

Neural Rendering and Its Hardware Acceleration: A Review

XINKAI YAN^{1,3}, JIETING XU¹, YUCHI HUO^{1,2}, and Hujun Bao^{1,2}

¹The State Key Laboratory of CAD&CG,Zhejiang University,Hangzhou 310058,China

²Zhejiang Lab,Hangzhou 311121,China

³Jiangsu Key Laboratory of ASIC Design (Wuxi),Wuxi 214153,China

ABSTRACT Neural rendering is a new image and video generation method based on deep learning. It combines the deep learning model with the physical knowledge of computer graphics, to obtain a controllable and realistic scene model, and realize the control of scene attributes such as lighting, camera parameters, posture and so on. On the one hand, neural rendering can not only make full use of the advantages of deep learning to accelerate the traditional forward rendering process, but also provide new solutions for specific tasks such as inverse rendering and 3D reconstruction. On the other hand, the design of innovative hardware structures that adapt to the neural rendering pipeline breaks through the parallel computing and power consumption bottleneck of existing graphics processors, which is expected to provide important support for future key areas such as virtual and augmented reality, film and television creation and digital entertainment, artificial intelligence and metaverse. In this paper, we review the technical connotation, main challenges, and research progress of neural rendering. On this basis, we analyze the common requirements of neural rendering pipeline for hardware acceleration and the characteristics of the current hardware acceleration architecture, and then discuss the design challenges of neural rendering processor architecture. Finally, the future development trend of neural rendering processor architecture is prospected.

INDEX TERMS Neural rendering, hardware acceleration, NRPU, MLP, ray marching, hash table.

I. INTRODUCTION

RENDERING refers to the process in computer graphics of generating a 2D image from a 3D scene. In contrast, inverse rendering is the reverse process, involving the reconstruction of a 3D scene from a series of input 2D images. Traditional rendering and inverse rendering have mature software stacks and are supported by graphics processing units (GPUs), making them applicable in various scenarios such as gaming, film production, visual exhibitions, as well as in the creation and generation of 3D content. However, traditional rendering and inverse rendering still suffer from issues such as reliance on manual operation by professionals and complexity in specialized software, making it challenging to meet the demands of professionals in other industries for the generation and visualization of 3D content.

With the rapid rise and development of deep neural networks, significant progress has been made in various fields of artificial intelligence, including computer vision, image processing, and natural language processing. Consequently, the integration of deep learning with computer graphics has naturally become a hot research direction. On one hand, leveraging data-driven automated design methods based on deep neural networks can replace manual design by professionals, while also utilizing the prior features extracted from a large amount of training data to accelerate the rendering process. On the other hand, methods based on deep learning can

integrate the entire inverse rendering process into gradient-based optimization, thereby achieving end-to-end training of deep neural networks. This integration results in faster inverse rendering and more robust outcomes.

Neural rendering, as a fusion approach between deep learning and computer graphics, has emerged as a pivotal concept. Eslami *et al.* [21] first introduced the notion of neural rendering by incorporating a Generative Query Network (GQN). This team employed GQN, which takes varying numbers of images and their corresponding camera parameters as input. Subsequently, it encodes complete scene information into a vector, utilizing this vector as input for the generative network to produce correctly occluded views. GQN innovatively learned a potent neural renderer from data without the need for human-annotated data. This renderer is capable of generating accurate, novel perspective scene images, although its rendering capabilities are limited. Nevertheless, it has inspired a plethora of subsequent work. Tewari *et al.* [97] provided a definition for neural rendering: Neural rendering is a method for generating images and videos based on deep learning. It combines deep neural network models with the physical knowledge of computer graphics to obtain controllable and realistic scene models, enabling control over scene attributes such as lighting, camera parameters, and pose. Compared to other areas of deep learning, the focal point of neural rendering is the fusion of physical/mathematical

knowledge from traditional rendering with the design of neural network structures. Furthermore, it integrates knowledge from the field of computer graphics into the design of neural networks. Generally, neural rendering views neural networks as a universal function approximation tool. It achieves this by training on real-world scene data and constructing loss functions to simulate the approximation of physical/mathematical laws. Therefore, unlike traditional rendering, the quality of the neural network's function approximation directly impacts the rendering quality. This implies that the quantity, distribution, and quality of the training input data are closely related to the final rendering effect.

II. APPLICATIONS OF RENDERING

Rendering refers to the process of generating an image or sequence of images from a 2D or 3D model (or scene) using computer software. In the context of computer graphics, rendering transforms the data in a virtual environment into a visual representation that can be displayed on a screen or printed. This process involves calculations for lighting, shading, texture mapping, and other visual elements to create a realistic or stylized final image. Recently, much progress has been seen in vanilla rendering and rendering plus AI.

Vanilla rendering technology focused on exploring efficient representation of 3D assets and the calculation of light transportation among 3D worlds. These works include Wang et al. [107] explores directed bounding box reconstruction for solid mesh modeling, editing and rendering scattering effects for translucent objects [106] or waters [58], interactive model generation system [27], [118], etc.

Deep learning technology enhances parts of the rendering process for higher quality and performance. Wang et al. [103] propose a biologically-based skin model for heterogeneous volume rendering, Zheng et al. [127] present neural light transform functions for global rendering pipeline on neural networks, enhance image resolution via neural super-resolution and frame prediction technologies [115], [130], [131], etc.

Deep learning-based Monte Carlo noise reduction by training a neural network denoiser through offline learning, it can filter noisy Monte Carlo rendering results into high-quality smooth output, greatly improving physics-based Availability of rendering techniques [38], common research includes predicting a filtering kernel based on g-buffer [7], using GAN to generate more realistic filtering results [117], and analyzing path space features Perform manifold contrastive learning to enhance the rendering effect of reflections [16], use weight sharing to quickly predict the rendering kernel to speed up reconstruction [22], filter and reconstruct high-dimensional incident radiation fields for unbiased reconstruction rendering guide [31], [37], etc.

The many-light rendering framework is an important rendering framework outside the path-tracing algorithm. Its basic idea is to simplify the simulation of the complete light path illumination transmission after multiple refraction and reflection to calculate the direct illumination from many virtual

light sources and provide a unified mathematical framework to speed up this operation [19], including how to efficiently process virtual point lights and geometric data in external memory [108], [109], how to efficiently integrate virtual point lights using sparse matrices and compressed sensing [35], [36], and how to handle virtual line light data in translucent media [34], sample important virtual point lights [40], use spherical Gaussian virtual point lights to approximate indirect reflections on glossy surfaces [32], and more.

Automatic optimization of rendering pipelines Apply high-quality rendering technology to real-time rendering applications by optimizing rendering pipelines. The research contents include automatic optimization based on quality and speed [55], [110], automatic optimization for energy saving [111], [126], LOD optimization for terrain data [53], automatic optimization and fitting of pipeline rendering signals [52], anti-aliasing [129], automatic shader simplification [33], [52], etc.

Some works use a physically-based process to guide the generation of data for single image reflection removal [47], propagating local image features in a hypergraph for image retrieval [4]–[6], managing 3D assets in [75], [83], [84], generating physically plausible grasp pose [50] or segmentation [120], automatic indoor lighter design [82], Seal3D [112] exploit rendering decomposition for interactive NeRF editing, using physical knowledge for 3D human reconstruction [13], [128], etc.

III. CURRENT STATUS REVIEW

A. RELATED REVIEWS

Eslami *et al.* [21] discuss where and how machine learning can improve classical rendering pipelines, as well as the data requirements for training neural rendering. The paper provides a brief overview of the fundamental knowledge in the fields of physical image generation and deep generative models, demonstrating the advantages and limitations of the hybrid approach of computer graphics and machine learning. The review further discusses the categorization of neural rendering methods, including control, computer graphics modules, explicit or implicit control, multimodal synthesis, and generality. It elaborates on applications related to neural rendering based on these categories, including new view synthesis, relighting, scene manipulation, and synthesis. The authors indicate that the purpose of this review is to enhance the research interest in neural rendering, thereby contributing to the development of the next generation of neural rendering and graphics applications.

In recent years, the surge in research related to Neural Radiance Fields (NeRF) [62] and its variants has confirmed the predictions of the DeepMind team. Tewari *et al.* [98] provide a comprehensive overview of different scene representations for neural rendering, briefly introducing the basic principles of classical rendering pipelines and machine learning components, with a focus on advanced aspects of combining classical rendering with learnable 3D representations. The literature cited in this review mainly focuses on NeRF and its vari-

ant optimization work in the past two years, encompassing applications ranging from free-viewpoint videos of rigid and non-rigid scenes to shape and material editing, light field and relighting, and digital human avatar generation. Additionally, Tewari *et al.* [98] also discuss the societal impacts brought about by neural rendering methods, including implications across industrial research and production, social ethics, and the environment.

In contrast to the aforementioned reviews, Wang *et al.* [105] specifically concentrate on the forward applications of neural rendering, i.e., rendering pipeline-related applications that combine deep neural networks with rendering components. The paper provides a brief introduction to the theoretical foundations of physics-based rendering and deep generative networks, focusing on how neural networks can replace or enhance the work of renderers in traditional rendering pipelines and the pros and cons of this combination. It covers general and specific methods for applications such as occlusion generation, volume and subsurface rendering, multi-scene representation rendering, global illumination rendering, direct illumination rendering, human-related rendering, and rendering post-processing. The authors assert the viewpoint that traditional graphics rendering pipelines can be partially or completely replaced by deep learning-enhanced rendering.

B. THE SCOPE OF THIS ARTICLE

Unlike the related reviews on the applications of neural rendering algorithms, this paper focuses more on the integration of neural rendering in both forward and inverse rendering aspects with traditional rendering pipelines or the construction of novel neural rendering pipelines, emphasizing the common hardware acceleration requirements. To provide a clearer understanding, this paper first briefly outlines the theoretical foundations of neural rendering and then introduces representative research achievements of neural rendering in forward rendering, inverse rendering, and post-processing applications. Subsequently, this paper meticulously analyzes the common hardware acceleration requirements for computing and storage in neural rendering applications, presents recent works related to hardware-accelerated neural rendering, and discusses the design challenges of neural rendering processor architectures.

IV. NEURAL RENDERING

Given a high-quality description of a scene, traditional rendering methods can produce realistic images of various complex real-world phenomena. However, constructing high-quality scene models requires a significant amount of manual work. In contrast, neural rendering methods take images corresponding to specific scene conditions (such as viewpoint, lighting, layout, etc.) as data inputs. Through training, these methods construct a scene representation based on a “neural network” and render this representation under new scene attributes, thereby synthesizing new images. The “deep neural networks” in neural rendering approximate real functions

through learning, enabling the resulting scene representation to be optimized for high-quality new images without being constrained by simple scene modeling approximations.

A. THEORETICAL FOUNDATIONS

This section discusses the theoretical foundations of neural rendering work. It first introduces the physics-based rendering methods in traditional computer graphics and then outlines the approach of generative models based on deep neural networks and the representation of 3D scenes.

1) Physics-Based Rendering

Ray Tracing is the traditional graphics pipeline models the formation of images as a physical process in the real world: photons emitted from light sources interact with objects in the scene, resulting in a bidirectional scattering distribution function (BSDF) determined by geometric and material properties, which is then captured by the camera. This process, known as light transport, can be represented using the classic rendering equation [45]:

$$L_o(p, \omega_o, \lambda, t) = L_e(p, \omega_o, \lambda, t) + L_r(p, \omega_o, \lambda, t) \quad (1)$$

Where L_o represents the radiance emitted from a surface, it is a function of position p , light direction $\vec{\omega}$, wavelength λ , and time t . L_e denotes the radiance emitted directly from the surface, while L_r accounts for the interaction of incident light with surface reflectance.

$$L_r(p, \omega_o, \lambda, t) = \int_{\Omega} f_r(p, \omega_i, \omega_o, \lambda, t) L_i(p, \omega_i, \lambda, t) (\omega_i \times n) d\omega_i \quad (2)$$

In (2), n represents the surface normal, i denotes the incident light direction, L_i represents the incident radiance, f_r signifies the Bidirectional Scattering Distribution Function (BSDF), and Ω denotes the hemisphere surrounding the surface point. The classical rendering equation is an integral equation, and the most accurate approximation is based on Monte Carlo algorithms, simulating light path traversal through the scene via sampling.

Forward rendering is the process of transforming a scene, including the camera, lighting, surface geometry, and materials, into a simulated camera image. Computer graphics provides various approximations for the rendering equation, with the two most common forward rendering methods being rasterization and ray tracing. Rasterization, a forward process, essentially converts the 3D scene into triangles, then projects them onto the screen using perspective projection, thereby transforming the 3D representation of triangles into a 2D representation. Ray tracing, on the other hand, is a recursive process where rays are cast from image pixels into a virtual scene, simulating reflection and refraction by recursively casting new rays from intersection points. Currently, mainstream GPUs primarily accelerate the rendering process based on rasterization, as it is parallel computation-friendly, suitable for hardware acceleration, and exhibits good memory coherence. However, due to the superior effects of ray tracing

in global illumination and other complex lighting scenarios (such as depth of field and motion blur), hardware manufacturers have recently added ray tracing hardware acceleration structures to their GPUs, thereby supporting ray tracing in real-time rendering pipelines [72], [101]. This allows for a blend of rasterization and ray tracing to achieve better rendering effects.

Inverse rendering, the process of estimating different model parameters (such as camera, geometry, material, and light parameters) from real data to generate new views or edit materials or lighting, addresses the problem of constructing a 3D scene that generates a given image. Due to mathematical complexity or computational expense, predefined physical models or data structures used in classical rendering do not always accurately reproduce all features of real-world physical processes. This constitutes a primary challenge in inverse rendering. Deep neural networks, however, can statistically approximate the physical processes of inverse rendering, leading to output data that more closely resembles training data, thereby enabling a more accurate representation of real-world effects.

2) Deep Neural Networks

Traditional computer graphics methods focus on modeling scenes using physical/mathematical formulas and simulating light transport to generate images. Deep neural networks, however, address this problem from a statistical probability perspective by learning the image distribution in the real world.

In the field of machine learning, the multilayer perceptron (MLP) is commonly utilized as a universal function approximator. It is a traditional fully connected neural network that takes spatial coordinates as input within the context of scene representation and produces corresponding output values. This type of network, also known as a coordinate-based neural network, serves as a function approximator for representing surfaces or volumes, as well as storing other properties. Since NeRF, a substantial amount of neural rendering work has adopted MLP as a function approximator.

Deep generative networks excel at generating random realistic images that resemble statistical data from the training set. While traditional generative adversarial networks (GANs) synthesize virtual images from random vectors resembling the training data, this is insufficient for scene rendering. Current neural rendering employs deep generative networks using perceptual distance for training or utilizes conditional generative adversarial nets (cGANs), both of which employ pre-trained auxiliary networks to define an effective training objective, resulting in improved generative networks and rendering effects.

Originating from the Transformer model proposed in natural language processing, the self-attention mechanism reduces the distance between any two positions in a sequence to a constant, thereby addressing long-range dependency issues. Through matrix calculations, it directly computes the correlation between each word without the need for passing

through hidden layers. This concept inspires and simplifies the task of neural rendering using MLP to learn the mapping from position to specific values, and it can also be used for geometric and appearance inference.

3) Three-Dimensional Scene Representation

To model objects in a 3D scene, various methods of scene representation have been proposed, mainly categorized into explicit and implicit representations. Both explicit and implicit methods can be used for surface and volume representations of scenes. Mainstream 3D data formats include depth, voxels [18], points, and meshes.

A depth map contains information related to the distance from the viewpoint to objects in the scene. The channel itself is similar to a grayscale image, where each pixel value represents the actual distance from the sensor to the object. Since a depth map contains information only about the foremost object surfaces in the scene, additional visibility calculations are unnecessary. Furthermore, depth maps are very friendly to neural networks and serve as a format well-suited for training and computation through neural networks.

Voxels are commonly used to represent volumes. They can store geometric occupancy, density values of scenes with volume effects (such as transparency), and the appearance of the scene [25]. Traditional rendering seldom utilizes voxels because it requires both visibility and shadow calculations. However, voxels are relatively friendly to neural rendering, as neural networks can use voxels to learn end-to-end rendering pipelines.

