IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024IMP0003

Publicized:2025/01/15

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

AMDIS: Amplitude Dissimilarity Reduced Reference IQA Metric for Neural Radiance Field

Ren TOGO[†a)], ***Member***, **Rintaro YANAGI**[††b)], **Masato KAWAI**[†c)], ***Nonmembers***, **Takahiro OGAWA**[†d)], ***and*** **Miki HASEYAMA**[†e)], ***Members***

SUMMARY This paper presents a novel reduced-reference image quality assessment (RR IQA) method from monocular dynamic scene images for neural radiance fields (NeRF). Despite recent advancement in NeRF, evaluating the performance of NeRF models remains challenging due to the difficulty associated with obtaining ground truth viewpoint images for dynamic scenes. Collecting such ground truth images for NeRF model evaluation typically requires capturing the target scene from multiple synchronized cameras, which is labor-intensive. To address this issue, we propose a novel RR IQA metric called amplitude-dissimilarity (AMDIS), which focuses on evaluating NeRF models without requiring ground truth viewpoint images. The key idea behind AMDIS is that the differences between two near-viewpoint images are mainly absorbed in the phase components.

Thus, AMDIS evaluates NeRF models by measuring the dissimilarity between the Fourier amplitude components of the training and synthesized images. Because AMDIS only uses the training and synthesized images, the corresponding ground truth viewpoint images are not required for the evaluation. The experimental results demonstrate that the proposed AMDIS is strongly correlated with major full-reference IQA methods that directly use ground truth viewpoint images.

key words: *Neural radiance fields, image quality assessment, reduced reference, amplitude dissimilarity.*

1. Introduction

Neural radiance fields (NeRF) [1], which learn 3D information from multiple images captured from various viewpoints, have been actively studied [2]–[5]. NeRF can synthesize novel viewpoint images by estimating the volume density σ and view-dependent radiance RGB at the same spatial position from a single continuous set of 5D coordinates (spatial position (x, y, z) and viewing direction (θ, ϕ)). Previous NeRF models have focused primarily on static scenes, such as benches, pianos, and plants [1], [6]. However, real-world applications often involve dynamic scenes, such as video content shared on platforms such as YouTube and TikTok. To effectively use these video contents, the capability to handle dynamic scenes is essential. However, applying NeRF to dynamic scenes remains challenging because it typically requires multiple synchronized cameras for accurate evaluation.

Conventional evaluation procedures for NeRF models require multiple viewpoint images in each scene for training and evaluation. Although these images can be obtained easily by moving a single camera in static scenes, capturing them in dynamic scenes requires the use of multiple synchronized cameras. Therefore, dynamic scenes require significantly more effort to prepare ground truth viewpoint images than static scenes, making it challenging to effectively evaluate NeRF models. Given the importance of dataset size for NeRF models, it is highly desirable to establish a novel image quality assessment (IQA) metric that can accurately evaluate NeRF performance while minimizing data preparation effort.

Setting up multiple synchronized cameras to capture different viewpoint images requires significant effort. Ideally, it would be more efficient to evaluate NeRF models using only dynamic scenes from a single camera. Conventional evaluation metrics for NeRF models have been developed based on full-reference (FR) IQA methods [7]–[9]. In contrast, reduced-reference (RR) IQA methods have also been explored in the field of IQA [10]. FR IQA calculates evaluation scores by comparing the entire information of a synthesized image against the corresponding ground truth viewpoint image. However, RR IQA calculates an evaluation score by focusing on the partial elements of the synthesized and ground truth viewpoint images [10]. We found that if only viewpoint-invariant elements in the images can be evaluated, images from different viewpoints can serve as substitutes for ground truth viewpoint images in NeRF evaluation. In other words, an RR IQA method focusing on viewpoint-invariant elements can be used to evaluate the synthesized viewpoint images using only a single moving camera. Since the amplitude component in the Fourier domain is robust to viewpoint translation [11], [12], using RR IQA based on this component enables the evaluation of NeRF without requiring ground truth viewpoint images.

