

NeRF-NQA: No-Reference Quality Assessment for Scenes Generated by NeRF and Neural View Synthesis Methods

Qiang Qu , Hanxue Liang, Xiaoming Chen *, Yuk Ying Chung , and Yiran Shen *



Humans	✓		✓		✓
PSNR, SSIM, LPIPS		✓	✓		✓
VMAF, FovVideoVDP		✓	✓		✓
Proposed NeRF-NQA	✓			✓	✓

Fig. 1: **Which NVS-generated scene (left or right) is better?** The areas manifesting significant blur and artifacts are demarcated with red boxes for enhanced visibility. In each instance, image quality assessment methods (PSNR, SSIM, LPIPS) and video quality assessment methods (VMAF, FovVideoVDP) diverge from human evaluations. Remarkably, the decisions from proposed quality assessment method exhibit strong concordance with human subjective perception.

Abstract—Neural View Synthesis (NVS) has demonstrated efficacy in generating high-fidelity dense viewpoint videos using a image set with sparse views. However, existing quality assessment methods like PSNR, SSIM, and LPIPS are not tailored for the scenes with dense viewpoints synthesized by NVS and NeRF variants, thus, they often fall short in capturing the perceptual quality, including spatial and angular aspects of NVS-synthesized scenes. Furthermore, the lack of dense ground truth views makes the full reference quality assessment on NVS-synthesized scenes challenging. For instance, datasets such as LLFF provide only sparse images, insufficient for complete full-reference assessments. To address the issues above, we propose NeRF-NQA, the first no-reference quality assessment method for densely-observed scenes synthesized from the NVS and NeRF variants. NeRF-NQA employs a joint quality assessment strategy, integrating both viewwise and pointwise approaches, to evaluate the quality of NVS-generated scenes. The viewwise approach assesses the spatial quality of each individual synthesized view and the overall inter-views consistency, while the pointwise approach focuses on the angular qualities of scene surface points and their compound inter-point quality. Extensive evaluations are conducted to compare NeRF-NQA with 23 mainstream visual quality assessment methods (from fields of image, video, and light-field assessment). The results demonstrate NeRF-NQA outperforms the existing assessment methods significantly and it shows substantial superiority on assessing NVS-synthesized scenes without references. An implementation of this paper are available at <https://github.com/VincentQu/NeRF-NQA>.

Index Terms—Perceptual Quality Assessment, Quality of Experience (QoE), Immersive Experience, No-Reference Quality Assessment, Novel View Synthesis, 3D Reconstruction, Neural Radiance Fields (NeRF)

1 INTRODUCTION

The synthesis of photorealistic free views plays a pivotal role in enhancing user experiences in Virtual Reality (VR) and Augmented Reality (AR) [1, 47, 59]. Such realistic rendering immerses users deeply into the VR or AR environment, making it easier for them to engage in the virtual content [11, 45]. In AR, the seamless integration of virtual objects with real-world scenes is vital, and photorealistic rendering ensures that these virtual elements appear natural and believable. In VR, efficient view synthesis techniques can generate these “realistic” views without

extensive data storage for every perspective, optimizing application performance [36, 53]. The adaptability of these views to real-world lighting conditions ensures that virtual objects reflect, refract, and cast shadows realistically, enhancing the immersive experience.

However, the synthesis of photorealistic free views from limited RGB images collected from sparse viewpoints remains a pivotal challenge in the field of image-based rendering [7, 12, 21]. Recently, Neural View Synthesis (NVS) via implicit representations has emerged as a promising research field, with techniques such as Neural Radiance Fields (NeRF) [27] and its variants [3, 10, 48, 49, 60] gaining considerable attention for their exceptional fidelity and robustness. However, the quality assessment of NVS-generated scenes presents a complex task as it necessitates a comprehensive evaluation encompassing various dimensions, such as spatial fidelity and smoothness across consecutive views. This complexity is further amplified in immersive VR/AR environments, where users have the liberty to perceive NVS-generated scenes from unrestricted viewpoints [47, 59].

Current quality assessment protocols for NVS-generated scenes are typically based on full-reference image quality assessment methods, such as PSNR, SSIM [56], and LPIPS [64], on a subset of hold-out views. Nevertheless, these methods are primarily tailored for images, thus may not adequately capture the comprehensive and immersive quality of NVS-generated scenes as perceived by human observers. Figure 1 illustrates such examples where the quality assessment methods diverge from human assessments. Along with the issue above, the

• Qiang Qu and Yuk Ying Chung are with the School of Computer Science, the University of Sydney, Australia. E-mail: {vincent.qu, vera.chung}@sydney.edu.au.
 • Hanxue Liang is with the Department of Computer Science and Technology, the University of Cambridge, United Kingdom. E-mail: hl589@cam.ac.uk.
 • Xiaoming Chen is with the School of Computer and Artificial Intelligence, Beijing Technology and Business University, China. E-mail: xiaoming.chen@bjtu.edu.cn.
 • Yiran Shen is with the School of Software, Shandong University, China. E-mail: yiran.shen@sdu.edu.
 • Xiaoming Chen * and Yiran Shen * are the corresponding authors

Manuscript received 4 October 2023; revised 17 January 2024; accepted 24 January 2024. Date of publication 4 March 2024 on IEEE Transactions on Visualization and Computer Graphics; date of current version 15 April 2024. Digital Object Identifier: 10.1109/TVCG.2024.3372037

absence of ground truth views from diverse viewpoints makes the comprehensive quality assessment even more challenging. For example, existing datasets like LLFF [26] and DTU [16] provide only images from sparse views, and even the datasets with reference videos [4, 17, 20] are often limited to fixed capturing paths, rendering them inadequate for assessing NVS methods with full-reference methods, as NVS is able to generate unlimited views.



Fig. 2: NVS-generated scenes can be conceptualized from two perspectives: views (left) and points (right). From the perspective of views, a scene can be perceived as an ensemble of views originating from diverse viewpoints. From the perspective of points, a scene can be perceived as a collection of surface points where each surface point can be observed from multiple angles.

To design a quality assessment method tailored for NVS-generated scenes, it is imperative to understand the foundation of these scenes. NVS-generated scenes can be conceptualized from two perspectives: views and points. Intuitively, an NVS-generated scene can be perceived as an ensemble of views originating from diverse viewpoints, as illustrated on the lefthand side of Figure 2. Predominant quality assessment methods, such as PSNR and SSIM [56], are the view-centric approaches. These methods evaluate the quality of an NVS-generated scene by comparing, subsequently averaging the quality scores across all views to derive a final assessment. However, the view-centric methodologies have two inherent limitations. First, as previously highlighted, there is often a scarcity or complete absence of reference views in real-world scenarios. Second, a mere aggregation of quality scores might not accurately reflect the scene’s quality as perceived by the human visual system. An alternative perspective treats an NVS-generated scene as a collection of surface points, as depicted on the righthand side of Figure 2. Given that each surface point can be observed from different view angles, the angular quality can be meticulously evaluated by scrutinizing the visual patterns associated with each surface point from varied orientations.

To bridge the gap, we introduce NeRF-NQA, the first no-reference quality assessment framework for synthesized scenes with dense viewpoints. NeRF-NQA adopts a joint assessment approach consisting of both viewwise and pointwise assessment modules. The viewwise module evaluates the spatial quality of individual synthesized views, while the pointwise module focuses on the angular quality of individual scene surface point. Upon extensive evaluation and comparison against 23 established visual quality assessment methods, NeRF-NQA demonstrates superior performance on assessing the quality of NVS-synthesized scenes and better matches the perceptual judgement of human compared with existing assessment methods. The results shown in Figure 1 illustrates the alignment of NeRF-NQA with human perceptual judgments, in contrast to mainstream image and video quality assessment methods. While our research primarily focuses on scenes synthesized by NeRF or other NVS variants, it is important to note that NeRF-NQA is designed with versatility in mind. It is applicable to any novel viewpoint synthesis method that provides dense viewpoints. Our focus on NVS is driven by its capability to reconstruct high-quality scenes, offering a wide array of viewpoints and rich diversity. The primary contributions of this research are as follows:

- We propose the first no-reference quality assessment method for synthesized scenes with dense viewpoints, considering the limited availability or absence of reference views in NVS-synthesized views.

- We propose a joint quality assessment strategy, integrating both viewwise and pointwise approaches, to assess the quality of NVS-generated scenes. The viewwise approach focuses assessing the overall spatial quality of individual synthesized views and their inter-view consistency, while the pointwise approach focuses on the angular qualities, making it the pioneering approach on evaluating the quality of individual scene surface points and their compound inter-point quality.

- To achieve accurate quality assessment without reference, we design a deep learning-based model for NeRF-NQA. Our extensive evaluation, comparing NeRF-NQA against 23 well-established visual quality assessment methods, clearly demonstrates its superiority over these traditional approaches by a substantial margin.

