TensorGRAF: Tensorial Generative Radiance Field

Pin-Chieh Yu[1], Der-Lor Way[2] and Zen-Chung Shih[1]
[1]Institute of Multimedia Engineering, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[2]Department of New Media Art, Taipei National University of the Arts, Taipei 112, Taiwan
pinjayyu@gmail.com, adlerway@gmail.com, zcshih@cs.nycu.edu.tw

***Abstract*—We propose using a voxel grid as the explicit representation of the radiance field, combining a shallow network to interpret the spacial features. The voxel is further decomposed into axis-align feature vectors using the tensor decomposition technique. Therefore, the space complexity of synthesis is reduced from $O(n^3)$ to $O(n)$. We also benefit from the mature 2D generative adversarial network and utilize the network structure in our 1D feature vector generator.**

I. INTRODUCTION

Recently, 3D-aware generative method that based on neural radiance field are widely explored. However, the inherent characteristics of volume rendering and deep neural networks produce poor training and execution speeds. However, instead of direct 3D generation, recent 3D-aware GANs have shown preliminary success on multi-view-consistent image synthesis. This category of GANs combines 3D-structure aware generators, differentiable renderers, and adversarial training processes to capture 3D information. The most iconic pioneers are models that based on Neural Radiance Fields (NeRF) [1]. NeRFs are robust to single scene representation, but they are slow to query the coordinates and thus impracticable for high-resolution image generation. Some researches take advantage of 2D upsampling networks to enhance the image quality, but doing so damages the 3D consistency of scenes.

II. METHOD AND EXPERIMENTS

Our goal is to embrace the efficiency of the voxel representations, but at the same time escape from memory loading. Inspired by TensoRF [2], we model the voxelated radiance fields as a 4D tensor. We then apply traditional CP decomposition by factorizing the tensor into three rank-one components. The space complexity of our proposed generator is thus reduced to one dimension, providing a better scalability than previous works. Our final model attains comparable image fidelity to recent state-of-the-art 3D-aware GANs, and reduces one-third of memory usage while training. The full network architecture is shown in Figure 1. G denotes the feature generator, S denotes the 2D super resolution network, and D denotes the discriminator, respectively. $f \in R^{32}$ and $\sigma \in R$ are the feature vector and the density decoded by the feature decoder.

Figure 2 demonstrates the multi-view consistency result of our network. Our proposed method can capture the underlying 3D information from the dataset and produce multi-view consistent results without direct 3D supervision. We also compare the run-time rendering speed of our model in Table 1. The speed is evaluated in millisecond per image. Because our network is roughly twice as fast as EG3D's, there is a decent trade-off between synthesis quality and rendering speed.

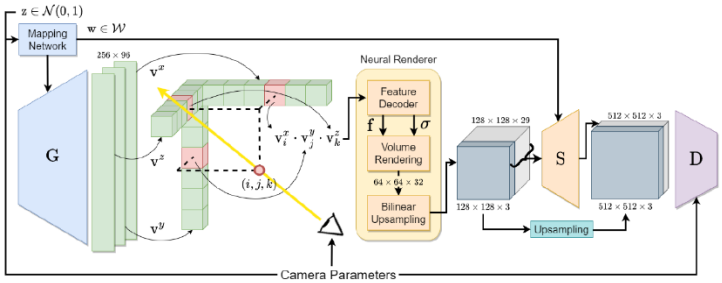

Figure 1: Network Architecture

Figure 2: Experimental results of Multi-View Consistency.

Table 1: Comparison with previous methods.

	FFHQ	AFHQ	SP
π-GAN	85	47	608
GIRAFFE	31.5	16.1	5
EG3D	4.8	3.9	27
VoxGRAF	14.4	9.6	200
Our Method	12.7	6.8	14

ps: The FID for FFHQ and AFHQ Cats, the rendering speed per image.

III. CONCLUSIONS

Our approach leverages both the expressiveness of the implicit network and the run-time efficiency of the explicit voxel grid. To further reduce the memory cost introduced by the explicit voxel grid representation, we utilize tensor decomposition to factorize the voxel into three 2D vector components. Our generator adopts the network structure of the state-of-the-art style-based 2D generator, and thus inherits the expressiveness of the structure. To avoid 3D inconsistency caused by the 2D super-resolution network, we augment the input channels of our discriminator to 6 channels to pass in both the rendering result of raw volume rendering and the neural rendering.

REFERENCE

[1] 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 European Conference on Computer Vision (EECV), 2020.

[2] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "TensoRF: Tensorial radiance fields," in European Conference on Computer Vision (ECCV), 2022.