This paper proposes a novel RR IQA-based metric, amplitude-dissimilarity (AMDIS), for evaluating NeRF models. The proposed AMDIS assumes that the phase component in the Fourier domain primarily absorbs differences between the training and synthesized viewpoint images. As shown in Fig. 1, while the conventional NeRF metrics require a ground truth viewpoint image from the same viewpoint as the synthesized image, AMDIS enables the evaluation of NeRF models without requiring such ground truth viewpoint

[†]Hokkaido University
[††]National Institute of Advanced Industrial Science and Technology (AIST)
a) E-mail: togo@lmd.ist.hokudai.ac.jp
b) E-mail: rintaro.yanagi@aist.go.jp
c) E-mail: kawai@lmd.ist.hokudai.ac.jp
d) E-mail: ogawa@lmd.ist.hokudai.ac.jp
e) E-mail: mhaseyama@lmd.ist.hokudai.ac.jp

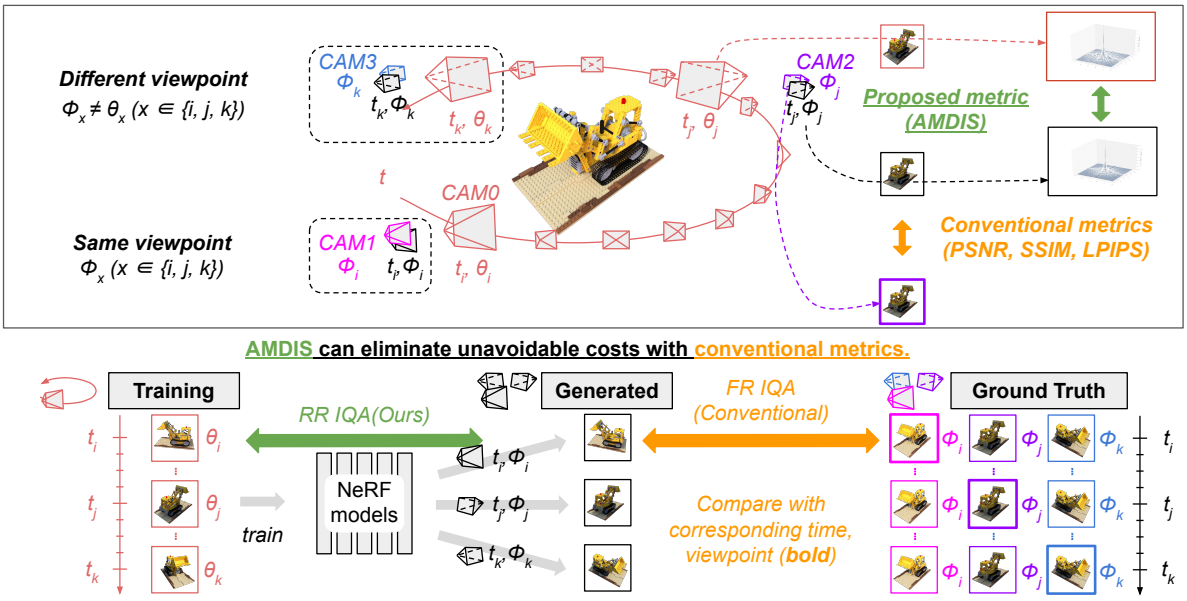


Fig. 1 Difference between the proposed amplitude-dissimilarity (AMDIS) and conventional metrics. The target scene is captured using CAM0 from viewpoints θ_i to θ_k at time t_i to t_k (red). NeRF learns from the captured scene to synthesize novel viewpoints ϕ_i to ϕ_k (near viewpoints θ_i to θ_k). In the conventional FR IQA (orange), the synthesized images are evaluated based on the novel viewpoint images captured from corresponding viewpoint cameras (pink, purple, and blue). This means that multiple synchronized cameras (CAM1, CAM2, and CAM3) are required for conventional evaluation. Unlike conventional metrics, the proposed RR IQA "AMDIS" (green) is designed to handle the Fourier amplitude components of images without requiring additional cameras for evaluation.

images. AMDIS can be used to evaluate various dynamic scenes without requiring multiple ground truth viewpoints.

The contributions of this paper are summarized as follows.

- We propose a novel RR IQA metric, AMDIS, for evaluating the performance of NeRF models in dynamic scenes. The proposed AMDIS can accurately measure the quality of the novel viewpoint synthesized images without requiring the ground truth viewpoint images.
- The experimental results demonstrate that the proposed AMDIS is strongly correlated with FR IQA metrics using ground truth viewpoint images.

