CoordNet: Data Generation and Visualization Generation for Time-Varying Volumes via a Coordinate-Based Neural Network

Jun Han and Chaoli Wang, *Senior Member, IEEE*

Abstract—Although deep learning has demonstrated its capability in solving diverse scientific visualization problems, it still lacks generalization power across different tasks. To address this challenge, we propose CoordNet, a single coordinate-based framework that tackles various tasks relevant to time-varying volumetric data visualization without modifying the network architecture. The core idea of our approach is to decompose diverse task inputs and outputs into a unified representation (i.e., coordinates and values) and learn a function from coordinates to their corresponding values. We achieve this goal using a residual block-based implicit neural representation architecture with periodic activation functions. We evaluate CoordNet on data generation (i.e., temporal super-resolution and spatial super-resolution) and visualization generation (i.e., view synthesis and ambient occlusion prediction) tasks using time-varying volumetric data sets of various characteristics. The experimental results indicate that CoordNet achieves better quantitative and qualitative results than the state-of-the-art approaches across all the evaluated tasks.

Index Terms—Volume visualization, implicit neural representation, data generation, visualization generation

1 Introduction

In recent years, the scientific visualization community has witnessed the power of deep learning in processing various visualization tasks [36]. Examples include data generation [10], [23], [42] and visualization generation [5], [14], [39]. However, these proposed neural network models only fit in one particular task without the generalization capability across different tasks, limiting their usefulness. Yet model generalization over tasks represents an essential step toward general artificial intelligence and makes deep learning-based solutions more practical. Therefore, in this paper, we propose a single learning-based architecture that processes diverse tasks, including data generation and visualization generation.

Designing this generalized framework poses several challenges. First, a unified data formulation should be introduced to represent diverse data types, including volumes and images, which could be time-varying (i.e., data pattern changes across time). This ensures the framework uses consistent inputs rather than customized headers to process different data types. Second, unlike previous single-task solutions, where various network structures (e.g., encoder, decoder, and encoder-decoder) and specific designs can be proposed to solve one particular task, a single network structure capable of handling different tasks needs to be established. Third, the framework should be flexible to fit into various data resolutions with high quality. For example, synthesizing images with different resolutions (e.g., 256, 512, and $1,024$).

- *J. Han is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong 518172, China. E-mail: hanjun@cuhk.edu.cn.*
- *C. Wang is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, U.S.A. E-mail: chaoli.wang@nd.edu.*

To address these challenges, we present CoordNet (i.e., coordinate-based neural network), a single deep learning approach for handling diverse scientific visualization tasks. We formulate various data types as a unified representation, i.e., a set of coordinates and their values. Then, we utilize an encoder-decoder based *implicit neural representation* (INR) independent of the data resolution to learn the mapping from coordinates to values. Specifically, the encoder extracts a dense representation from the coordinate, and the decoder predicts the value at the coordinate from the representation. Utilizing task-driven objective functions, CoordNet can learn this mapping effectively and accurately. Once CoordNet is trained, it can explore different coordinate spaces to produce the corresponding values based on the required task. For example, spatial coordinates for spatial super-resolution, temporal coordinates for temporal super-resolution, view coordinates for view synthesis, etc.

We qualitatively and quantitatively evaluate our approach on four tasks (two data generation tasks and two visualization generation tasks) using several data sets with various characteristics. The two data generation tasks are *temporal super-resolution* (TSR) and *spatial super-resolution* (SSR). The two visualization generation tasks are *view synthesis* (VS) and *ambient occlusion prediction* (AOP). For each task, we compare CoordNet against the state-of-the-art approaches. Our results show that CoordNet achieves better visual quality in direct volume rendering and isosurface rendering. It yields better quantitative scores using the data-, image-, and surface-level metrics.

The contributions of CoordNet are summarized as follows. First, our work is the first generalized framework for processing diverse scientific visualization tasks. Unlike previous deep learning solutions tailored for a single application, CoordNet handles different tasks without changing the network architecture. Second, instead of leveraging con-

volutional neural networks (CNNs), we design a powerful, lightweight, yet simple architecture based on INR with periodic activation functions. Third, we comprehensively evaluate our approach on four tasks: TSR, SSR, VS, and AOP.

2 Related Work

This section discusses the related works of deep learning in volume visualization, implicit neural representation, and using a single model for processing diverse tasks.

Deep learning in volume visualization. Deep learning has swept the scientific visualization community in various tasks, including data generation and visualization generation. Here we review related work in volume visualization. For data generation, Han and Wang designed generative adversarial networks (GANs) for generating super-resolution for time-varying volumetric data at temporal [7] and spatial [8] domains. Han et al. [10] later presented an end-to-end solution for generating spatiotemporal volumes. Lu et al. [23] introduced a multilayer perceptron to compress volumetric data. For visualization generation, Berger et al. [2] explored the volume rendering image space through different transfer functions and view parameters using generative models. He et al. [14] designed a generative framework to synthesize rendering images by exploring the parameter spaces (e.g., view, ensemble, and isovalue). Engel and Ropinski [5] built a 3D U-Net to predict *local ambient occlusion* (LAO) volume given an intensity volume and transfer function. Weiss et al. [38] applied a CNN to upscale isosurface rendering images by predicting the visual representations (e.g., normal and depth). Han and Wang [9] developed a GAN-based solution for volume completion that synthesizes missing subvolumes using the adversarial and volumetric losses.

All the above works are task-specific; namely, they are tailored only for one application, which does not have the capability to handle different tasks. Therefore, instead of designing task-specific solutions, we propose a single framework (i.e., CoordNet) to process diverse tasks.

INR. INR or coordinate-based representation aims to parameterize a signal as a continuous function that maps the domain of the signal (i.e., coordinate) to the value at that coordinate (e.g., RGB color of an image). Sitzmann et al. [34] introduced a scene representation network that encodes geometry and appearance to reconstruct objects. Mildenhall et al. [25] utilized a fully-connected deep network to predict color and density values given 3D spatial locations and view parameters. Sitzmann et al. [32] combined INR and meta-learning to learn the 3D shape space. Chan et al. [3] utilized neural representation and neural volume rendering to synthesize images under different views. Guo et al. [6] learned object-centric neural scattering functions to synthesize photorealistic scenes.

Instead of using INR for object reconstruction and shape learning, we leverage this technique to solve data generation and visualization generation tasks in scientific visualization.

Single model for processing diverse tasks. Hashimoto et al. [12] designed a joint many-task model for learning multiple natural language processing tasks. McCann et al. [24] formulated ten natural language processing tasks as question answering over a context and proposed a long short-term memory-based framework to solve these tasks. Kaiser et al. [18] presented an encoder-decoder network that solves translation, image captioning, speech recognition corpus, and English parsing tasks. Lu et al. [22] leveraged ViLBERT to learn four different vision and language-related tasks on large-scale data sets. Pramanik et al. [28] introduced OmniNet to perform the tasks of part-of-speech tagging, image captioning, visual question answering, and video activity recognition.

Our work differs from the above ones. Instead of focusing on the multiple classification and detection tasks in computer vision and natural language processing, CoordNet aims to tackle diverse data generation and visualization generation tasks in scientific visualization.

3 Background: INR

CNNs utilize weight-sharing kernels to extract hidden representations by accumulating information from neighborhoods. The applications of CNNs include data generation [7], [11] and visualization generation [2], [14]. However, there are several disadvantages associated with CNNs. First, the performance heavily relies on the receptive fields. Therefore, CNNs may not process high resolution (e.g., 512^3) data well since modern GPU memory cannot afford such a resolution. Second, CNNs treat the signals as discrete ones (e.g., volumes are discrete grids of voxels). That means the convolutional operation only performs on these grids rather than in arbitrary locations. Third, the design of CNNs is task-dependent. For example, VS requires a pure decoder structure since the network aims to transform 1D representations into 2D representations (i.e., view parameters (ϕ and θ) to the corresponding images) [14]. In contrast, TSR follows an encoder-decoder framework [7], which encodes the input time steps into hidden features and decodes the learned representations to the intermediate time steps.

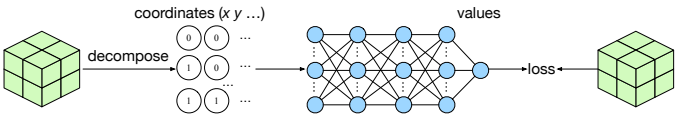


Fig. 1: Overview of CoordNet. Taking coordinates as input, CoordNet predicts the values at these coordinates and computes the difference between the prediction and GT.

INR leverages multiple fully-connected layers to map coordinates to their corresponding values. Compared with CNNs, INR offers several benefits. First, as a coordinate-based representation, INR is not coupled with spatial resolution, which means it can process data with arbitrary resolution. Second, INR regards the signal as continuous functions, interpolating data in arbitrary parameter spaces (e.g., spatial and temporal). Third, INR operates on different tasks by building an encoder-decoder framework. For instance, both TSR and VS can be formulated as a mapping from coordinate to value. Namely, $(x, y, z, t) \rightarrow v$ for TSR and $(x, y, \theta, \phi) \rightarrow (r, g, b)$ for VS.

4 CoordNet

This section first provides an overview of our proposed approach, then introduces the network architecture, training and inference details, and finally discusses the objective function.

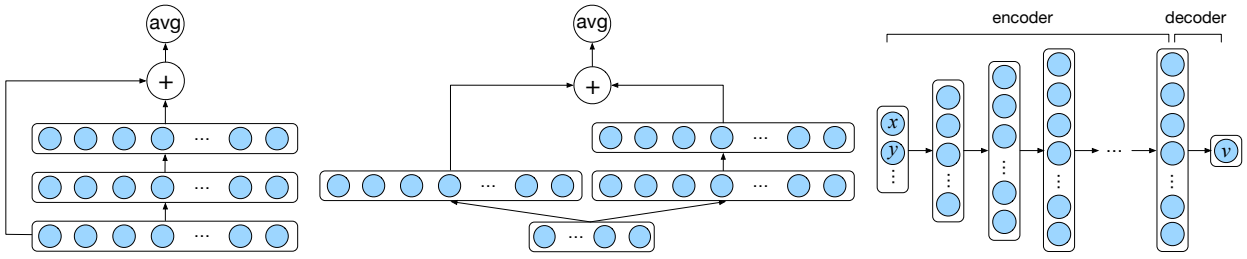


Fig. 2: (a) A traditional SIREN-based residual block, where the input dimension equals the output dimension. (b) A SIREN-based residual block, where the input dimension does not equal the output dimension. (c) The architecture of CoordNet is composed of multiple SIREN-based residual blocks.

4.1 Overview

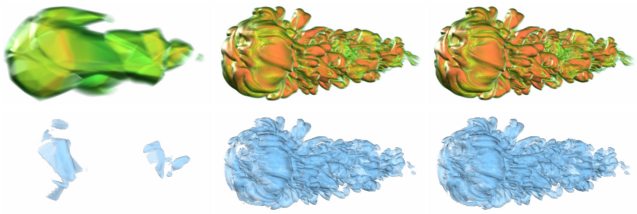
Given a set of data $\mathbf{D} = \{\mathbf{D}_0, \mathbf{D}_1, \cdots, \mathbf{D}_{N-1}\}$, where N is the total number of data collections, we represent each $\mathbf{D}_i$ as a set of coordinates and their values. Namely, $\mathbf{D}_i = \{\mathbf{C}_i, \mathbf{V}_i\}$, where $\mathbf{C}_i = \{(x_0^i, y_0^i, \cdots), (x_1^i, y_1^i, \cdots), \cdots\}$ and $\mathbf{V}_i = \{(v^i[x_0^i, y_0^i, \cdots]), (v^i[x_1^i, y_1^i, \cdots]), \cdots\}$. As an example, for an image, $\mathbf{C}$ represents the pixel locations, and $\mathbf{V}$ is the pixel values. The goal of CoordNet is to learn a mapping from coordinates to values, i.e., $f(x, y, \cdots) = v$. As shown in Figure 1, given a data $\mathbf{D}_i$, we first decompose the data into a set of coordinates $(x, y, \cdots)$. CoordNet accepts these coordinates as input and produces the values at these coordinates. After prediction, the difference between the predicted value and its ground truth (GT) is measured, and the parameters of CoordNet are updated by backpropagation. Once trained, CoordNet accepts unseen coordinates to predict the corresponding values. With this coordinate-based formulation, CoordNet solves different tasks based on diverse input coordinates. We summarize the four evaluated tasks in Table 1. For both TSR and SSR tasks, the input to CoordNet is (x, y, z, t). The difference is that CoordNet needs to explore the temporal coordinate (t) or spatial coordinates (x, y, z) during inference. For VS, CoordNet produces new rendering images given unseen view parameters (θ, ϕ). For AOP, CoordNet generates LAO values given the new temporal coordinate (t) and opacity parameter (o). Please refer to the Appendix for the detailed discussion on applying CoordNet to these tasks.