A point cloud is a set of elements in Euclidean space, where the continuous surface of a scene can be discretized using a point cloud. Each element of the point cloud represents a sample point (x, y, z) on the surface. Additionally, point clouds can also represent volume (such as storing opacity or density values). In neural rendering, point clouds can store some learnable features [1], [114], thereby projecting and fusing individual features into 2D feature maps, mainly applied in denoising, stylization, and novel view synthesis [114]–[48]. However, due to the association of point cloud data with the quantity, order, and sampling of surface points, it is relatively unfriendly to neural networks.

Polygon mesh represents a piecewise linear approximation of a surface and is widely used as a standard representation in most graphic editing tools in traditional computer graphics. To be compatible with traditional rendering pipelines, many neural rendering methods use mesh representation.

4) Neural Scene Representation

In the context of neural rendering, using neural networks to approximate the surface or volume representation function of a scene is referred to as neural scene representation. Both surface and volume representations can store additional information, such as color or radiance.

The signed distance function (SDF) is the most commonly used implicit surface representation. It returns the shortest distance from any point in space to the object surface. All

points with a distance of 0 represent the object surface, points with a distance less than 0 represent the interior of the object, and points with a distance greater than 0 represent the exterior.

Neural implicit surfaces [15], [76], [116], [133]–[135] in neural rendering are obtained by combining the signed distance function with Multilayer Perceptron (MLP) and are used for shape modeling. MLP maps continuous coordinates to signed distance values. This representation method has been widely applied in neural scene representation and rendering. Neural implicit surface representation has many advantages, such as high memory efficiency and theoretically being able to represent geometry at infinite resolution. Similar to neural implicit surfaces, neural voxels represent voxels using neural networks instead of using voxel grids to store features or other variables. In neural rendering, coordinate-based neural networks are generally used to model scenes volumetrically. Since MLP networks can be used to parameterize volumes, neural voxel representation may be more efficient than explicit voxel grid representation.

Volume rendering [18] is based on ray casting and is mainly used for rendering non-rigid objects such as clouds and smoke. It divides the process of photon-particle interaction into four types: absorption, emission, external scattering, and internal scattering. NeRF [62] and related works [60]–[8] have demonstrated good performance in learning scene representation from multi-view input data using volume rendering, which can be utilized in neural rendering-based inverse rendering frameworks.

B. APPLICATION

This section briefly introduces the latest or representative applications of neural rendering technology in the three categories of forward rendering, inverse rendering, and post-processing.

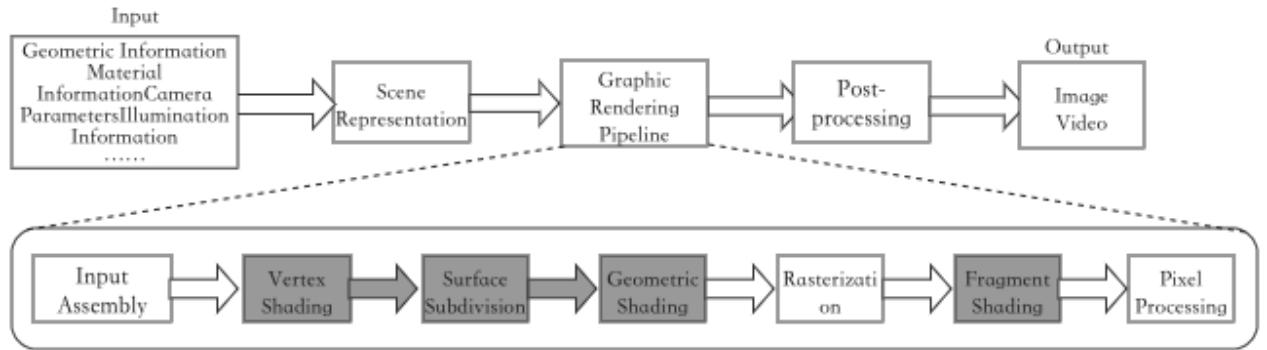
1) Forward Rendering

Fig. 1 illustrates the traditional forward rendering pipeline, which consists of three stages: scene representation, graphics rendering pipeline, and post-processing. The role of neural rendering technology in forward rendering typically involves enhancing various sub-modules of the traditional rendering process, such as voxel scene representation and global illumination.

Table I presents relevant research on neural rendering in various rendering sub-modules and the input data types for the networks. Voxel-based Scene Representation. Fig. 2 depicts the end-to-end neural voxel rendering process implemented by Render Net [67]. Initially, it transforms the input voxel grid, camera pose, and light source position into camera coordinates. Subsequently, a 3D convolutional neural network (3D CNN) converts the camera coordinates into a 4D tensor (H, W, D, C) representing neural voxels. Following this, a projection unit transforms the neural voxels into a 3D tensor representing neural pixels (with matrix dimensionality transformation for the depth dimension D and feature channel dimensionality C). Finally, a 2D CNN network performs deconvolution op-

erations to render the projected neural voxels into images. Neural voxel renderer [81] proposes a deep learning-based rendering method to map voxelized scenes to high-quality images. The authors design two neural renderers, NVR and NVR+, for rendering scenes. NVR and Render Net are quite similar, with the distinction that NVR's projection unit uses a 2-layer MLP to process light positions, thereby encoding lighting information. NVR+ is discussed in Section V-B2. Point Cloud-based Scene Representation. Neural point-based graphics [2] introduces a method for modeling the appearance of real scenes based on point clouds. The authors use the original point cloud as the geometric representation of the scene and enhance each point cloud through encoding local geometry and appearance, using pre-trained neural descriptors. The article uses Z-Buffer to rasterize points at several resolutions as input data for the neural rendering network, employing neural descriptors as pseudo-colors. The network architecture is similar to U-Net, rendering images through rasterization. This model adapts to new scenes by optimizing the parameters of the neural rendering network and propagating the loss function through backpropagation of the neural descriptors. Neural point cloud [20] further proposes a neural point cloud rendering pipeline using multi-plane projection (MPP). The authors facilitate the automatic learning of the visibility of 3D points by projecting 3D points into a layered volume within the camera frustum. The network framework consists of two modules: voxelization based on multiple planes and multi-plane rendering. The voxelization module uniformly divides the 3D space of the camera frustum into small voxels based on the image size and a predefined number of planes, aggregating each small frustum and generating a multi-plane 3D representation. The plane rendering module, as the rendering network, predicts a 4-channel output (RGB + blending weight) for each plane, ultimately blending all planes based on the blending weights to produce the rendered image. Neural light transport [87] proposes a method using neural networks to learn light transport in static and dynamic 3D scenes. Fig. 3 illustrates the specific model structure of NLT. The neural rendering network architecture primarily comprises three modules: light transport layer, transformation and projection layer, and image synthesis layer. The light transport layer uses MLP to achieve the transformation from feature-rich point clouds to processed features, while the image synthesis layer employs a U-Net-like architecture with residual blocks to achieve image rendering. In comparison with existing 2D image-domain methods, the proposed approach supports inference in both 3D and 2D spaces, thereby achieving global illumination effects and 3D scene geometry processing. Additionally, this model can produce realistic rendering of static and dynamic scenes.

Based on network-based scene representation. Differentiable volumetric rendering (DVR) [69] proposes a differentiable rendering formula for implicit shape and texture representation. The authors devise an occupancy network that assigns occupancy probabilities to each point in 3D space, using isosurface extraction techniques to extract object

**FIGURE 1. Pipeline of forward rendering.****TABLE 1. The Application of Neural Rendering in Forward Rendering**

Stage Classification	Submodule Classification	Work	Input Data Type	Generalization
Scene Representation	Voxel-based	render net [67]	Voxel Camera Pose Light Source Position	Generalized
		neural voxel renderer [81]	Voxel Camera Pose Light Source Position	Generalized
	Point Cloud-based	neural point-based graphic [2]	Point Cloud Descriptors Camera Pose	Generalized
		neural point cloud rendering [20]	Point Cloud Camera Pose	Specialized
		neural light transport [87]	Point Cloud Scene Features	Specialized
	Network-based	differentiable volumetric [69]	Pixel	Specialized
		neural implicit surfaces [119]	Image Set/SDF	Specialized
Global Illumination	Indirect Illumination	neural light transport [87]	Point Cloud Scene Features	Specialized
		deep illumination [99]	Normal Map Indirect Illumination Diffuse Map Depth Map	Generalized
		neural control variates [65]	Camera Pose Probability Density	Generalized
	Direct Illumination	neural screen space rendering [94]	Material Geometric Shape Lighting	Generalized
Spatial Effects	Ambient Shadows	deep shading [66]	Position Normal Reflectance Color	Generalized
	Ambient Shadows	AOGAN [85]	Position Normal	Generalized

surfaces, while directly learning implicit shape and texture representations from RGB images using texture fields [73]. Yariv *et al.* [119] define the volume density function as the cumulative distribution function (CDF) of the Laplacian, applied to Signed Distance Function (SDF) representations. The authors argue that this simple density representation has three advantages: first, it provides useful inductive biases for learning geometry in the neural volume rendering pro-

cess; second, it aids in constraining approximation errors in opacity, leading to precise sampling of observed light rays; and third, it allows effective unsupervised disentanglement of shape and appearance in voxel rendering. The network framework consists of two Multilayer Perceptrons (MLPs): the first MLP is used to approximate geometric SDF and 256-dimensional global geometric features, while the second MLP is used to represent the scene radiance field. Global

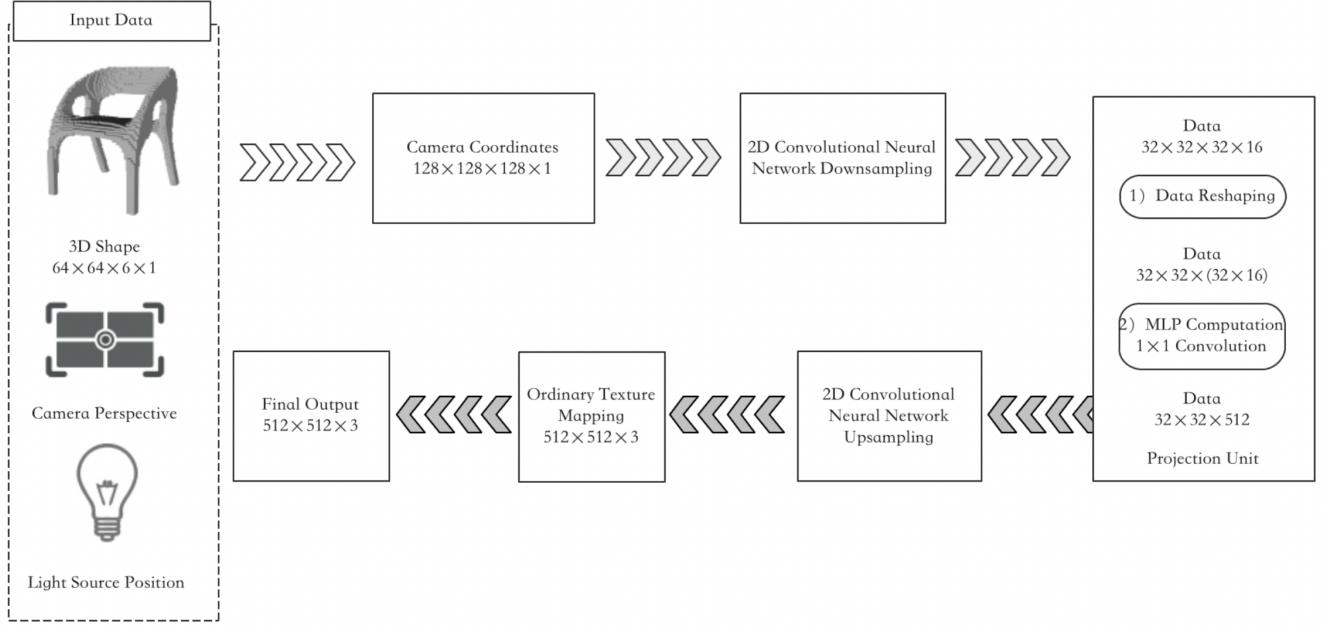


FIGURE 2. The network architecture of render net.

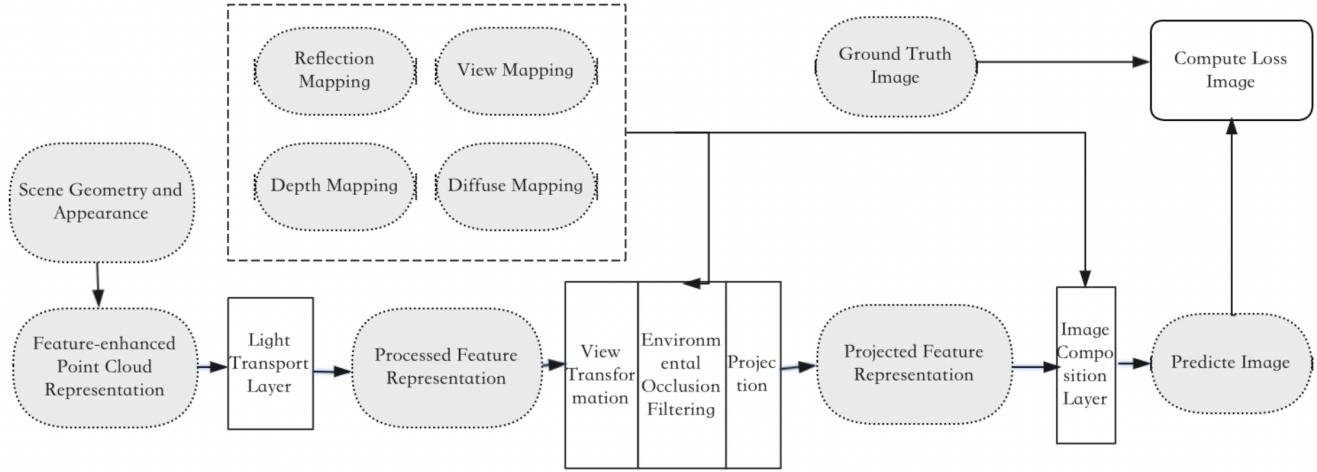


FIGURE 3. The model architecture of neural light transport.

illumination. Deep illumination (DI) [99], in real-time applications, approximates global illumination using conditional Generative Adversarial Networks (cGAN). The authors propose a mapping network from the G-buffer (depth, normal, and diffuse albedo) and direct illumination to any global illumination, where the generative model can be used to learn the density estimation of advanced lighting models from the screen-space buffer to the 3D environment. Neural control variates (NCV) [65] propose a method to reduce unbiased variance in parameter Monte Carlo integration. The authors use a neural network to learn a function close to the rendering equation, a neural sampler to generate sampling probabilities,

and a neural network to infer the solution of the integral equation. The neural direct-illumination renderer (NDR) [94] can render direct illumination images of any geometric shape with opaque materials under distant lighting. Fig. 4 illustrates the specific model structure of NDR. First, the screen-space buffer (material, geometric shape, and lighting) is input into a CNN network to learn the mapping from pixel buffer to rendered image, then NDR predicts the shadow map and combines it with the irradiance map to render the image.

Environment Shading. Deep shading [66] introduces a novel technique for generating various rendering effects, including environment shading using deferred shading buffers

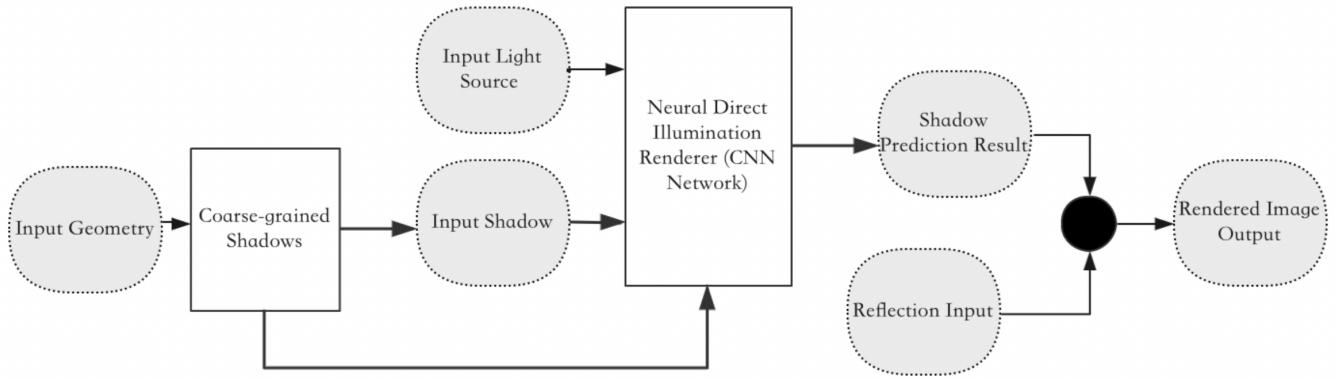


FIGURE 4. The model architecture of neural direct-illumination renderer.

and convolutional neural networks. The authors employ deep neural networks as renderers, taking deferred shading buffers as input to produce specific rendering outcomes. AOGAN [85] presents an end-to-end generative adversarial network for producing realistic environment shading. The authors explore the significance of perceptual loss in the generation model’s accuracy for environment shading by combining VGG-structured perceptual loss with GAN-structured adversarial loss. Additionally, they introduce a self-attention module, taking position and normal shading buffers as input, to enhance the training’s generalization.