2. Related Work

2.1 NeRF for Monocular Dynamic Scene and Its Evaluation

NeRF aims to synthesize photorealistic 3D scenes from a sparse set of 2D images. Refer to Appendix for a basic explanation of NeRF. Conventional NeRF methods have demonstrated remarkable results in generating high-quality 3D scenes; these approaches have been improved by focusing on rendering procedure [2], [3], computational efficiency [13], [14], editability [15]–[17], and application to dynamic scene [4], [18]–[21]. In the NeRF field, significant efforts have been made to enhance rendering quality. However, methodologies for accurately evaluating such quality, especially in dynamic scenes, remain notably underdeveloped. To address this issue, this study focuses on developing robust evaluation metrics specifically tailored to dynamic scene environments.

In static scenes, training and ground truth viewpoint images can be captured using a single camera, as the target scene remains unchanged over time. However, dynamic scenes, the position and shape of objects can change over time. Therefore, it is necessary to use multiple synchronized cameras to capture ground truth viewpoint images. This requirement significantly makes it difficult to evaluate NeRF in dynamic scenes. Chen et al. [18] evaluated the performance of NeRF models using dynamic scenes captured from multiple cameras at various viewpoints. In contrast, D-NeRF [4] employed dynamic 3D graphic data rather than real-world scenes as a dataset, thereby allowing for preparation of desired ground truth viewpoint images to evaluate the learned NeRF model. Although these metrics address the evaluation of dynamic scenes, they do not resolve the challenge of preparing large datasets with real dynamic scenes. In contrast, our approach focuses on evaluating NeRF models using real dynamic scenes captured without requiring multiple synchronized cameras from different viewpoints. Our metric is designed to enable the preliminary evaluation of various dynamic scenes without multiple viewpoints, such as monocular dynamic scene images.

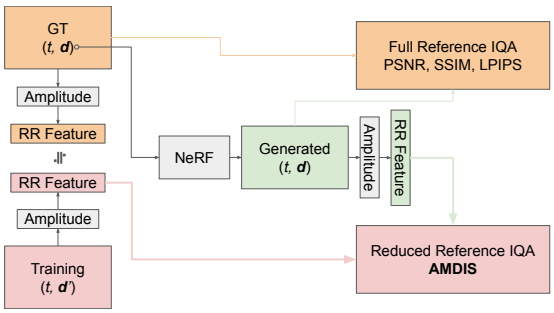


Fig. 2 Overview of conventional and proposed evaluation metric. The ground truth viewpoint image and synthesized viewpoint image have the same viewpoint **d** at the same time t. However, the training viewpoint image is a different viewpoint **d′**.

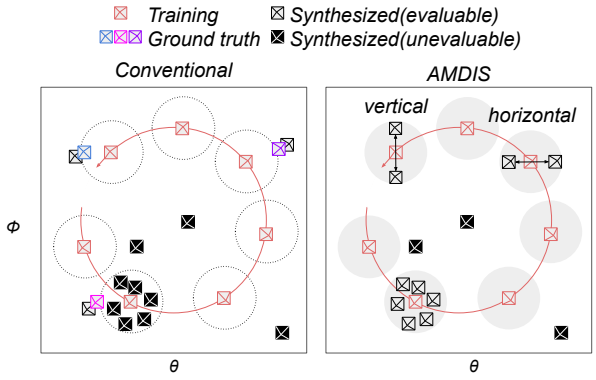


Fig. 3 Evaluable range difference between conventional metrics and proposed AMDIS metric. We consider the viewpoint range of the NeRF models as the polar coordinates θ and ϕ at a certain radius. Conventional metrics can only evaluate the viewpoints (red, pink, and purple) of cameras prepared in advance. In contrast, the proposed AMDIS can evaluate any viewpoints in the gray circle that are shifted horizontally or vertically the training viewpoint images (red).