TABLE 1: Input and output for each task.

task	input	output
TSR [7]	x, y, z, t	voxel value
SSR [8]	x, y, z, t	voxel value
VS [14]	x, y, θ, ϕ	pixel value
AOP [5]	x, y, z, t, o	LAO value

4.2 Network Architecture

SIREN. The building block of CoordNet is SIREN [33]. It is a fully-connected layer followed by $\sin(\omega x)$ as the activation function, where ω is a hyperparameter. In this paper, we set ω to 30, as recommended by Sitzmann et al. [33]. Compared with using other activation functions, such as ReLU, SIREN has the following advantages. First, the training process is more stable. Second, the gradient of sinusoidal activations exists almost everywhere, while others will be close to zero in some particular regions. This means using sinusoidal activations can speed up network convergence. Third, using sinusoidal functions fits complex signals better in both data and gradient spaces [33], as shown in Figure 3.

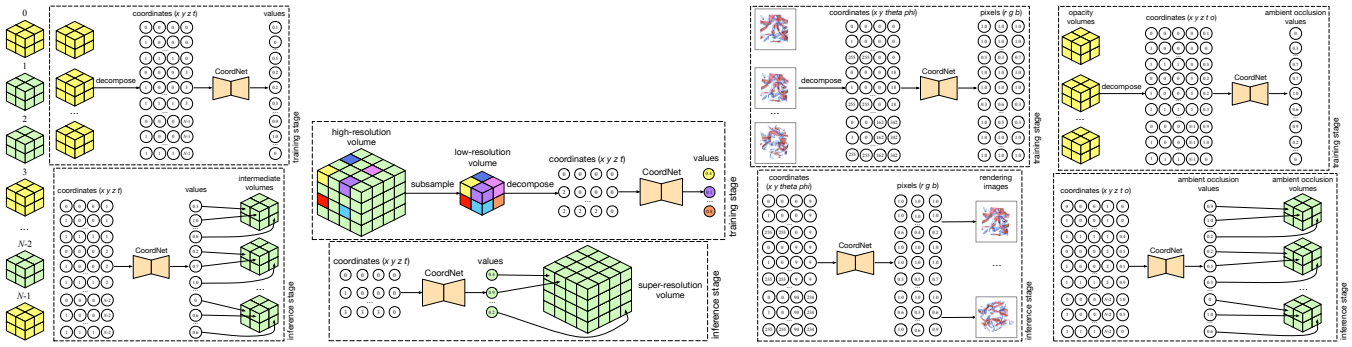


(a) ReLU (b) Sine (c) GT

Fig. 3: Comparison of different activation functions via volume rendering results for the TSR task using the argon bubble data set. Top: data. Bottom: gradient.

SIREN-based residual block. We utilize residual blocks [13] to increase network depth for performance improvement. The original residual block design requires the input and output to have the same dimension. To tackle the case where the input dimension does not equal that of the output, we add one more SIREN layer in the residual block to ensure the input dimension is consistent with the output dimension. The demonstration is shown Figure 2 (b). Figure 2 (a) shows the traditional residual block structure. Furthermore, after adding the input and output from the residual block, we average the result (i.e., divide it by 2). The rationale behind this operation is explained as follows. We use SIREN as the basic block, and the output range of SIREN is $[-1, 1]$. Without averaging, the value computed by the residual block lies in $[-2, 2]$, which does not belong to the data range of sinusoidal activation and makes the training unstable. Utilizing these residual blocks, we can build a network with tens or even hundreds of layers to boost the performance and improve gradient propagation.

CoordNet. The architecture of CoordNet is sketched in Figure 2 (c). It accepts a coordinate with k components as input and predicts the corresponding value at that coordinate. Note that different data can be decomposed with various coordinates. For example, the coordinate of time-varying volumetric data is (x, y, z, t) and the coordinate of the rendered image under different time steps and view parameters is (x, y, θ, ϕ). In general, the design of CoordNet follows an encoder-decoder structure. We first map k coordinates to m neurons in the encoder by applying one residual block. After



(a) TSR (b) SSR (c) VS (d) AOP

Fig. 4: Overview of the training and inference stages of CoordNet for the (a) TSR, (b) SSR, (c) VS, and (d) AOP tasks.

TABLE 2: Network parameter details of CoordNet, where k, m, d, and p are the dimension of input coordinates, number of initial neurons, network depth, and output dimension, respectively.

structure	layer	# input neurons	# output neurons
encoder	residual block	k	m
	residual block	m	$2m$
	residual block	$2m$	$4m$
	residual block $\times d$	$4m$	$4m$
decoder	residual block	$4m$	p

that, two additional residual blocks are followed to increase the number of neurons in the hidden layers to expand the network width. Finally, we leverage d residual blocks to learn the representation of the input coordinates. During design, we can set a large number of neurons (i.e., m) and increase network depth (i.e., d) to guarantee that CoordNet has enough capacity to learn a dense representation of coordinates. After encoding, we apply one residual block to generate the value at the coordinate by decoding the learned representation. For volumetric data, the output has only one component (i.e., scalar value). For an image, it has three components (i.e., R, G, B). The parameter detail is listed in Table 2. In this paper, we set m and d to 64 and 10, respectively. The studies of m and d are discussed in the Appendix.

4.3 Network Training and Inference

In Figure 4 (a), we show the training and inference processes of CoordNet for the TSR task without supervision. During training, CoordNet accepts the coordinates (x, y, z, t) from the sampled volumes as input and outputs their corresponding voxel values. During inference, CoordNet loops all candidate coordinates (i.e., the unobserved time steps) and predicts the voxel values. Figure 4 (b) describes how CoordNet is trained and inferred in the SSR task in an unsupervised way. The low-resolution data is obtained by directly subsampling the high-resolution data. That is, given an upscaling factor s, for each subvolume in high-resolution data with $s \times s \times s$ (there is no overlap among these subvolumes), we sample one voxel. During training, CoordNet is optimized by only using the subsampled coordinates and voxel values. After that, CoordNet goes through all coordinates and outputs the voxel values. In Figure 4 (c), we sketch the training and inference stages of our approach for the VS task. During training, the coordinates (x, y, θ, ϕ) are fed into CoordNet, and the pixel values at these coordinates are predicted. During inference, we go through different view parameters to produce the corresponding images. Figure 4 (d) shows how our method is trained and inferred for the AOP task. During training, CoordNet accepts the positional and opacity parameters and produces the LAO values. Once trained, CoordNet takes new temporal and opacity values to predict the corresponding LAO value at each voxel.

4.4 Objective Function

To optimize CoordNet, we use mean squared error, which is defined as

$$\mathcal{L} = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{|\mathbf{C}_i|}\sum_{(x,y,\cdots)\in\mathbf{C}_i}||f(x,y,\cdots)-v||_2, \qquad (1)$$

where N is the number of training samples, v is the value at the coordinate $(x, y, \cdots)$, and $||\cdot||_2$ is L_2 norm. Note that adversarial [11], [14] and feature [7], [14] losses cannot be applied to optimize CoordNet since the output from CoordNet is a single voxel or pixel, while adversarial and feature losses are calculated based on a subvolume.

5 Results

In this section, we describe the evaluated tasks and data sets, provide optimization details and evaluation metrics, and finally show each task's quantitative and qualitative results.

Task description. We briefly describe the evaluated tasks for CoordNet as follows.

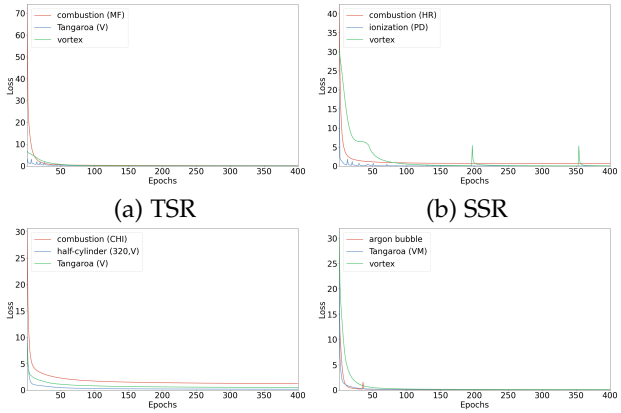
- TSR [7] interpolates the missing time steps in time-varying volumetric data through sparsely sampled time steps.
- SSR [8] produces super-resolution time-varying volumes using the corresponding low-resolution ones.
- VS [14] synthesizes rendering images under different view parameters (i.e., θ and ϕ).
- AOP [5] predicts the LAO volume given an opacity volume generated by intensity volume and transfer function.

Optimization details. We tested CoordNet using the data sets reported in Table 3. The half-cylinder is an ensemble data set with three Reynolds numbers (320, 640, 6400). PyTorch was used for implementation. Both training and inference were performed on a single NVIDIA TESLA

TABLE 3: The dimensions of each data set.

data set	variable	dimension ($x \times y \times z \times t$)
argon bubble	intensity	$320 \times 128 \times 128 \times 100$
combustion	MF, HR, CHI	$480 \times 720 \times 120 \times 100$
earthquake	intensity	$256 \times 256 \times 96 \times 598$
half-cylinder [29]	velocity magnitude (VM) vorticity (V)	$640 \times 240 \times 80 \times 100$
ionization [40]	H2, PD, T	$600 \times 248 \times 248 \times 100$
Tangaroa [27]	velocity magnitude (VM) vorticity (V)	$300 \times 180 \times 120 \times 150$
vortex	vorticity	$128 \times 128 \times 128 \times 90$

P100 GPU. The input coordinate and output value are scaled to $[-1, 1]$ to match the value range of $\sin(\cdot)$. We initialized parameters following Sitzmann et al. [33] and utilized the Adam optimizer [19] for parameter update. The batch size is set as $32,000$ coordinates. The learning rate is started as 10^{-5} with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and decayed with 10^{-6} using L_2 regularization [21]. Each task is trained independently. We set the training epochs to 300 for all tasks since the network has been converged, as shown in Figure 5. All these hyperparameter settings are determined based on experiments.



(a) TSR (b) SSR

(c) VS (d) AOP

Fig. 5: Loss curves of each task using different data sets.

Evaluation metrics. We utilize the data-level metric *peak signal-to-noise ratio* (PSNR), image-level metric *learned perceptual image patch similarity* (LPIPS) [41], and surface-level metric *chamfer distance* (CD) [1] to evaluate the quality of synthesized data. Using AlexNet [20] as a backbone, LPIPS computes a weighted average of the activations at hidden layers to predict relative image similarities, which correlate well with perceptual judgments. CD measures the bidirectional overall node-wise distance between two isosurfaces extracted from synthesized and ground-truth data. For PSNR, the higher values are better, while for LPIPS and CD, the lower values are better.

In reference to the GT results, we compare CoordNet results against the state-of-the-art approaches in each task. The supplementary video provides the frame-to-frame comparison results.

5.1 Task 1: TSR

Baselines. We compare CoordNet against three interpolation approaches for the TSR task.