2) Inverse Rendering

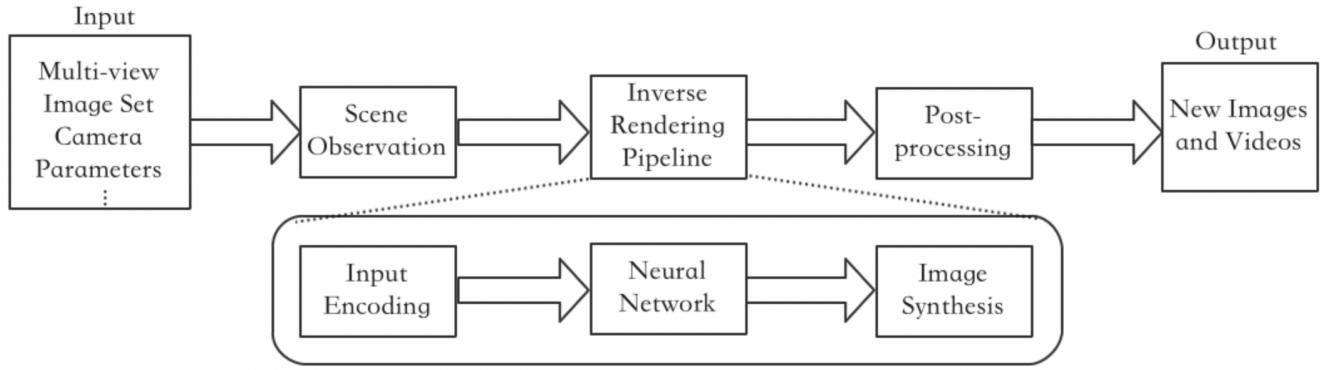
Fig. 5 illustrates the schematic of the inverse rendering pipeline, divided into three stages: scene observation, inverse rendering pipeline, and post-processing. The role of neural rendering techniques in inverse rendering typically involves utilizing deep neural networks for tasks such as novel view synthesis, relighting, and material editing. Table. II provides an overview of relevant research in neural rendering for inverse rendering, along with corresponding types of generative networks and 3D scene representation types.

Novel View Synthesis based on 3D Voxel Representation. The neural voxel renderer NVR+ network proposed by neural voxel renderer [81] extends the NVR network by incorporating a splatting network and a neural re-rendering network. The splatting network initially synthesizes images by splitting the colored voxel centers in the target view and then passes this image, along with the feature vector generated by NVR, to a 2D convolutional encoder. The final result is processed by the neural re-rendering network U-Net to generate the output image. Deep-voxels [89] encode the visual appearance of 3D scenes, achieving end-to-end inverse rendering. It first employs a 2D U-Net network to extract 2D feature maps, then utilizes differentiable lifting layers to raise 2D features to 3D feature volumes. Subsequently, a 3D U-Net processes the feature volume after fusion, mapping the feature volume to the camera coordinate systems of two target views through differentiable reprojection layers. An occlusion network calculates the soft visibility of each voxel, and finally, a 2D U-Net

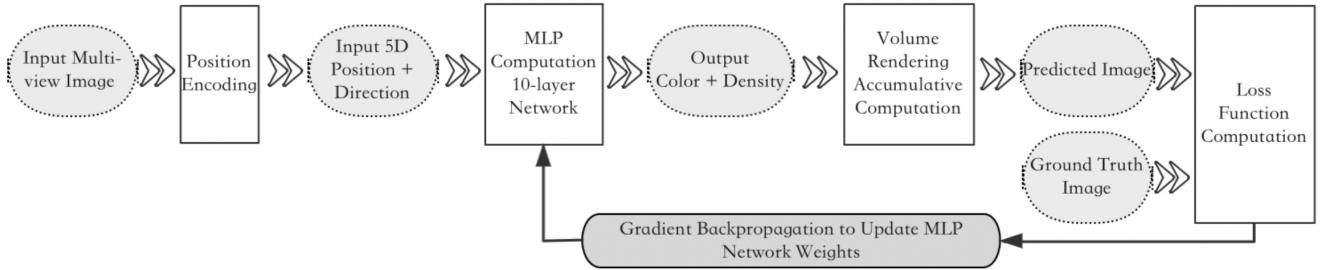
rendering network generates two final output images. Neural volume [59] introduces a learning-based approach to represent dynamic objects, employing an encoder-decoder network to convert input images into 3D voxel representations while using differentiable ray-marching operations, thus enabling end-to-end training. The innovation of neural volume lies in the fact that any input image can serve as supervision through re-rendering loss, eliminating the need for explicit reconstruction or target tracking. The emergence of Neural Radiance Fields (NeRF) [62], based on Multi-Layer Perceptron (MLP) networks for scene representation, marks a breakthrough in the application of realistic novel view synthesis within a single scene. Fig. 6 illustrates the NeRF model structure and its differentiable rendering process. NeRF directly applies voxel rendering to synthesize images from the MLP. The MLP maps positions and directions to volume density and color. Subsequently, it optimizes the MLP weights based on pixel-level rendering losses from input images to represent each new input scene. The primary innovation of NeRF lies in the introduction of positional encoding, enabling effective differential compression of scenes during the optimization process, thereby achieving higher resolution compared to discrete 3D voxel representations without increasing the number of MLP weights. However, the computation of color and volume density for individual points within NeRF introduces significant computational overhead, resulting in slow rendering speeds. Additionally, NeRF suffers from poor generalization, supports only static scenes, and requires extensive sample training.

Several efforts [57]–[24] [121] have focused on optimizing the NeRF architecture to improve rendering speeds. AutoInt [56] and Light Field Networks [88] enhance rendering speeds by training the MLP itself. Müller *et al.* [64] optimize NeRF training speed by introducing multi-resolution hash encoding to alleviate MLP training burdens, thereby accelerating training speeds.

Further enhancements based on NeRF have been explored. Pumarola *et al.* [77] and Li *et al.* [54] have improved NeRF for static scenes. D-NeRF [77] introduces time as an input, dividing the learning process into two main stages: en-

**FIGURE 5.** The pipeline of inverse rendering.**TABLE 2.** The Application of Neural Rendering in Inverse Rendering

Functionality Classification	Work	Generation Model	3D Scene Representation	Input Data Type
New View Synthesis (Voxel-based)	neural voxel renderer [81] deep voxels [89] neural volumes [59]	U-Net U-Net None	Voxel Voxel Voxel	Voxel Voxel Image
New View Synthesis (NeRF-based)	NeRF [62], autoint [56] neural sparse voxel fields [57], kiloNeRF [79] baking NeRF [28], fastNeRF [24], GeoNeRF [42], plenoclouds [121], light field networks [88], instant neural graphics primitives [64] D-NeRF [77], neural scene flow fields [54] GIRAFFE [68] pixelNeRF [122] IBRnet [104], common objects in 3D [80]	None None None None CNN None None	MLP Neural Voxel + Grid Grid MLP Neural Feature Field Neural Voxel Neural Voxel	Image Camera Pose Image Camera Pose Image Camera Pose Image Camera Pose Image Camera Pose Image Camera Pose Image Camera Pose Image Camera Pose Image Camera Pose
New View Synthesis and Scene Reconstruction (SDF-based)	deepSDF [76] I2-SDF [134] scene representation networks [90]	CNN MLP CNN	SDF SDF SDF	Image Image Image Camera Pose Image Camera Pose Image Camera Pose
Relighting and Material Editing	I2-SDF [134] neural reflectance fields [10] neural light transport [125] NeLF [93]	MLP CNN MLP MLP	SDF Neural Voxel Neural Voxel Neural Feature Vector	Image Image Image Image Camera Pose Light Parameters Image Light Parameters Portrait Camera Pose

**FIGURE 6.** The model architecture and differentiable render procedure of NeRF.

coding scenes into canonical space and mapping canonical representations to deformed scenes at specific times, utilizing fully connected networks. Meanwhile, neural scene flow fields [54] model dynamic scenes as time-varying continuous functions of appearance, geometry, and 3D scene motion. In recent years, numerous efforts have focused on optimizing the generalization of NeRF. GIRAFFE [68] learns local features for each pixel in 2D images and projects these features into 3D points, thereby generating a universal and rich point representation. GeoNeRF [42] utilizes Transformers to render and infer geometry and appearance, while using volume rendering to capture image details, thus achieving a NeRF-based generalized realistic novel view synthesis method. PixelNeRF [122] aggregates features across multiple views within a voxel rendering framework, thereby rendering in a NeRF-like manner to produce color and density fields using MLP. When trained on multiple scenes, they learn scene priors for reconstruction, enabling high-fidelity scene reconstruction from a few viewpoints. IBRNet [104] introduces Ray-Transformer networks to learn a universal view interpolation function, supporting generalization to new scenes. NeRFormer [80] employs Transformer networks to reconstruct objects with few viewpoints.

Based on SDF for new view synthesis. DeepSDF [76] proposes a learnable continuous Signed Distance Function (SDF) representation, enabling high-quality shape representation, interpolation, and completion from partial and noisy 3D input data. Scene Representation Networks (SRNs) [90] primarily consist of an MLP scene representation module, a ray-stepping LSTM module, and a pixel generator module. The authors achieve end-to-end training of a set of images without explicit supervision based on the SDF implicit function, enabling new view synthesis, few-shot reconstruction, shape and appearance interpolation. I2-SDF [134] recovers basic shape, incident radiance, and material from multi-view images based on an overall neural SDF framework. By employing surface-based differentiable Monte Carlo ray tracing and radiance primitive segmentation, the neural radiance field is decomposed into spatially varying material within the scene, demonstrating superior quality in indoor scene reconstruction, new view synthesis, and scene editing.

Relighting and material editing. Neural Reflectance Fields [10] propose the first extension of NeRF to achieve relighting, representing the scene as a volume density field, surface normals, and Bidirectional Reflectance Distribution Functions (BRDFs) field, allowing scene rendering under arbitrary lighting conditions. The method evaluates the amount of incoming light rays reflected from a particle at each 3D position to the camera using predicted surface normals and BRDFs, reducing MLP computation by training only on images illuminated by a single-point light co-located with the camera. Neural Reflectance and Visibility Fields [125] train an MLP to approximate light source visibility for any input 3D position and 2D incident light direction. The visible light MLP requires only one query per incident light direction, enabling the recovery of a scene's relighting model from

images with pronounced shadows and self-occlusion effects. NeLF proposes a human portrait view synthesis and relighting system, using a neural network to predict the light transport field in 3D space and generating human portraits from the predicted neural light transport field in new environmental lighting, achieving simultaneous view synthesis and relighting for given multi-view human portraits.

3) reprocessing

In the realm of real-time rendering, post-processing stands as one of the common techniques due to the significant challenges posed by higher resolutions, refresh rates, and more realistic effects. Traditional post-processing methods have relied on linear interpolation techniques such as bilinear interpolation algorithms [61], bicubic interpolation algorithms [46], among others. However, neural rendering employs neural networks in the post-processing stage to alleviate the burden on the rendering pipeline. Super-resolution and frame interpolation techniques operate at lower resolutions or frame rates within the rendering pipeline and restore the target resolution and frame rate through deep learning methods. Post-processing methods based on deep learning leverage convolutional operations to automatically extract image details, address low-level visual issues, and generate higher quality high-resolution images (e.g., denoising, deblurring). For instance, they use residual networks to enhance inter-layer connections, transmitting low-resolution image feature information to deeper layers to mitigate gradient vanishing and feature loss issues. Additionally, they employ perceptual loss and adversarial loss information from multiple generators in generative adversarial networks to enhance the realism of high-resolution images.

In 2019, NVIDIA introduced the Deep Learning Super Sampling (DLSS) technology, now updated to version 3.0 [71], which utilizes deep learning-based super-resolution techniques to construct clearer, higher-resolution images while reducing pixel rendering. The DLSS convolutional autoencoder frame generator receives four input data: the current game frame, the previous game frame, the optical flow field generated by the motion vector accelerator, and game engine data (e.g., motion vectors and depth). For each frame, the DLSS frame generation network determines how to use information from the game motion vectors, optical flow field, and subsequent game frames to generate intermediate frames. Zhang *et al.* [123] employed a hybrid network architecture of CNN and Transformer to explicitly extract motion and appearance information uniformly, reducing the computational complexity of inter-frame attention while retaining low-level structural details. Additionally, Bi *et al.* [9] proposed a method to enhance the visual realism of low-quality synthesized images. The authors initially aimed at physics-based rendering, learning to predict accurate shadows in a supervised manner, and then used an improved cycleGAN network [136] to further enhance the realism of texture and shadows.

V. HARDWARE ACCELERATION

As neural rendering technology becomes widely applied across various domains of real-time rendering, the computational demands for neural rendering increase with the complexity of the neural networks used. Therefore, the exploration of how to better accelerate neural rendering applications at the hardware level has become a research hotspot. Among the 6 papers accepted in the emerging vision/graphics/AR-VR subtopic of the International Symposium on Computer Architecture (ISCA) in 2023, 4 are related to hardware acceleration for neural rendering. This trend underscores the growing importance of hardware acceleration structures for neural rendering.

Hardware architects should systematically consider the following three questions:

- The common requirements and performance bottlenecks of neural rendering tasks.
- Whether existing general-purpose processors (CPUs), graphics processing units (GPUs), or general-purpose graphics processing units (GPGPUs) can meet the computational and memory access requirements of neural rendering tasks, and whether emerging artificial intelligence processors (neural processing units, NPUs/tensor processing units, TPUs) can provide the operators or special hardware support required for neural rendering tasks.
- If the answer is negative, then whether neural rendering requires dedicated hardware acceleration support.

A. COMMON ALGORITHMS

Neural rendering tasks employ various types of neural network algorithms as well as some typical algorithms from traditional computer graphics. This section analyzes the common requirements of the neural rendering techniques introduced in the section on the hardware acceleration.

1) Neural Networks

Since neural rendering is a deep learning-based technology, the deep neural networks utilized by neural rendering algorithms form the core of each algorithm. Table. III summarizes the types of neural networks used and the architectural components of the backbone networks employed in the works introduced in Section V. The residual layer (resblk) is a computational layer where the input of the first unit is linearly stacked with the output of the second unit and then activated, mitigating the vanishing gradient problem. The concatenation layer (concat) connects multiple input data to one output, achieving data concatenation and allowing for alterations in data dimensions and an increase in the number of channels.

Multilayer perceptrons (MLPs) mainly consist of fully connected layers (FC) and activation layers. Convolutional Neural Networks (CNNs) are primarily composed of convolution layers, activation layers, batch normalization layers (BN), pooling layers, and fully connected layers. Generative Adversarial Networks (GANs) constitute a system comprising two models, a generator (G), and a discriminator (D), with

the main network architecture of both being convolutional layers. U-Net, an end-to-end encoder-decoder structure, primarily comprises convolutional layers in the encoder and deconvolutional layers in the decoder. Transformer, fundamentally an encoder-decoder structure, is characterized by the self-attention mechanism as its most crucial component, requiring two matrix-matrix multiplications. Subsequently, the multi-head attention module concatenates and linearly transforms the outputs of H self-attention modules to obtain the final feature matrix. The feature matrix then undergoes normalization residual (add and norm) layers before being fed into a feed-forward neural network (FFN), which consists of a network with 2 FC layers. The primary common operators in deep neural networks, as shown in Table. IV, include convolution/deconvolution, matrix-matrix multiplication, matrix-vector multiplication, pooling/unpooling, element-wise multiplication/addition operations, activation functions, and normalization. Operators such as convolution and matrix-matrix multiplication represent the most computationally intensive operations in neural networks, necessitating specialized hardware units for acceleration, such as systolic arrays, dot products, and multi-level accumulation trees. These hardware units demonstrate superior acceleration effects compared to general-purpose processors utilizing single instruction multiple data (SIMD) and graphics processors utilizing single instruction multiple thread (SIMT) data parallelism. For operations like pooling/unpooling and activation functions, although they do not contribute significantly to computational load, their inherently serial execution characteristics occupy the program's critical path runtime, thus requiring dedicated hardware units for acceleration, such as parallel look-up tables, transcendental function/interpolation special operation units. As for element-wise operations, although they do not entail significant computational load, they demand high memory bandwidth. While not mandating the design of specialized hardware computing acceleration units, they necessitate optimized memory access paths to enhance operation efficiency, such as designing dedicated on-chip network structures or optimizing memory units to support multiple data transfer modes.