2 RELATED WORK

Quality assessment methods can be broadly classified into full-reference and no-reference, contingent upon the dependence of reference media [39]. Full-reference methods necessitate complete access to the reference media during quality score prediction. In contrast, no-reference methods determine quality without referencing the original media. The no-reference methods, while more intricate in design, are better suited for practical scenarios [40]. This is particularly relevant for NVS quality assessment, given that some datasets, like LLFF, offer sparse images, rendering them inadequate for full-reference evaluations [26]. Therefore, our research emphasizes no-reference quality assessment.

Image Quality Assessment. The domain of image quality assessment is extensively studied. A plethora of full-reference methods for 2D images, such as PSNR, SSIM [56], MS-SSIM [58], IW-SSIM [57], VIF [44], FSIM [63], GMSD [61], VSI [62], DSS [2], HaarPSI [41], MDSI [33], LPIPS [64], PieAPP [37], and DISTs [9], have been delineated in literature. PSNR is a prevalent objective quality assessment method that quantifies the quality of reconstructed images by comparing the maximum possible power of a signal to the power of corrupting noise, with a higher PSNR indicating a closer resemblance to the original image. In contrast, SSIM evaluates the perceptual quality of images by considering changes in structural information, luminance, and texture, providing a more comprehensive understanding of perceived image quality [56]. VIF gauges image quality by considering the mutual information shared between the reference and the distorted image, offering a nuanced assessment by accounting for characteristics of the human visual system [62]. Lastly, the LPIPS employs deep learning techniques to measure perceptual differences between images, capturing intricate visual discrepancies that traditional methods might overlook [64]. No-reference image quality assessment methods include the likes of BRISQUE [29], NIQE [31], and CLIP-IQA [52]. As one of the most popular, BRISQUE leverages the scene statistics of locally normalized luminance coefficients to measure potential reductions in "naturalness" due to distortions [29].

Video Quality Assessment. Beyond standard images, quality assessment methodologies exist for alternative visual media formats such as videos. Leading video quality methods encompass STRRED [46], VIIDEO [30], VMAF [22], and FovVideoVDP [24]. STRRED focuses on the structural retention in videos, offering insights into the preservation of inherent video patterns post-processing [46]. The VIIDEO, on the other hand, is a no-reference video quality assessment method that relies solely on the video being evaluated, utilizing intrinsic statistical regularities observed in natural videos [30]. VMAF, or Video Multi-Method Assessment Fusion, combines multiple algorithms to predict video quality, aligning closely with human perception by considering factors like texture, luminance, and motion [22]. Meanwhile, FovVideoVDP is a sophisticated method tailored for video quality assessment, taking into account the viewer’s field of view to provide a more contextual evaluation [24]. The video quality assessment techniques are adaptable to NVS, given that the synthesized view sequence can be analogously interpreted as a video.

Light-Field Quality Assessment. Besides videos, light-field images are another media format that contains unique angular dimension for visual content. For light-field quality assessment, cutting-edge methods such as ALAS-DADS [39] and LFACon [40] are the-state-of-the-arts

in the field. ALAS-DADS is a pioneering no-reference light-field image quality assessment method designed for immersive media services. It introduces the light-field depthwise separable convolution for efficient spatial feature extraction and the light-field anglewise separable convolution to capture both spatial and angular features, ensuring a comprehensive yet efficient quality assessment [39]. LFACOn, on the other hand, addresses light-field imaging’s unique challenges by introducing the “anglewise attention” concept. This approach integrates a multihead self-attention mechanism into the angular domain of light-field images. With innovative attention kernels like anglewise self-attention, grid attention, and central attention, LFACOn effectively gauges light-field image quality while optimizing computational efficiency [40]. Those light-field quality assessment methodologies are compatible with NVS, as the synthesized views can be systematically rearranged into a light-field subview matrix, aligned with the respective camera poses of the views.

NVS Quality Assessment. Presently, the evaluation of NVS methods or NeRF variants predominantly employs full-reference image quality methods [3, 10, 27, 48, 49, 60], which involve comparing the test set with the synthesized set image by image. In particular, PSNR and SSIM are the predominant image similarity methods, while LPIPS stands out as the leading perceptual deep-learned quality assessment method. In this work, we assess the efficacy of the aforementioned quality assessment methods, including image, video, and light-field evaluation methods, to establish a performance benchmark for NVS quality assessment.

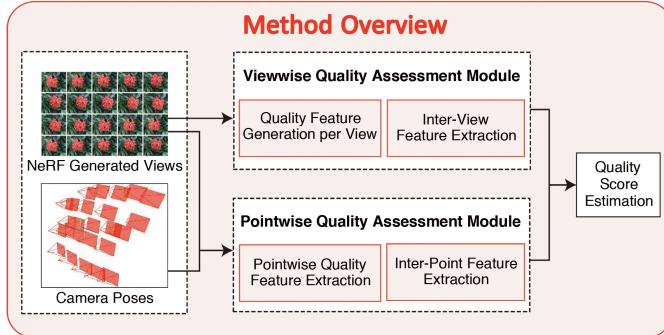


Fig. 3: Overview of the Proposed NVS Quality Assessment Framework.

3 METHODOLOGY

3.1 Overview of NeRF-NQA

As depicted in Figure 3, the architecture of NeRF-NQA is principally divided into three major components: the Viewwise Quality Assessment Module, the Pointwise Quality Assessment Module, and the Quality Score Estimation Module.

The Viewwise Quality Assessment Module is designed to evaluate the spatial quality of scenes generated from NVS. This module ingests the synthesized views and undergoes two primary stages: Quality Feature Generation per View and Inter-View Feature Extraction. The output consists of viewwise quality features that encapsulate the spatial characteristics of the scene (detailed in Section 3.2).

The Pointwise Quality Assessment Module aims to capture angular quality features that are challenging for the Viewwise Module to assess. Both NVS-generated views and their corresponding camera poses are taken as input and processed through a sequence of operations, including Pointwise Quality Feature Extraction and Inter-Point Feature Extraction, to yield pointwise quality features (detailed in Section 3.3).

Finally, the Quality Score Estimation Module employs a Multi-Layer Perceptron (MLP) to fuse the viewwise and pointwise features generated by the preceding modules, resulting in the final quality scores to offer a comprehensive assessment of the NVS scene. The intentional use of the MLP fusion aims to highlight the effectiveness of our proposed features. This fusion, common in representation learning as shown in references [6, 13], allows us to demonstrate the strength and discriminative power of extracted features without the interference of complex fusion techniques.

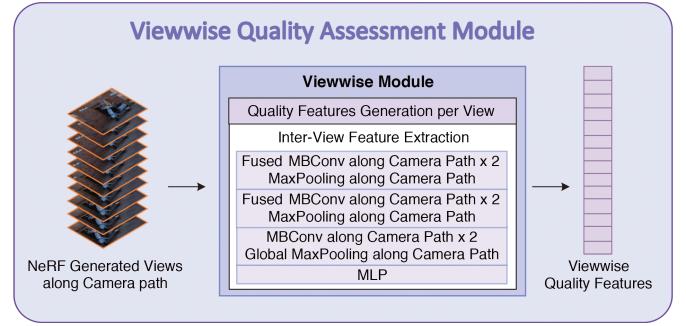


Fig. 4: The Structure of the Viewwise Quality Assessment Module.

3.2 Viewwise Quality Assessment

The quality of NVS-generated scene is intrinsically influenced by the quality of each synthesized view. After generating the quality features of individual views, it is imperative to holistically evaluate the final quality, factoring in the interrelation of these views. Given that NVS outcomes typically follow a camera trajectory, an intuitive approach is to analyze the quality features along this path.

Based on this concept, we introduce a viewwise quality assessment module, as depicted in Figure 4. The module starts with an initial block, Quality Features Generation per View [28], to individually assess the synthesized views to produce quality features for each view. Then, the Inter-View Feature Extraction block extracts features along the camera path, with the model structure inspired by EfficientNetV2 [50]. Specifically, it integrates two repeated sets of two (Fused) MBCConv layers (as per [42, 50]) combined with MaxPooling, two standalone MBCConv layers [42] followed by global MaxPooling and a MLP. The MBCConv employs the inverted bottleneck structure [42] and depthwise convolutional layers [14] to enhance memory efficiency. Additionally, a squeeze-and-excitation unit [15] is integrated within the MBCConv to recalibrate channel-wise feature responses adaptively. The fused MBCConv variant replaces depthwise convolutional layers with standard ones, proven to be more efficacious for larger spatial dimensions [50]. All layers operate coherently along the camera path, allowing the viewwise module to integrate inter-view quality features. This ensures that the quality of each view is evaluated in conjunction with its neighboring views. The global MaxPooling layer ensures the module’s compatibility with view sequences of varying lengths.

