Localised-NeRF: Specular Highlights and Colour Gradient Localising in NeRF

Dharmendra Selvaratnam, Dena Bazazian
University of Plymouth, Faculty of Science and Engineering
School of Engineering, Computing and Mathematics (SECaM)
dharmendra.selvaratnam@postgrad.plymouth.ac.uk; dena.bazazian@plymouth.ac.uk

Abstract

Neural Radiance Field (NeRF) based systems predominantly operate within the RGB (Red, Green, and Blue) space; however, the distinctive capability of the HSV (Hue, Saturation, and Value) space to discern between specular and diffuse regions is seldom utilised in the literature. We introduce Localised-NeRF, which projects the queried pixel point onto multiple training images to obtain a multi-view feature representation on HSV space and gradient space to obtain important features that can be used to synthesise novel view colour. This integration is pivotal in identifying specular highlights within scenes, thereby enriching the model's understanding of specular changes as the viewing angle alters. Our proposed Localised-NeRF model uses an attention-driven approach that not only maintains local view direction consistency but also leverages image-based features namely the HSV colour space and colour gradients. These features serve as effective indirect priors for both the training and testing phases to predict the diffuse and specular colour. Our model exhibits competitive performance with prior NeRF-based models, as demonstrated on the Shiny Blender and Synthetic datasets. The code of Localised-NeRF is publicly available [1].

1. Introduction

The journey of image synthesis has evolved significantly from traditional methods to cutting-edge Neural Radiance Fields (NeRF). Initially, image synthesis relied heavily on voxels [6, 30] and meshes [7, 37], which are 3D models composed of cubes and polygons, respectively. These methods were instrumental in creating structured, yet often rigid and computationally intensive representations of 3D scenes. As technology progressed, the desire for more realistic and dynamically lit environments led to the development of NeRF, a technique introduced by Mildenhall [28]. NeRF [28] represents a paradigm shift in image synthesis, using a neural network to model the volumetric scene function. This approach allows for generating highly detailed and photorealistic images from novel viewpoints, surpassing the limitations of traditional voxel and mesh-based methods. NeRF's ability to interpolate and reconstruct complex scenes with intricate lighting and materials, marks a significant advancement in the field of computer graphics and virtual reality [8, 20, 50]. In the pursuit of advancing NeRF for the photorealistic rendering of complex view-dependent phenomena, particularly reflections on glossy surfaces, various methodologies have been explored. Prior efforts, such as those by [44], have sought to enhance the depiction of reflective surfaces through modifications to the parameters within volumetric rendering equations. Further, several approaches have emerged, ranging from the parameterisation of pixels as a linear amalgamation of basis functions derived via neural networks [49], to the redefinition of the NeRF function with conditional dependencies on surface position and orientation in observation space [51],

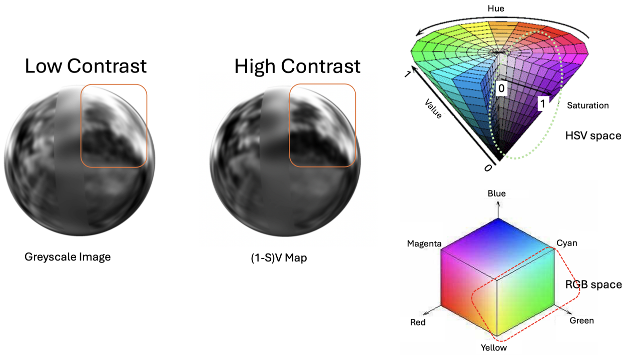


Figure 1. Dark colours and light shiny colours have a higher distinction in (1- Saturation) value map obtained through HSV space when compared with the greyscale image obtained from RGB space. HSV space based on the value and saturation parameters is capable of distinguishing the glossy regions from other non-shiny surfaces whereas RGB space does not differentiate it similarly (note the light green dotted circle area in the HSV cone and red dotted square area in the RGB cube).

[1] https://github.com/Dharmendra04/Localised-NeRF

and the integration of informed priors for optimised rendering [46]. More recently, [21] proposed a novel framework incorporating a learnable Gaussian directional encoding to adeptly model view-dependent effects under near-field illumination, offering an alternative to conventional environment map techniques akin to Ref-NeRf [44].

In this paper, we introduce our Localised-NeRF technique, which diverges from existing strategies by formulating a mechanism that can synthesise a specular feature indicating the level of specularity of each point. By doing this, rather than contributing towards the physical equation of rendering as mentioned by earlier literature, we direct our research towards localising the specular highlight spots that can be used by the final colour prediction model to represent accurate specularities. This preliminary step is inspired by techniques employed in medical imaging, notably for reducing specular reflections in colonoscopy video frames [1], thereby demonstrating the cross-disciplinary applicability of such filtering processes. As shown in Figure 1 by opting for the HSV colour space over the traditional RGB model, we further enhance the efficacy of this specular filtering phase, where the HSV space can clearly distinguish the glossy regions with other non-shiny surfaces.

Our proposed methodology for rendering non-Lambertian surfaces employs attention-based transformers to assimilate multiple viewpoints, thereby accurately delineating colour transitions and specular highlights. This multifaceted strategy is designed to surmount the inherent constraints of NeRF in rendering such surfaces. Utilising an attention mechanism, our model discerns local specularity and derives colour information from the pixel colours and gradients of proximate source views. The inclusion of colour gradients aids the model in identifying regions of edge significance. A zero gradient across all RGB channels indicates the absence of an edge or pronounced intensity variations. Given that our model synthesises target view features from multi-view inputs, these gradient nuances are essential for precise colour prediction.