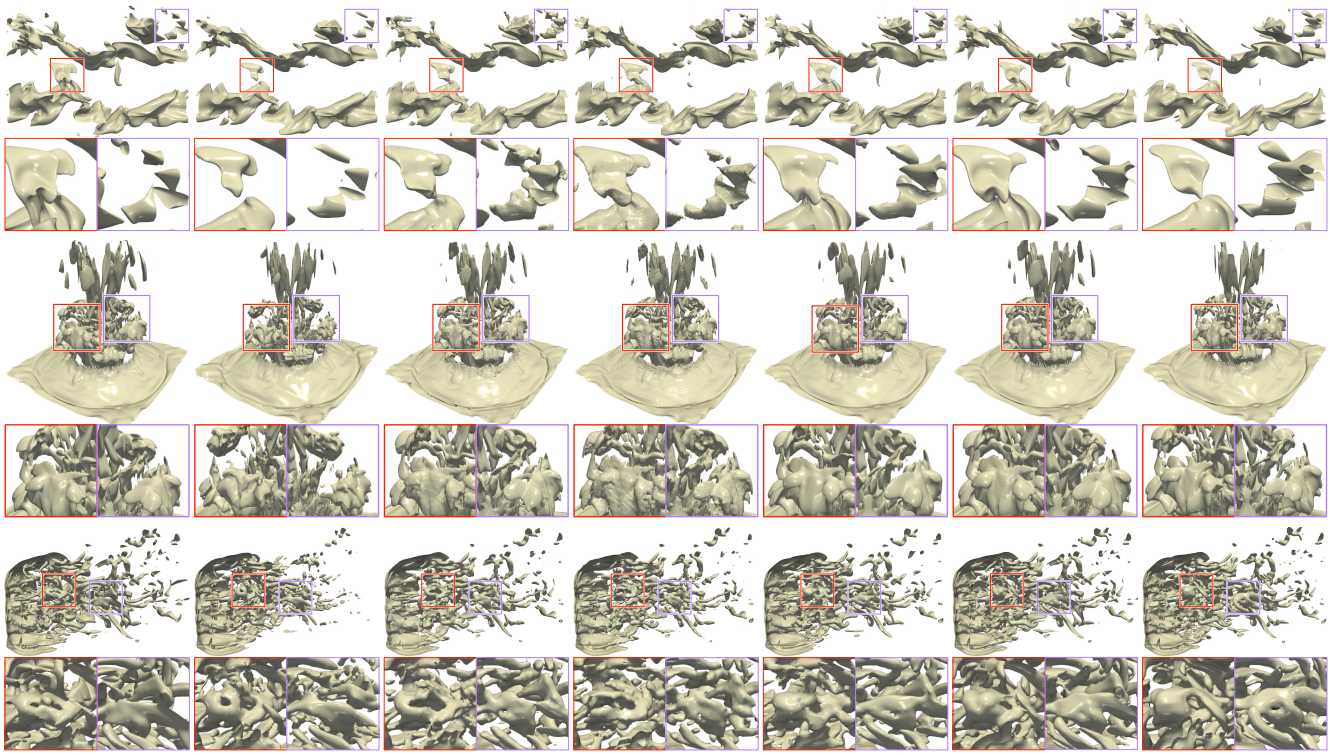
- Linear interpolation (LERP): LERP is a traditional approach that linearly interpolates the intermediate time steps.
- SloMo: SloMo [17] is a CNN-based solution for frame interpolation. It utilizes convolutional layers for feature learning, average pooling for downsampling, and trilinear interpolation for upsampling.
- TSR-TVD [7]: TSR-TVD is a recurrent generative framework for interpolating intermediate time steps with supervision. We add skip connection between the encoder and decoder to improve the performance.

Quantitative and qualitative analysis. We compare isosurface rendering results among LERP, SloMo, TSR-TVD, and our method using the combustion (MF), ionization (H2), and Tangaroa (V) data sets, as shown in Figure 6. Overall, CoordNet outperforms the state-of-the-art approaches in the evaluated data sets when comparing the isosurface rendering images. For example, for the ionization (H2) data set, LERP produces fewer detailed isosurfaces due to the linear change assumption. SloMo and TSR-TVD introduce noise and artifacts because of the limited capability of upscaling modules in CNN (e.g., deconvolution). Table 4 reports average PSNR, LPIPS, and CD values. In general, CoordNet achieves the best scores, with three exceptions out of 15 comparisons.

TABLE 4: Average PSNR (dB), LPIPS, and CD for the TSR task with an interpolation interval of 3. The chosen isovalues of each data for computing CD are 0.1, -0.4, -0.7, -0.9, and -0.75, respectively. Note that the selected time steps for the earthquake data set are non-uniform, which does not meet the assumption for using SloMo and TSR-TVD.

data set	method	PSNR ↑	LPIPS ↓	CD ↓
combustion (MF)	LERP	29.42	0.238	2.24
	SloMo	36.50	0.152	1.10
	TSR-TVD	37.34	0.162	0.98
	CoordNet	**37.82**	**0.127**	**0.96**
earthquake	LERP	41.05	**0.086**	1.67
	SloMo	—	—	—
	TSR-TVD	—	—	—
	CoordNet	**42.85**	0.107	**1.62**
half-cylinder (640, V)	LERP	35.22	0.056	6.99
	SloMo	44.27	0.019	1.25
	TSR-TVD	**46.29**	0.017	**1.04**
	CoordNet	42.38	**0.012**	1.34
ionization (H2)	LERP	36.52	0.183	1.27
	SloMo	47.11	0.112	0.57
	TSR-TVD	48.22	0.116	0.48
	CoordNet	**49.05**	**0.110**	**0.45**
Tangaroa (V)	LERP	39.01	0.097	1.09
	SloMo	44.92	0.042	0.54
	TSR-TVD	45.08	0.043	0.52
	CoordNet	**45.32**	**0.041**	**0.50**

Evaluation of interpolation interval. We investigate the performance of CoordNet under different interpolation intervals. The isosurface rendering results are displayed in Figure 7. With interpolating 3 time steps, both LERP and CoordNet produce results close to GT, but CoordNet preserves more details at the top-right corner. When interpolating 5 and 7 time steps, the isosurface generated by CoordNet is more similar to GT. In addition, we also observe that CoordNet recovers even more details for the corresponding isosurface with an interpolation interval of 7, compared with the isosurface generated by LERP under the interpolation interval of 5. Table 5 reports average PSNR and LPIPS values



(a) left obs. (b) LERP (d) SloMo (d) TSR-TVD (e) CoordNet (f) GT (g) right obs.

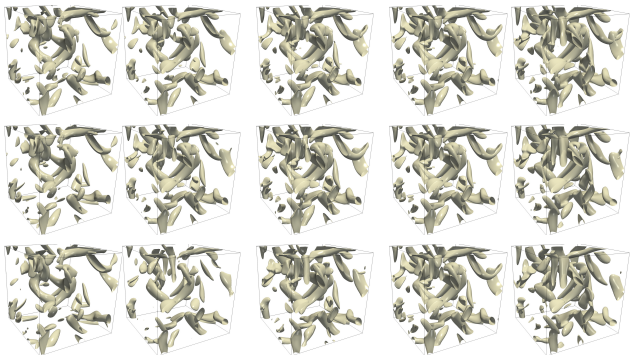
Fig. 6: Isosurface rendering results for the TSR task with an interpolation interval of 3. Top to bottom: combustion (MF), ionization (H2), and Tangaroa (V). The chosen isovalues are 0.1, -0.9, and -0.75. respectively. Top to bottom: the interpolated time steps are 95, 75, and 147. Left and right observations are the sampled time steps used for interpolation or network training. From top to bottom, they are 93 and 97, 73 and 77, 145 and 149. The quantitative scores are reported in Table 4.

under different settings, which confirms the effectiveness of CoordNet for the TSR task.

TABLE 5: Average PSNR (dB) and LPIPS for the TSR task under different interpolation intervals using the vortex data set.

interval	method	PSNR ↑	LPIPS ↓
3	LERP	30.45	0.165
	CoordNet	**38.92**	**0.066**
5	LERP	26.86	0.248
	CoordNet	**30.73**	**0.159**
7	LERP	24.68	0.304
	CoordNet	**27.01**	**0.236**

Non-uniform sampling. Unlike TSR-TVD, which assumes the selected time steps are uniformly sampled, our approach interpolates intermediate time steps through non-uniform sampling. We apply an importance-driven approach [37] to non-uniformly select 50 time steps from the total of 598 time steps using the earthquake data set. Table 4 reports average PSNR and LPIPS values. Although LERP produces a lower LPIPS value, CoordNet achieves a higher PSNR value. In terms of visual quality, the volume rendering results are displayed in Figure 8. As we can observe, compared with LERP, CoordNet produces closer rendering results compared with GT, for example, the epicenter (i.e., the red and blue parts) and the boundary of the earthquake.



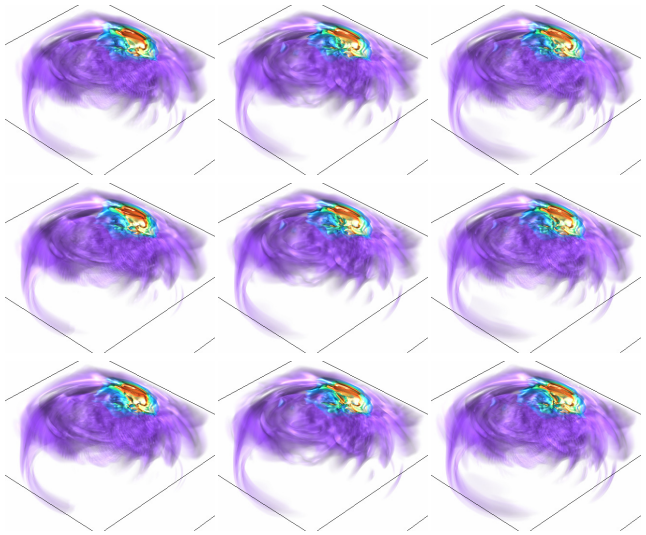
(a) l. obs. (b) LERP (c) CoordNet (d) GT (e) r. obs.

Fig. 7: Isosurface rendering results for the TSR task under different interpolation intervals (top to bottom: 3, 5, and 7) using the vortex data set. The chosen isovalue is -0.1. The displayed time step is 54. From top to bottom, left and right observations are 52 and 56, 51 and 57, 50 and 58.

5.2 Task 2: SSR

Baselines. We compare our method against three interpolation approaches for the SSR task.

- Bicubic interpolation (BI): BI is a traditional approach for spatial upscaling. We use reflective padding in BI during upscaling.
- ESPCN [31]: ESPCN is a CNN-based solution for SSR. It consists of several convolutional layers and



(a) LERP (b) CoordNet (c) GT

Fig. 8: Volume rendering results for the TSR task using the earthquake data set. Top to bottom: time steps 88, 89, and 90. We interpolate 548 time steps from the 50 non-uniformly selected time steps.

TABLE 6: Average PSNR (dB), LPIPS, and CD for the SSR task with an upscaling factor of $4\times$. The chosen isovalues for computing CD are 0.4, -0.4, -0.3, and -0.1, respectively.

data set	method	PSNR ↑	LPIPS ↓	CD↓
combustion (HR)	BI	40.07	0.090	0.71
	ESPCN	35.68	0.151	1.10
	SSR-TVD	39.77	0.140	1.55
	CoordNet	**42.25**	**0.064**	**0.57**
ionization (PD)	BI	42.28	0.098	2.98
	ESPCN	42.22	0.147	1.48
	SSR-TVD	42.91	0.114	1.36
	CoordNet	**50.26**	**0.028**	**0.30**
ionization (T)	BI	40.76	0.108	0.76
	ESPCN	38.55	0.184	1.24
	SSR-TVD	36.23	0.201	1.81
	CoordNet	**43.26**	**0.069**	**0.42**
vortex	BI	32.37	0.188	1.39
	ESPCN	36.52	0.109	0.90
	SSR-TVD	**39.56**	**0.072**	0.52
	CoordNet	36.64	0.089	**0.40**

two pixel shuffle layers that upscale four times along each dimension.

- SSR-TVD [8]: SSR-TVD is a supervised deep learning solution for the SSR task. It leverages a generator to upscale low-resolution volumes. Besides, a spatial and a temporal discriminator are designed to judge the realness of the synthesized data in the spatial and temporal dimensions, respectively.

Quantitative and qualitative analysis. We compare volume rendering results among BI, ESPCN, SSR-TVD, and CoordNet using the combustion (HR), ionization (T), and vortex data sets, as shown in Figure 9. CoordNet achieves the best performance compared to the other approaches in terms of visual quality. For example, for the combustion (HR) data set, BI presents artifacts in the zoom regions because of its simple interpolation mechanism, and ESPCN and SSR-TVD show fewer details due to the limited receptive fields for large volumetric data. Table 6 reports average PSNR, LPIPS, and CD values. Out of 12 comparisons, CoordNet achieves the best scores in all but one case.

Evaluation of upscaling factor. We study the performance of our approach under different upscaling factors. The volume rendering results are shown in Figure 10. With an upscaling factor of $2\times$, both approaches achieve similar rendering results, but CoordNet produces a sharpened image closer to GT. When upscaling $4\times$ and $8\times$, BI cannot faithfully recover the volumes, for example, the middle branching structure. However, CoordNet still preserves the volume with high fidelity. The overall shape can be reconstructed even with upscaling $8\times$. Table 7 shows average PSNR and LPIPS values for different upscaling factors. These values confirm the effectiveness of CoordNet for the SSR task.

TABLE 7: Average PSNR (dB) and LPIPS for the SSR task under different upscaling factors using the ionization (PD) data set.

factor	method	PSNR ↑	LPIPS ↓
$2\times$	BI	53.45	0.032
	CoordNet	**53.57**	**0.019**
$4\times$	BI	42.28	0.131
	CoordNet	**50.26**	**0.028**
$8\times$	BI	36.32	0.256
	CoordNet	**41.15**	**0.095**

SSR-TVD performance variation. We find the performance of SSR-TVD degrades when the volume resolution is large. This is because, for large volumes, SSR-TVD cannot have enough receptive fields due to GPU memory limitations. Besides, we find BI outperforms SSR-TVD for the ionization (T) data set in terms of PSNR and LPIPS. The possible reason is that the distribution shifts more from the early to later time steps compared with other data sets.