2) Ray Marching and Volume Rendering

Ray marching refers to the incremental progression of a ray from its origin, halting at each step to perform computations before continuing until it approaches the endpoint. This method is commonly employed in implicit surface, point cloud, or voxel rendering. Many neural rendering tasks, particularly those akin to Neural Radiance Fields (NeRF), utilize ray marching for rendering, replacing traditional ray tracing or rasterization. Table. V enumerates the relevant works employing ray marching techniques.

In general, the ray marching in neural rendering tasks is mainly executed in three steps: first, determining the number of rays to be generated; then, establishing the stepping points for the current ray propagation; and finally, querying an MLP network based on the coordinates to calculate the color and

TABLE 3. The Architecture of Neural Network in Neural Rendering Application

Neural Network Type	Work	Residual Layer	Concatenation Layer
MLP	GIRAFFE [68],neural light transport [87],neural relighting [125],NeLF [93]	required	required
	deepSDF [76],pixelNeRF [122],IBRnet [104], neural radiosity [26], common objects in 3D [80],NeRF [62],NeRF in the wild [60], render net [67] ,neural voxel renderer [81],neural volumes [59] neural sparse voxel fields [57],kiloNeRF [79],baking neural radiance fields [28], fastNeRF [24],plenoclouds [121],autoint [56],light field networks [88], instant neural graphics primitives [64],neural scene flow fields [54], scene representation networks [90],extracting motion and appearance [123] ,instant 3D [51]	not required	required
CNN	GIRAFFE [68], render net [67],neural voxel renderer [81] [87],texture fields [73], neural reflectance fields [10],extracting motion and appearance [123]	required	required
	neural volumes [59],neural point cloud rendering [20], neural screen space rendering [94],deep shading [66]	not required	required
	neural control variates [65],AOGAN [85],pixelNeRF [122], scene representation networks [90]	required	not required
	neural light-transport field [93]	not required	not required
GAN	deep illumination [99]	not required	required
	AOGAN [85]	not required	not required
U-Net	deep voxels [89],neural point-based graphics [2],deep illumination [99]	not required	required
Transformer	IBRnet [104],common objects in 3D [80],GeoNeRF [42], extracting motion and appearance [123],gen-NeRF [23]	required	required

TABLE 4. The Neural Network Operator Requirement of Neural Rendering Application

Operator	Main Purpose	Operator Features	Hardware Acceleration Requirement	Neural Rendering Process
Convolution/Deconvolution	Feature extraction Data upscaling Image synthesis	High computational demand	Tensor/matrix computation units	Training/Rendering
Matrix-Matrix Multiplication	Attention mechanism Coordinate transformation Feature mapping	High computational demand	Tensor/matrix computation units	Training/Rendering
Matrix-Vector Multiplication	Backpropagation Attention mechanism Probability generation	High computational demand	Tensor/matrix computation units	Training/Rendering
Pooling/Unpooling (Max, Average)	Dimensionality reduction of data downsampling Increasing receptive field	Located on critical path	General computation units Execution pipelines	Training/Rendering
Element-wise Operations (Multiplication, Addition)	Residual modules, Bias	High memory access demand	General computation units Data transfer units	Training/Rendering
Activation Functions (softmax, ReLU, tanh, sigmoid, etc.)	Probability generation Function approximation	Located on critical path	Table lookup units Execution pipelines	Training/Rendering
Normalization	Gradient vanishing	Located on critical path	General computation units Execution pipelines	Training/Rendering

TABLE 5. The Ray Marching in Neural Rendering Application

Work	Type	Purpose	Neural Rendering Process
Neural Volumes [59]	Accumulation-Ray Marching	Render Voxel	Rendering
NeRF [62] and its variants [28]– [80]	Ray-marching-Ray Stepping	Render Voxel	Rendering
Common Objects in 3D [80]	Radiance Absorption-Ray Stepping	Render Transparency Field	Rendering
Scene Representation Networks [90]	Differential-Ray Stepping	Render Voxel	Training

writing it back to the pixel. The core computation in ray marching is the GetDist function, which takes the coordinates of a point in space as input and returns the distance from this point to the nearest object in space. Typically, the SDF function is used to calculate the distance from a point in space to a spherical surface, as shown in (3):

$$\Phi_{\text{circle}}(x) = \|x - c\| - r \quad (3)$$

In (3), x represents the coordinates of a point in space, c denotes the center coordinates of the sphere, and r represents the radius of the sphere. When the function value is greater

than 0, it indicates that the coordinates are outside the various objects in the scene. When the function value is less than 0, it indicates that the coordinates are inside the various objects in the scene. When the function value equals 0, it indicates that the coordinates are on the surfaces of the various objects in the scene. Because the calculation of the 3D coordinates in (3) can be geometrically transformed into a vector modulus problem, the calculation of ray marching can be transformed into vector multiplication, square root, and scalar addition. Hence, the computational demand for ray marching is a general computational unit, without the need for specialized

hardware acceleration units.

Neural rendering tasks, especially NeRF-like tasks, require the use of traditional volume rendering calculations. This involves computing the final color of the sample point by applying the colors and density results obtained from querying the MLP network through the volume rendering formula. The discrete volume rendering calculation formula is represented as (4):

$$\hat{C}(r) = \sum_{k=1}^N T_k (1 - \exp(-\sigma_k (t_{k+1} - t_k))) c_k \quad (4)$$

$$T_k = \exp \left(- \sum_{j=1}^{k-1} \sigma_j (t_{j+1} - t_j) \right) \quad (5)$$

From (4), it can be inferred that the computational requirements for volume rendering consist of general computational units and lookup table units, aligning with the hardware acceleration unit requirements for neural network computations.

3) Other Considerations

In addition to neural networks, ray marching, and volume rendering calculations, linear interpolation is also a common requirement. Table. VI presents the usage of interpolation calculations in neural rendering work. Linear interpolation calculations can be applied not only in end-to-end rendering pipelines for neural rendering but also in neural rendering pipelines combined with traditional post-processing pipelines. The computational demand for linear interpolation involves general computational units and special function units, aligning with the hardware acceleration requirements of pooling layers and pointwise operation layers in ray marching and neural networks. Furthermore, the introduction of multi-resolution hash encoding in Instant NGP [64] has proven to accelerate the training process of NeRF-like neural rendering tasks. Hash encoding issues can be transformed into hardware look-up tables, aligning with the hardware acceleration requirements of activation function operators in neural networks.

B. REQUIREMENTS ANALYSIS

Based on the common requirements discussed in Section V-A, this section separately discusses the neural rendering task in terms of training and inference processes.

1) Training Process

The training process of the neural rendering forward task is fundamentally similar to the training processes of other deep learning applications. It involves the use of the backpropagation algorithm for training, with data types being single-precision floating-point or half-precision floating-point data. The general steps include initially pretraining the model in a supervised or unsupervised manner on a large dataset and then fine-tuning the pretrained model to adapt to specific downstream tasks on a test set. The training process of the

neural inverse rendering task slightly differs from training processes based on deep learning applications. Fig. 7 presents a comparison of the training processes for NeRF [62], Instant NGP [64], and Instant 3D [51]. In the original NeRF framework, training for a single scene typically requires convergence over 100,000 to 300,000 iterations. Each batch processing scale, i.e., the number of ray sampling points per pixel, amounts to 192 points/pixel multiplied by 4,096 pixels. Therefore, each iteration necessitates 786,432 MLP queries, resulting in a training computational load approaching 10^{17} . Utilizing a single NVIDIA V100 GPU, the training duration typically requires approximately 1 to 2 days. Instant NGP achieves comparable training effects to the original NeRF in a matter of 5 seconds on the same hardware platform through algorithmic optimizations. Li *et al.* [51] expedited the NeRF training process through a collaborative design of software and hardware. They proposed the Instant 3D algorithm based on Instant NGP and its corresponding hardware acceleration structure. By decoupling the color and density branches of the embedded grid without compromising reconstruction quality, they compressed NeRF computations to enhance training efficiency. Consequently, they achieved a training time of 1.6 seconds and a training power consumption of 19,000 watts per scene, with virtually unchanged quality on AR/VR devices. This represents a reduction in training time by 41 to 248 times compared to traditional NeRF-like applications. Subsequent work on neural rendering training has predominantly adopted hash encoding as a means of training acceleration.

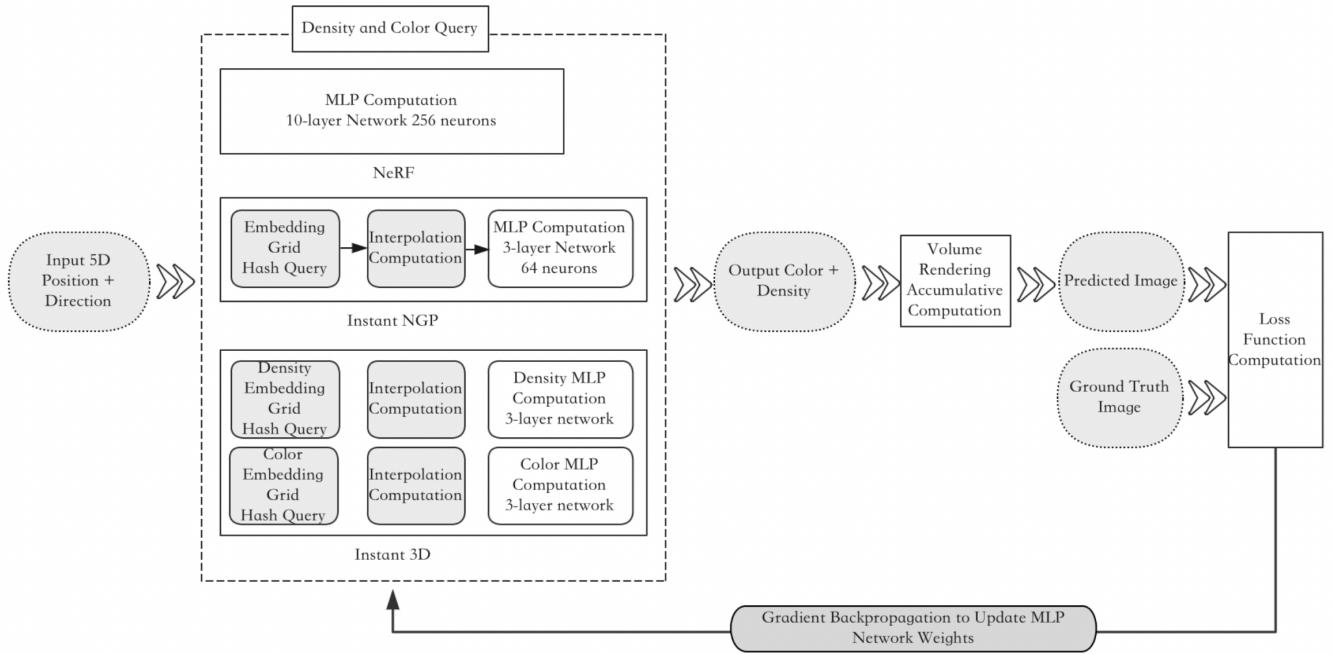
Numerous comprehensive works have already addressed training methodologies for various types of neural networks. Zhou *et al.* [132] discussed supervised and unsupervised learning methods for Convolutional Neural Networks (CNNs). Wang *et al.* [113] examined the primary challenges faced in single-machine and distributed training, providing an overview of optimization algorithms and representative achievements across research branches. Additionally, Zhang *et al.* [124] and Yin *et al.* [41] respectively explored the training aspects of Transformer network architectures in image and video tasks. Therefore, this paper does not delve into the general training issues of neural networks.

2) Rendering Process

The rendering process for neural rendering tasks closely resembles the inference process of other deep learning applications. The data types involved are typically half-precision floating point or integer. The general steps include input encoding, neural network queries, and finally, pixel output generation. The rendering process primarily employs neural networks for feature extraction, coordinate transformation, and image synthesis functions. Algorithms utilizing MLP coordinate projection combined with CNN generation networks as the rendering framework require substantial convolutional and matrix computational power. For instance, Render Net employs a 1-layer $32 \times 32 \times 512$ scale MLP network for coordinate projection and an 8-layer CNN network for generating $512 \times 512 \times 3$ RGB images. In neural rendering algorithms

TABLE 6. The Interpolation in Neural Rendering Application

Work	Interpolation Type	Data Type	Purpose	Neural Rendering Process
deep voxels [89]	3-linear	Voxel Feature	Voxel Feature Interpolation	Rendering
Instant NGP [64]	d-linear	Voxel	Hash Encoding Interpolation	Training
deep voxels [89],neural point-based graphics [2], deep illumination [99]	Bi-linear	Feature	Deconvolution Interpolation	Rendering
Traditional Post-processing	Bi-linear/Bi-cubic	Pixel	Pixel Interpolation	Rendering

**FIGURE 7.** The render process versus of NeRF, instant NGP and instant 3D.

using U-Net as the generation model, the inference process necessitates the use of a pre-trained U-Net for image generation, requiring significant convolutional and general computational power. For instance, as depicted in Fig. 8, the rendering generation process of Neural Light Transport utilizes downsampling with a stride of 2 using convolutions, followed by linear interpolation for upsampling, and incorporates 3 skip connections (original U-Net has 4 skip connections) to merge information for image synthesis. Skip connection layers can also be considered as a form of residual layer, fundamentally involving element-wise operations. Additionally, the skip connection and concatenation layers in Fig. 8 require storing input images and two intermediate feature maps, thereby increasing memory access overhead. In summary, the hardware acceleration requirements for the forward rendering process in neural rendering applications encompass matrix operations, convolutional operations, and general computational capabilities. The specific computational and memory requirements for different tasks are determined by the voxel count in the target scene and the resolution of the rendered images.

In the neural inverse rendering application, the rendering process primarily employs neural networks for input encod-

ing, scene querying, and ray marching algorithms for ray tracing. Taking the original NeRF as an example, the authors selected 64 points per ray (for coarse scenes) and 128 points per ray (for fine scenes), totaling 192 ray marching points. For each ray marching point, the MLP network, as depicted in Fig. 9, first encodes the 3D coordinate vector into a 60-dimensional vector for input and processes it through an MLP network comprising 8 FC layers (where the 5th layer requires re-inputting the 60-dimensional vector to enhance coordinate information), with each layer having 256 channels. This process yields an output of a 256-dimensional feature vector and σ . Subsequently, the 256-dimensional feature vector is concatenated with a 24-dimensional vector generated from the input encoding of the camera viewing direction and is then fed into an FC layer with 256 channels followed by an FC layer with 128 channels to compute the RGB color. To achieve high-quality rendering, the authors set the number of rays per image to be 762k, requiring 150 to 200 million ray marching and MLP network computations to render a single frame, taking 30s on an NVIDIA V100 GPU to render one frame.