We proposed using two transformers to enhance the accuracy of the multi-view Stereo algorithm. The transformers leverage the HSV and RGB colour spaces to produce initial specular and RGB colours. We were inspired by [48] and followed a similar approach to Localised-NeRF, which aggregates multi-view consistency features, such as saturation and value, in the HSV colour space, RGB and Colour gradients in RGB space. By employing a transformer model as opposed to a multilayer perceptron (MLP), we have efficiently encapsulated the interdependencies among adjacent features along the ray. This capability to discern relational nuances presents a significant advantage inherent to attention-based mechanisms, as opposed to the isolated treatment of points characteristic of conventional NeRF methodologies. Lastly, our approach integrates a 4-dimensional specular feature vector contingent upon volume density, which is instrumental in distilling essential specular attributes, significantly influenced by the spatial positioning of points, as delineated in [13]. Our contributions in Localised-NeRF can be summarised as follows:

1. To the best of our knowledge, Localised-NeRF pioneers the utilisation of HSV space within the Neural Radiance Field framework, capitalising on its ability to distinguish high-specularity points more effectively than RGB space. This technique notably enhances the precision in identifying specular highlight regions within an image, facilitating their subsequent accurate colouration.
2. Localised-NeRF exhibits robustness not only to colour variations within its view space but also demonstrate locality awareness via gradient space.

The contribution of this paper is not merely in surpassing the benchmarks of current state-of-the-art NeRF-based techniques. The significant aspects of our work include the incorporation of HSV and gradient space to tackle the challenges of accurate specular colour representation. Given that these features are readily available, our work paves the way for research into exploiting existing source image characteristics. The results of our proposed method are promising and represent the initial steps towards developing reliable approaches for using the HSV space to identify specularities. Our work opens new research avenues in NeRF-based approaches and poses novel research questions, highlighting potential areas for further improvement.

2. Related Work

Novel View Synthesis. View synthesis, the creation of images from novel, unobserved camera viewpoints, has evolved from simple light field interpolation techniques [10] for densely captured scenes to sophisticated methods for sparsely captured images that reconstruct 3D geometry for novel view rendering [5]. Recent advancements include NeRF [28], which employs a coordinate-based neural representation for photorealistic synthesis by simulating light interaction within a scene. Subsequent models have extended NeRF's application to dynamic scenes [33], avatar animation [34], and phototourism [26], by focusing on improving view-dependent appearance and geometry's smoothness. Additionally, efforts have been made to generalise NeRF to unseen objects and scenes using local feature projection [43], improving generalisation through initial weight adjustments [40], and disentangling shape and texture for enhanced scene reconstruction [16]. These developments underline a significant shift towards neural rendering techniques for efficient and high-fidelity novel view synthesis across various applications. Usually, the above-mentioned methods are used to colour each 3d point along a ray and use volume rendering to obtain the final colour of each pixel. In contrast, [13] used a per-pixel colouring approach to obtain a specular colour map, that can added directly to the diffuse colour map to obtain the final colour. Inspired by this approach and aiming to alleviate the inaccuracies in specular maps produced therein, we exploited the HSV space to locate the specularities more efficiently and accurately.

Specular Effects in Neural Radiance Field. The exploration of specular effects within the domain of Neural Radiance Fields (NeRF) has led to various advancements aimed at improving the depiction of glossiness and reflections. A seminal contribution in this area is Ref-NeRF [44], which innovatively applies the concept of reflection direction for parameterising the NeRF rendering equation, deviating from conventional environmental mapping techniques used in computer graphics to simulate reflections by mapping the environment onto a spherical or cubic surface. This approach is embodied in a view-directional Multilayer Perceptron (MLP) within Ref-NeRF, further complemented by the integration of Integrated Positional Encoding (IDE) [3] to meld surface roughness with reflection direction, utilising spherical harmonic functions for representation [29]. Given its reliance on precise surface normals, Ref-NeuS [46] builds upon Ref-NeRF by incorporating normal priors, which can be efficiently sourced from tools like Open3D [54]. In an evolution of encoding methods, [21] adopted Gaussian-based encoding over IDE, reporting enhanced outcomes. These methodologies predominantly adopt the Phong reflection model [35] for BRDF representation, yet historically overlooked the Fresnel effect—a gap recently addressed by [41], who differentiated between translucent and reflective surfaces to refine specular rendering, leveraging an attention mechanism for this purpose [31]. The realm of dynamic scenes also saw progress with efforts like [51], which augmented specular reflections dynamically. Reflecting principles from the field of computer graphics, the surface colour is often delineated through a blend of diffuse, specular, and ambient components. This concept has led researchers to employ both specular and diffuse colours in generating the final scene colour [13, 24]. Furthermore, NeX [49] innovated in rendering shiny surfaces by employing basis functions within a multi-plane image framework [55], marking a significant advancement in efficiently capturing specular effects. In our approach, we stand alone in contrast to other methods, by pioneering in using the techniques used in medical imaging to [39] separate high specular highlight from the image and allow the attention-based model to learn about the variation of the specularity along the ray. Especially our approach is also the pioneer in using HSV space with NeRF to exploit the saturation and the Value which is highly sensitive to these specularities [1, 2, 9, 25, 32].

