pp. 259–261, Jan. 2017.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of 3rd International Conference on Learning Representations (ICLR)*, 2015.