FastNeRF [24] splits the sequential coordinate encoding MLP network and the ray direction MLP network in NeRF

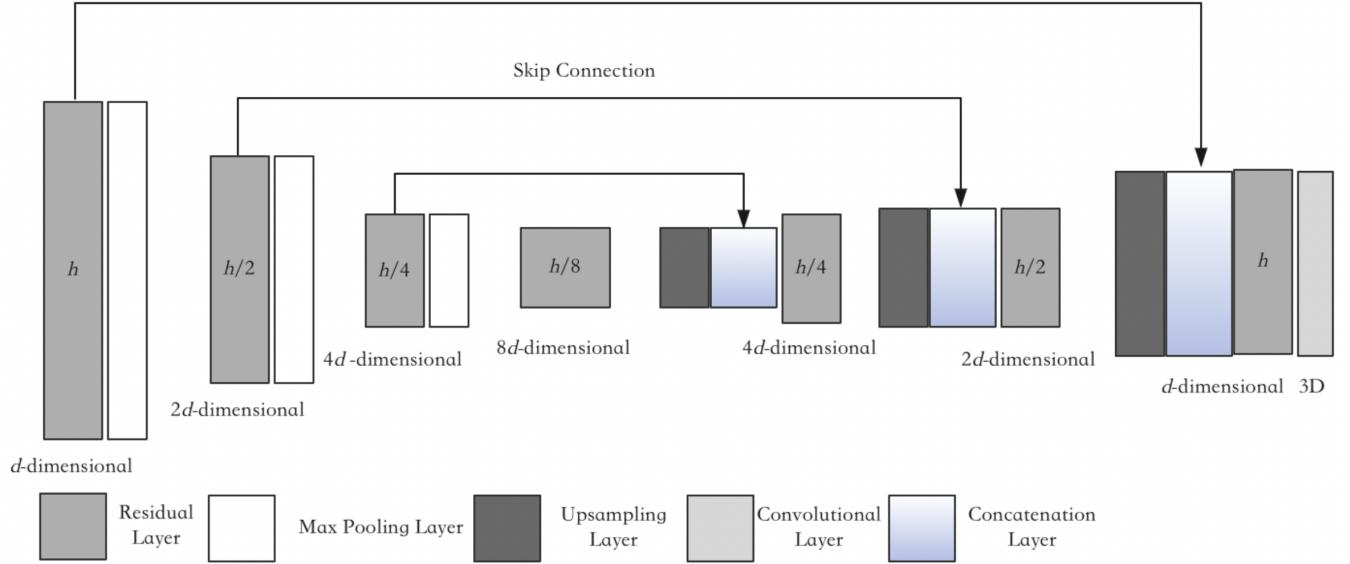


FIGURE 8. The network architecture of image synthesis layer.

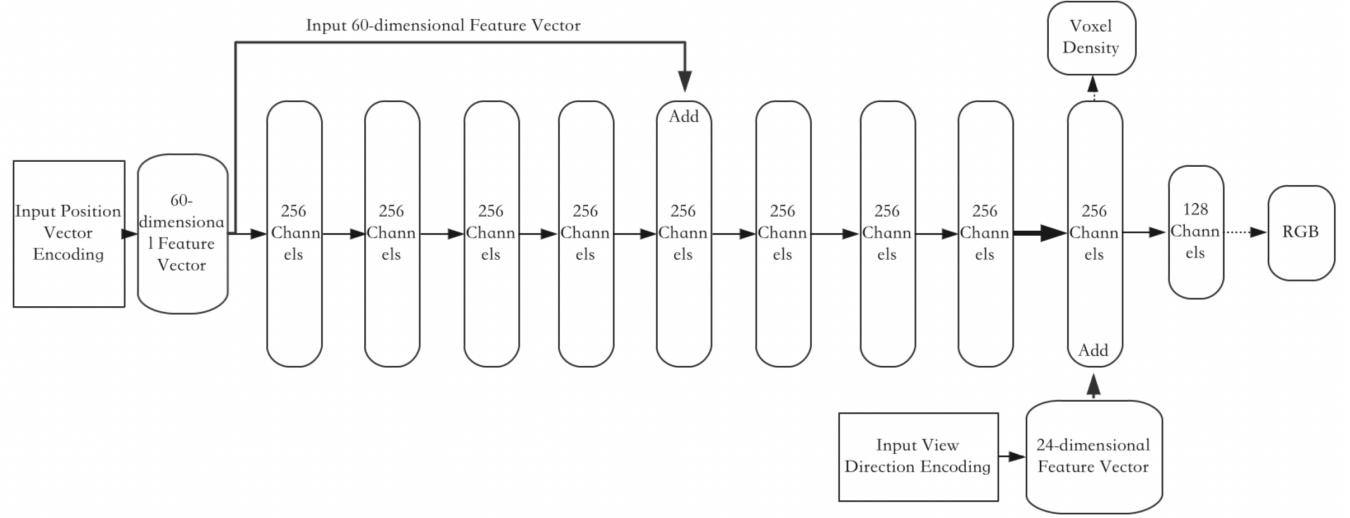


FIGURE 9. The FC network architecture of NeRF.

into two parallel MLP networks for accelerated rendering. One network generates d -dimensional neural radiance volume outputs using an 8-layer position-aware MLP network, while the other produces 1-dimensional directional feature outputs using a 4-layer direction-aware MLP network. Experimental results on an NVIDIA RTX 3090 GPU show that FastNeRF is 3000 times faster than the original NeRF in rendering speed. Notably, the cache occupancy of FastNeRF's MLP network weights is 54GB, which can be placed in the GPU's on-chip cache to enhance rendering speed. MobileNeRF [14] decomposes NeRF's MLP network into an encoder (first 8 layers) and a decoder (last 2 layers), pre-storing

the output of the encoder and only executing the inference calculation of the decoder to achieve real-time NeRF neural rendering on mobile devices. However, MobileNeRF also faces several issues, such as high requirements for the training dataset (otherwise, numerous voids may appear) and lengthy training times (requiring training in 3 steps).

The R2L model [102] replaces NeRF with a neural light field NeLF to accelerate rendering speed. Due to the model's direct output of RGB and the absence of volumetric density learning and opacity synthesis steps, predicting pixel colors only requires a single forward pass of the rays, making R2L significantly faster in rendering compared to NeRF. However,

NeLF is more challenging to train compared to NeRF. To address this, the authors employed an 88-layer deep MLP architecture with residual layers as the mapping function, resulting in longer rendering delays. Additionally, NeLT requires NeRF to generate pseudo datasets to aid in training. MobileR2L [12], based on the R2L model, proposed an architecture optimization tailored for mobile devices. In contrast to MobileNeRF, it does not require pre-storing intermediate data, making it more suitable for hardware platforms with limited computational capabilities. However, the method proposed in the paper leads to significant memory issues when handling large ray sizes. In summary, Fig. 10 illustrates the pipeline of the neural inverse rendering task, where each sampling point on every ray independently executes this pipeline, thus naturally possessing high parallelism that can be accelerated by hardware. Hardware acceleration requirements for the inverse rendering process in neural rendering applications include input encoding queries, ray marching, matrix operations, convolution operations, and general computational capabilities. Compared to forward rendering, inverse rendering requires higher computational power and storage. The ray marching, encoding queries, volume rendering, neural network computations, rendering resolution, and the number of generated rays are closely related. Obtaining high-quality images entails significant computational and memory overhead. Consequently, achieving faster input encoding queries and MLP network computations becomes a key focus of hardware acceleration.

3) Bottleneck Analysis

From Section V-B1, it can be inferred that the training processes of neural forward rendering and inverse rendering applications exhibit characteristics of high computational load and demanding memory bandwidth. In comparison to inverse rendering applications, neural forward rendering tasks demonstrate lower computational and memory requirements. For instance, considering the deep illumination model [99] which generates 3-channel 256×256 pixel images using a 16-layer neural network, a single NVIDIA P500 GPU requires 3 hours. Conversely, for neural inverse rendering tasks such as those exemplified by NeRF [62], a single NVIDIA V100 GPU may require 1 to 2 days for a single scene. Therefore, the training processes for neural rendering tasks, especially for inverse rendering applications, necessitate substantial computational and storage resources. Typically, the training of neural rendering processes is performed offline, often running on hardware platforms at the level of cloud data centers or server-grade hardware platforms, as the computational and storage capabilities of device-level hardware platforms are insufficient to support such training tasks. From the section on rendering process of requirements analysis in hardware acceleration, it can be observed that in the rendering process of neural forward rendering and inverse rendering applications, neural forward rendering applications typically achieve real-time rendering at high resolutions, while inverse rendering applications highlight performance bottlenecks. Muhammad *et*

al. [63] evaluated four algorithms (NeRF, NSDF, GIA, NVR) on an NVIDIA RTX3090 GPU (35.58 TFLOPS@FP16) using a multi-resolution hash encoding method [64]. Rendering a single frame of size 1920×1080 required total runtimes of 231 ms, 27.87 ms, 2.12 ms, and 6.32 ms for the respective algorithms, which is deemed unacceptable for real-time rendering. Notably, the subtasks of embedded grid hash encoding and MLP queries are the most time-consuming performance bottlenecks in neural rendering applications, jointly accounting for approximately 70% of the total duration. Fu *et al.* [23] assessed NeRF synthesis and deep voxels tasks, reporting that they required 28 s and 13 s, respectively, on an NVIDIA RTX 2080Ti desktop-grade GPU, whereas on an NVIDIA Jetson TX2 edge-grade GPU, they required 1000 s and 380 s, respectively. Here, the runtimes of the subtasks involving MLP queries and ray transformer calculations constituted approximately 70% to 90% of the total runtime. Therefore, embedded grid hash encoding, MLP queries, and ray transformer calculations are the primary targets for hardware platform acceleration. Li *et al.* [51] designed the Instant 3D accelerator for 3D reconstruction and neural rendering tasks akin to NeRF. The Instant 3D accelerator, illustrated in Fig. 14, primarily consists of three parts: MLP units, grid cores, and I/O interfaces. The authors mapped the embedded grid hash queries and MLP queries from the Instant 3D algorithm to the Instant 3D accelerator for hardware pipelining acceleration. Additionally, they designed a forward read mapper (FRM) to merge multiple memory read requests and a backward update merger (BUM) to combine multiple grid updates into a single update, thus enhancing the on-chip SRAM array utilization, minimizing SRAM write counts and power consumption, and supporting various grid sizes required by the Instant 3D algorithm. Fu *et al.* [23] devised the Gen-NeRF accelerator for real-time, generalizable NeRF-like tasks. The overall architecture of the Gen-NeRF accelerator, as depicted in Fig. 15, mainly comprises a software-defined rendering engine, workload scheduler, and on-chip memory. The authors employed epipolar geometry inference to expedite the target workload and designed the rendering engine architecture to optimize the software-to-hardware data flow mapping, maximizing data reuse among rays.

C. PLATFORM STATUS

Common hardware processing platforms for graphic applications include general-purpose processing units, graphics processing units, and domain-specific architectures (DSAs) represented by tensor processors and neural processors [29]. Currently, these hardware processing platforms all provide support for neural networks and are applicable to neural rendering applications. This section respectively introduces their current status in neural network acceleration and conducts an adaptability analysis for neural rendering applications.

1) General-Purpose Processor (CPU)

As the most widely used hardware processing platform, general-purpose processors (CPUs) mainly focus on three

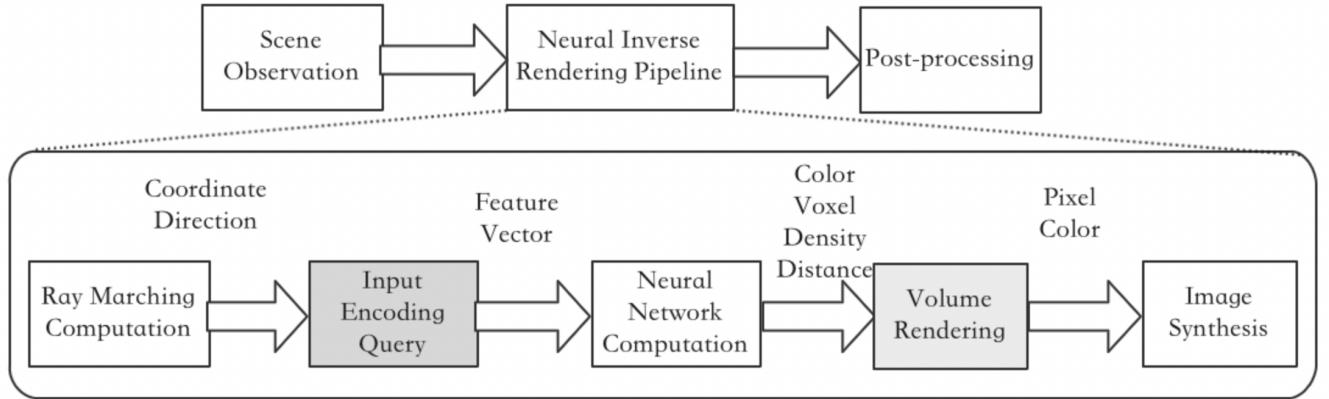


FIGURE 10. Pipeline of inverse rendering.

$$D = \begin{matrix} & \begin{matrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{matrix} & \times & \begin{matrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{matrix} & + & \begin{matrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{matrix} \end{matrix}$$

Matrix A Matrix B Matrix C

FIGURE 11. The MAC operation in tensor core.

aspects of hardware support for neural network acceleration:

- increasing neural network extension instruction sets;
- introducing neural network-specific data types;
- incorporating neural network-specific acceleration components. Table. VII presents the acceleration support for neural networks across different CPU platforms.

The IBM Power10 processor [100] introduces a new set of Matrix Multiply Assist (MMA) instructions and corresponding MMA compute units in the latest Power-ISA 3.1 version. It directly performs numerical linear algebra operations on small matrices while supporting half-precision data operation instructions. These instructions can accelerate compute-intensive kernels such as matrix multiplication, convolution, discrete Fourier transform, and artificial intelligence workloads including neural networks. When facing neural network workloads, the addition of MMA units in the POWER10 processor can improve energy efficiency at the single-core level by 2.6 times compared to POWER9.

Dojo [95] introduces the microarchitecture of Tesla's high-throughput general-purpose processor for AI training. The instruction width of the general processing cores in the DOJO

processing node is 64 b, while the instruction width of the vector/matrix coprocessor is 64 B. Coprocessor instructions are sent to the vector scheduling module for processing through the scalar scheduling module interface. The memory access path width for the general processing cores is $8B \times 2$, while for the vector/matrix coprocessor, it is $64B \times 3$. The general processing cores and coprocessors use SRAM space instead of data cache for data interaction. Therefore, in DOJO, the general processing cores are only responsible for general address calculation and logical computation, while the calculation tasks for neural network workloads are handled by the coprocessor.

Both AMD and Intel have respectively added the Vector Neural Network (VNN) instruction set and corresponding hardware acceleration units for accelerating neural network workload computation in their latest EPYC Milan processors and Xeon 4th Gen processors. They have also added Scatter/Gather memory access instruction extensions to accelerate memory access for neural network workloads. Additionally, AMD, Intel, and ARM have all added support for the BF16 data type in their latest CPUs.

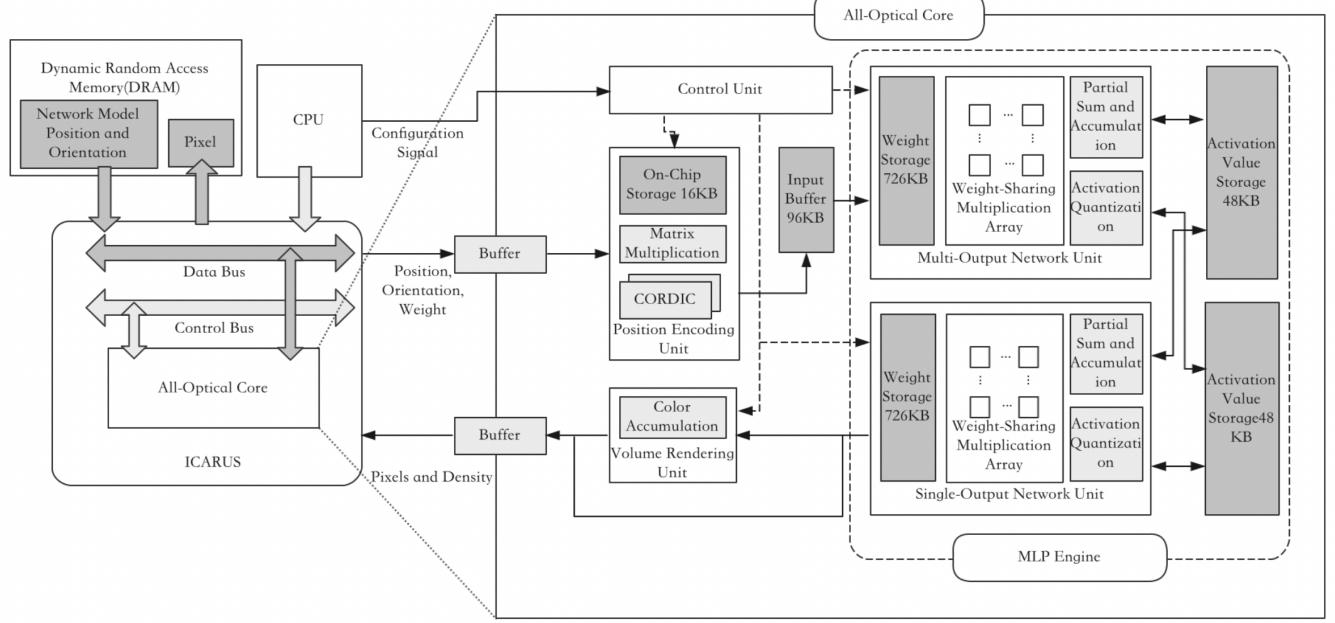


FIGURE 12. The architecture of ICARUS.