Attention-based NeRF. In the evolving landscape of NeRF, the integration of attention mechanisms has emerged as a critical avenue for enhancing the fidelity of novel view synthesis. The concept of applying an attention mechanism across the viewing direction is inspired by Pixel NeRF [52], which pioneered the integration of multi-view feature consistency into NeRF models utilising single input images [11, 15]. Generalisable NeRF Transformer [47] leverages transformers for neural scene representation and rendering by first aggregating multi-view geometry data along Epipolar line to predict features, and then decoding the features to render novel views through attention-guided ray marching [42]. This approach completely proved that a scene can be reconstructed for novel view synthesis by using only attention mechanism, and eliminated physical-based rendering techniques [36]. TransNeRF [45] represents another leap forward, employing an attention mechanism to decode intricate relationships between an arbitrary number of source views into a unified scene representation, thus addressing the local consistency often overlooked by MLP-based NeRFs. This method demonstrates superior performance, especially in scenarios with significant viewpoint shifts. ViewFormer [19], by contrast, proposes a 2D-only method focusing on efficiency and rapid training, using a novel branching attention mechanism for both neural rendering and camera pose estimation [38], offering competitive results without explicitly reasoning in 3D. The Vision Transformer [22] method takes a distinctive approach by reducing input complexity to a single unposed image, merging global and local features to synthesise novel views with rich detail, surpassing existing methods in rendering quality. Lastly, NeRF-AD [4] introduces an attention-based disentanglement module for talking face synthesis [14], highlighting the potential of attention mechanisms in producing highly realistic facial animations driven by audio cues. Collectively, these advancements underscore the transformative impact of attention-aware techniques in NeRF [28], paving the way for more expressive and computationally efficient models in the domain of novel view synthesis. GeoNeRF [17] introduces a geometry reasoning stage combined with a Transformer-based rendering process that adeptly handles occlusions and synthesises photorealistic images from cascaded cost volumes, showcasing the model's adaptability to both Synthetic [28] and Real datasets [27]. Similarly, ABLE-NeRF [41] departs from traditional volumetric rendering constraints, incorporating a self-attention [12] framework along rays and leveraging learnable embeddings to capture scene-specific view-dependent effects, significantly diminishing the blurriness in glossy surfaces and achieving state-of-the-art results in rendering translucency. As these previously stated models used attention-based mechanisms to leverage different properties we proposed a transformer-based model in gradient space and HSV space to capture local information and specularity property. Our proposed approach contributes to the literature on using attention models in Neural Radiance

Fields by introducing a transformer-based model in gradient and HSV space.

3. Methodology

In pursuit of demonstrating the efficacy of utilising the HSV space within a NeRF-based model to efficiently highlight specular reflections, we encountered several challenges in modelling this network. A primary hurdle was the extraction of HSV features, given the infeasibility of utilising target view source pixels during rendering. To address this, we developed a method to aggregate adjacent source view features, encompassing both HSV and gradient information, as delineated in Section 3.2, drawing on the approach described by Wang et al. [48].

Additionally, to accurately predict the final colour, it was imperative to supplement the residual colour with an initial colour. This necessitated the integration of multi-view colour and gradient details from adjacent source view images, as detailed in Section 3.2. The model's performance was further enhanced by the incorporation of point-based specular features, as explicated in Section 3.4, which markedly improved image quality. Ultimately, by amalgamating all extracted features, our model successfully generated pixel-wise residual colour, as presented in Section 3.5.

For those new to the field, we begin with a brief introduction to the basics of NeRF to lay the groundwork for understanding our contributions.

3.1. NeRF Premilinaries

In the context of neural rendering, accurately modelling view-dependent effects such as reflections is pivotal for achieving high-fidelity scene reconstruction. Neural Radiance Fields (NeRF) offer a compelling foundational framework by leveraging a dual MLP architecture that encodes both the scene's density and colour. The rendering process for NeRF synthesises the perceived colour $C(\mathbf{o}, \mathbf{d})$ from a viewpoint $\mathbf{o}$ in direction $\mathbf{d}$ through the equation:

$$C(\mathbf{o}, \mathbf{d}) = \sum_i w_i c_i, \tag{1}$$

with weights w_i derived from the transmittance and density along the ray path. These weights are formulated as:

$$w_i = \exp\left(-\sum_{j<i} \sigma_j (t_{j+1} - t_j)\right) \left(1 - e^{-\sigma_i (t_{i+1} - t_i)}\right), \tag{2}$$

and the optimisation objective is to minimise the L2 norm of the difference between the predicted and ground truth pixel colours:

$$\mathcal{L} = \sum_{\mathbf{o},\mathbf{d}} \left\| C(\mathbf{o}, \mathbf{d}) - C_{\text{gt}}(\mathbf{o}, \mathbf{d}) \right\|^2. \tag{3}$$

3.2. Localising Specular Highlights Using Transformer

The incorporation of multiple views in our methodology provides a critical advantage by mitigating ambiguities inherent to single-view reconstructions and enriching the scene understanding. our versatile framework accommodates an arbitrary number of views, thus enhancing reconstruction fidelity.