2.2 Image Quality Assessment

Conventionally, the evaluation NeRF models is based the concept of IQA. IQA can be categorized into three types: full-reference (FR) [7]–[9], [22], non-reference (NR) [23]–[25], and reduced-reference (RR) [26]–[28]. FR IQA evaluates the quality of the synthesized images by referring the corresponding ground truth viewpoint images. For example, structural similarity (SSIM) [7] calculates similarity in terms of pixel, contrast, and structure from synthesized and ground truth viewpoint images. In contrast, NR IQA evaluates the quality of the synthesized images using only the information extracted from the images. Although NR IQA can calculate evaluation scores from only synthesized images, it cannot evaluate whether the NeRF model learns the real scene because it does not compare novel viewpoint images with the ground truth viewpoint images.

Unlike FR and NR IQAs, RR IQA evaluates the quality of the synthesized images using the partial elements of the ground truth viewpoint images. In this study, we focus on RR IQA metrics. By assuming that differences between the training and synthesized viewpoint images are primarily captured by the phase component in the Fourier domain, we propose a viewpoint-independent RR IQA metric that focuses on the amplitude component in the Fourier domain. The proposed metric evaluates NeRF models without requiring ground truth viewpoint images. Further details are presented in the next section.

3. Proposed Metric

The proposed AMDIS metrics evaluate the quality of novel viewpoint images synthesized using NeRF models without requiring ground truth viewpoint images. In NeRF models, novel viewpoint images are synthesized by vertically and horizontally shifting from the training viewpoint image on the spherical surface. Therefore, if we can ignore the components affected by the horizontal and vertical movements, the preliminary assessment of NeRF models without preparing ground truth viewpoint images becomes feasible. Motivated by this assumption, as shown in Fig. 2, AMDIS compares the amplitude component in the Fourier domain between different viewpoint images.

Conventional metrics can only evaluate the synthesized viewpoint images with the corresponding ground truth viewpoint images. In contrast, AMDIS can evaluate the synthesized viewpoint images from the viewpoint near the training viewpoint images as shown in Fig. 3.

Let X and Z be the training and synthesized images from viewpoints, respectively. Here, an image $\mathrm{F} \in \{\mathrm{X}, \mathrm{Z}\}$ has the size of width M and height N, and the pixel of the (m, n) element is denoted as $\mathrm{F}_{m,n}$. Let the horizontal and vertical spatial frequencies as k and l, respectively. The Fourier coefficients $\hat{\mathrm{F}}_{k,l}$ can be calculated using the discrete Fourier transform from $\mathrm{F}_{m,n}$. Additionally, the amplitude component $\bar{\mathrm{F}}_{k,l}$ of $\hat{\mathrm{F}}_{k,l}$ can be calculated as follows:

$$\bar{\mathrm{F}}_{k,l} = \sqrt{\Re(\hat{\mathrm{F}}_{k,l})^2 + \Im(\hat{\mathrm{F}}_{k,l})^2}, \tag{1}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ donote the real and imaginary parts of the Fourier component, respectively. Here, since the amplitude component of each image is not affected by the vertical and horizontal movements, the proposed AMDIS calculates the evaluation scores by comparing the amplitude component of the training viewpoint image X and the synthesized viewpoint image Z as follows:

$$\mathrm{AMDIS} = \frac{1}{MN} \|\bar{\mathrm{X}}_{k,l} - \bar{\mathrm{Z}}_{k,l}\|_F^2, \tag{2}$$

where $\|\cdot\|_F^2$ represents the Frobenius norm. By using AMDIS in Eq. (2), the RR evaluation in NeRF becomes feasible.

The advantages of AMDIS lie in its simplicity and interpretability. This system represents the first attempt to propose an RR IQA metric specifically tailored to the properties of NeRF, based on the robust and well-known Fourier shift theorem. In this study, we aim to develop a comprehensive framework for more accurate and diverse evaluations of

3D scenes by combining the proposed AMDIS with existing metrics. This approach is expected to enable more detailed and multifaceted assessments of 3D scenes.