5.3 Task 3: VS

Baselines. We compare CoordNet against LERP, NeRV [4], and InSituNet [14] for the VS task. NeRV is a CNN for video compression. Here, we adopt this architecture for the VS task, where view parameters are input to generate rendering images with different resolutions. InSituNet accepts the view parameters as input and produces the corresponding rendering images using a set of convolutional and fully-connected layers. Adversarial and perceptual losses are computed during training. The original version generates images with 256 resolution. One additional upscaling block is inserted into InSituNet to synthesize images with 512 image resolution. For $1,024$ image resolution, we insert two additional upscaling blocks. Since InSituNet does not consider temporal input, for a fair comparison, we do not input the temporal coordinate (t) to CoordNet. We uniformly sample the view parameters to generate 200 rendering images for training. We also produce 600 rendering images under new view parameters for evaluation.

Quantitative and qualitative analysis. In Figure 11, we compare the visual quality of the images synthesized by InSituNet and our method using the combustion (CHI), half-cylinder (320,V), Tangaroa (V), and vortex data sets. As the

(a) BI (b) ESPCN (c) SSR-TVD (d) CoordNet (e) GT

Fig. 9: Volume rendering results for the SSR task with an upscaling factor of $4\times$. Top to bottom: combustion (HR), ionization (T), and vortex. The quantitative values are shown in Table 6.

rendering images indicate, CoordNet preserves the shape and texture better than the other methods. For instance, for the combustion (CHI) data set, the image generated by CoordNet preserves the better shape and lighting details, while those produced by NeRV and InSituNet are more blurry and contain noise. Table 8 reports the average PSNR and LPIPS values for each data set. Although LERP achieves better PSNR values, our approach produces lower LPIPS values and better visual quality. On the other hand, we find that all deep learning solutions produce more or less blurry results compared with GT. One possible cause is the visual components in the rendering images with different opacities, posting challenges for the networks to learn.

5.4 Task 4: AOP

Baselines. We compare CoordNet against Hernell et al. [15], V2V [11], and deep volumetric ambient occlusion (DVAO) [5] for the AOP task. V2V is a GAN for variable translation. Here, we use this architecture to produce the LAO volume given the opacity volume. DVAO is a 3D U-Net with Mish activation function [26] based solution that accepts the intensity volume and transfer function as input

TABLE 8: Average PSNR (dB) and LPIPS for the VS task under 256 image resolution.

data set	method	PSNR ↑	LPIPS ↓
combustion (CHI)	LERP	**33.11**	0.094
	NeRV	24.19	**0.088**
	InSituNet	22.30	0.096
	CoordNet	25.87	0.108
half-cylinder (320, V)	LERP	**39.19**	0.040
	NeRV	31.41	0.035
	InSituNet	29.75	0.033
	CoordNet	34.29	**0.016**
Tangaroa (V)	LERP	**36.69**	0.055
	NeRV	26.52	0.053
	InSituNet	27.23	0.045
	CoordNet	31.40	**0.032**

(a) BI (b) CoordNet (c) GT

Fig. 10: Volume rendering results for the SSR task under different upscaling factors using the ionization (PD) data set. Top to bottom: $2\times$, $4\times$, and $8\times$.

and outputs the LAO volume. It is optimized using a 3D structural dissimilarity index and the mean squared error. We randomly sample 30% data for training and 70% for inference. Following Engel and Ropinski [5], we use Monte Carlo simulation with 196 rays and 10% of the volume diameter as a radius restriction to generate GT LAO volumes. We also cast 196 rays when applying Hernell et al.

Quantitative and qualitative analysis. In Figure 12, we compare volume rendering results of LAO volumes among Hernell et al., V2V, DVAO, and CoordNet using argon bubble, half-cylinder (6400, VM), Tangaroa (VM), and vortex data sets. CoordNet generally achieves closer rendering results compared with GT, as the zoom regions indicate. Refer to the Appendix for the volume rendering results with LAO. Table 9 reports average PSNR and LPIPS values for the AOP task. Again, CoordNet achieves the best PSNR and LPIPS values, except LPIPS for the vortex data set.

5.5 Comparison of Deep Learning Approaches

In Table 10, we report the average training and inference time of each method for different tasks. The time to train and infer using CoordNet depends on the data or image resolution, the number of input coordinates, and the number of output values. Specifically, VS with a $1,024$ image resolution requires the longest training time, while AOP takes the shortest time to optimize. In Table 10, we summarize the supervision of different methods for the evaluated tasks. CoordNet can perform data generation tasks in an unsupervised manner while it requires supervision for visualization generation tasks. The data generation tasks aim to learn the data itself, which can be operated without supervision, while visualization generation tasks require additional information. For example, VS needs view parameters rather than only relying on the positional coordinates to synthesize rendering images. Furthermore, we comprehensively compare CoordNet against the state-of-the-art approaches in other aspects, as shown in Table 10. While CoordNet requires a longer training time than these approaches, it still provides the following advantages. (1) It processes both 3D and 2D data. (2) It can produce arbitrary data resolutions (e.g., interpolating arbitrary time steps in the TSR task, generating super-resolution volumes with arbitrary upscaling factors, and synthesizing rendering images with arbitrary resolutions). This is impossible for CNN-based networks because convolutional operations only allow an integer upscaling factor (e.g., 2 or 4) instead of a floating-point one (e.g., 1.5 or 3.1). (3) CoordNet can tackle diverse tasks without modifying network architectures. (4) CoordNet only takes around 6MB for model storage, while others require tens or hundreds of MB.

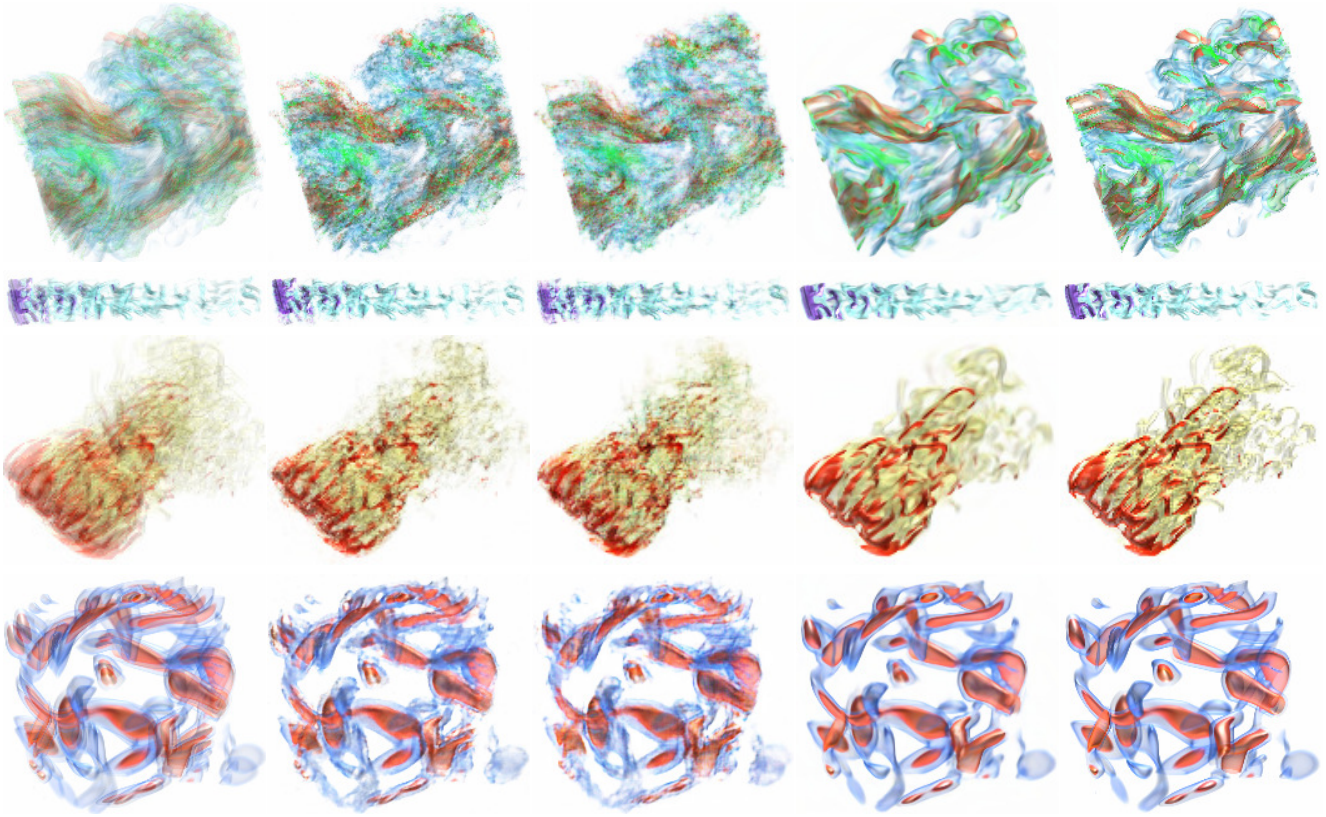
TABLE 9: Average PSNR (dB) and LPIPS (rendering of LAO volume) for the AOP task.

data set	method	PSNR ↑	LPIPS ↓
argon bubble	Hernell et al.	29.86	0.015
	V2V	42.17	0.014
	DVAO	27.76	0.053
	CoordNet	**47.16**	**0.008**
earthquake	Hernell et al.	30.42	0.081
	V2V	25.61	0.077
	DVAO	16.07	0.142
	CoordNet	**33.31**	**0.074**
half-cylinder (6400, VM)	Hernell et al.	20.24	0.062
	V2V	24.83	0.029
	DVAO	17.37	0.032
	CoordNet	**30.23**	**0.027**
Tangaroa (VM)	Hernell et al.	20.90	0.106
	V2V	30.77	0.050
	DVAO	18.41	0.190
	CoordNet	**41.90**	**0.032**
vortex	Hernell et al.	23.66	0.222
	V2V	30.92	0.165
	DVAO	30.06	**0.153**
	CoordNet	**31.94**	0.154

6 Discussion and Limitations

What tasks can CoordNet not tackle? To demonstrate its effectiveness, we validate CoordNet on four data generation and visualization generation tasks. However, CoordNet cannot tackle tasks requiring global data information, for example, viewpoint recommendation [30]. This is because such a task needs to consider the whole image rather than a few pixels to predict the corresponding viewpoint. Another task that CoordNet cannot tackle is ensemble generation. That is, given ensemble simulation parameters, we aim to synthesize the corresponding data. We have used the ensemble fluid data set with $8,000$ Reynolds numbers [16] to produce new ensemble data. However, the quality is unsatisfactory. A potential reason is that the ensemble space is much more complex to explore than spatial and temporal spaces.

Limitations. Although CoordNet processes diverse data generation and visualization generation tasks without modifying network architecture and outperforms existing learning-based solutions quantitatively and qualitatively, it still has two limitations. (1) *Training time*: Compared with



(a) LERP (b) NeRV (c) InSituNet (d) CoordNet (e) GT

Fig. 11: Comparison of synthesized volume rendering images under 256 image resolution for the VS task. Top to bottom: combustion (CHI), half-cylinder (320,V), Tangaroa (V), and vortex. The quantitative metrics are listed in Table 8.

TABLE 10: Comparison of training time, inference time, supervision, and others for different deep learning methods. A '—' indicates that the corresponding method cannot be applied. The model size is in MB.