During training and inference time, with multiple views available, we utilize the predefined camera poses. We represent the i^{th} input image as $I^{(i)}$ and denote its camera's transformation matrix, which maps coordinates from the world space to the camera's specific view space, as follows:

$$P^{(i)} = \begin{pmatrix} R^{(i)} & t^{(i)} \end{pmatrix} \tag{4}$$

Given a ray from a novel target camera that intersects a point x in space and has a direction d, we apply the transformation from world space to the camera space of each input view [52]. This process yields the transformed position and direction for the point as observed from the i^{th} camera's viewpoint:

$$x(i) = P^{(i)} \cdot x, \quad d(i) = R^{(i)} \cdot d \tag{5}$$

We can then transform these camera space coordinates to pixel coordinates and find the corresponding features for the queried point for each selected input source camera view.

After extracting the HSV features for the corresponding source views, view angles between the point of the projected source camera can be obtained in the world coordinate system using $t^{(i)}$ and x from the following equation:

$$\text{Viewing Angle} = \arccos\left(\frac{(X, Y, Z) \cdot t^{(i)}}{\|(X, Y, Z)\| \|t^{(i)}\|}\right)$$

Here, X,Y, and Z refer to the coordinates of the queried point x.

More importantly, We formulated obtained HSV colours, to a single channel using the below equation, which represents the specularity levels of each pixel [1]. A higher value means, high specularity in that pixel.

$$HSV(\mathbf{P})) = (1 - S)V \tag{6}$$

For each point, the extracted HSV features from the corresponding source views—as well as the viewing angles, are transformed into a higher-dimensional space through the application of a small MLP. This process is uniformly applied to both the features and the derived viewing angles for each source view associated with the points. As a result, this approach generates two distinct sequences: one representing the high-dimensional feature space and the other delineating the space of viewing directions. We employed

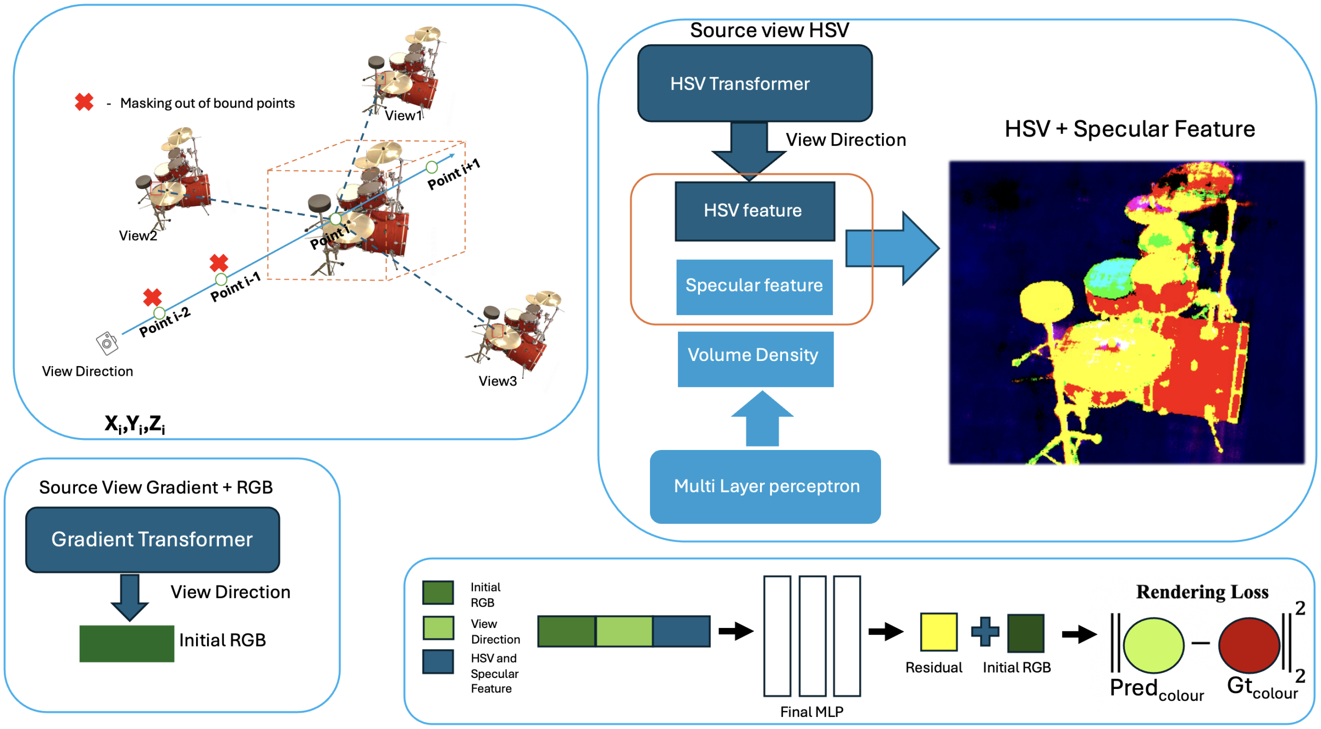


Figure 2. **End-to-End training Pipeline:** We obtained multi-view HSV and Colour gradients from adjacent source views for a target view from the training set. First, we used an HSV Feature-based transformer mechanism to attend weights and then concatenated the obtained HSV features with the specular feature vector obtained from a large MLP. Additionally, we predicted an initial colour by using a self-attention-based transformer mechanism using source view colour and its gradient. Finally, a small MLP is used to obtain these features, initial colour, and view direction to output the residual colour. The addition of the initial colour with the residual colour will give the final colour. The top left side of the figure illustrates we only query the point if it is inside the bounds of the scene.