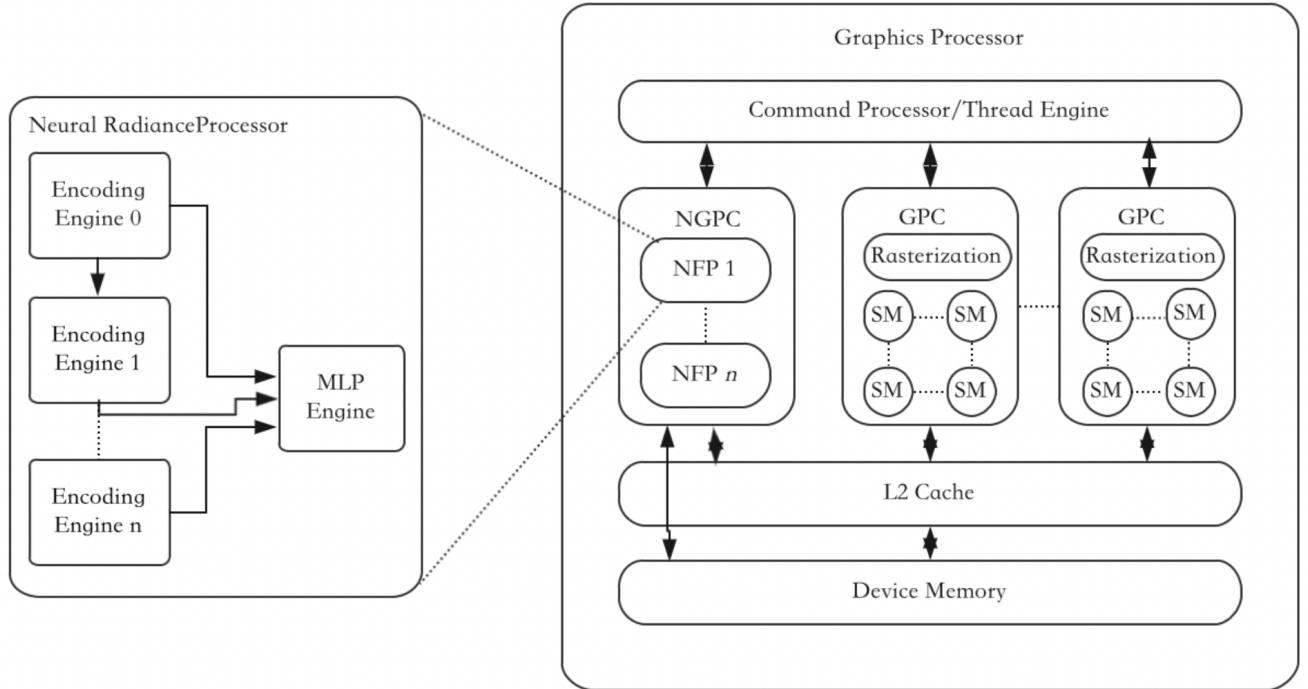


FIGURE 13. The architecture of NGPC.

In both server-level and desktop/mobile mainstream CPU platforms, architectural-level optimizations for neural network applications have been implemented. The addition of neural network compute units and corresponding instruction

sets to CPU platforms has three main benefits:

- no impact on other parts of the original architecture, minimal impact on the CPU microarchitecture;
- CPUs can reduce the power consumption impact of

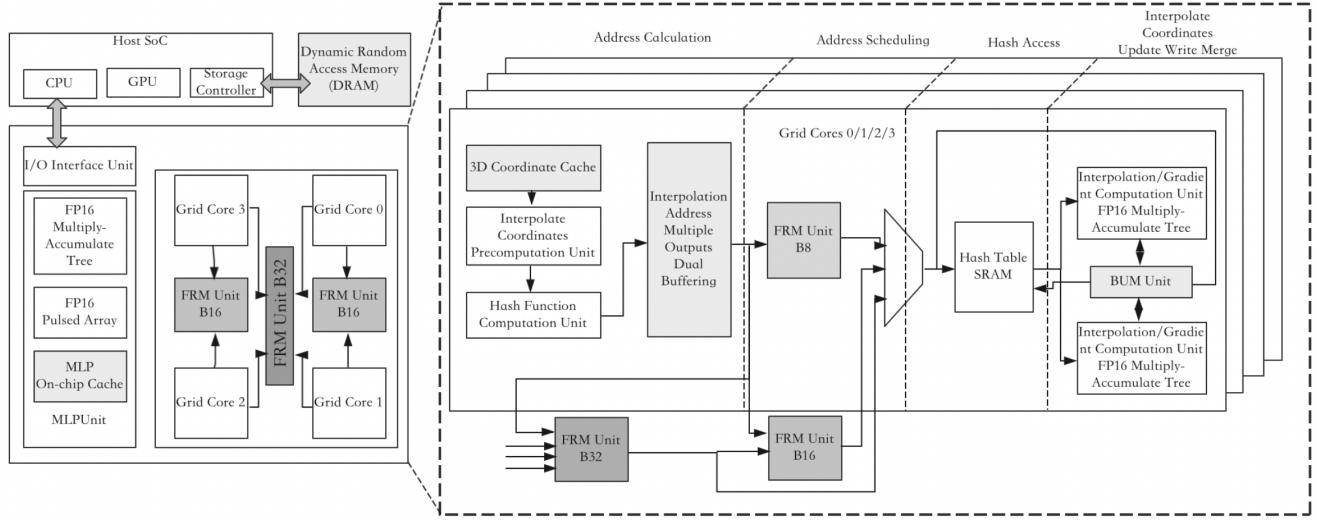


FIGURE 14. The architecture of instant 3D accelerator.

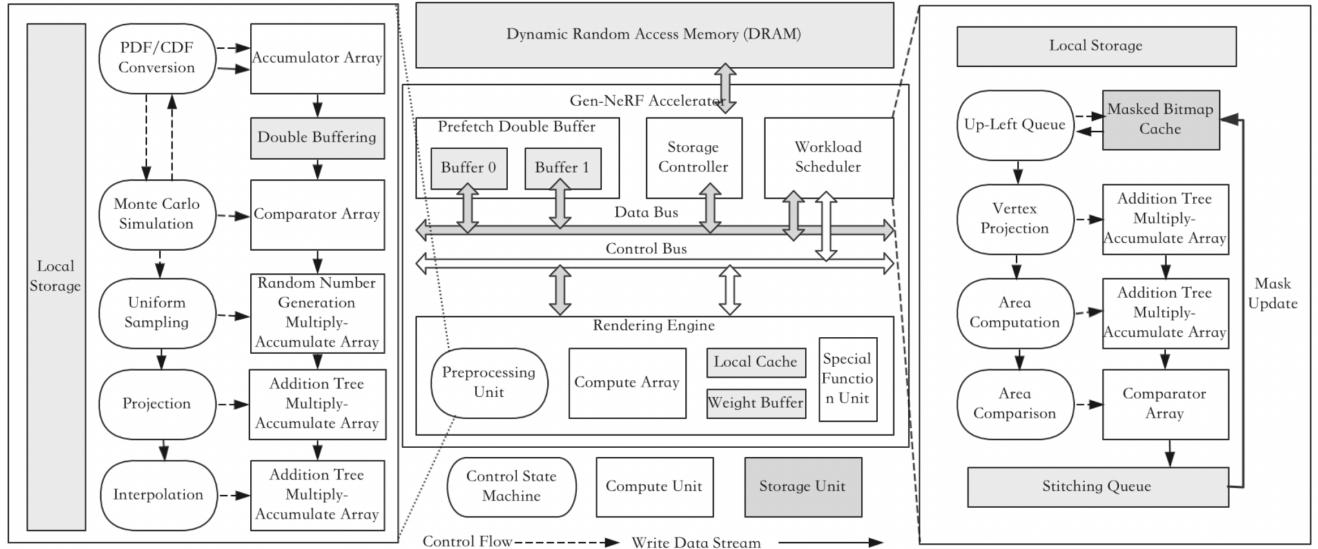


FIGURE 15. The architecture of Gen-NeRF accelerator.

TABLE 7. The Acceleration for NN in CPU

Platform	Instruction Set	Data Type	Hardware Unit	Scene Orientation
IBM Power10 [100]	Power-ISA 3.1 MMA Instruction Extension	FP16	MMA Computing Unit	Inference/Training
Tesla DOJO [95]	64B Specialized Instruction Set	BF16 CFP8	Vector/Matrix Coprocessing Unit Scatter/Gather Unit	Inference/Training
AMD EPYC	Scatter/Gather Instruction BF16 Instruction VNN (Vector Neural Network) Instruction	BF16 FP16 INT8	Mul MAC Unit	Inference/Training
NVIDIA Grace	BF16 Instruction Scatter/Gather Instruction	BF16 INT8	MatMul Unit	Inference/Training
Intel Xeon 4th Gen	TMUL Instruction VNN Instruction	BF16 INT8	AMX Unit	Inference/Training

- neural network units through gated clocks or dynamic switching;
- they can inherit the original programming model and possess a certain level of versatility.

2) Graphics Processing Unit (GPU)

The Graphics Processing Unit (GPU) or General-Purpose Graphics Processing Unit (GPGPU) possesses significantly greater computational and storage resources than the CPU, along with powerful floating-point calculation capabilities. Their architecture is exceptionally well-suited for handling large-scale, highly parallel, computation-intensive tasks. Since its inception in 1998, the GPU has undergone rapid development, evolving from a dedicated platform for graphics tasks to encompassing hardware processing platforms for high-performance general computing, graphics processing, and deep learning computations. Currently, the GPU/GPGPU stands as the most widely deployed hardware platform for neural network applications, with many deep learning frameworks based on GPU platform interfaces. Hardware support for accelerating neural networks primarily focuses on the following four aspects:

- Chip architectures suitable for neural network computation.
- Introduction of dedicated data types for neural networks.
- Inclusion of specialized acceleration components for neural networks.
- Addition of dedicated memory access units for neural networks.

Table. VIII provides an overview of the acceleration support for neural networks across different GPU/GPGPU platforms.

NVIDIA, starting from the Kepler architecture, made substantial alterations to its GPU architecture, transitioning from a graphics processing-oriented architecture to a general-purpose computing architecture. In the Volta architecture, NVIDIA introduced tensor cores for the first time to accelerate neural network tasks. Each Tensor Core can perform mixed-precision operations for $4 \times 4 \times 4$ matrix multiplication and accumulation, as depicted in Fig. 11. In its latest H100 GPU, the peak performance for half-precision floating point (FP16) has reached 1000 TFLOPS. Furthermore, NVIDIA has sequentially incorporated dedicated data types for neural network training/inference, such as TF32, BF16, FP8, and introduced Tensor Memory Access (TMA) units for asynchronous read/write acceleration of tensor data, further expediting neural network tasks. These technological transformations have garnered significant attention from numerous industrial and academic researchers.

AMD has also integrated matrix cores and FP16/BF16 data types into the architecture of its GPGPU series chips to support accelerated neural network tasks. Intel, in its upcoming Ponte Vecchio XMX chip architecture, has incorporated matrix engines and control flow processing prefetch units to support accelerated neural network tasks.

In traditional GPU architectures, NVIDIA has focused on accelerating applications such as neural graphics, metaverse,

and digital humans. In its latest GeForce RTX 4090 chip, it collaboratively utilizes ray tracing cores (RTCore), tensor cores, and general-purpose computing cores to realize 3D scene enhancement based on deep learning and the DLSS technology introduced in Section IV-B3. AMD has also incorporated matrix cores and ray tracing accelerators into its high-end GPU series chips to support relevant field applications. Additionally, the Special Function Unit (SFU) within the GPU can accelerate transcendental functions (such as reciprocal, square root, power functions, logarithms, trigonometric functions). Since the hardware acceleration of SFU essentially employs a method combining table lookup with double interpolation, it can also support hash encoding table lookup algorithms.

3) Domain-Specific Architecture (DSA)

With the continuous advancement of deep learning technologies, the demand for underlying hardware computational power in upper-layer application scenarios has experienced explosive growth. John *et al.* [29] proposed Domain-Specific Architecture (DSA) as a response through collaborative software-hardware design. DSA is a type of architecture more inclined towards hardware-centric, custom-designed structures for specific problem domains. DSA can provide significant performance (and energy efficiency) gains for the domain, being more aligned with the application requirements compared to general-purpose processors. It achieves better performance when accelerating specific applications, while also possessing a certain level of programmability. Thus, compared to Application-Specific Integrated Circuits (ASIC), DSA offers greater flexibility and functional coverage.

Neural Processing Units (NPUs) or Tensor Processing Units (TPUs) are DSA architectures tailored for neural networks. Several commercial NPU/TPU hardware platforms have emerged. Reuther *et al.* [86] have compiled and summarized commercially available neural network accelerators with peak performance and power consumption figures. Table. IX lists some NPU/TPU platforms and their support for neural network acceleration. NPU/TPU architectures are specifically designed for neural networks, primarily targeting deep neural networks and convolutional neural networks, supporting the limited operators and data types required for neural network computations. NPUs/TPUs designed for training typically exhibit higher computational capabilities and abundant memory resources, efficiently supporting tensor computation acceleration. NPUs/TPUs designed for inference, on the other hand, are usually constrained in terms of computational capability and memory bandwidth, with limited resources allocated for non-linear operations such as activation function processing, while other computational capabilities are relatively deficient.

The NePU, Neural Photonics Processing Unit, is a Domain-Specific Architecture (DSA) designed for tasks related to Neural Radiance Fields (NeRF) and is currently in the early stages of research. ICARUS [78] represents a specific

TABLE 8. The Acceleration for NN and Ray in GPU/GPGPU

Platform	Peak Compute /FLOPS	Data Type	Hardware Accelerator Unit	Memory Access Accelerator Unit	Ray Tracing Accelerator Unit	Scene Orientation
NVIDIA H100 [17]	500T@TF32 1000T@FP16 1000T@BF16 2000T@FP8	TF32 FP16 BF16 FP8	Tensor Core	Tensor Memory Accelerator Unit TMA	None	Training/Inference/ High Performance
AMD MI250 [92]	383T@FP16 383T@BF16 383@INT8	FP16 BF16	Matrix Core	None	None	Training/Inference/ High Performance
Intel Ponte Vecchio XMX [39]	419T@TF32 839T@FP16 839T@BF16 1678@INT8	TF32 FP16 BF16	Matrix Engine	Control Flow Processing Prefetch Unit CSP	None	Training/Inference/ High Performance
Biren BR100 [30]	512T@TF32 1024T@BF16 2048T@INT8	TF32 BF16	Vector Unit	Tensor Data Storage Accelerator Unit TDA	None	Training/ Inference
NVIDIA GeForce RTX 4090 [70] AMD RADEON™ RX 6950 XT [3]	660.6T@FP8 191T@ray 47.31T@FP16	FP8 FP16	Tensor Core Matrix Core	None	Ray Tracing Unit Ray Tracing Accelerator	Inference/Graphic Inference/Graphic

TABLE 9. The Acceleration for NN in NPU/TPU

Platform	Supported Network Types	Data Types	Acceleration Structure	Scene-oriented
TPUv4 [43]	MLP CNN RNN Transformer	BF16 INT8	Pulsed Array	Training/Inference
TPUv4i [44]	MLP CNN RNN Transformer	BF16 INT8	Pulsed Array	Inference
Cambricon MLU370 [11]	MLP CNN RNN Transformer	FP32 FP16 BF16 INT16/8/4	Dot Product Tree	Training/Inference
Baidu Kunlun Chip 2nd Generation [74]	CNN Transformer GEMM	FP16 INT8	Multiply-Accumulate Array	Training/Inference
ARM Ethos-U55 [91]	CNN RNN	INT8	Multiply-Accumulate Unit	Inference

architecture tailored for NeRF rendering. Fig. 12 illustrates the overall architecture of ICARUS, which, together with the CPU and memory, forms the rendering system. The NeRF rendering process executed by ICARUS involves the following steps:

- storing input such as network models, encoding frequencies, and positional directions in memory, which is controlled by the CPU during runtime to provide input to ICARUS;
- distributing input data through on-chip buses to corresponding target plenoptic cores;
- loading network models, position, and directional data into ICARUS for NeRF processing (from the Position

Encoding Unit (PEU) to the Multi-Layer Perceptron (MLP) engine and then to the Volumetric Rendering Unit (VRU)), where a group of rays (including all stepping points of the ray) is processed by the same plenoptic core;

- streaming out the final rendered pixel colors via the data bus.