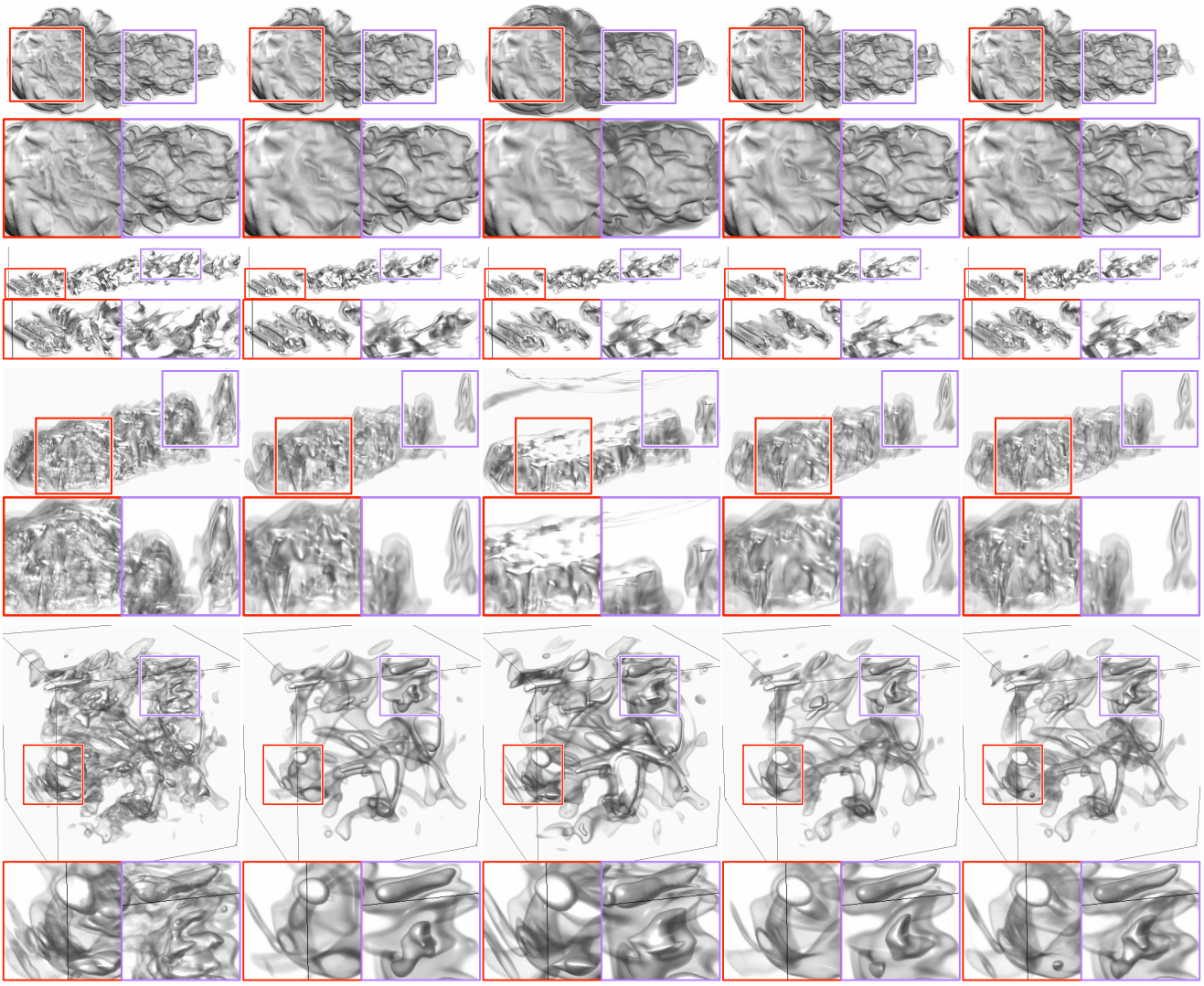
		SloMo	TSR-TVD	ESPCN	SSR-TVD	NeRV	InSituNet	V2V	DVAO	CoordNet
training time (days)	TSR	$0.3 \sim 1.46$	$1 \sim 3$	—	—	—	—	—	—	$0.5 \sim 3.5$
	SSR	—	—	$0.05 \sim 0.4$	$1 \sim 4$	—	—	—	—	$1 \sim 4$
	VS	—	—	—	—	$0.04 \sim 0.14$	$0.5 \sim 2$	—	—	$1.5 \sim 5$
	AOP	—	—	—	—	—	—	$0.2 \sim 1.25$	$0.3 \sim 1.5$	$0.5 \sim 2$
inference time (seconds)	TSR	$0.32 \sim 80.45$	$0.41 \sim 100.33$	—	—	—	—	—	—	$4.89 \sim 258.14$
	SSR	—	—	$0.35 \sim 127.16$	$0.52 \sim 1200.67$	—	—	—	—	$4.67 \sim 259.02$
	VS	—	—	—	—	$0.08 \sim 2.05$	$0.10 \sim 2.56$	—	—	$0.12 \sim 4.13$
	AOP	—	—	—	—	—	—	$0.42 \sim 60.76$	$0.33 \sim 76.53$	$4.56 \sim 26.89$
supervision	TSR	supervised	supervised	—	—	—	—	—	—	unsupervised
	SSR	—	—	supervised	supervised	—	—	—	—	unsupervised
	VS	—	—	—	—	supervised	supervised	—	—	supervised
	AOP	—	—	—	—	—	—	supervised	supervised	supervised
others	data type	3D	3D	3D	3D	2D	2D	3D	3D	**3D+2D**
	arbitrary res.	no	no	no	no	no	no	no	no	**yes**
	task	TSR	TSR	SSR	SSR	VS	VS	AOP	AOP	**all four**
	model size	224.60	41.40	10.70	50.05	169.28	166.64	36.10	143.90	**5.68**

CNN-based solutions, which treat one *volume* or *image* as a training sample, CoordNet regards each *coordinate* as a training sample. This treatment significantly increases the number of training samples, leading to a longer training time, as shown in Table 10 (2) *Image generation quality*: Although CoordNet outperforms LERP, NeRV, and InSituNet in the VS task, the synthesized images are still somewhat blurry and present artifacts, which may require improvement for image analysis purposes. (3) *Data range recovery:* Our method cannot recover the inferred data to their original data range for certain tasks such as TSR where the resolved time steps are not simulated (so no minimum and maximum values can be obtained for rescaling). This may prevent domain scientists from analyzing data in some specific applications.

7 Conclusions and Future Work

We have presented CoordNet, a simple yet versatile deep learning framework for tackling diverse tasks in scientific visualization. Through building an INR-based neural network, CoordNet processes different data generation and visualization generation tasks in both 2D and 3D cases. Moreover, the data generation pipeline fits the in-situ scenarios well. That is, the simulation data can be sparsely stored and recovered during post-processing. Compared with the state-of-the-art approaches in each task, CoordNet produces higher-quality results, both qualitatively and quantitatively.

The future work of CoordNet includes two directions. (1) *Multi-task analysis for better network initialization:* Currently, we train each task from scratch. In the future, we



(a) Hernell et al. (b) V2V (c) DVAO (d) CoordNet (e) GT

Fig. 12: Volume rendering results of LAO volumes for the AOP task. Top to bottom: argon bubble, half-cylinder (6400, VM), Tangaroa (VM), and vortex. The quantitative scores are displayed in Table 9.

would like to explore better initialization algorithms (e.g., meta learning [35]) by considering the relationship among different tasks. (2) *Acceleration for coordinate-based networks:* CoordNet takes a significant amount of time to optimize, which is not comparable to CNN-based solutions. We plan to speed up the training process by utilizing hash tables to build a dictionary from coordinates.

Acknowledgements

This research was supported in part by the start-up fund UDF01002679 of the Chinese University of Hong Kong, Shenzhen, Shenzhen Science and Technology Program ZDSYS20211021111415025, and the U.S. National Science Foundation through grants IIS-1455886, CNS-1629914, DUE-1833129, IIS-1955395, IIS-2101696, and OAC-2104158. The authors would like to thank the anonymous reviewers for their insightful comments.

References

[1] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 659–663, 1977.

[2] M. Berger, J. Li, and J. A. Levine. A generative model for volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1636–1650, 2019.

[3] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021.

[4] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava. NeRV: Neural representations for videos. In *Proceedings of Advances in Neural Information Processing Systems*, 2021.

[5] D. Engel and T. Ropinski. Deep volumetric ambient occlusion. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1268–1278, 2021.

[6] M. Guo, A. Fathi, J. Wu, and T. Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020.

[7] J. Han and C. Wang. TSR-TVD: Temporal super-resolution for time-varying data analysis and visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):205–215, 2020.

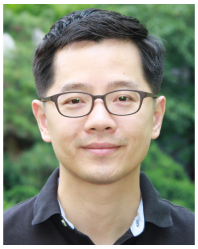
[8] J. Han and C. Wang. SSR-TVD: Spatial super-resolution for time-varying data analysis and visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(6):2445–2456, 2022.

[9] J. Han and C. Wang. VCNet: A generative model for volume completion. *Visual Informatics*, 6(2):62–73, 2022.

[10] J. Han, H. Zheng, D. Z. Chen, and C. Wang. STNet: An end-to-

end generative framework for synthesizing spatiotemporal super-resolution volumes. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):270–280, 2022.
[11] J. Han, H. Zheng, Y. Xing, D. Z. Chen, and C. Wang. V2V: A deep learning approach to variable-to-variable selection and translation for multivariate time-varying data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1290–1300, 2021.
[12] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, 2017.
[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
[14] W. He, J. Wang, H. Guo, K.-C. Wang, H.-W. Shen, M. Raj, Y. S. G. Nashed, and T. Peterka. InSituNet: Deep image synthesis for parameter space exploration of ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):23–33, 2020.
[15] F. Hernell, P. Ljung, and A. Ynnerman. Local ambient occlusion in direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 16(4):548–559, 2009.
[16] J. Jakob, M. Gross, and T. Günther. A fluid flow data set for machine learning and its application to neural flow map interpolation. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1279–1289, 2021.
[17] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.
[18] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.
[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference for Learning Representations*, 2015.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference for Learning Representations*, 2019.
[22] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.
[23] Y. Lu, K. Jiang, J. A. Levine, and M. Berger. Compressive neural representations of volumetric scalar fields. *Computer Graphics Forum*, 40(3):135–146, 2021.
[24] B. McCann, N. S. Keskar, C. Xiong, and R. Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.
[25] 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 European Conference on Computer Vision*, pages 405–421, 2020.
[26] D. Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019.
[27] S. Popinet, M. Smith, and C. Stevens. Experimental and numerical study of the turbulence characteristics of airflow around a research vessel. *Journal of Atmospheric and Oceanic Technology*, 21(10):1575–1589, 2004.
[28] S. Pramanik, P. Agrawal, and A. Hussain. OmniNet: A unified architecture for multi-modal multi-task learning. *arXiv preprint arXiv:1907.07804*, 2019.
[29] I. B. Rojo and T. Günther. Vector field topology of time-dependent flows in a steady reference frame. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):280–290, 2019.
[30] N. Shi and Y. Tao. CNNs based viewpoint estimation for volume visualization. *ACM Transactions on Intelligent Systems and Technology*, 10(3):27:1–27:22, 2019.
[31] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.
[32] V. Sitzmann, E. R. Chan, R. Tucker, N. Snavely, and G. Wetzstein. MetaSDF: Meta-learning signed distance functions. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.
[33] 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.
[34] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Proceedings of Advances in Neural Information Processing Systems*, 2019.
[35] M. Tancik, B. Mildenhall, T. Wang, D. Schmidt, P. P. Srinivasan, J. T. Barron, and R. Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021.
[36] C. Wang and J. Han. DL4SciVis: A state-of-the-art survey on deep learning for scientific visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2022. Accepted.
[37] C. Wang, H. Yu, and K.-L. Ma. Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1547–1554, 2008.
[38] S. Weiss, M. Chu, N. Thuerey, and R. Westermann. Volumetric isosurface rendering with deep learning-based super-resolution. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):3064–3078, 2021.
[39] S. Weiss, M. Işik, J. Thies, and R. Westermann. Learning adaptive sampling and reconstruction for volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(7):2654–2667, 2022.
[40] D. Whalen and M. L. Norman. Ionization front instabilities in primordial H II regions. *The Astrophysical Journal*, 673:664–675, 2008.
[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
[42] Z. Zhou, Y. Hou, Q. Wang, G. Chen, J. Lu, Y. Tao, and H. Lin. Volume upscaling with convolutional neural networks. In *Proceedings of Computer Graphics International*, pages 38:1–38:6, 2017.

Jun Han is an assistant professor of data science at the Chinese University of Hong Kong, Shenzhen. He obtained a Ph.D. degree in computer science and engineering from the University of Notre Dame in 2022. Before that, he received a BS degree in software engineering and an MS degree in computer software and theory in 2014 and 2017. Both degrees are from Xidian University. His current research focuses on applying deep learning techniques to solve scientific visualization and human-computer interaction problems.



Chaoli Wang is a professor of computer science and engineering at the University of Notre Dame. He received a Ph.D. degree in computer and information science from The Ohio State University in 2006. Dr. Wang's main research interest is data visualization, particularly on the topics of time-varying multivariate data visualization, flow visualization, information-theoretic algorithms, graph-based techniques, and deep learning solutions for big data analytics. He is an associate editor of IEEE Transactions on Visualization and Computer Graphics.