a method akin to that described by [48] for aggregating these multi-view features into a higher-dimensional representation. Our technique incorporates a masking strategy for points falling outside the scene's bounds or obscured from all source views, thereby enhancing computational efficiency and reducing storage demands. Furthermore, to encapsulate global scene characteristics, we augment the initial HSV features with variance and covariance metrics computed across the source views. This approach facilitates a more nuanced and high-dimensional representation of the HSV features, enriching the model's perceptual understanding of the scene.

For a set of features (will take as $\mathbf{B}$ for simplicity) and camera parameters $\mathbf{C}$, the multi-head attention mechanism is defined as:

$$\text{MHA}(\mathbf{P}, \mathbf{B}, \mathbf{C}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n)\mathbf{W}^O \tag{7}$$

where each head is computed as :

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{Q}_i(\mathbf{P})\mathbf{K}_i(\mathbf{B}, \mathbf{C})^T}{\sqrt{d_{k_i}}}\right)\mathbf{V}_i(\mathbf{B}, \mathbf{C}) \tag{8}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value functions of the attention mechanism, respectively, and d_k is the scaling factor to normalise the dot products.

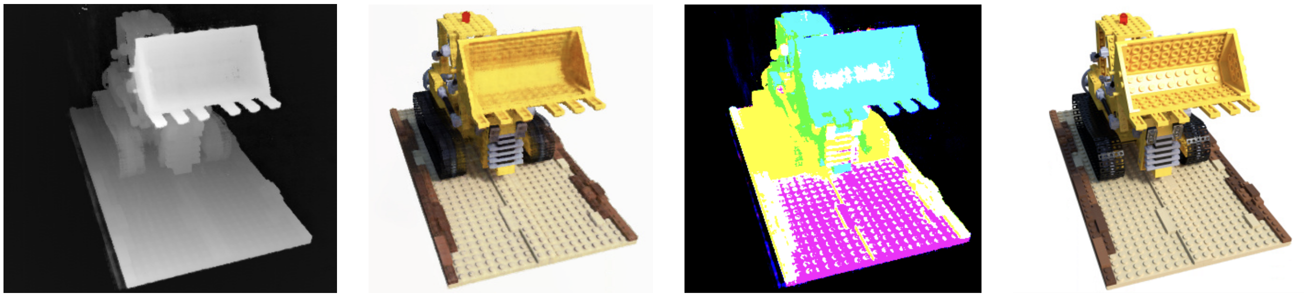
The attention-weighted feature representations $\mathbf{F}$ for each point are then obtained by:

$$\mathbf{F}(\mathbf{P}) = \text{CA}(\mathbf{P}, \mathbf{B}, \mathbf{C}) \cdot \mathbf{W}, \tag{9}$$

where $\mathbf{W}$ is a weight matrix learned during training. This allows the model to allocate weights depends on the specular highlights level. The obtained weights are then decoded using a small MLP to produce a 3-dimensional feature vector representing the HSV space. This method ensures a more robust detection of specular highlights compared to the traditional RGB space, where luminance is intertwined with colour information, potentially leading to less precise highlight localisation.

3.3. Utilising Colour Gradients Space

In computer vision, the gradient of an image is a fundamental concept used to identify the directional change in the intensity or colour of an image. Mathematically, the image gradient is a two-dimensional vector containing the partial



Depth Map

Residual Colour

Specular Feature

Ground truth

Figure 3. Obtained Depth map, Residual colour produced from the specular feature + HSV feature, and 4-dimensional specular feature vector.

derivatives of the image intensity function with respect to its x and y coordinates. Given an image I, the gradient at each image point (x, y) is denoted as ∇I and is defined as:

$$\nabla I(x, y) = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right) \tag{10}$$

where $\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$ represent the intensity changes in the horizontal and vertical directions, respectively.

The magnitude and direction of the gradient vector can be computed as:

$$|\nabla I| = \sqrt{\left(\frac{\partial I}{\partial x} \right)^2 + \left(\frac{\partial I}{\partial y} \right)^2} \tag{11}$$

$$\theta = \arctan \left(\frac{\frac{\partial I}{\partial y}}{\frac{\partial I}{\partial x}} \right) \tag{12}$$

These equations to an image I yield gradient components I_x and I_y for horizontal and vertical directions, enabling the computation of gradient magnitude and direction [18]. In our model, instead of applying gradients to a greyscale image, we utilise the gradient on each of the RGB channels to derive gradients within each RGB space which outputs a 6 dimension feature vector. This approach offers significant advantages over using greyscale, as our objective is not edge detection through this algorithm, but rather to monitor shifts in colour within the RGB spectrum.

Finally, we output a 3-dimensional initial colour using the colour gradient and actual colours of adjacent source views using our gradient transformer. This will enable the initial colour to capture the directional shifts in colour that signal the presence of identical objects (Zero Gradient) or sharp transitions (High Gradient).