3.3 Pointwise Quality Assessment

While the viewwise module adeptly captures the spatial quality of synthesized views, it encounters challenges in encapsulate the important angular quality explicitly. The angular quality, often delineated as the experience of observing a consistent location from varied angles [40], can be contextualized in NVS scenes as viewing a singular surface point from diverse viewpoints. To encapsulate the angular quality inherent in the NVS scenes, we introduce the pointwise quality assessment module. This module is one of the key technical contribution of NeRF-NQA and its detailed design is shown in Figure 5. The module commences by accepting NeRF synthesized views and camera poses, subsequently sampling sparse surface points via COLMAP [43]. For each point, we compute pointwise quality features, elaborated in the subsequent paragraph. These high-dimensional pointwise quality features undergo further refinement in a feature extraction block, which distills the features per point and diminishes their dimensionalities. This block comprises four 3D convolutional layers followed by a MLP. Subsequently, an Inter-Point Feature Extraction block is designed by employing PointNet [38] to extract inter-point quality features based on the spatial positioning of the points within the scene.

Pointwise Quality Feature Calculation. To encapsulate the angular quality inherent to NVS scenes, we introduce the Pointwise Normalized Spherical Gradient map (PNSG) as the foundational pointwise quality features. The essence of PNSG lies in computing the gradient of pixel values observed from different viewpoints targeting at an identical surface point. The intricate procedures underpinning the pointwise

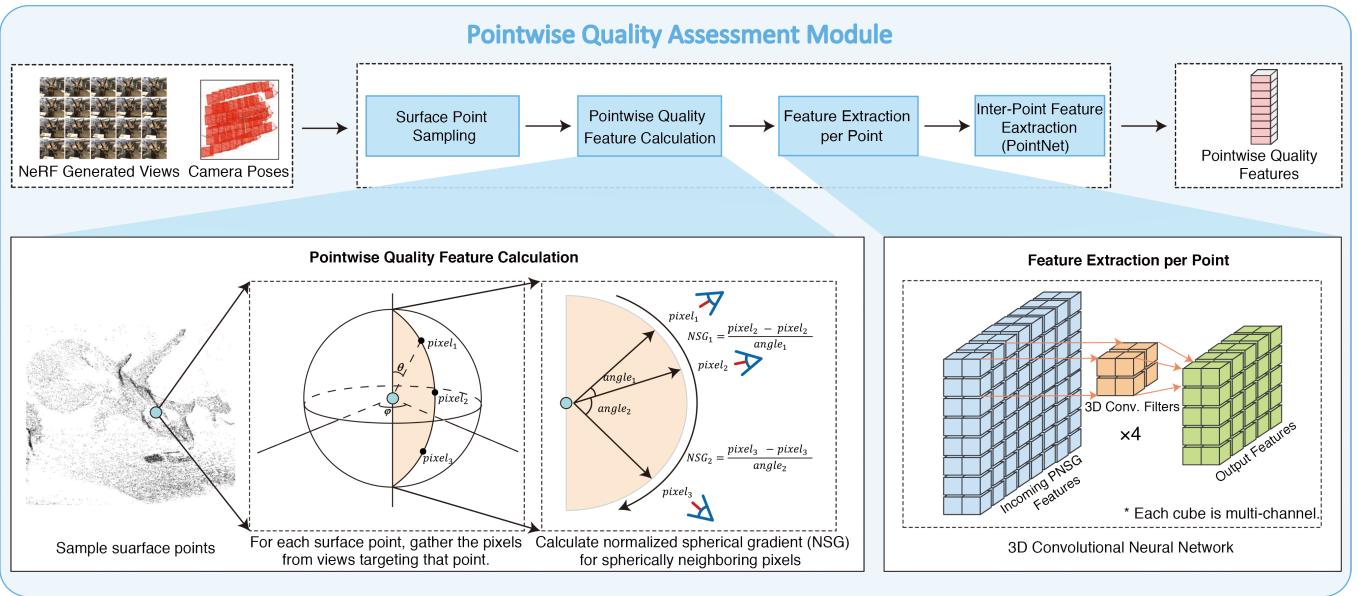


Fig. 5: The Detailed Architecture of the Pointwise Quality Assessment Module.

quality feature calculation are delineated on the lower-left quadrant of Figure 5. For each sampled surface point, we collate pixels from views targeting at the point. Subsequently, we compute the normalized spherical gradients (NSG) for spherically adjacent pixels. As depicted in the figure, for two pixels in proximity, the NSG is derived as the variance in pixel values normalized by the angular difference. Formally, let o denotes a surface point, and x_i, x_j are two pixels observing that point, NSG can be obtained by,

$$NSG(x_i, x_j) = \frac{I(x_i) - I(x_j)}{\angle x_i o x_j}, \quad (1)$$

where $I(x_i)$ and $I(x_j)$ denote the corresponding pixel values (i.e., vectors of RGB values), and $\angle x_i o x_j$ signifies the angular disparity between the two points.

Formal Definition of PNSG. The PNSG is derived as an aggregation of NSG values. Consider a set of n surface points, denoted as $\{P_i\}_{i=0}^{n-1}$, for which we aim to compute the PNSG. For a given surface point P_i , we can collate pixels from all synthesized views targeting at that point, given the respective camera poses. Each pixel is associated with both its viewpoint position in 3D space and its RGB values. Subsequently, we transform the pixel positions from Cartesian to spherical coordinates, using the surface point as the spatial origin. This transformation allows us to represent the view direction of each pixel using azimuthal and polar angles.

Initially, we compute the NSG along the azimuthal axis by partitioning the polar axis into b evenly spaced bins, represented as $\{B_i\}_{i=0}^{b-1}$. Pixels are then grouped into the nearest bins. For a specific azimuthal bin B_i containing m_i pixels, we arrange the pixels by their azimuthal angles, denoted as $B_i = \{x_j^i\}_{j=0}^{m_i-1}$. We then compute the NSG for each adjacent pair of pixels, resulting in b bins of NSG along the azimuthal axis, represented as NSG_{azi} .

$$NSG_{azi} = \{\{NSG(x_j^i, x_{j+1}^i)\}_{j=0}^{m_i-1}\}_{i=0}^{b-1}. \quad (2)$$

In a similar vein, we compute the NSG along the polar axis, denoted as NSG_{pol} . The PNSG for the surface point P_i is then represented as $\{NSG_{azi}^i, NSG_{pol}^i\}$. The cumulative PNSG for the entire scene is defined as:

$$PNSG = \{\{NSG_{azi}^i, NSG_{pol}^i\}_{i=0}^{n-1}\}. \quad (3)$$

From the derivations presented, it is apparent that the PNSG captures the dynamics within the angular domain, serving as a feature set for evaluating the angular quality inherent to NVS scenes.

Table 1: Ablation study on effectiveness of NeRF-NQA variants (with or without pointwise module) with quantitative evaluation (RMSE/SRCC) across the Fieldwork, LLFF, and Lab datasets. For each row, the best results are highlighted in bold.

NeRF-NQA Variant	Fieldwork		LLFF		Lab	
	RMSE ↓	SRCC ↑	RMSE ↓	SRCC ↑	RMSE ↓	SRCC ↑
w/o Pointwise	0.9202	0.9343	0.8856	0.7412	1.0033	0.8076
w/ Pointwise	1.1969	0.9701	0.5909	0.9023	0.6337	0.8628

4 EXPERIMENTS

4.1 Datasets for Evaluation

We evaluate and compare our proposed NeRF-NQA with existing quality assessment methods on three NVS datasets: Lab [23], LLFF [26], and Fieldwork [23]. Lab dataset features 6 real scenes captured in a lab setting with a 2D gantry, facilitating both horizontal and vertical camera movements. Training views were taken on a uniform grid, and reference videos ranged from 300 to 500 frames [23]. LLFF dataset comprises 8 real scenes captured via a handheld cellphone, each with sparse test views (20-30 images) [26]. Poses for these images were computed using the COLMAP structure from motion [43]. Fieldwork dataset contains 9 real scenes from outdoor urban areas and indoor museum spaces. These scenes are challenging due to intricate backgrounds, occlusions, and varying lighting. Reference videos typically have around 120 frames with diverse trajectories [23]. For each dataset, we randomly designate four scenes for testing, while the remaining scenes are allocated for training. To mitigate overfitting, we conduct ten rounds of random surface sampling on every scene, effectively augmenting both the training and testing samples tenfold.

4.2 Perceptual Quality Labels

The perceptual quality labels are derived from subjective experiments by Liang et al. [23]. They engaged 39 color-normal volunteers, with each participant completing 4–5 batches of comparisons using ASAP [25]. The results, scaled from pairwise comparisons, were articulated in Just-Objectionable-Difference (JOD) units via the Thurstone Case V observer model [35]. The JOD scores are offset by reference scores and thus predominantly negative values. A JOD score of 0 indicates undistorted quality. Higher JOD values suggest better quality perceived by human visual systems.