APPENDIX

1 ADDITIONAL RESULTS

1.1 VS

Evaluation of image resolution. In Figure 1, we evaluate the capability of CoordNet, NeRV, and InSituNet in generating images with different resolutions using the vortex data set. Both CoordNet and InSituNet produce satisfactory results under 256 image resolution. However, taking a closer comparison, the image generated by InSituNet includes noise, and the features are not preserved well, for example, at the bottom region. Using the resolutions of 512 and $1,024$, CoordNet is the clear winner, while InSituNet does not produce acceptable results. This is because InSituNet only has hundreds of training images, and most GAN-based architectures do not have enough capacity to generate high-resolution images (e.g., 512 and $1,024$) [1], [2], [3], [4]. Besides qualitative analysis, Table 1 reports average PSNR and LPIPS values. Compared with NeRV and InSituNet, CoordNet achieves the best PSNR and LPIPS values under different image resolutions.

TABLE 1: Average PSNR (dB) and LPIPS for the VS task under different image resolutions using the vortex data set.

resolution	method	PSNR ↑	LPIPS ↓
256	NeRV	20.96	0.144
	InSituNet	19.38	0.162
	CoordNet	**22.84**	**0.083**
512	NeRV	20.89	0.201
	InSituNet	19.50	0.190
	CoordNet	**23.30**	**0.105**
1,024	NeRV	19.75	0.255
	InSituNet	20.36	0.193
	CoordNet	**23.74**	**0.129**

Additional results. Figure 2 displays the synthesized images under different view parameters using the Tangaroa (V) data sets. As these images show, CoordNet preserves the overall shapes and details under diversified view parameters.

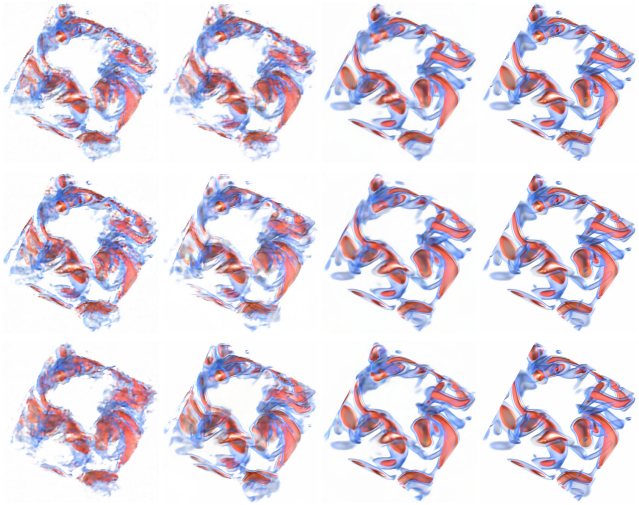
1.2 AOP

Figure 3 shows the volume rendering results with LAO. The difference image is displayed in the top-left corner for each approach. As the difference images indicate, CoordNet produces fewer differences than other methods.

1.3 TSR

Unsupervised time interpolation. Since CoordNet treats the coordinates as a continuous function; it can interpolate an arbitrary number of intermediate time steps, which is impossible with TSR-TVD. We produce six non-integer time steps between two neighboring time steps and compare their temporal coherence with LERP. The isosurface rendering results are shown in Figure 4. We can observe how the isosurfaces smoothly grow (refer to the red ellipses) and merge (refer to the blue ellipses) based on the rendering results generated by CoordNet, while LERP does not exhibit such smooth temporal variations.

Volume rendering results. Figure 5 shows the volume rendering results among TSR-TVD, CoordNet, and GT. For the combustion (MF) data set, TSR-TVD does not produce the green part at the bottom-left corner and the yellow part at the top-right corner well, while CoordNet preserves those details. For the half-cylinder (6400,V) and ionization (H2) data sets, both methods produce similar rendering results compared with GT. But taking a close comparison, the image produced by TSR-TVD contains more artifacts.



(a) NeRV (b) InSituNet (c) CoordNet (d) GT

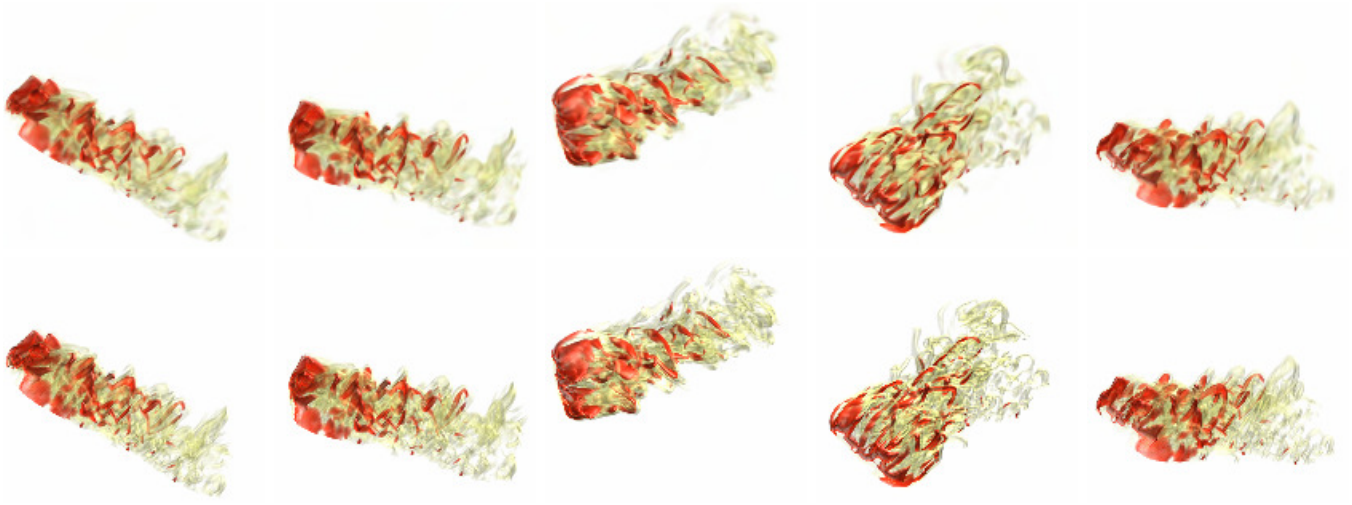
Fig. 1: Comparison of synthesized volume rendering images for the VS task under different image resolutions using the vortex data set. Top to bottom: 256, 512, and $1,024$ image resolutions.

Slice of volume rendering results. Figure 6 shows a slice of volume rendering results for the TSR task. These results indicate the sharpness of the synthesized data generated by CoordNet.

Discussion. Compared with TSR-TVD, CoordNet achieves better visual quality (direct volume rendering and isosurface rendering) and better quantitative scores. Besides, CoordNet has the following advantages. (1) The interpolation process is unsupervised, which means CoordNet does not require to see the complete subsequence of early time steps for training. (2) Given two time steps, CoordNet can synthesize arbitrary numbers of time steps with coherent and high-quality results, while TSR-TVD needs to perform this recursively (i.e., the synthesized time steps are fed into TSR-TVD to produce new time steps), and the performance cannot be guaranteed due to error accumulation in the recursive process. (3) CoordNet can operate in non-uniform sampling cases, while TSR-TVD only assumes the time steps are selected uniformly.

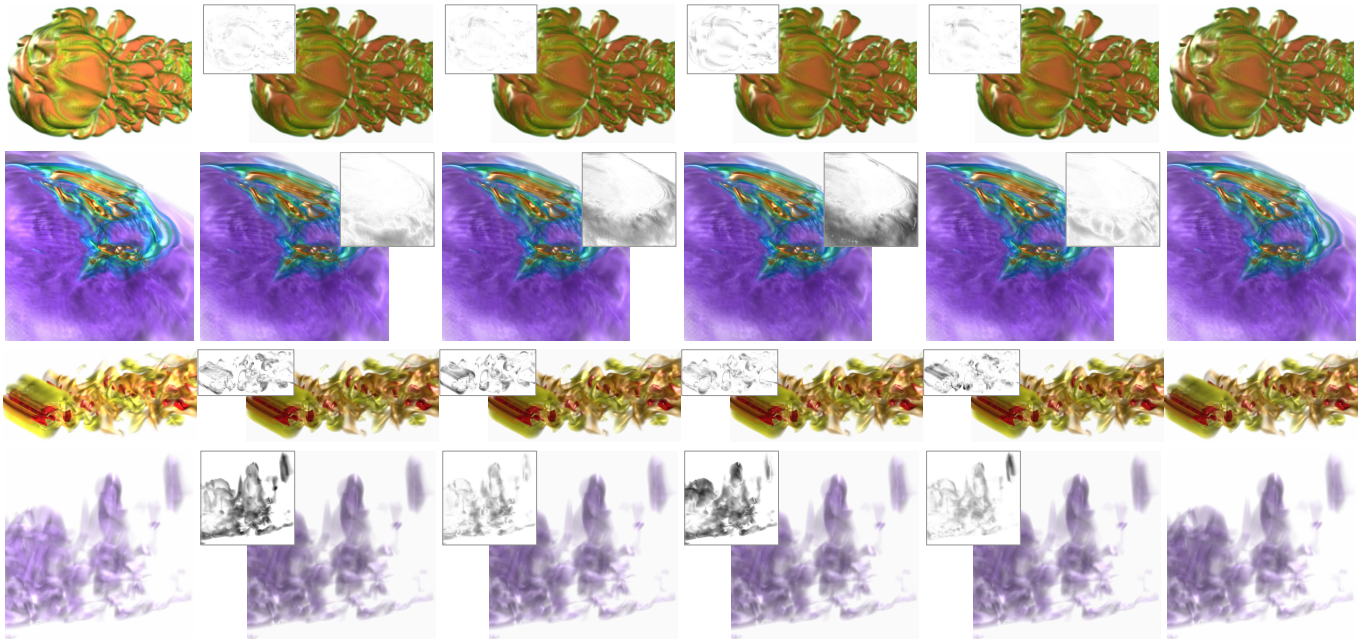
1.4 SSR

Unsupervised space interpolation. Because CoordNet processes the SSR task without supervision, it can produce higher-resolution volumes. That is, we can assume the original volumes (e.g., 128^3) are subsampled from higher-resolution volumes (e.g., 512^3), utilize these original volumes to train CoordNet, and inferCoordNet to synthesize higher-resolution ones. We use the vortex ($128 \times 128 \times 128$) and ionization (PD) ($600 \times 248 \times 248$) data sets to train CoordNet and produce volumes with higher-resolution (i.e., vortex with $512 \times 512 \times 512$ and ionization (PD) with



(a) $(9, 45)$ (b) $(27, 18)$ (c) $(45, 324)$ (d) $(90, 297)$ (e) $(135, 306)$

Fig. 2: Volume rendering results for the VS task under different view parameters (θ, ϕ) using the Tangaroa (V) data set. Top: CoordNet. Bottom: GT.



(a) GT wo. LAO (b) Hernell et al. (c) V2V (d) DVAO (e) CoordNet (f) GT w. LAO

Fig. 3: Zoom-in volume rendering with LAO results for the AOP task. Top to bottom: argon bubble, earthquake, half-cylinder (6400, VM), and Tangaroa (VM)

.

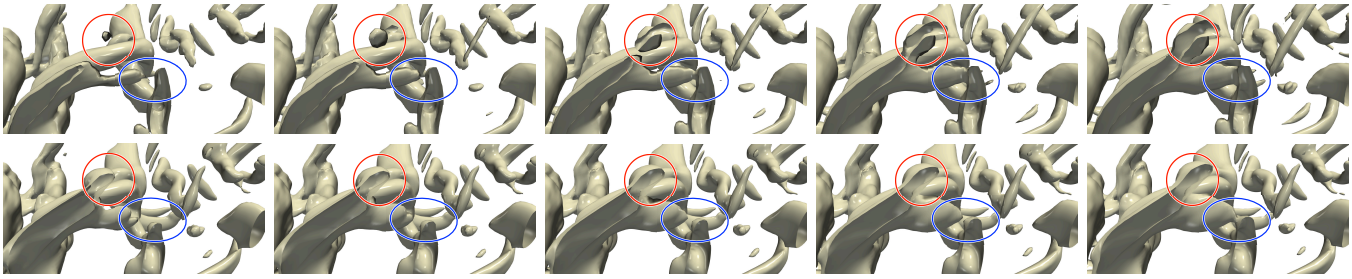
$2400 \times 992 \times 992$). We compare our results against BI. As displayed in Figure 7, CoordNet produces sharper results with fewer artifacts compared with BI (refer to the arrows in the images).

Additional results. Figure 8 displays the volume rendering results of the argon bubble, earthquake, and Tangaroa (VM) data sets. Compared with BI, CoordNet produces closer results in both shape and texture.