The author enhances hardware acceleration efficiency by transforming the operations of the Fully Connected (FC) layer of the MLP computation into approximable Reconfigurable Multiple Constant Multiplications (RMCs), thereby reducing approximately one-third of hardware complexity (including computational load and parameter capacity) compared to

traditional multiply-accumulate computations. This enables the entire NeRF pipeline to be completed internally within the plenoptic core, with all weight parameters stored in the plenoptic core's SRAM without the need for data exchange with DRAM. ICARUS, fabricated on a 40nm process node, occupies a chip area of 16.5 mm^2 , achieves an operating frequency of 400 MHz, and requires 45.75 seconds to render an image of 800×800 resolution (with 192 sampling points for ray stepping), consuming 282.8mW. While the ICARUS architecture represents a certain level of innovation, there still exists a significant gap to achieve real-time rendering tasks for NeRF.

The Neural Graphics Processing Cluster (NGPC) [63] integrates a Neural Fields Processor (NFP) into the traditional GPU architecture to accelerate neural graphics applications. The overall architecture of NGPC is depicted in Fig. 13. The NGPC containing NFP is embedded as a hardware unit parallel to the Graphics Processing Cluster (GPC) within the GPU architecture to achieve on-chip heterogeneous computing. The NFP comprises an encoding engine (enc engine) and an MLP engine, which respectively accelerate input encoding and MLP computations. However, GPC units are required to handle hash encoding query computations. Through experimentation, the author determined that NGPC configurations containing 8, 16, and 32 NFP units deliver performance improvements of 12.94x, 20.85x, and 33.73x, respectively, compared to the RTX 3090 platform when executing NeRF, NSDF, GIA, and NVR applications. The author indicates that NGPC can achieve 30 FPS rendering at 4K ultra-high-definition resolution for NeRF applications (without explicitly specifying the quantity of NFP units). Estimations for the 7nm process node indicate that an NGPC-8 with a single NFP unit incurs approximately a 4.52% increase in GPU area and a 2.75% increase in power consumption. NGPC presents a method for achieving heterogeneous acceleration of neural rendering based on GPU architecture, leveraging the GPU's existing general computing resources such as ray stepping and transcendental functions. However, this approach may lead to further hardware overhead and potentially exacerbate the occurrence of dark silicon phenomena in dedicated neural rendering scenarios.

D. DESIGN CHALLENGES

In summary, there are still certain design challenges for current hardware platforms when handling neural rendering applications:

- Mainstream CPU platforms have been specifically designed and optimized for operations such as convolutions and matrix-matrix multiplications required by neural networks. However, for neural rendering tasks, CPU platforms are relatively constrained by neural network computational power and general-purpose computing capabilities, making them more suitable for neural forward rendering tasks or the inferencing portion of inverse rendering. Furthermore, due to the lack of rapid table lookup capabilities (activation functions,

hash encoding), completing end-to-end high-definition real-time neural inverse rendering tasks poses significant challenges for CPU platforms.

- Currently, leading manufacturers such as AMD and NVIDIA have bifurcated their GPU chips into two distinct branches: GPGPU and traditional GPU platforms, catering separately to neural network/high-performance applications and traditional graphics applications. GPGPU platforms typically possess extremely high neural network computational power, high memory bandwidth, and on-chip storage capacity. However, they have omitted units such as rasterization, rendering output, and ray tracing found in traditional GPUs. GPGPU platforms can effectively accelerate neural network and volume rendering computations, well suiting the training requirements of neural rendering tasks. Nonetheless, they slightly lack specialized computational support for ray marching, hash code table lookup, and similar tasks. On the other hand, traditional GPU platforms, built upon traditional rasterization and ray tracing units, have augmented neural network computational power and higher memory performance. They can adapt well to the demands of neural forward rendering tasks but are unable to meet the acceleration requirements of neural inverse rendering tasks. Additionally, the development paths of these two product branches have increased manufacturers' research and development costs and system complexity.
- For DSA platforms, NPUs/TPUs are specific architectures designed for neural networks, primarily targeting deep neural networks and convolutional neural networks. Most hardware supports only the operations and data types required for neural network computations, with on-chip computational resources primarily dedicated to accelerating tensor computations, with a small portion allocated for non-linear operations such as activation function processing. Their general computing capabilities are relatively limited. Therefore, for neural rendering applications, due to the lack of computational power for ray marching, hash encoding, volume rendering computations, and the limitations in supported operations and data types, NPUs/TPUs hardware platforms may only be able to support a small number of neural rendering tasks. Research on various NePU architectures is mainly focused on NeRF-like neural rendering tasks and has achieved a certain degree of architectural breakthrough. However, they are commonly implemented as supplementary acceleration platforms for CPUs and GPUs, with certain limitations in specific scenarios. For instance, ICARUS [78] and the Instant 3D accelerator [51] require the use of a host SoC, while NGPC [63] adds complexity to GPU internal task scheduling and data flow mapping.

VI. RESEARCH AND DEVELOPMENT TRENDS

This section provides a prospective analysis of the research and development directions for neural rendering systems from the perspectives of neural rendering applications and hardware acceleration architectures.

A. NEURAL RENDERING APPLICATIONS

After several years of development, neural rendering has demonstrated remarkable capabilities in scene representation, real-time global illumination, novel view synthesis, relighting, and spatial illumination. However, challenges such as generalization, scalability, and multimodal synthesis persist. Neural rendering applications have played a pivotal role in the advancement of Augmented Reality (AR) and Virtual Reality (VR). It is believed that neural rendering can entirely replace various submodules in traditional graphics rendering pipelines, such as surface subdivision or rasterization, enabling end-to-end neural forward rendering pipelines and the development of graphics applications based on neural rendering. In inverse rendering applications, current neural rendering works primarily focus on simple objects and relatively uncomplicated composite scenes. Challenges remain in extending methods developed based on single objects to large-scale scenes [60], complex environmental scenarios [96], or dynamic scenes [49]. Additionally, new view synthesis tasks still rely on large-scale multi-view datasets or are limited to training for specific objects, highlighting the importance of enhancing the generality of neural voxel representations across scenes or with a limited number of views. Multimodal learning applications of neural rendering imply simultaneous processing of semantic, textual, auditory, and visual signals. Rendering dynamic interactions and voice-face matching for digital humans pose significant challenges. Meeting consumer demands for neural rendering involves delineating functionalities and interaction modes for cloud-based and mobile-based applications, achieving lightweight and cost-effective deployment.

B. NEURAL RENDERING PROCESSORS

In light of the high parallelism and specificity of neural rendering tasks, as well as the differences between neural rendering tasks and traditional graphics tasks, and the current hardware platform's limitations in handling neural rendering tasks, the research and development of Neural Rendering Processors (NRPU) to accelerate neural rendering applications represent a crucial developmental trend.

1) Collaborative Design

The collaborative design of software and hardware stands as a core approach in architectural design. It involves simplifying hardware implementation by partitioning system tasks into software and hardware functionalities, designing hardware acceleration units to address software algorithm bottlenecks, and seeking a balance between the two—a pursuit of paramount importance for designers. Neural network systems, exemplified by TPU, have provided us with much inspi-

ration and experiential lessons in collaborative software and hardware design. For the future collaborative design of neural rendering systems, attention must be paid to the following issues:

- 1) The design objectives encompass neural rendering applications, a full-stack toolchain based on neural rendering pipelines for graphic editing, and neural rendering processors. This aims to innovate existing graphics processors and graphic editing tools, effectively supporting future key development areas such as virtual reality, augmented reality, film and television production, digital entertainment, artificial intelligence, and the metaverse.
- 2) Optimization methods should be determined by deployment scenarios. For instance, both Instant 3D [51] and GEN-NeRF [23] represent collaborative designs for NeRF-type applications based on AR/VR devices. As AR/VR devices operate within computational, memory, and power constraints, their collaborative design objectives focus on reducing the computational or storage overhead of algorithms and hardware while striving to maintain algorithm accuracy within an acceptable range or, if possible, lowering accuracy under such constraints.
- 3) Reasonable division of sub-functionality between software and hardware is essential. For instance, in NeRF-type neural rendering tasks, hardware acceleration is employed for sub-functions such as hash table computation, MLP computation, and volume rendering computation, while in RenderNet-type neural rendering tasks, hardware acceleration is utilized for sub-functions such as neural network computation or projection computation. On the other hand, software handles sub-functions like coordinate preprocessing, positional encoding, and ray-sampling point determination.
- 4) Hardware architecture design should fully consider algorithm maturity and hardware design cycles, retaining a certain degree of design margin and flexibility. Jouppi *et al.* [44] proposed ten crucial experiential lessons in TPU design, highlighting the importance of providing dynamic margins for DSAs to remain effective throughout their entire lifecycle, as well as the significance of flexibility in DSA design, as evidenced by the rapid replacement of a large number of LSTMs by Transformer-type applications.

2) Scene Segmentation

In anticipation of potential deployment scenarios for future neural rendering systems, the design of neural rendering processors is discussed separately for cloud and device scenarios:

- Integrated Training and Rendering NRPU is tailored for cloud scenarios, deployed as neural rendering processors (accelerator cards) in data centers. In this setup, neural rendering applications accomplish neural rendering network training and rendering in the cloud or handle the re-computational workload during rendering. This processor can independently execute rendering tasks or transmit streams or specific encoded files over the network to be completed for final rendering on end-device platforms.

- Rendering-Specific NRPUs are aimed at device scenarios, deployed within mobile devices or AR/VR devices as neural rendering processors (system-on-chip). In this context, neural rendering applications carry out image rendering at the device end. The rendering-specific NRPUs may receive streams from the cloud or specific encoded files, and, through dedicated hardware acceleration, synthesize images and render output.

3) Architecture Design

The integrated NRPUs for rendering and training needs to possess powerful neural network computing capabilities and memory access performance similar to GPGPU/high-end NPU, effectively supporting the training of neural rendering networks. Simultaneously, it also requires robust specialized neural rendering computational capabilities for operations such as transcendental function processing, hash encoding table processing, and ray marching processing. Designing a programming model and architecture to meet the training and rendering data flow of neural rendering applications, as well as increasing the parallelism of the hardware acceleration units for neural rendering while designing performance-matched memory access and data transmission paths and on-chip storage space, will be a significant challenge within the constraints of transistor capacity and process technology.

Rendering-specific NRPUs need to determine the operator units, peak computing performance, storage capacity, and bandwidth required by the hardware based on specific neural rendering algorithms. Additionally, it necessitates data flow and performance analysis of competition for memory access bandwidth and network transmission bandwidth among NRPUs and other on-chip system hardware units (such as CPU, ISP, encoding/decoding units, etc.). Efficiently mapping the neural rendering pipeline to the on-chip system to enhance on-chip data reuse and reduce memory access bandwidth requirements, as well as reducing hardware operating power consumption and design costs while ensuring high-quality real-time neural rendering, presents a significant challenge in device-end platforms constrained by power consumption, area, and price.

VII. CONCLUSION

With the rapid rise and development of deep neural networks, neural rendering, as a fusion method of deep learning and computer graphics, has emerged. Neural rendering is an image and video generation method based on deep learning, combining deep neural network models with the physical knowledge of computer graphics to obtain controllable and realistic scene models, enabling control of scene attributes such as lighting, camera parameters, and posture. In recent years, neural rendering has made rapid progress, with applications ranging from enhancing forward rendering effects using deep neural networks to new view synthesis, shape and material editing, relighting, and avatar generation in inverse rendering. This paper introduces representative research achievements of neural rendering in forward rendering, in-

verse rendering, and post-processing applications, analyzes in detail the common hardware acceleration requirements of neural rendering applications, and finally discusses the design challenges of neural rendering processor architecture according to scenarios.

REFERENCES

- [1] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020.
- [2] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020.
- [3] AMD. Amd compare graphics specifications. 2023.
- [4] G. An, Y. Huo, and S.-E. Yoon. Hypergraph propagation and community selection for objects retrieval. *Advances in Neural Information Processing Systems*, 34, 2021.
- [5] G. An, W. J. Kim, S. Yang, R. Li, Y. Huo, and S.-E. Yoon. Towards content-based pixel retrieval in revisited oxford and paris. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20507–20518, 2023.
- [6] G. An, J. Seon, I. An, Y. Huo, and S.-E. Yoon. Topological ransac for instance verification and retrieval without fine-tuning. *arXiv preprint arXiv:2310.06486*, 2023.
- [7] S. Bako, T. Vogels, B. McWilliams, M. Meyer, J. Novák, A. Harvill, P. Sen, T. Derose, and F. Rousselle. Kernel-predicting convolutional networks for denoising monte carlo renderings. *ACM Trans. Graph.*, 36(4):97–1, 2017.
- [8] C. Bao, Y. Zhang, B. Yang, T. Fan, Z. Yang, H. Bao, G. Zhang, and Z. Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20919–20929, 2023.
- [9] S. Bi, K. Sunkavalli, F. Perazzi, E. Shechtman, V. G. Kim, and R. Ramamoorthi. Deep cg2real: Synthetic-to-real translation via image disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2730–2739, 2019.
- [10] S. Bi, Z. Xu, P. Srinivasan, B. Mildenhall, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020.
- [11] Cambicon. Cambicon mlu370 chip. 2023.
- [12] J. Cao, H. Wang, P. Chemerys, V. Shakhrai, J. Hu, Y. Fu, D. Makoviichuk, S. Tulyakov, and J. Ren. Real-time neural light field on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8328–8337, 2023.
- [13] A. Chen, X. Wang, K. Shi, S. Zhu, B. Fang, Y. Chen, J. Chen, Y. Huo, and Q. Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2752–2758. IEEE, 2023.
- [14] Z. Chen, T. Funkhouser, P. Hedman, and A. Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16569–16578, 2023.
- [15] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.
- [16] I.-Y. Cho, Y. Huo, and S.-E. Yoon. Weakly-supervised contrastive learning in path manifold for monte carlo image reconstruction. *ACM Transactions on Graphics (TOG)*, 40(4):38–1, 2021.
- [17] J. Choquette. Nvidia hopper gpu: Scaling performance. In *2022 IEEE Hot Chips 34 Symposium (HCS)*, pages 1–46. IEEE Computer Society, 2022.
- [18] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.
- [19] C. Dachsbacher, J. Krivánek, M. Hašan, A. Arbree, B. Walter, and J. Novák. Scalable realistic rendering with many-light methods. In *Computer Graphics Forum*, volume 33, pages 88–104. Wiley Online Library, 2014.
- [20] P. Dai, Y. Zhang, Z. Li, S. Liu, and B. Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference*