These experiments encompassed ten representative NVS methods, showcasing a variety of models with both explicit and implicit geometric representations, different rendering models, and optimization

Table 2: **Quantitative evaluation of various quality assessment methods across the Fieldwork, LLFF, and Lab datasets**, using measures such as RMSE, SRCC, PLCC, and OR. For each column, the best results are highlighted in bold, with the last row indicating the enhancement relative to the second-best result. The results for full-reference video quality assessment methods are marked as “–” for the LLFF dataset due to the absence of ground-truth videos.

Type	Method	Fieldwork				LLFF				Lab			
		RMSE ↓	SRCC ↑	PLCC ↑	OR ↓	RMSE ↓	SRCC ↑	PLCC ↑	OR ↓	RMSE ↓	SRCC ↑	PLCC ↑	OR ↓
FR-IQA	PSNR	3.4726	0.8941	0.8609	0.0000	1.0871	0.4058	0.3931	0.0000	1.0633	0.6090	0.5250	0.0000
	SSIM	2.4503	0.9371	0.9345	0.0000	1.0815	0.4359	0.4077	0.0450	1.0689	0.5171	0.3840	0.1700
	MS-SSIM	2.8353	0.9337	0.9236	0.0175	1.0961	0.3898	0.3838	0.0400	1.1061	0.4973	0.2942	0.1475
	IW-SSIM	2.4700	0.9447	0.9348	0.0000	1.0846	0.4801	0.4674	0.0150	1.2668	0.5330	0.2622	0.1500
	VIF	3.3400	0.9206	0.9300	0.0000	1.1744	0.1403	0.1303	0.0000	1.1971	0.5433	0.3328	0.0000
	FSIM	3.0174	0.9332	0.9304	0.0000	1.0872	0.4502	0.4250	0.0050	1.1181	0.5239	0.3182	0.1275
	GMSD	3.4804	0.9270	0.9068	0.0000	1.0697	0.4542	0.4473	0.0050	1.0909	0.5257	0.3554	0.0000
	VSI	3.7439	0.7077	0.7583	0.0025	1.1790	0.1703	0.1746	0.0025	1.3812	0.2198	0.1851	0.0200
	DSS	2.4960	0.9293	0.8930	0.0000	0.9216	0.6166	0.6077	0.0250	1.1328	0.5477	0.4338	0.0000
	HaarPSI	2.9704	0.9412	0.9298	0.0000	1.0612	0.4772	0.4651	0.0000	1.2917	0.5485	0.3418	0.0000
	MDSI	3.1141	0.9362	0.9109	0.0000	1.0579	0.4554	0.4581	0.0000	1.1171	0.5499	0.4224	0.0000
	LPIPS	2.6378	0.8894	0.9256	0.0000	1.1561	0.1532	0.2242	0.0275	1.1342	0.4010	0.3572	0.0450
	PieAPP	2.8222	0.9006	0.8527	0.0000	1.1331	0.3289	0.2941	0.0500	0.9576	0.7442	0.6315	0.0250
	DISTS	2.3891	0.9215	0.9383	0.0300	1.0841	0.4215	0.3628	0.0150	1.0604	0.4400	0.4164	0.0175
NR-IQA	BRISQUE	3.8765	-0.3824	-0.4039	0.0050	1.2805	0.1243	0.1160	0.0000	1.1461	0.3850	0.3307	0.0000
	NIQE	3.7584	0.1934	0.2077	0.0700	1.5258	-0.0767	-0.0740	0.0000	1.5213	-0.3272	-0.2698	0.0000
	CLIP-IQA	2.9339	0.7096	0.7412	0.0000	1.6296	-0.1653	-0.1586	0.0000	1.6787	-0.4358	-0.3698	0.0000
VQA	STRRED	1.7864	0.9416	0.8676	0.1000	—	—	—	—	1.0835	0.5338	0.4690	0.0175
	VMAF	3.7054	0.9142	0.8987	0.0000	—	—	—	—	1.1395	0.5077	0.3200	0.1050
	FovVideoVDP	3.8254	0.4979	0.4967	0.0100	—	—	—	—	1.3689	0.1280	0.1193	0.0050
	VIIVIDEO	3.7445	0.2902	0.3640	0.0000	1.2384	0.3213	0.3077	0.0000	1.0526	0.5677	0.5134	0.0000
LFIQA	ALAS-DADS	2.8179	0.5223	0.6299	0.1050	1.2504	0.5498	0.4691	0.0225	1.3447	0.3367	0.3936	0.0000
	LFACon	3.1131	0.4606	0.5628	0.0950	0.9074	0.5573	0.6466	0.3100	0.7918	0.5870	0.7309	0.0500
	NeRF-NQA	1.1969	0.9701	0.9804	0.0000	0.5909	0.9023	0.8858	0.0000	0.6337	0.8628	0.8720	0.0000
	Boost v.s. 2nd Best	+33.0%	+0.025	+0.042	—	+34.9%	+0.286	+0.239	—	+20.0%	+0.119	+0.141	—

strategies. NeRF [27] introduces a neural volumetric representation optimized for image-based scene reconstruction and the synthesis of novel views. Mip-NeRF [3] offers a multiscale representation tailored for anti-aliasing in view synthesis. Both DVGO [49] and Plenoxels [10] employ hybrid representations, streamlining the training and rendering processes. NeX [60] leverages multi-plane images combined with trainable basis functions, specifically designed to render view-dependent effects in forward-facing scenes. LFNR [48] adopts a light-field representation, incorporating an epipolar constraint to enhance the rendering process. Furthermore, both IBRNet [55] and GNT [54] are built upon the NeRF model and promote the generalizability. For IBRNet and GNT, both cross-scene models (GNT-C and IBRNet-C) and scene-specific models (GNT-S and IBRNet-S) were tested.

4.3 Training Setup

The model was trained utilizing the ADAM optimizer [19], over 200 epochs with a batch size of 10. It is designed as a generalized model, which, post-training, is capable of operating across diverse scenes without necessitating scene-specific fine-tuning. The weights, established during this initial training phase, are maintained consistently. In other words, once the model is trained, it is supposed to be proficiently applied to unseen scenes across different datasets. The computational experiments were conducted on a desktop equipped with an AMD 5950X processor, an RTX 3090 GPU, and 32GB of RAM, operating on Windows 10. The implementation replied on the PyTorch [34].

4.4 Metrics to Evaluate the Quality Assessment Methods

In the realm of quality assessment, several metrics are commonly employed to quantify the performance of quality assessment methods [39, 40]. Among these, the Root Mean Square Error (RMSE) [8] serves as a standard measure of the differences between predicted and ground-truth values, with lower RMSE values indicating more accurate predictions. The Spearman Rank Order Correlation Coefficient (SRCC) [65] assesses the strength and direction of the monotonic relationship between the predicted and ground-truth scores. Higher SRCC values signify a stronger correlation and, consequently, better performance. Similarly, the Pearson Linear Correlation Coefficient (PLCC) [8] evaluates the linear correlation between the predicted and

actual quality scores. A PLCC value closer to 1 indicates a strong positive linear correlation, thereby suggesting that the quality assessment algorithm is highly accurate in its predictions. Additionally, the Outlier Ratio (OR) is another important metric that is often calculated using statistical methods such as Tukey’s fences [51]. OR measures the proportion of data points that deviate significantly from the rest of the data distribution, providing insights into the robustness of the algorithm against anomalies or extreme values. Lower OR values are indicative of fewer outliers and thus suggest a more reliable and consistent performance. Collectively, these metrics provide a comprehensive evaluation on the quality assessment method’s performance in terms of both accuracy and correlation with human perceptual judgments.

4.5 Ablation Study on the Design of NeRF-NQA Model

As elaborated in Section 3, the Pointwise Module is specifically designed to capture angular quality features that are inherently difficult for the Viewwise Module for assessment. To empirically validate the efficacy of the Pointwise Module, we construct two variants of NeRF-NQA: one incorporating the Pointwise Module and the other excluding it. Comparative performance metrics for these variants are presented in Table 1. Our experimental findings reveal that the NeRF-NQA variant with the Pointwise Module consistently outperforms its counterpart across nearly all evaluation criteria, with the exception of RMSE on the Fieldwork dataset, where the results are closely aligned. Notably, in the LLFF dataset, the fully-equipped NeRF-NQA demonstrates a 33.3% reduction on RMSE and a 0.1611 increase on SRCC. These outcomes substantiate the utility and effectiveness of the Pointwise Module, thereby justifying its inclusion in subsequent experiments.