Slice of volume rendering results. Figure 9 shows a slice of volume rendering results for the SSR task. These results indicate that CoordNet preserves the sharpness and smoothness of the synthesized data.

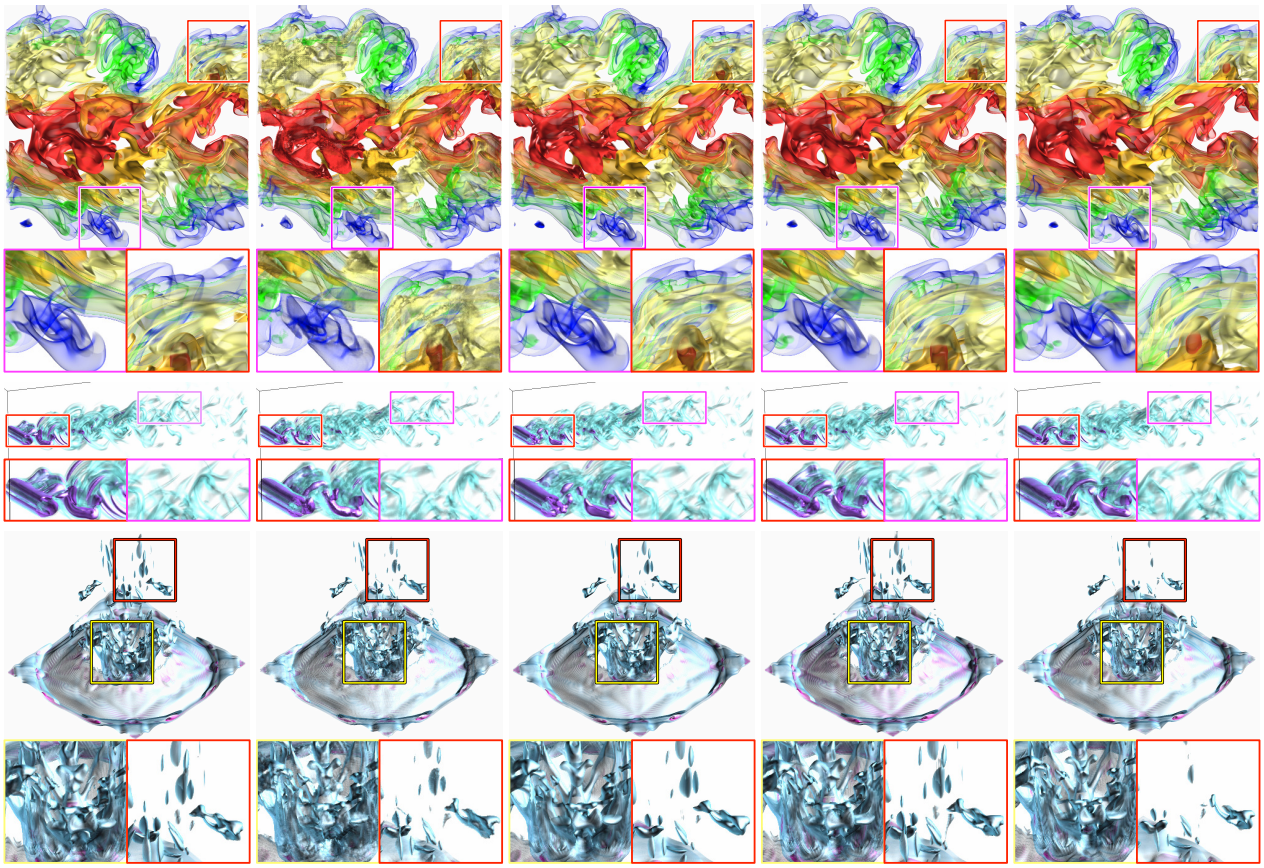
Isosurface rendering results. Figure 10 shows the isosurface rendering results among SSR-TVD, CoordNet, and GT. Both SSR-TVD and CoordNet produce close isosurface results of the combustion (HR) data set compared with GT, but SSR-TVD misses some isosurfaces at the top-right corner. For the ionization (PD) data set, SSR-TVD extracts the isosurfaces with artifacts and does not preserve the isosurface's shape in the feature region. For the vortex data set, CoordNet generates more similar isosurfaces compared to GT. For example, SSR-TVD cannot reconstruct the isosurfaces at the top-left corner.

Discussion. Compared with SSR-TVD, CoordNet achieves better visual quality and similar quantitative val-



(a) time step 95.167 (b) time step 95.333 (c) time step 95.500 (d) time step 95.667 (e) time step 95.833

Fig. 4: Zoom-in isosurface rendering results for the TSR task using the half-cylinder (640,V) data set. Top: LERP. Bottom: CoordNet. We generate 576 time steps from sparsely sampled 25 time steps. The chosen isovalue is -0.7.



(a) left observation (b) TSR-TVD (c) CoordNet (d) GT (e) right observation

Fig. 5: Volume rendering results for the TSR task with an interpolation interval of 3. Top to bottom: combustion (MF), half-cylinder (640,V), and ionization (H2). The displayed time steps are 95, 75, and 75, respectively. From top to bottom, left and right observations are 93 and 97, 73 and 77, 73 and 77.

ues. Still, CoordNet offers the following benefits. (1) The upscaling operation is completed in an unsupervised fashion. This means we do not need to store high- and low-resolution pairs for optimization. (2) CoordNet can upscale high resolution (e.g., $600 \times 248 \times 248$) to higher resolution (e.g., $2400 \times 992 \times 992$).

2 Hyperparameter Study

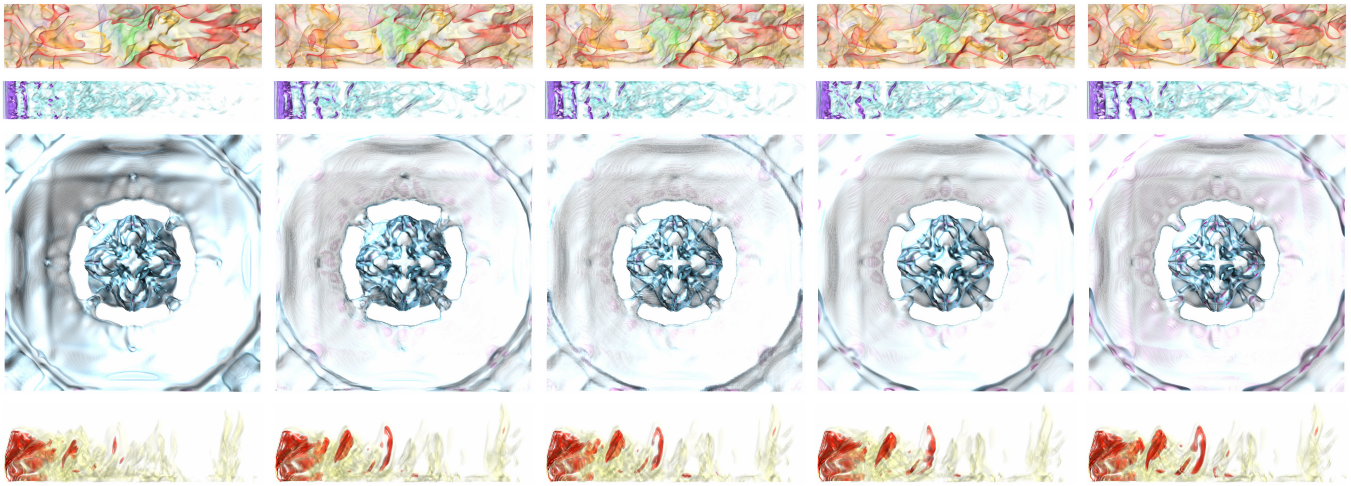
We further study the hyperparameters of CoordNet in the following aspects.

2.1 Sample Size (N)

To study the impact of sample size, we train CoordNet using different N for the TSR task. Table 2 reports the average PSNR, LPIPS, and training time under different sample sizes. The average PSNR and LPIPS are improved as we sample more voxels. However, the improvement becomes marginal as the sample size reaches 128K. In addition, as shown in Figure 13, the quality of rendering results benefits from the larger sample size. However, once N reaches 256K and 512K, the performance degrades since CoordNet begins to overfit the training data. Therefore, we suggest that the sample size should be 128K.

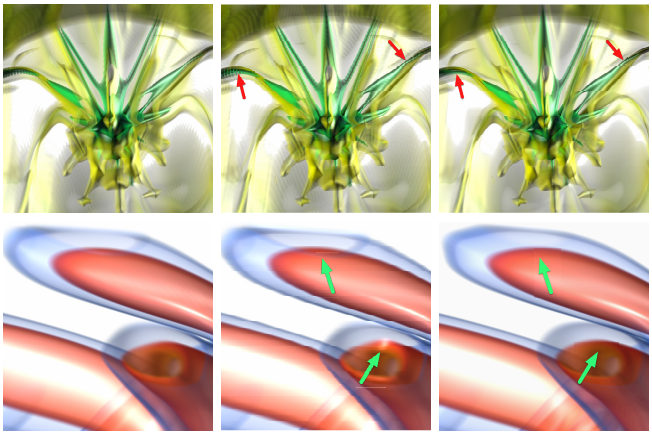
2.2 Number of Initial Neurons (m)

We optimize CoordNet using different numbers of m for the SSR task to determine an appropriate number of initial



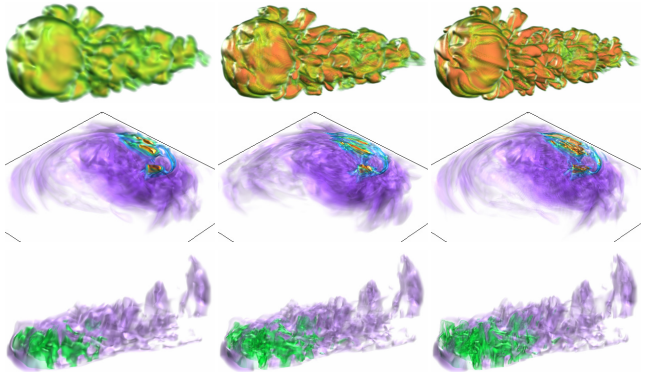
(a) LERP (b) SloMo (c) TSR-TVD (d) CoordNet (e) GT

Fig. 6: Slice of volume rendering results for the TSR task with an interpolation interval of 3. Top to bottom: combustion (MF), half-cylinder (640,V), ionization (H2), and Tangaroa (V). The displayed time steps are 95, 75, 75, and 147, respectively.



(a) input (b) BI (c) CoordNet

Fig. 7: Zoom-in volume rendering results for the SSR task. Top: ionization (PD). Bottom: vortex. For the ionization (PD) data set, we generate $2400 \times 992 \times 992$ volumes from $600 \times 248 \times 248$ ones. For the vortex data set, we generate $512 \times 512 \times 512$ volumes from $128 \times 128 \times 128$ ones.



(a) BI (b) CoordNet (c) GT

Fig. 8: Volume rendering results for the SSR task with an upscaling factor of $4\times$. Top to bottom: argon bubble, earthquake, and Tangaroa (VM).

neurons. Table 3 reports the average PSNR, LPIPS, training time, and model size under different numbers of m. In general, the average PSNR and LPIPS can be improved if a larger number of neurons is set. However, it takes longer to train, and more parameters need to be saved. Moreover, as shown in Figure 11, the quality of the rendering result is the best with 64 initial neurons. Beyond that, CoordNet could jump into overfitting, which decreases the performance. Therefore, we suggest that the number of initial neurons should be 64.