3.4. Volume Density and Specular Feature Vector

We used a traditional-based MLP to predict the volume density, but instead of predicting the RGB, we predicted a specular feature with a 4-dimensional vector inspired by [13]. The volume density predicted with the feature vector gives a much better image quality when compared with the separate attention-based mechanism for volume density and specular features. Thus this approach is obtained after experimenting with the model with different approaches. This proves that though the HSV feature encapsulates the specularity level of each point, the actual 3D point-related features are requisite in addition to accurately obtaining the final residual colour.

3.5. Final RGB colour Calculation

After obtaining the initial colour from the gradient-based transformer and the specular feature from the HSV transformer, we concatenate the obtained HSV features, initial colour, and the 4-dimensional specular feature vector. Finally, we add the view direction encoding to produce a view-dependent residual colour using a small MLP network. The right side of Figure 2 and 3rd image on Figure 3 illustrate the mentioned mechanism and obtained residual colour. Finally by adding the initial colour with this residual colour final colour is produced.

4. Implementation

Our Implementation is based on using Jax Implementation of NeRF similar to [13]. Our strategy has a multi-head attention mechanism with Multilayer Perceptrons (MLPs) to enhance the rendering of specular highlights within HSV and colour gradient spaces. The end to end network was trained for $250k$ iterations, with a learning rate initially set at 2×10^{-4} and annealed to 2×10^{-6} after a 2500 iteration warm-up. The optimisation utilised Adam with β_1 set to 0.9 and β_2 to 0.999. Each scene, comprising 25 training sample images as candidate views, was trained on GeForce RTX 4090 GPUs, leveraging a batch size of 64 rays, reflecting the volumetric properties influenced by view direction.

Localised-NeRF incorporates distinct multi-head and self-attention mechanisms to process HSV and gradient features independently. Feature aggregation from adjacent source views is executed through a two-layer MLP, encompassing 64 and 32 channels, respectively, to encode the acquired features. For HSV attributes, four attention heads are employed in conjunction with a dual-layer feed-forward network. Gradient features are analogously subjected to a self-attention mechanism paired with a two-layer feed-forward network. After the attention weights computation, a decoder MLP with two layers of 16 and 3 channels is utilised. Density prediction and the generation of a four-dimensional feature vector are facilitated by a substantial MLP with 512 channels. Conclusively, a smaller MLP comprising 16 channels across two layers is deployed to yield the terminal residual colour output.

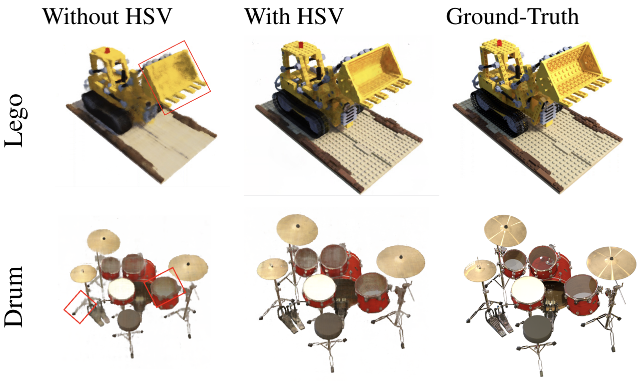


Figure 4. First column - Localised-NeRF without HSV feature (Ablating HSV transformer), Second column - Localised-NeRF model with HSV feature, Third column - Ground Truth image. Lego and Drum scenes from synthetic dataset [28]. The red bounding box in the images without HSV features has been significantly enhanced by the model which considers HSV features.

To bolster the dependability of our model across both the training and testing phases, we adopted a strategy to train the model to obtain different sequences of source views based on the Euclidean distance and viewing angle difference. This facilitated the selection of a set of 25 viewing cameras that were nearest to the queried camera, all the while ensuring the exclusion of the queried training view from the source views. For each queried point, we organised the angles between the source view camera and the queried camera, following the methodology outlined in [48]. These sequences delineated the camera direction projecting the pixel of a queried point towards that point, which was then leveraged by our attention models. This identical approach was applied during the testing phase also. Hence, we avoided the model relying on a particular set of images.

5. Results

Table 1. Comparison of PSNR and SSIM Metrics on Synthetic Dataset [27]. Our results (Localised-NeRF) are compared with and without HSV features. The top first and second results are shown in bold and underlined respectively.

Method	PSNR ↑	SSIM ↑
PhySG [53]	20.60	0.861
VolSDF [3]	27.96	0.932
Mip-NeRF [3]	33.09	0.961
Ref-NeRF [23]	33.99	0.966
ABLE-NeRF [41]	**35.02**	**0.975**
Localised-NeRF (without HSV)	29.94	0.945
Localised-NeRF (with HSV)	33.25	0.969

Table 2. Comparison of PSNR and SSIM Metrics on Shiny-Blender Dataset [44]. Our results (Localised-NeRF) are compared with and without HSV features. The top first and second results are shown in bold and underlined respectively.