4.6 Comparison with Other Quality Assessment Methods

Our benchmarking considered prevalent full-reference image quality assessment metrics (FR-IQA) such as PSNR, SSIM [56], MS-SSIM [58], IW-SSIM [57], VIF [44], FSIM [63], GMSD [61], VSI [62], DSS [2], HaarPSI [41], MDSI [33], LPIPS [64], PieAPP [37], and DIST [9], along with no-reference image quality assessment metrics (NR-IQA) such as BRISQUE [29], NIQE [31], and CLIP-IQA [52]. We also included video quality assessment methods (VQA) such as STRRED [46],

Table 3: **Comparative analysis of quality assessment methods for various evaluated scenes**, using RMSE and SRCC. The penultimate column presents the rankings of NeRF-NQA. The final column delineates either the enhancement achieved by NeRF-NQA over the second-leading method (when NeRF-NQA is top-ranked) or the difference relative to the foremost method (if NeRF-NQA doesn't achieve the best score). The results of VMAF are “-” for Flower, Fortress, Horns, and Room because these scenes have no ground-truth videos.

NVS Scene	Evaluation	PSNR	SSIM	LPIPS	BRISQUE	VMAF	VIIDEO	LFACon	NeRF-NQA	Rank	Against Best Alt. Method
Dinosaur	RMSE ↓	2.9468	2.5588	2.5790	3.4200	3.2933	3.2472	2.9987	0.5898	1	+77.0%
	SRCC ↑	0.9522	0.9344	0.9338	-0.2168	0.9319	0.6352	0.4659	0.9735	1	+0.021
Elephant	RMSE ↓	1.5342	1.6444	1.5850	2.0242	1.6906	1.8351	1.3404	0.9576	1	+28.6%
	SRCC ↑	0.8930	0.8402	0.6311	-0.2853	0.8604	0.4167	0.6514	0.9567	1	+0.064
Naiad-Sta.	RMSE ↓	2.8302	1.8489	2.1174	3.3172	3.2272	3.8593	2.7868	1.5221	1	+17.7%
	SRCC ↑	0.9488	0.9540	0.8505	-0.0667	0.8184	-0.3570	0.4978	0.9535	2	-0.001
Vespa	RMSE ↓	5.4027	3.3683	3.7661	5.7716	5.5498	5.2229	4.4956	1.4659	1	+56.5%
	SRCC ↑	0.9449	0.9621	0.9530	-0.0949	0.9122	-0.4887	0.3489	0.9708	1	+0.009
Flower	RMSE ↓	1.2168	1.2335	1.4285	1.2765	-	1.3719	0.9715	0.4967	1	+48.9%
	SRCC ↑	0.4239	0.3089	-0.3086	0.3279	-	0.1380	0.5171	0.9756	1	+0.458
Fortress	RMSE ↓	0.9892	0.8335	0.8957	1.3640	-	1.0133	0.6981	0.6458	1	+7.5%
	SRCC ↑	0.3870	0.7396	0.6562	-0.1220	-	0.2178	0.6585	0.8916	1	+0.152
Horns	RMSE ↓	0.9626	0.9777	1.0308	1.2459	-	1.1435	0.9492	0.4264	1	+55.1%
	SRCC ↑	0.7518	0.6839	0.2275	0.1698	-	0.4256	0.5398	0.8859	1	+0.134
Room	RMSE ↓	1.1583	1.2274	1.2003	1.2316	-	1.3848	0.9807	0.7422	1	+24.3%
	SRCC ↑	0.3704	0.4322	0.2589	-0.3193	-	-0.0756	0.4285	0.7995	1	+0.367
CD-Occ.	RMSE ↓	0.9937	1.1628	1.1442	1.0021	0.9256	0.8736	0.5789	0.4982	1	+13.9%
	SRCC ↑	0.1293	0.0756	0.1718	-0.4688	0.0196	-0.1964	0.7807	0.8439	1	+0.063
Animals	RMSE ↓	1.3983	1.3806	1.3976	1.2730	1.2836	1.2849	0.7894	0.9927	2	-20.5%
	SRCC ↑	-0.0865	0.0263	0.0480	-0.3974	0.0052	-0.4333	0.7883	0.7890	1	+0.001
Metal	RMSE ↓	0.7097	0.4987	0.4562	0.4904	0.8212	0.7927	0.6188	0.3854	1	+15.5%
	SRCC ↑	0.5773	0.0930	0.0206	0.1244	0.1881	0.2620	0.4390	0.4751	2	-0.102
Toys	RMSE ↓	1.0374	1.0312	1.2941	1.5456	1.4194	1.1788	1.0802	0.4733	1	+54.1%
	SRCC ↑	0.6605	0.2962	0.3089	-0.1676	0.4893	0.3961	0.4146	0.7574	1	+0.097

Table 4: **Comparative analysis of quality assessment methods for different NVS methods**, using RMSE and SRCC. The last two columns show NeRF-NQA's ranking and its performance relative to the top or second-best method.

NVS Method	Evaluation	PSNR	SSIM	LPIPS	BRISQUE	VMAF	VIIDEO	LFACon	NeRF-NQA	Rank	Against Best Alt. Method
DVGO	RMSE ↓	1.1971	1.3403	1.4128	1.4384	1.5433	1.3659	0.9998	0.6837	1	+31.6%
	SRCC ↑	0.4763	0.4317	0.2963	-0.1425	0.1293	0.4086	0.4711	0.8537	1	+0.377
GNT-C	RMSE ↓	3.7748	2.6914	3.0733	4.1168	3.9801	3.9746	3.4138	0.6789	1	+74.8%
	SRCC ↑	0.6023	0.8206	0.6963	0.1707	-0.1342	0.3088	0.4432	0.9604	1	+0.140
GNT-S	RMSE ↓	3.1913	2.2837	2.3795	3.4402	3.3729	3.2597	2.7248	1.0084	1	+55.8%
	SRCC ↑	0.5641	0.7092	0.8658	0.3071	0.1773	0.4936	0.6731	0.9229	1	+0.057
IBRNet-C	RMSE ↓	2.8699	2.1395	2.3615	3.0050	3.0304	3.0659	2.6500	0.7860	1	+63.3%
	SRCC ↑	0.8464	0.8922	0.8832	0.5059	0.2711	0.3542	0.3833	0.9398	1	+0.048
IBRNet-S	RMSE ↓	1.8191	1.3968	1.4331	2.0398	2.1254	2.0002	1.2924	0.8834	1	+31.6%
	SRCC ↑	0.8623	0.8841	0.7367	0.5380	0.2084	0.2214	0.7325	0.9189	1	+0.035
LFNR	RMSE ↓	1.6693	1.3496	1.2549	1.5739	1.7614	1.6298	1.2313	0.7623	1	+38.1%
	SRCC ↑	0.2339	0.3743	0.8081	0.6265	0.1286	0.3687	0.7098	0.9668	1	+0.159
MipNeRF	RMSE ↓	1.0777	1.0727	1.1959	2.0036	1.6776	1.7545	1.0997	1.0721	1	+0.1%
	SRCC ↑	0.5050	0.3566	0.1638	-0.2746	0.4199	0.4742	0.0622	0.5664	1	+0.061
NeRF	RMSE ↓	1.9513	1.1090	1.2401	2.2956	2.2111	2.1238	1.9637	0.9019	1	+18.7%
	SRCC ↑	0.7639	0.8048	0.7820	0.2147	0.6347	0.5441	0.5653	0.9326	1	+0.128
NeX	RMSE ↓	1.2453	1.2616	1.2307	1.5399	1.1071	1.4861	1.0716	0.8470	1	+21.0%
	SRCC ↑	0.5811	0.6754	0.5128	-0.0095	0.7404	0.3016	0.4478	0.8184	1	+0.078
Plenoxel	RMSE ↓	1.0900	1.0692	1.0700	1.3264	1.4421	1.1829	0.7990	0.8204	2	-2.6%
	SRCC ↑	0.3497	0.3211	0.4910	-0.2480	-0.0534	0.3419	0.6916	0.8570	1	+0.165