- on Computer Vision and Pattern Recognition*, pages 7830–7839, 2020.
- [21] S. A. Eslami, D. Jimenez Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.
- [22] H. Fan, R. Wang, Y. Huo, and H. Bao. Real-time monte carlo denoising with weight sharing kernel prediction network. In *Computer Graphics Forum*, volume 40, pages 15–27. Wiley Online Library, 2021.
- [23] Y. Fu, Z. Ye, J. Yuan, S. Zhang, S. Li, H. You, and Y. Lin. Gen-nerf: Efficient and generalizable neural radiance fields via algorithm-hardware co-design. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–12, 2023.
- [24] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021.
- [25] G. Greger, P. Shirley, P. M. Hubbard, and D. P. Greenberg. The irradiance volume. *IEEE Computer Graphics and Applications*, 18(2):32–43, 1998.
- [26] S. Hadadian, S. Chen, and M. Zwicker. Neural radiosity. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021.
- [27] Z. He, R. Wang, W. Hua, and Y. Huo. An interactive image-based modeling system. *arXiv preprint arXiv:2203.14441*, 2022.
- [28] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021.
- [29] J. L. Hennessy and D. A. Patterson. A new golden age for computer architecture. *Communications of the ACM*, 62(2):48–60, 2019.
- [30] M. Hong and L. Xu. Biren br100 gppu: Accelerating datacenter scale ai computing. In *2022 IEEE Hot Chips 34 Symposium (HCS)*, pages 1–22. IEEE Computer Society, 2022.
- [31] Y. Huo. Extension-adaptive sampling with implicit radiance field. *arXiv preprint arXiv:2202.00855*, 2022.
- [32] Y. Huo, S. Jin, T. Liu, W. Hua, R. Wang, and H. Bao. Spherical gaussian-based lightcuts for glossy interreflections. In *Computer Graphics Forum*, volume 39, pages 192–203. Wiley Online Library, 2020.
- [33] Y. Huo, S. Li, Y. Yuan, X. Chen, R. Wang, W. Zheng, H. Lin, and H. Bao. Shadertransformer: Predicting shader quality via one-shot embedding for fast simplification. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.
- [34] Y. Huo, R. Wang, T. Hu, W. Hua, and H. Bao. Adaptive matrix column sampling and completion for rendering participating media. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016.
- [35] Y. Huo, R. Wang, S. Jin, X. Liu, and H. Bao. A matrix sampling-and-recovery approach for many-lights rendering. *ACM Transactions on Graphics (TOG)*, 34(6):1–12, 2015.
- [36] Y. Huo, R. Wang, X. Liu, and H. Bao. Sparse sampling and completion for light transport in vpl-based rendering. *arXiv preprint arXiv:2202.12567*, 2022.
- [37] Y. Huo, R. Wang, R. Zheng, H. Xu, H. Bao, and S.-E. Yoon. Adaptive incident radiance field sampling and reconstruction using deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 39(1):1–17, 2020.
- [38] Y. Huo and S.-e. Yoon. A survey on deep learning-based monte carlo denoising. *Computational Visual Media*, 7(2):169–185, 2021.
- [39] H. Jiang. Intel’s ponte vecchio gpu: Architecture, systems & software. In *2022 IEEE Hot Chips 34 Symposium (HCS)*, pages 1–29. IEEE Computer Society, 2022.
- [40] S. Jin, R. Wang, W. Zheng, W. Hua, and Y. Huo. A virtual point light generation method in close-range area. *arXiv preprint arXiv:2203.11484*, 2022.
- [41] Y. Jiong, Z. Zedong, and G. Yuhan. A survey on visual language pre-training. *Journal of Software*, 34(5):0–0, 2022.
- [42] M. M. Johari, Y. Lepoittevin, and F. Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022.
- [43] N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, et al. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–14, 2023.
- [44] N. P. Jouppi, D. H. Yoon, M. Ashcraft, M. Gottsch, T. B. Jablin, G. Kurian, J. Laudon, S. Li, P. Ma, X. Ma, et al. Ten lessons from three generations shaped google’s tpuv4i: Industrial product. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 1–14. IEEE, 2021.
- [45] J. T. Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986.
- [46] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.
- [47] S. Kim, Y. Huo, and S.-E. Yoon. Single image reflection removal with physically-based training images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5164–5173, 2020.
- [48] G. Kopanas, J. Philip, T. Leimkühler, and G. Drettakis. Point-based neural rendering with per-view optimization. In *Computer Graphics Forum*, volume 40, pages 29–43. Wiley Online Library, 2021.
- [49] D. Lee, M. Lee, C. Shin, and S. Lee. Dp-nerf: Deblurred neural radiance field with physical scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12396, 2023.
- [50] H. Li, X. Lin, Y. Zhou, X. Li, J. Chen, and Q. Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint*, 2022.
- [51] S. Li, C. Li, W. Zhu, B. Yu, Y. Zhao, C. Wan, H. You, H. Shi, and Y. Lin. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–13, 2023.
- [52] S. Li, R. Wang, Y. Huo, W. Zheng, W. Hua, and H. Bao. Automatic band-limited approximation of shaders using mean-variance statistics in clamped domain. In *Computer Graphics Forum*, volume 39, pages 181–192. Wiley Online Library, 2020.
- [53] S. Li, C. Zheng, R. Wang, Y. Huo, W. Zheng, H. Lin, and H. Bao. Multi-resolution terrain rendering using summed-area tables. *Computers & Graphics*, 95:130–140, 2021.
- [54] Z. Li, S. Niklaus, N. Snavely, and O. Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.
- [55] Y. Liang, Q. Song, R. Wang, Y. Huo, and H. Bao. Automatic mesh and shader level of detail. *IEEE Transactions on Visualization and Computer Graphics*, 2022.
- [56] D. B. Lindell, J. N. Martel, and G. Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14556–14565, 2021.
- [57] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, and C. Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.
- [58] Y. Liu, Y. Huo, L. Zhu, M. Jin, H. Zhang, S. Li, and W. Hua. Conical emission induced by the filamentation of femtosecond vortex beams in water. *Applied Sciences*, 13(22):12435, 2023.
- [59] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.
- [60] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.
- [61] E. H. Meijering, K. J. Zuiderveld, and M. A. Viergever. Image reconstruction by convolution with symmetrical piecewise nth-order polynomial kernels. *IEEE transactions on image processing*, 8(2):192–201, 1999.
- [62] 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. *Communications of the ACM*, 65(1):99–106, 2021.
- [63] M. H. Mubarik, R. Kanungo, T. Zirr, and R. Kumar. Hardware acceleration of neural graphics. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–12, 2023.
- [64] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.
- [65] T. Müller, F. Rousselle, A. Keller, and J. Novák. Neural control variates. *ACM Transactions on Graphics (TOG)*, 39(6):1–19, 2020.
- [66] O. Nabach, E. Arabadzhyska, D. Mehta, H.-P. Seidel, and T. Ritschel. Deep shading: convolutional neural networks for screen space shading. In *Computer graphics forum*, volume 36, pages 65–78. Wiley Online Library, 2017.
- [67] T. H. Nguyen-Phuoc, C. Li, S. Balaban, and Y. Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. *Advances in neural information processing systems*, 31, 2018.
- [68] M. Niemeyer and A. Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–

- 11464, 2021.
- [69] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020.
- [70] NVIDIA. Nvidia ada gpu architecture. 2023.
- [71] NVIDIA. Nvidia dlss 3. 2023.
- [72] NVIDIA. Rtx technology. 2023.
- [73] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019.
- [74] J. Ouyang, X. Du, Y. Ma, and J. Liu. 3.3 kunlun: A 14nm high-performance ai processor for diversified workloads. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pages 50–51. IEEE, 2021.
- [75] H. Park, Y. Huo, and S.-E. Yoon. Meshchain: Secure 3d model and intellectual property management powered by blockchain technology. In *Computer Graphics International Conference*, pages 519–534. Springer, 2021.
- [76] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deep sdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- [77] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.
- [78] C. Rao, H. Yu, H. Wan, J. Zhou, Y. Zheng, M. Wu, Y. Ma, A. Chen, B. Yuan, P. Zhou, et al. Icarus: A specialized architecture for neural radiance fields rendering. *ACM Transactions on Graphics (TOG)*, 41(6):1–14, 2022.
- [79] C. Reiser, S. Peng, Y. Liao, and A. Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.
- [80] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.
- [81] K. Rematas and V. Ferrari. Neural voxel renderer: Learning an accurate and controllable rendering tool. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5417–5427, 2020.
- [82] H. Ren, H. Fan, R. Wang, Y. Huo, R. Tang, L. Wang, and H. Bao. Data-driven digital lighting design for residential indoor spaces. *ACM Transactions on Graphics*, 42(3):1–18, 2023.
- [83] H. Ren, H. Zhang, J. Zheng, J. Zheng, R. Tang, Y. Huo, H. Bao, and R. Wang. Minervas: Massive interior environments virtual synthesis. In *Computer Graphics Forum*, volume 41, pages 63–74. Wiley Online Library, 2022.
- [84] H. Ren, H. Zhang, J. Zheng, J. Zheng, R. Tang, Y. Huo, H. Bao, and R. Wang. Supplementary material for minervas: Massive interior environments virtual synthesis. 2022.
- [85] L. Ren and Y. Song. Aogan: A generative adversarial network for screen space ambient occlusion. *Computational Visual Media*, 8(3):483–494, 2022.
- [86] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner. Ai and ml accelerator survey and trends. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–10. IEEE, 2022.
- [87] P. Sanzenbacher, L. Mescheder, and A. Geiger. Learning neural light transport. *arXiv preprint arXiv:2006.03427*, 2020.
- [88] V. Sitzmann, S. Reznikov, B. Freeman, J. Tenenbaum, and F. Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.
- [89] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.
- [90] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [91] A. Skillman and T. Edso. A technical overview of cortex-m55 and ethos-u55: Arm's most capable processors for endpoint ai. In *2020 IEEE Hot Chips 32 Symposium (HCS)*, pages 1–20. IEEE Computer Society, 2020.
- [92] A. Smith and N. James. Amd instinctTM mi200 series accelerator and node architectures. In *2022 IEEE Hot Chips 34 Symposium (HCS)*, pages 1–23. IEEE Computer Society, 2022.
- [93] T. Sun, K.-E. Lin, S. Bi, Z. Xu, and R. Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. *arXiv preprint arXiv:2107.12351*, 2021.
- [94] C. Suppan, A. Chalmers, J. Zhao, A. Doronin, T. Rhee, S. Lee, S. Zollmann, M. Okabe, and B. Wünsche. Neural screen space rendering of direct illumination. *Pacific Graphics Short Papers, Posters, and Work-in-Progress Papers*, 2021.
- [95] E. Talpes, D. Williams, and D. D. Sarma. Dojo: The microarchitecture of tesla's exa-scale computer. In *2022 IEEE Hot Chips 34 Symposium (HCS)*, pages 1–28. IEEE Computer Society, 2022.
- [96] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.
- [97] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020.
- [98] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022.
- [99] M. M. Thomas and A. G. Forbes. Deep illumination: Approximating dynamic global illumination with generative adversarial network. *arXiv preprint arXiv:1710.09834*, 2017.
- [100] B. W. Thompto, D. Q. Nguyen, J. E. Moreira, R. Bertran, H. Jacobson, R. J. Eickemeyer, R. M. Rao, M. Goulet, M. Byers, C. J. Gonzalez, et al. Energy efficiency boost in the ai-infused power10 processor. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 29–42. IEEE, 2021.
- [101] W. Trina. Truly global illumination: ray tracing for the masses. 2023.
- [102] H. Wang, J. Ren, Z. Huang, K. Olszewski, M. Chai, Y. Fu, and S. Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision*, pages 612–629. Springer, 2022.
- [103] Q. Wang, F. Luan, Y. Dai, Y. Huo, H. Bao, and R. Wang. A biophysically-based skin model for heterogeneous volume rendering.
- [104] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.
- [105] Q. Wang, Z. Zhong, Y. Huo, H. Bao, and R. Wang. State of the art on deep learning-enhanced rendering methods. *Machine Intelligence Research*, 20(6):799–821, 2023.
- [106] R. Wang, W. Hua, Y. Huo, and H. Bao. Real-time rendering and editing of scattering effects for translucent objects. *arXiv preprint arXiv:2203.12339*, 2022.
- [107] R. Wang, W. Hua, G. Xu, Y. Huo, and H. Bao. Variational hierarchical directed bounding box construction for solid mesh models. *arXiv preprint arXiv:2203.10521*, 2022.
- [108] R. Wang, Y. Huo, Y. Yuan, K. Zhou, W. Hua, and H. Bao. Implementation details of gpu-based out-of-core many-lights rendering.
- [109] R. Wang, Y. Huo, Y. Yuan, K. Zhou, W. Hua, and H. Bao. Gpu-based out-of-core many-lights rendering. *ACM Transactions on Graphics (TOG)*, 32(6):1–10, 2013.
- [110] R. Wang, X. Yang, Y. Yuan, W. Chen, K. Bala, and H. Bao. Automatic shader simplification using surface signal approximation. *ACM Transactions on Graphics (TOG)*, 33(6):1–11, 2014.
- [111] R. Wang, B. Yu, J. Marco, T. Hu, D. Gutierrez, and H. Bao. Real-time rendering on a power budget. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.
- [112] X. Wang, J. Zhu, Q. Ye, Y. Huo, Y. Ran, Z. Zhong, and J. Chen. Seal-3d: Interactive pixel-level editing for neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17683–17693, 2023.
- [113] G. Z. Wang Endong, Yan Ruidong. A survey of distributed training system and its optimization algorithms. 2023.
- [114] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.
- [115] Z. Wu, C. Zuo, Y. Huo, Y. Yuan, Y. Peng, G. Pu, R. Wang, and H. Bao. Adaptive recurrent frame prediction with learnable motion vectors. In

- SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.
- [116] Z. Xie, R. Xie, R. Li, K. Huang, P. Qiao, J. Zhu, X. Yin, Q. Ye, W. Hua, Y. Huo, et al. Holistic inverse rendering of complex facade via aerial 3d scanning. *arXiv preprint arXiv:2311.11825*, 2023.
 - [117] B. Xu, J. Zhang, R. Wang, K. Xu, Y.-L. Yang, C. Li, and R. Tang. Adversarial monte carlo denoising with conditioned auxiliary feature modulation. *ACM Trans. Graph.*, 38(6):224–1, 2019.
 - [118] Y. Yang, R. Wang, and Y. Huo. Rule-based procedural tree modeling approach. *arXiv preprint arXiv:2204.03237*, 2022.
 - [119] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.
 - [120] X. Yin, D. Min, Y. Huo, and S.-E. Yoon. Contour-aware equipotential earning for semantic segmentation. *IEEE Transactions on Multimedia*, 2022.
 - [121] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021.
 - [122] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.
 - [123] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, and L. Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023.
 - [124] H. Zhang, T. Wang, M. Li, Z. Zhao, S. Pu, and F. Wu. Comprehensive review of visual-language-oriented multimodal pretraining methods. *Journal of Image and Graphics*, 27(9):2652–2682, 2022.
 - [125] X. Zhang, S. Fanello, Y.-T. Tsai, T. Sun, T. Xue, R. Pandey, S. Orts-Escalano, P. Davidson, C. Rhemann, P. Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021.
 - [126] Y. Zhang, R. Wang, Y. Huo, W. Hua, and H. Bao. Powernet: Learning-based real-time power-budget rendering. *IEEE Transactions on Visualization and Computer Graphics*, 2021.
 - [127] C. Zheng, Y. Huo, S. Mo, Z. Zhong, Z. Wu, W. Hua, R. Wang, and H. Bao. Nelt: Object-oriented neural light transfer. *ACM Transactions on Graphics*, 2023.
 - [128] F. Zhong, R. Wang, Y. Huo, and H. Bao. Normal and visibility estimation of human face from a single image. *arXiv preprint arXiv:2203.04647*, 2022.
 - [129] Y. Zhong, Y. Huo, and R. Wang. Morphological anti-aliasing method for boundary slope prediction. *arXiv preprint arXiv:2203.03870*, 2022.
 - [130] Z. Zhong, G. Chen, R. Wang, and Y. Huo. Neural super-resolution in real-time rendering using auxiliary feature enhancement. *Journal of Database Management (JDM)*, 34(3):1–13, 2023.
 - [131] Z. Zhong, J. Zhu, Y. Dai, C. Zheng, G. Chen, Y. Huo, H. Bao, and R. Wang. Fusesr: Super resolution for real-time rendering through efficient multi-resolution fusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023.
 - [132] F.-Y. Zhou, L.-P. Jin, J. Dong, et al. Review of convolutional neural network. 2017.
 - [133] J. Zhu, Y. Huo, Q. Ye, F. Luan, J. Li, D. Xi, L. Wang, R. Tang, W. Hua, H. Bao, et al. Supplementary material: I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdbs.
 - [134] J. Zhu, Y. Huo, Q. Ye, F. Luan, J. Li, D. Xi, L. Wang, R. Tang, W. Hua, H. Bao, et al. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdbs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12489–12498, 2023.
 - [135] J. Zhu, F. Luan, Y. Huo, Z. Lin, Z. Zhong, D. Xi, R. Wang, H. Bao, J. Zheng, and R. Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022.
 - [136] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

• • •