Method	PSNR ↑	SSIM ↑
PhySG [53]	26.21	0.921
Mip-NeRF [3]	29.21	0.942
Ref-NeRF (without normal) [44]	30.91	0.936
Ref-NeRF [23]	**35.96**	0.967
ABLE-NeRF [41]	33.88	**0.969**
Localised-NeRF (without HSV)	30.93	0.949
Localised-NeRF (with HSV)	33.79	0.964

We conducted an experiment using two datasets: the NeRF synthetic dataset [28] and the Shiny Blender dataset of Ref NeRF [44]. The NeRF synthetic [28] dataset consists of objects that have 100 training views and 200 test views at a resolution of 800 × 800. The views were sampled either on the upper hemisphere or the full sphere. On the other hand, this dataset contains objects with complex geometry but is limited in terms of material variety, with most scenes being largely Lambertian. For our experiment, we used the Shiny Blender [44] dataset, which consists of six different glossy objects rendered in Blender under conditions similar to NeRF's dataset [28]. Each scene in this dataset has 100 training and 200 testing images, which contain more reflection and glossy effects to showcase our model potential. We compared our models with other previous models which used attention mechanisms by using multi-view features through SSIM and PSNR metrics. SSIM measures image quality by comparing structural similarity. PSNR assesses reconstruction quality by calculating the ratio of signal power to noise power. From Table 1 and Table 2, it is evident that our model was outperformed by both the Ref-NeRF [44] and ABLE-NeRF [41] models across the datasets. This suggests that Localised-NeRF could enhance its capabilities by incorporating parameterisation based on reflection direction to predict RGB colour or by transitioning towards a non-physically based rendering approach. Despite this, it performed comparably to previous models, demonstrating its proficiency in producing high-fidelity novel synthetic images.

Table 3. Ablation study made by changing the number of input views to 10,15 and 25. Drum and Lego scene from Synthetic dataset [28] and Ball scene from Shiny Blender dataset [44].

	PSNR ↑			**SSIM ↑**		
Scene	10	15	25 (Final)	10	15	25 (Final)
Ball	28.95	30.01	33.21	0.919	0.932	0.965
Lego	27.12	29.98	33.75	0.923	0.942	0.961
Drum	26.16	29.67	32.21	0.919	0.938	0.953

Table 4. Ablation study made by removing: HSV features, Initial colour obtained from gradient transformer, and 4-dimensional Specular feature. Final: All features are integrated. Scene: Ball scene obtained from Shiny Blender dataset [44] and Lego and Drum obtained from Synthetic dataset [28].

	PSNR ↑			**SSIM ↑**		
	Ball	Lego	Drum	Ball	Lego	Drum
HSV feature	27.23	28.17	26.88	0.912	0.909	0.896
Initial Colour	29.89	29.91	28.16	0.931	0.928	0.917
Specular Feature	30.09	30.15	29.90	0.948	0.943	0.938
Final	33.21	33.75	32.21	0.965	0.961	0.953

6. Ablation Studies

Figure 4 presents a comparison of Localised-NeRF without the HSV feature. It is apparent that the incorporation of localised specularity, obtained via the HSV transformer, markedly enhances the image quality and reduces blurriness. The observed data clearly demonstrate that, upon the removal of HSV features, the designated areas within both the Lego and Drum scenes undergo a substantial loss of geometric integrity. This underscores the benefits of utilising HSV features and leveraging the HSV space to accentuate specular reflections in the model. Furthermore, we have compared Localised-NeRF with and without the gradient transformer and the four-dimensional specular feature as shown in Table 4. It appears that, in contrast to other features such as colour gradients and specular features derived from the MLP, HSV features significantly aid the model in synthesising high-quality images. It can be seen all the HSV, gradient and specular features are essential components of Localised-NeRF to make it robust to produce better novel view synthesis From Table 3, we have provided an ablation on how the number of source views impacts the image quality metrics, and as usual, a high number of images leads to better image quality, due to the storage and training time concerns, we fixed the source views with 25, and if more resource available, it can be further increased to obtain a better performance.

7. Limitations

Our presented technique represents an initial foray into the ambitious endeavour of predicting specular effects within images utilizing the HSV space. Despite its innovative approach, our method is not without limitations. A significant constraint is its dependency on the availability of consistent, high-quality multi-view training data, which may not always be readily accessible. Furthermore, the complexity of the implemented attention mechanism escalates computational and storage requirements, which could adversely affect training and inference efficiency, particularly in environments with limited resources. Rather than focusing on the extraction of general features blindly and by specifically identifying useful feature spaces such as HSV for specular reflections, we have demonstrated the potential for making models more robust and their application more feasible. Modelling this approach without reliance on physically-based rendering and investigating its applicability for real-time rendering presents an intriguing and promising avenue for future research.

8. Conclusion

NeRF-based models, traditionally reliant on RGB space, encounter difficulties in identifying specularities within scenes containing non-Lambertian materials. Although previous works have explored the parameterisation of the NeRF equation and employed non-physical-based rendering methods, there has been a lack of consideration for utilising available image features to enhance specular effects. The current study introduces a novel and pioneering approach, termed Localised-NeRF, which exploits the HSV colour space to accurately identify specular locations without necessitating any complicated re-parameterisation of the radiance field. Specifically, this method leverages the transformer architecture to produce enhanced specular features that accurately represent the specular levels of each pixel. These features can subsequently be utilised in the final colour prediction phase to precisely locate the specularity levels. Furthermore, the incorporation of colour gradients across all RGB channels enables the model to encode information not only about the colour attributes of a point but also about the variations in intensity across different source views. To evaluate the generalization of the proposed approach, we trained an independent model that sampled the nearest cameras for a test sample in the testing dataset. We found that the performance of our approach is comparable to that of previous multi-view feature-dependent models. Overall, the results of the present study demonstrate the efficacy of our approach in producing novel views and its potential to be used for a variety of applications in computer vision such as product visualisation. We posit that Localised-NeRF stands as a pioneering effort in localising specular effects within novel view synthesis, thereby stimulating further interest among researchers to explore and leverage specific image-based features. This exploration aims to uncover their potential advantages in predicting the radiance field.