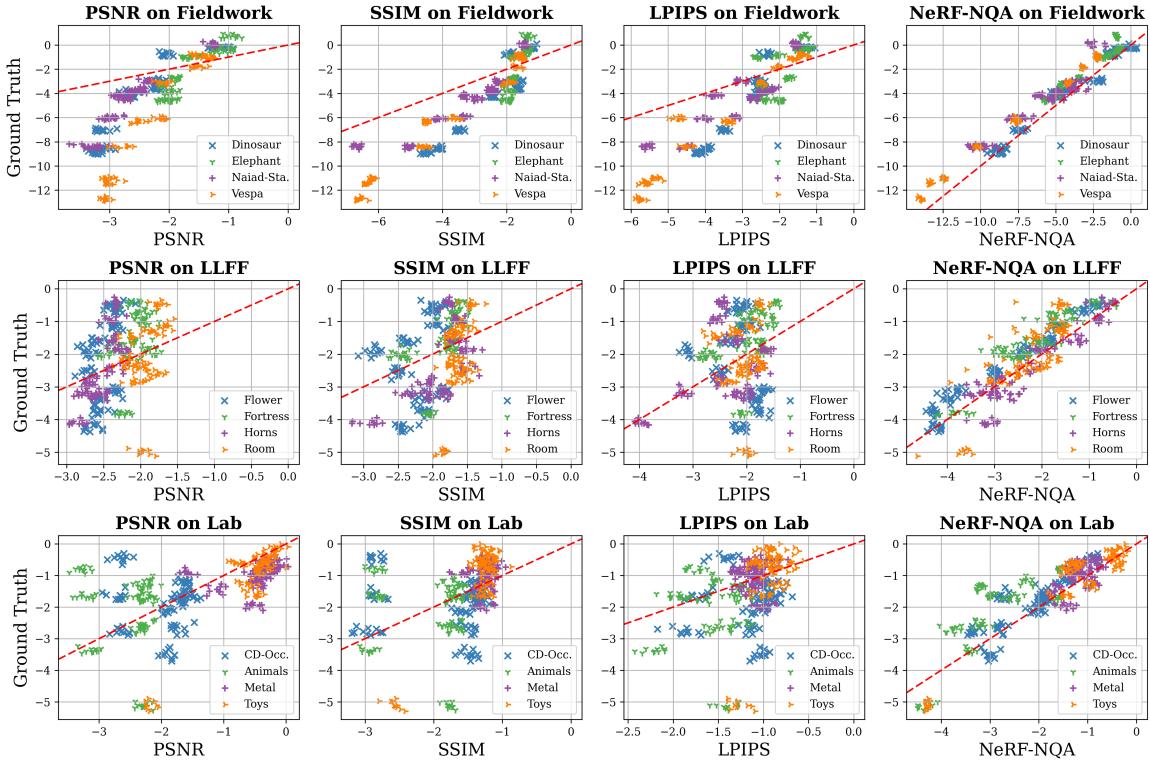


Fig. 6: Scatter plots illustrating the correlation between ground truth JOD and estimation made by the most widely used metrics for NVS (i.e., PSNR, SSIM, and LPIPS) and the proposed NeRF-NQA across the Fieldwork, LLFF, and Lab datasets. Distinct symbols and colors denote various scene. Each subplot features a red line representing the ideal prediction trajectory (i.e., ground truth == metric estimation). Notably, proximity of data points to this red line signifies superior metric performance.

VIIDEO [30], VMAF [22], and FovVideoVDP [24], and two state-of-the-art light-field quality assessment methods (LFIQA) including ALAS-DADS [39], and LFACon [40]. A detailed discussion of these quality metrics can be found in Section 2.

As delineated in Table 2, the evaluation results (RMSE, SRCC, PLCC and OR) of the 24 quality assessment methods (including NeRF-NQA) are presented across the Fieldwork, LLFF, and Lab datasets. The best results for each column are accentuated in boldface, while the bottom row quantifies the relative improvement over the second-best method. Upon examination of the table, it is evident that the proposed NeRF-NQA method consistently outshines all other benchmarked methods. Specifically, on the Fieldwork dataset, NeRF-NQA exhibits a remarkable 33% improvement on RMSE compared to the second-best method. In the LLFF dataset, NeRF-NQA demonstrates a 34.9% enhancement on RMSE, a 0.286 increment on SRCC, and a 0.239 rise on PLCC over the second-best results. On the Lab dataset, NeRF-NQA achieves significant gains: a 20.0% improvement in RMSE, a 0.119 increase in SRCC, and a 0.141 uptick in PLCC.

In summary, NeRF-NQA remarkably surpasses all other quality assessment methods, encompassing well-established FR-IQA methods such as PSNR, SSIM, and LPIPS, NR-IQA methods like BRISQUE and NIQE, recent advancements like CLIP-IQA, as well as VQA methods including VMAF and FovVideoVDP, and the-state-of-the-art LFIQA methods such as ALAS-DADS and LFACon.

4.7 Evaluation on Different Scenes

To analyze the efficacy of the evaluated quality assessment methods across different scenes, we present the scene-wise performance statistics in Table 3. As depicted in the table, the evaluation results of NeRF-NQA are consistently superior than others across a diverse array of NVS scenes. Specifically, NeRF-NQA attains the lowest RMSE values in 11 out of 12 scenes and the highest SRCC values in 10 out of 12 scenes. For RMSE, NeRF-NQA exhibits substantial improvements on scenes such as Dinosaur, Vespa, Horns, and Toys with enhance-

ments of 77.0%, 56.5%, 55.1%, and 54.1%, respectively, compared to the second-best methods. Similarly, in terms of SRCC, NeRF-NQA demonstrates remarkable advantages on scenes like Flower, Fortress, and Room, improving SRCC by 0.458, 0.152, and 0.367, respectively, against the second-best methods.

Figure 6 presents scatter plots contrasting ground truth quality scores with estimations from widely-used NVS methods (i.e., PSNR, SSIM, and LPIPS) as well as the proposed NeRF-NQA, across the evaluated datasets. Each subplot includes a red line, symbolizing the ideal prediction trajectory where ground truth scores are equivalent to estimations. The closeness of data points to this red line serves as an indicator of the method’s predictive accuracy. Upon scrutinizing the first row of Figure 6 pertaining to the Fieldwork dataset, it becomes evident that conventional methods like PSNR, SSIM, and LPIPS tend to produce biased estimations, particularly overestimating quality scores in scenes such as Dinosaur, Naiad-Sta., and Vespa. In contrast, NeRF-NQA effectively mitigates such biases across all scenes. Further analysis of the second and third rows of Figure 6 reveals that NeRF-NQA’s estimations are remarkably more concentrated and closely aligned with the red line, representing the ideal prediction trajectory, compared to other methods. This underscores that NeRF-NQA not only estimates with reduced bias but also with lower variance and a significantly diminished presence of outliers across all NVS scenes.

Figure 7 presents three illustrative NVS scenes generated by three different NVS methods to qualitatively demonstrate the efficacy of NeRF-NQA. Accompanying each scene are the ground truth Just-Noticeable Differences (JOD) scores, along with estimations from PSNR, SSIM, LPIPS, and NeRF-NQA. The results compellingly indicate that NeRF-NQA’s estimations are in close alignment with the ground truth JOD scores. For example, in the Dinosaur scene, which is characterized by pronounced blur and artifacts, conventional methods such as PSNR, SSIM, and LPIPS significantly overestimate the JOD score. In contrast, NeRF-NQA’s estimation stands at -8.6999, remarkably close to the ground truth score of -8.7655. A similar pattern is observed in the

Scene [Dinosaur] Synthesized by NVS method (GNT-C)				
 Ground Truth JOD: -8.7655				
Metric	PSNR	SSIM	LPIPS	NeRF-NQA
Estimation	-3.3529	-4.4325	-4.2286	-8.6999
Scene [Naiad-Sta.] Synthesized by NVS method (Nex)				
 Ground Truth JOD: -3.6308				
Metric	PSNR	SSIM	LPIPS	NeRF-NQA
Estimation	-2.5225	-2.4639	-2.2143	-3.5821
Scene [Horns] Synthesized by NVS method (MipNeRF)				
 Ground Truth JOD: -0.4222				
Metric	PSNR	SSIM	LPIPS	NeRF-NQA
Estimation	-2.4713	-1.7241	-2.3615	-0.4164

Fig. 7: Illustration of sample scenes generated via various NVS methods, accompanied by the corresponding ground truth JOD, which is underlined for emphasis. The estimations derived from NeRF-NQA are highlighted in bold and compared against those obtained from prevalent metrics for NVS, namely PSNR, SSIM, and LPIPS.

subsequent examples; while other methods either overestimate the JOD score in the Naiad-Sta. scene or underestimate it in the Horns scene, NeRF-NQA consistently produces estimations that closely approximate the ground truth JOD scores.

4.8 Evaluations on Different NVS Methods

Table 4 lists the performance of the evaluated quality assessment methods in different NVS methods. As delineated in Table 4, NeRF-NQA consistently outperforms other quality assessment methods across a diverse range of NVS methods. Specifically, NeRF-NQA achieves the most best RMSE values in 9 of the 10 evaluated NVS methods and the highest SRCC values across all NVS methods. In the context of RMSE, NeRF-NQA manifests significant performance gains in methods such as GNT-C, GNT-S, IBRNet-C, and LFNR, registering improvements of 74.8%, 55.8%, 63.3%, and 38.1%, respectively, when compared to the second-best performing metrics. Likewise, with respect to SRCC, NeRF-NQA exhibits pronounced advantages in methods like DVGO, NeRF, and Plenoxel, enhancing SRCC values by 0.377, 0.128, and 0.165, respectively, relative to the next best-performing metrics.

For a more comprehensive understanding, the line charts presented in Figure 8 visualize the performance of NeRF-NQA in terms of RMSE and SRCC metrics, juxtaposed with prevalent NVS methods such as PSNR, SSIM, and LPIPS across various NVS methods. Thus, the figure apparently show that NeRF-NQA excels among its competing methods for quality assessment across different NVS methods.