2.3 Choice of Network Depth (d)

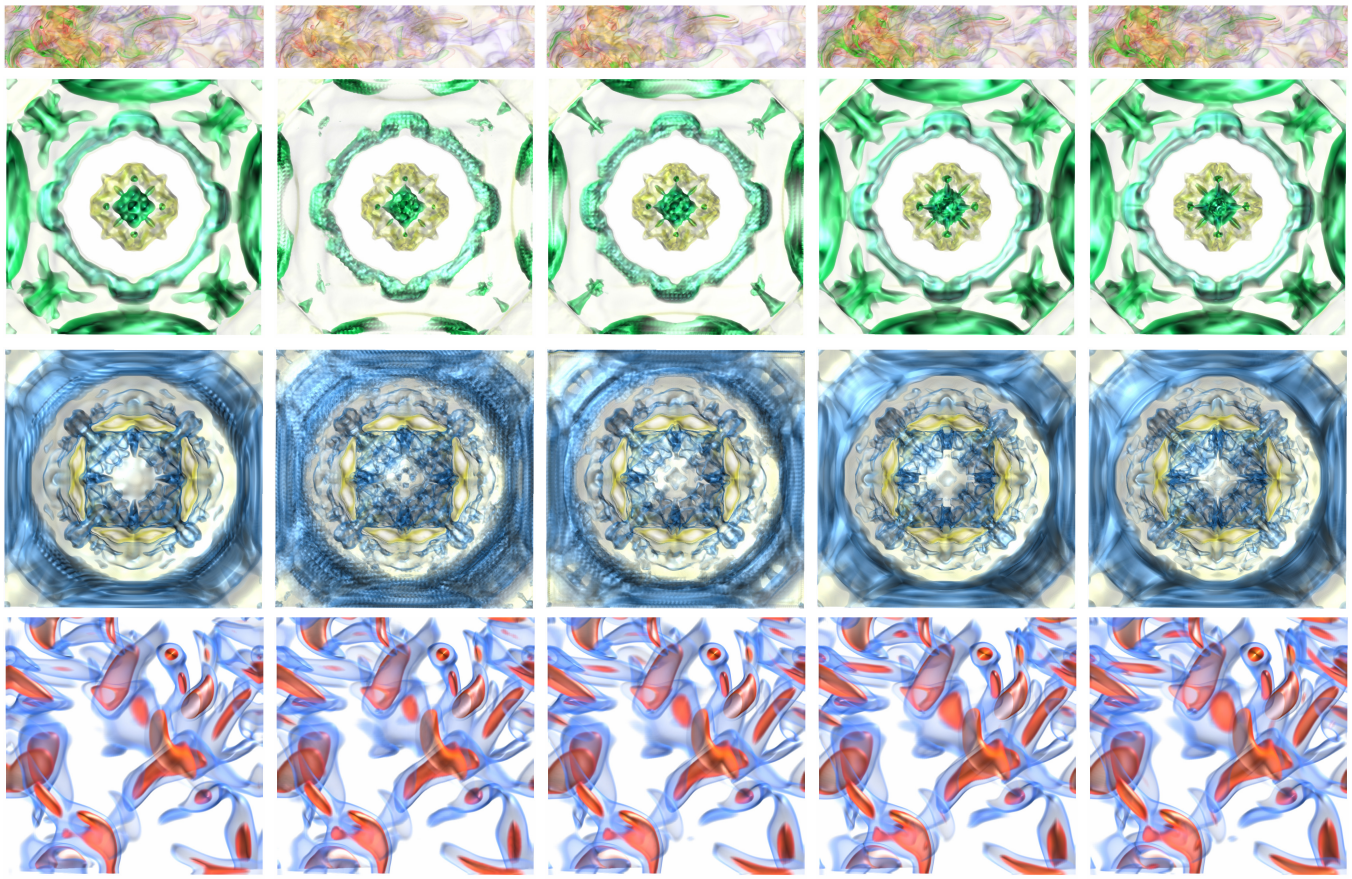
To choose an appropriate network depth, we apply different d to train CoordNet for VS task under 512 image resolution. As displayed in Figure 12, we can observe that as d increases, the result can be improved. However, there is no significant difference between $d = 10$ and $d = 15$. In Table 4, we report the average PSNR, LPIPS, and model size under different d. The quantitative metrics are better as d gets larger. However, the increment is small when d changes from 10 to 15. Thus, we choose the network depth as 10 for CoordNet.

TABLE 2: Average PSNR (dB), LPIPS values, and training time per epoch (in second) using the vortex data set under different numbers of sampled coordinates for the TSR task.

#coordinates	PSNR ↑	LPIPS ↓	train
32K	31.87	0.143	9.77
64K	35.56	0.103	20.41
128K	38.92	0.066	40.53
256K	39.68	0.058	90.55
512K	40.75	0.051	202.69

REFERENCES

[1] M. Berger, J. Li, and J. A. Levine. A generative model for volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1636–1650, 2019.



(a) BI (b) ESPCN (c) SSR-TVD (d) CoordNet (e) GT

Fig. 9: Slice of volume rendering results for the SSR task with an upscaling factor of $4\times$. Top to bottom: combustion (HR), ionization (PD), ionization (T), and vortex.

TABLE 3: Average PSNR (dB), LPIPS values, training time per epoch (in second), and model size (MB) using the half-cylinder (320, VM) data set under different numbers of initial neurons for the SSR task.

# neurons	PSNR ↑	LPIPS ↓	train	model
16	36.96	0.034	182.96	0.43
32	41.29	0.017	186.75	2.16
64	43.43	0.012	214.82	5.68
128	43.61	0.021	290.91	24.21

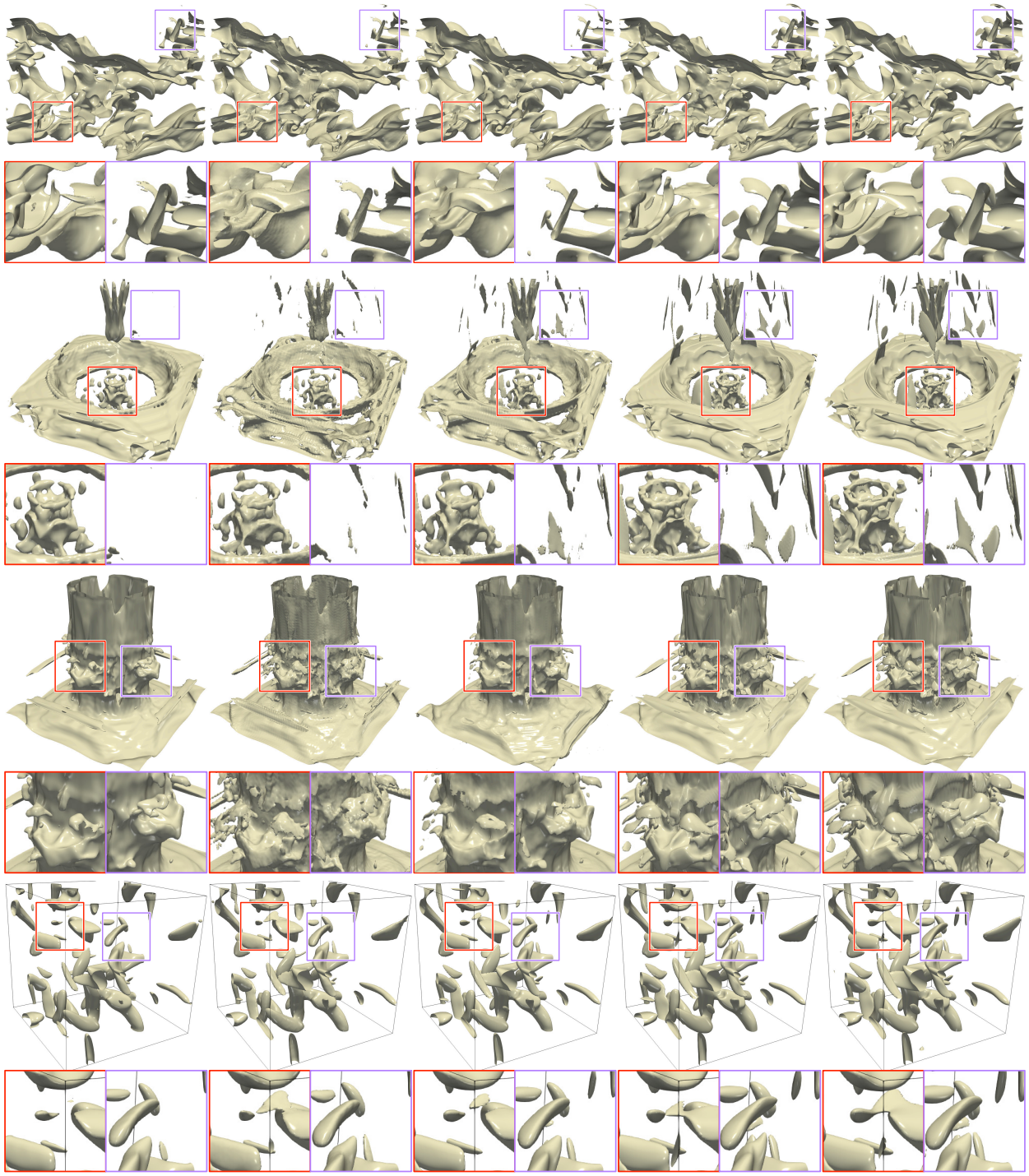
TABLE 4: Average PSNR (dB), LPIPS values, and model size (MB) using the combustion (CHI) data set under different network depths for the VS task. The training times across different depths are similar and therefore not reported here.

depth	PSNR ↑	LPIPS ↓	model
0	24.05	0.290	0.69
5	25.25	0.189	3.19
10	25.89	0.142	5.68
15	26.21	0.126	8.22

[2] W. He, J. Wang, H. Guo, K.-C. Wang, H.-W. Shen, M. Raj, Y. S. G. Nashed, and T. Peterka. InSituNet: Deep image synthesis for parameter space exploration of ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):23–33, 2020.

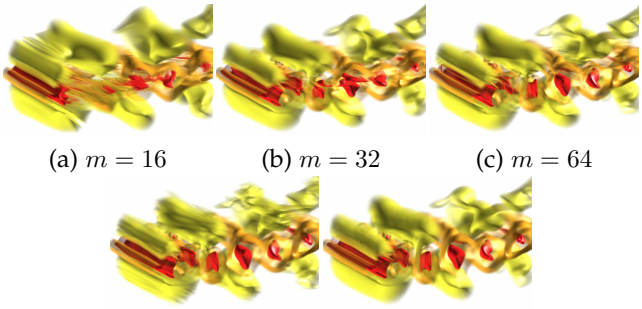
[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[4] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.



(c) BI (d) ESPCN (c) SSR-TVD (d) CoordNet (e) GT

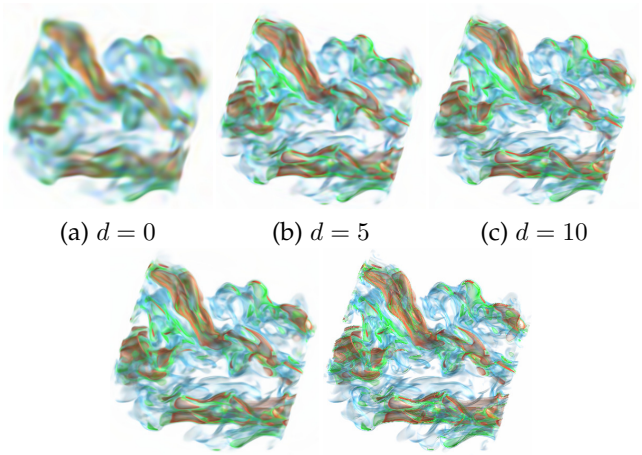
Fig. 10: Isosurface rendering results for the SSR task with an upscaling factor of $4\times$. Top to bottom: combustion (HR), ionization (PD), ionization (T), and vortex. The chosen isovalues are 0.4, -0.4, -0.3, and -0.1, respectively.



(a) $m = 16$ (b) $m = 32$ (c) $m = 64$

(d) $m = 128$ (e) GT

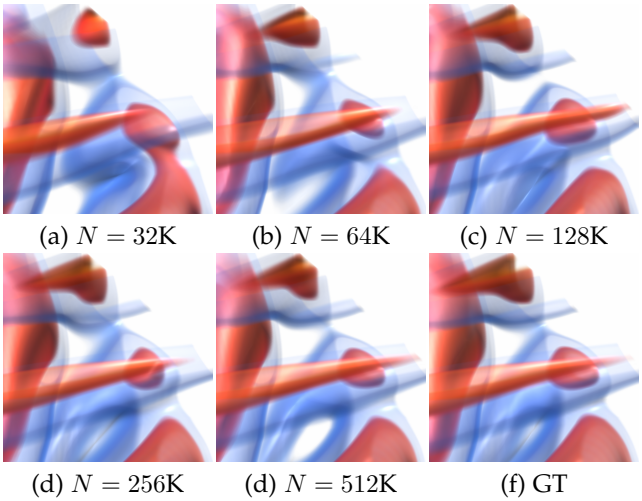
Fig. 11: Zoom-in volume rendering results for the SSR task under different numbers of initial neurons using the half-cylinder (320, VM) data set.



(a) $d = 0$ (b) $d = 5$ (c) $d = 10$

(d) $d = 15$ (e) GT

Fig. 12: Volume rendering results for the VS task under different network depths using the combustion (CHI) data set. The image resolution is 512.



(a) $N = 32$K (b) $N = 64$K (c) $N = 128$K

(d) $N = 256$K (d) $N = 512$K (f) GT

Fig. 13: Zoom-in volume rendering results for the TSR task under different sample sizes using the vortex data set.