References

[1] Mojtaba Akbari, Majid Mohrekesh, Kayvan Najariani, Nader Karimi, Shadrokh Samavi, and SM Reza Soroushmehr. Adaptive specular reflection detection and inpainting in colonoscopy video frames. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 3134–3138. IEEE, 2018. 2, 3, 4

[2] Samar M Alsaleh, Angelica I Aviles, Pilar Sobrevilla, Alicia Casals, and James K Hahn. Automatic and robust single-camera specular highlight removal in cardiac images. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 675–678. IEEE, 2015. 3

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3, 7

[4] Chongke Bi, Xiaoxing Liu, and Zhilei Liu. Nerf-ad: Neural radiance field with attention-based disentanglement for talking face synthesis. *arXiv preprint arXiv:2401.12568*, 2024. 3

[5] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 423–432. 2023. 2

[6] German KM Cheung, Takeo Kanade, J-Y Bouguet, and Mark Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, pages 714–720. IEEE, 2000. 1

[7] Hervé Delingette. General object reconstruction based on simplex meshes. *International journal of computer vision*, 32:111–146, 1999. 1

[8] Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022. 1

[9] Melanie Ganz, Xiaoyun Yang, and Greg Slabaugh. Automatic segmentation of polyps in colonoscopic narrow-band imaging data. *IEEE Transactions on Biomedical Engineering*, 59(8):2144–2151, 2012. 3

[10] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 453–464. 2023. 2

[11] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 3

[12] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022. 3

[13] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul 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. 2, 3, 6

[14] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 3

[15] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 3

[16] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[17] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3

[18] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2): 358–367, 1988. 6

[19] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 3

[20] Chaojian Li, Sixu Li, Yang Zhao, Wenbo Zhu, and Yingyan Lin. Rt-nerf: Real-time on-device neural radiance fields towards immersive ar/vr rendering. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, pages 1–9, 2022. 1

[21] Jiabao Li, Yuqi Li, Ciliang Sun, Chong Wang, and Jinhui Xiang. Spec-nerf: Multi-spectral neural radiance fields. *arXiv preprint arXiv:2310.12987*, 2023. 2, 3

[22] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 806–815, 2023. 3

[23] David B Lindell, Julien NP Martel, and Gordon 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. 7

[24] Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Cleannerf: Reformulating nerf to account for view-dependent observations. *arXiv preprint arXiv:2303.14707*, 2023. 3

[25] Stephane Mallat and Wen Liang Hwang. Singularity detection and processing with wavelets. *IEEE transactions on information theory*, 38(2):617–643, 1992. 3

[26] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel 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. 2

[27] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3, 7

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3, 7, 8

[29] Claus Müller. *Spherical harmonics*. Springer, 2006. 3

[30] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32 (6):1–11, 2013. 1

[31] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021. 3

[32] JungHwan Oh, Sae Hwang, JeongKyu Lee, Wallapak Tavanapong, Johnny Wong, and Piet C de Groen. Informative frame classification for endoscopy video. *Medical image analysis*, 11(2):110–127, 2007. 3

[33] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[34] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[35] Bui Tuong Phong. Illumination for computer generated pictures. In *Seminal graphics: pioneering efforts that shaped the field*, pages 95–101. 1998. 3

[36] Why Physically-Based Rendering. Physically-based rendering. *Procedia IUTAM*, 13:127–137, 2015. 3

[37] Simon Schreiberhuber, Johann Prankl, Timothy Patten, and Markus Vincze. Scalablefusion: High-resolution mesh-based real-time 3d reconstruction. In *2019 International conference on robotics and automation (ICRA)*, pages 140–146. IEEE, 2019. 1

[38] Yoli Shavit and Ron Ferens. Introduction to camera pose estimation with deep learning. *arXiv preprint arXiv:1907.05272*, 2019. 3

[39] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017. 3

[40] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 2

[41] Zhe Jun Tang, Tat-Jen Cham, and Haiyu Zhao. Ablenerf: attention-based rendering with learnable embeddings for neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16559–16568, 2023. 3, 7

[42] Lukasz Jaroslaw Tomczak. Gpu ray marching of distance fields. *Technical University of Denmark*, 8, 2012. 3

[43] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2

[44] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 1, 2, 3, 7, 8

[45] Dan Wang, Xinrui Cui, Septimiu Salcudean, and Z Jane Wang. Generalizable neural radiance fields for novel view synthesis with transformer. *arXiv preprint arXiv:2206.05375*, 2022. 3

[46] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022. 2, 3

[47] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all that nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 3

[48] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas 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. 2, 4, 5, 7

[49] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. 1, 3

[50] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 1

[51] Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8285–8295, 2023. 1, 3

[52] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo 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. 3, 4

[53] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 7

[54] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 3

[55] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3