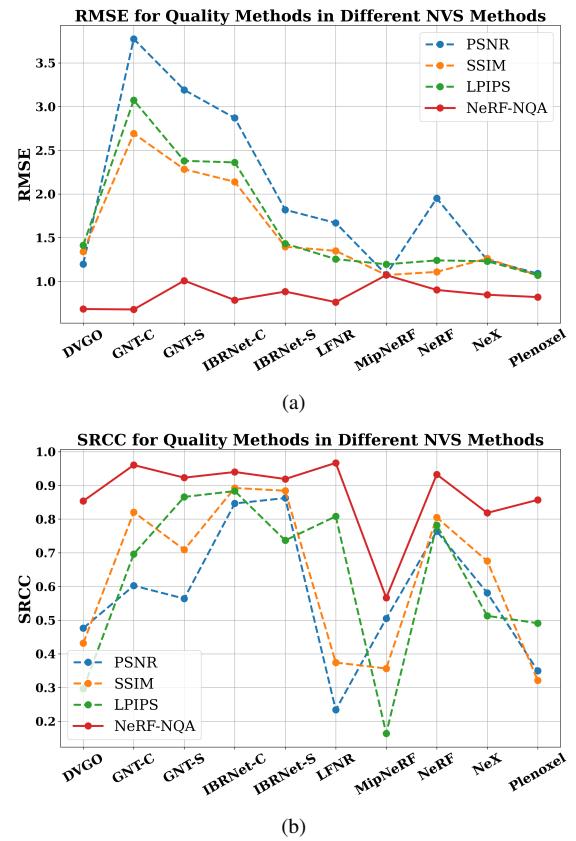


Fig. 8: Quantitative Evaluation of PSNR, SSIM, LPIPS and NeRF-NQA Across Various NVS Methods: (a) Line chart illustrating the RMSE ↓ performance for each NVS method; (b) Line chart depicting the SRCC ↑ values in relation to different NVS methods.

To rigorously evaluate the robustness of the proposed NeRF-NQA method in challenging scenarios, we have quantitatively assessed its performance across different scene types, as detailed in Table 5. This evaluation specifically targets special scenes traditionally deemed difficult for quality assessment, including those with complex shapes and specular objects, while also incorporating standard scenes for a comprehensive comparison. The empirical results consistently demonstrate that NeRF-NQA significantly surpasses competing methods, with a notably higher margin of improvement in special cases. This enhanced performance, particularly marked in complex and specular scenarios as evidenced in the last column of the table, underscores the method's robustness and efficacy. Such outcomes are likely attributable to the capacity of pointwise module on mitigating viewpoint dependency, affirming its integral role in the method's success.

4.9 Cross Dataset Evaluation

To substantiate the model's generalizability, cross-dataset evaluations were conducted, and the results are presented in Table 6. This table delineates the model's efficacy when trained on two datasets and subsequently tested on the third, exemplified by the 'Fieldwork' column, which reflects results from training on the LLFF and Lab datasets. The results reveal that, for the case of dataset independence, the proposed method consistently surpasses competing approaches, affirming its superior generalization capabilities. Notably, when the Fieldwork dataset is the test set, NeRF-NQA achieves a significant 35.0% improvement in RMSE over the second best method. Regarding the LLFF dataset, it exhibits a 13.2% enhancement in RMSE, along with increments of 0.290 in SRCC and 0.257 in PLCC. Similarly, for the Lab dataset, NeRF-NQA secures substantial advancements, bringing a 20.9% improvement in RMSE, a 0.133 increase in SRCC, and a 0.178 rise in PLCC, further evidencing its superior performance and generalization across diverse datasets.

Table 5: Comparative analysis of quality assessment methods for special cases including complex-shaped and specular scenes along with normal cases., using SRCC and PLCC metrics. Complex-shaped scenes include objects such as plants and skeletal specimens, while specular surfaces encompass scenes with specular reflections and transparent objects. The penultimate column presents the rankings of NeRF-NQA. The final column delineates either the enhancement achieved by NeRF-NQA over the second-leading method (when NeRF-NQA is top-ranked) or the difference relative to the foremost method (if NeRF-NQA doesn't achieve the best score).

Special Case	Evaluation	PSNR	SSIM	LPIPS	BRISQUE	VMAF	VIIDEO	LFACon	NeRF-NQA	Rank	Against Best Alt. Method
Complex Shapes	SRCC \uparrow	0.7093	0.6424	0.2842	0.0936	0.3027	0.3996	0.5076	0.9450	1	+0.236
	PLCC \uparrow	0.6916	0.6135	0.3628	0.1085	0.2931	0.4282	0.5974	0.9674	1	+0.276
Specular Surfaces	SRCC \uparrow	0.3302	0.1847	0.1616	-0.2457	0.1482	-0.0094	0.5702	0.7330	1	+0.163
	PLCC \uparrow	0.2473	0.2459	0.1438	-0.2311	0.1328	0.0890	0.6444	0.7962	1	+0.152
Normal Cases	SRCC \uparrow	0.7934	0.8740	0.7727	-0.1422	0.6615	-0.0528	0.5392	0.9431	1	+0.069
	PLCC \uparrow	0.7917	0.8226	0.7310	-0.1164	0.6793	-0.0102	0.5816	0.9591	1	+0.137

Table 6: Cross-dataset evaluation of various quality assessment methods. The table presents results from cross-dataset testing, where each method is trained on two datasets and tested on the third. For instance, results in the ‘Fieldwork’ column are derived from models trained on the LLFF and Lab datasets. For each column, the best results are highlighted in bold, with the concluding row indicating the enhancement relative to the second-best result. The results of full-reference video quality assessment methods are “-” for LLFF because this dataset has no ground-truth videos.

Type	Method	Fieldwork				LLFF				Lab			
		RMSE \downarrow	SRCC \uparrow	PLCC \uparrow	OR \downarrow	RMSE \downarrow	SRCC \uparrow	PLCC \uparrow	OR \downarrow	RMSE \downarrow	SRCC \uparrow	PLCC \uparrow	OR \downarrow
FR-IQA	PSNR	3.7940	0.8278	0.8323	0.0056	2.6261	0.1698	0.1813	0.0000	1.2514	0.6046	0.5452	0.0000
	SSIM	3.3938	0.8317	0.9026	0.0556	1.8391	0.2192	0.2161	0.0750	1.1192	0.5671	0.3983	0.1814
	MS-SSIM	3.5113	0.8255	0.8936	0.0711	1.8038	0.2125	0.2145	0.0737	1.1925	0.5637	0.3433	0.1714
	IW-SSIM	3.4253	0.8513	0.9075	0.0411	1.6957	0.2348	0.2262	0.0712	1.6057	0.5450	0.3202	0.1714
	VIF	3.8471	0.8399	0.8818	0.0000	2.6166	0.0434	0.0649	0.0000	1.6289	0.5328	0.3665	0.0000
	FSIM	3.6660	0.8509	0.9121	0.0422	1.6258	0.2887	0.2626	0.0750	1.2786	0.5700	0.3788	0.1414
	GMSD	3.7750	0.8462	0.8751	0.0067	2.1477	0.2438	0.2587	0.0187	1.1446	0.5676	0.4185	0.0000
	VSI	4.1231	0.5140	0.6343	0.0256	1.2846	0.2352	0.2669	0.0400	1.5947	0.3535	0.2868	0.0157
	DSS	3.3479	0.8799	0.8647	0.0000	1.9236	0.3962	0.3910	0.0250	1.1607	0.5827	0.4799	0.0000
	HaarPSI	3.5647	0.8790	0.8983	0.0011	2.0701	0.2317	0.2426	0.0213	1.8664	0.5426	0.3926	0.0000
	MDSI	3.6792	0.8614	0.8811	0.0122	2.1075	0.2719	0.2780	0.0250	1.3003	0.5591	0.4668	0.0000
	LPIPS	3.5493	0.8173	0.8806	0.0189	2.0290	0.1255	0.1944	0.0488	1.3633	0.3351	0.3333	0.0271
	PieAPP	3.5930	0.8626	0.8707	0.0300	3.4550	0.1613	0.2252	0.0250	0.8802	0.7266	0.6461	0.0429
	DISTS	3.4913	0.8554	0.9112	0.0622	1.4191	0.3393	0.3507	0.0275	1.0460	0.3167	0.3242	0.0157
NR-IQA	BRISQUE	4.1241	-0.1489	-0.1253	0.0022	1.8850	-0.0759	-0.0781	0.0000	2.6836	-0.3995	-0.4152	0.0000
	NIQE	4.2406	0.0296	0.0078	0.0033	1.4894	0.0649	0.0379	0.0000	1.6843	0.3635	0.2811	0.0000
	CLIP-IQA	4.5593	-0.4992	-0.5361	0.0000	1.5487	-0.0831	-0.0870	0.0000	1.5275	-0.4851	-0.3970	0.0000
VQA	STRRED	1.8680	0.9002	0.8778	0.0911	—	—	—	—	1.3699	0.5700	0.4939	0.0400
	VMAF	4.1449	0.8128	0.8283	0.0011	—	—	—	—	1.7342	-0.2734	-0.2234	0.0057
	FovVideoVDP	4.2092	0.6552	0.7067	0.0200	—	—	—	—	2.0782	-0.4587	-0.4000	0.0000
	VIIDEO	4.0178	0.2579	0.3029	0.0111	1.8884	0.3590	0.3817	0.0000	1.4648	0.3241	0.2732	0.0100
LFIQA	ALAS-DADS	2.6011	0.5655	0.6553	0.0700	1.1131	0.4886	0.4914	0.0000	1.0946	0.3126	0.2366	0.0000
	LFACon	2.4677	0.5662	0.6778	0.2800	1.1503	0.3700	0.4374	0.0537	0.8275	0.5708	0.6629	0.0900
	NeRF-NQA	1.2138	0.9125	0.9457	0.0444	0.9666	0.7785	0.7487	0.0000	0.6549	0.8600	0.8408	0.0000
	Boost v.s. 2nd Best	+35.0%	+0.012	+0.034	-0.044	+13.2%	+0.290	+0.257	—	+20.9%	+0.133	+0.178	—

5 LIMITATION

A notable limitation of the proposed method is its reliance on the sparse points generated by COLMAP [43] within the pointwise module. Despite this dependency, the empirical evidence presented in Table 5 suggests that this reliance does not markedly diminish the model’s performance, even in scenarios traditionally challenging for sparse point generation, such as scenes with complex shapes and specular surfaces. The sustained performance in these conditions indicates a degree of resilience to the noise inherent in COLMAP-generated sparse points. Future work will aim to address and potentially mitigate this dependency on COLMAP to further enhance the robustness and applicability of the proposed method. Due to time constraints, this research primarily focused on front-facing scenarios. As a result, perceptual scores for 360-degree scenes were not collected, and the methods related to 360-degree scenes were not tested. Additionally, the research did not encompass some of the latest advancements in NVS methods, including Instant-NGP [32], TensorRF [5], and 3D Gaussian Splatting [18]. Future work will aim to address these gaps by incorporating evaluations on 360-degree scenes and integrating a wider array of NVS methods to provide a more exhaustive analysis of the proposed approach.

6 CONCLUSION

In this paper, we introduce NeRF-NQA, an innovative quality assessment method to evaluate the quality of NVS-generated scenes without the dependency on reference views, addressing the prevalent challenges on scarce reference availability in NVS scenarios. NeRF-NQA adopts a joint quality assessment strategy, integrating both viewwise and pointwise assessment methodologies to facilitate a holistic evaluation of both the spatial fidelity and the intricate angular quality of the synthesized views. Empirical results underscore the pronounced superiority of NeRF-NQA in gauging the quality of NVS-generated views, outperforming extant quality assessment techniques for images, videos, and light fields. These findings accentuate the efficacy and robustness of NeRF-NQA as a pivotal instrument for discerning the perceptual quality of NVS-generated scenes.

ACKNOWLEDGMENTS

This work was supported in part by Beijing Natural Science Foundation (No. 4222003), National Natural Science Foundation of China (No. 62177001), and Shandong Provincial Natural Science Foundation (No. 2022HWYQ-040).

REFERENCES

- [1] D. Andersen and V. Popescu. An ar-guided system for fast image-based modeling of indoor scenes. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 501–502. IEEE, 2018. 1
- [2] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg. Image quality assessment based on dct subband similarity. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 2105–2109. IEEE, 2015. 2, 5
- [3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5470–5479, 2022. 1, 3, 5
- [4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2
- [5] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022. 9
- [6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 3
- [7] A. Colburn, A. Agarwala, A. Hertzmann, B. Curless, and M. F. Cohen. Image-based remodeling. *IEEE Transactions on Visualization and Computer Graphics*, 19(1):56–66, 2012. 1
- [8] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005. 5
- [9] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 2, 5
- [10] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5501–5510, 2022. 1, 3, 5
- [11] L. Gruber, T. Langlotz, P. Sen, T. Höherer, and D. Schmalstieg. Efficient and robust radiance transfer for probeless photorealistic augmented reality. In *2014 IEEE Virtual Reality (VR)*, pp. 15–20. IEEE, 2014. 1
- [12] S. Hauswiesner, M. Straka, and G. Reitmayr. Virtual try-on through image-based rendering. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1552–1565, 2013. 1
- [13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020. 3
- [14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3
- [15] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018. 3
- [16] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 406–413, 2014. 2
- [17] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pp. 402–418. Springer, 2022. 2
- [18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 9
- [19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [20] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2
- [21] L. Li and H.-W. Shen. Image-based streamline generation and rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(3):630–640, 2007. 1
- [22] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, et al. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2):2, 2016. 2, 7
- [23] H. Liang, T. Wu, P. Hanji, F. Banterle, H. Gao, R. Mantiuk, and C. Oztekin. Perceptual quality assessment of nerf and neural view synthesis methods for front-facing views. *arXiv preprint arXiv:2303.15206*, 2023. 4
- [24] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney. Fovvideodp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 2, 7
- [25] A. Mikhailuk, C. Wilmot, M. Perez-Ortiz, D. Yue, and R. K. Mantiuk. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2559–2566. IEEE, 2020. 4
- [26] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2, 4
- [27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020. 1, 3, 5
- [28] A. Mittal, A. K. Moorthy, and A. C. Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pp. 723–727. IEEE, 2011. 3
- [29] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 2, 5
- [30] A. Mittal, M. A. Saad, and A. C. Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 2, 7
- [31] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2, 5
- [32] 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. 9
- [33] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016. 2, 5
- [34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5
- [35] M. Perez-Ortiz and R. K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017. 4
- [36] C. Poullis, S. You, and U. Neumann. Rapid creation of large-scale photorealistic virtual environments. In *2008 IEEE Virtual Reality Conference*, pp. 153–160. IEEE, 2008. 1
- [37] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018. 2, 5
- [38] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017. 3
- [39] Q. Qu, X. Chen, V. Chung, and Z. Chen. Light field image quality assessment with auxiliary learning based on depthwise and anglewise separable convolutions. *IEEE Transactions on Broadcasting*, 67(4):837–850, 2021. 2, 3, 5, 7
- [40] Q. Qu, X. Chen, Y. Y. Chung, and W. Cai. Lfacon: Introducing anglewise attention to no-reference quality assessment in light field space. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2239–2248, 2023. 2, 3, 5, 7
- [41] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018. 2, 5
- [42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018. 3
- [43] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016. 3, 4, 9
- [44] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE*

- Transactions on image processing*, 15(2):430–444, 2006. 2, 5
- [45] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 1
- [46] R. Soundararajan and A. C. Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, 2012. 2, 5
- [47] S. Subramanyam, J. Li, I. Viola, and P. Cesar. Comparing the quality of highly realistic digital humans in 3dof and 6dof: A volumetric video case study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 127–136. IEEE, 2020. 1
- [48] M. Suhail, C. Esteves, L. Sigal, and A. Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8269–8279, 2022. 1, 3, 5
- [49] C. Sun, M. Sun, and H.-T. Chen. Direct voxel grid optimization: Superfast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5459–5469, 2022. 1, 3, 5
- [50] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pp. 10096–10106. PMLR, 2021. 3
- [51] J. W. Tukey et al. *Exploratory data analysis*, vol. 2. Reading, MA, 1977. 5
- [52] J. Wang, K. C. Chan, and C. C. Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 2555–2563, 2023. 2, 5
- [53] K. Wang, S. Peng, X. Zhou, J. Yang, and G. Zhang. Nerfcap: Human performance capture with dynamic neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1
- [54] P. Wang, X. Chen, T. Chen, S. Venugopalan, Z. Wang, et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 5
- [55] 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*, pp. 4690–4699, 2021. 5
- [56] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 2, 5
- [57] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on image processing*, 20(5):1185–1198, 2010. 2, 5
- [58] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2, pp. 1398–1402. IEEE, 2003. 2, 5
- [59] M. Whitlock, S. Smart, and D. A. Szafir. Graphical perception for immersive analytics. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 616–625. IEEE, 2020. 1
- [60] S. Wizadwongsu, P. Phongthawee, J. Yenpraphai, and S. Suwanjanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8534–8543, 2021. 1, 3, 5
- [61] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013. 2, 5
- [62] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014. 2, 5
- [63] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 2, 5
- [64] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 1, 2, 5
- [65] D. Zwillinger and S. Kokoska. *CRC standard probability and statistics tables and formulae*. CRC Press, 1999. 5