4. Performance Evaluation in Static Scene (Preliminary)

The aim of this study is to evaluate the effectiveness of the proposed AMDIS. We begin by testing its performance on static scenes. Specifically, we evaluated the translational component absorption associated with viewpoint changes in 3D objects by establishing specific reference viewpoints for each object. We employed the widely used NeRF Realistic Compositing Dataset [3] to examine realistic and detailed 3D content. The dataset comprises eight distinct Blender collections: Lego, ship, hotdog, material, ficus, chair, drums, and microphone, as shown in Fig. 4. The reference viewpoints were selected from the learning viewpoint images, and the line of sight was incrementally adjusted by 1° at a time to generate 360 rendering images from a complete 360° rotation. We used 13 different angular differences [1, 2, 3, 4, 5, 10, 15, 30, 45, 60, 90, 120, 180] for evaluation, allowing us to measure the sensitivity of the model to changes in viewpoint through comparisons between image pairs at these varying angles. For a chosen angular difference $\Delta\theta$ (e.g., 45°), the corresponding image pairs (e.g., (I_0, I_{45}), (I_1, I_{46}), ..., (I_{359}, I_{44})) were generated, and consistency evaluations were conducted across these 360 pairs. We used different evaluation metrics, including the peak signal-to-noise-ratio (PSNR), SSIM, learned perceptual image patch similarity (LPIPS), and our proposed AMDIS, to quantitatively measure visual similarity. The average SSIM scores of the image pairs were treated as the comprehensive evaluation score for each angular difference, and the results were used to analyze the viewpoint dependency of the objects.

The above mentioned performance in the static scene results are shown in Fig. 5. Note that the PSNR and SSIM scores are multiplied by a negative number (-) for an easy comparison. The figure shows that (-)PSNR, (-)SSIM, and AMDIS have lower scores than the general score, confirming that the proposed metric calculates small errors by absorbing errors in the phase component. For the change in the evaluation scores, we found that the scores increases as the distance between viewpoints increases across all metrics. Additionally, the variance of the proposed metric relative to the target scene is smaller than that of the other evaluation metrics, indicating that specific viewpoint differences (STEPs) and evaluation scores remain consistent regardless of the target scene.

5. Application to NeRF Synthesized Images in Dynamic Scene

5.1 Experimental Settings

Following the conventional NeRF models, four scenes (Balloon2, Jumping, Playground, and Skating) from Dynamic Scene Dataset [29] were used as the dataset. Each scene contains frames at 12 different time steps $\{t_1, t_2, \ldots, t_{12}\}$ from 12 viewpoints $\{\theta_1, \theta_2, \ldots, \theta_{12}\}$. The number of frames in each scene was 144. In this experiment, six models (NeRF [1], NeRF [1] + time, Yoon [30], NSFF [31], NR [32], and DynamicNeRF [18]) were used to evaluate the proposed AMDIS using dynamic scenes. We implemented each method using the open-source code provided by each author. Each method was trained from the 12 different viewpoints and time steps, namely, monocular dynamic scene images. The viewpoints that were not included in the training sets were used as evaluation set.

Fig. 4 Examples of the realistic synthetic dataset used in NeRF [3].

The effectiveness of the proposed AMDIS was confirmed by validating whether AMDIS has a sufficient correlation with conventional FR IQA metrics (hereinafter, referred to as ground truth metrics). The proposed AMDIS used only the training and synthesized viewpoint images from different viewpoints (RR setting), whereas the ground truth metrics used the synthesized and corresponding ground truth viewpoint images from the same viewpoints (FR setting). If the proposed AMDIS has a correlation with FR IQA metrics, it can be used as a preliminary metrics for NeRF models. For the ground truth metrics, we used PSNR, SSIM [7], and LPIPS [9] following the conventional NeRF evaluation [1]. Based on this approach, we confirmed the effectiveness of the proposed AMDIS by evaluating the correlation between AMDIS in the RR setting and PSNR, SSIM, and LPIPS in the FR setting.

5.2 Experimental Results

Tables 1, 2, and 3 show the experimental results for the six NeRF-based models obtained using PSNR, SSIM, and LPIPS as the ground truth metrics, respectively. And the absolute scores of correlation coefficient for each metric in the RR and FR settings are calculated. As shown in each result, AMDIS outperformed the other metrics in 38 out of 72 cases (53%). For each scene, AMDIS correlated more strongly with the FR metrics than the other RR metrics. These results confirmed that the proposed AMDIS is an effective evaluation metric for NeRF models that does not depend on the target scenes.

Next, Table 4 shows the absolute correlation coefficients

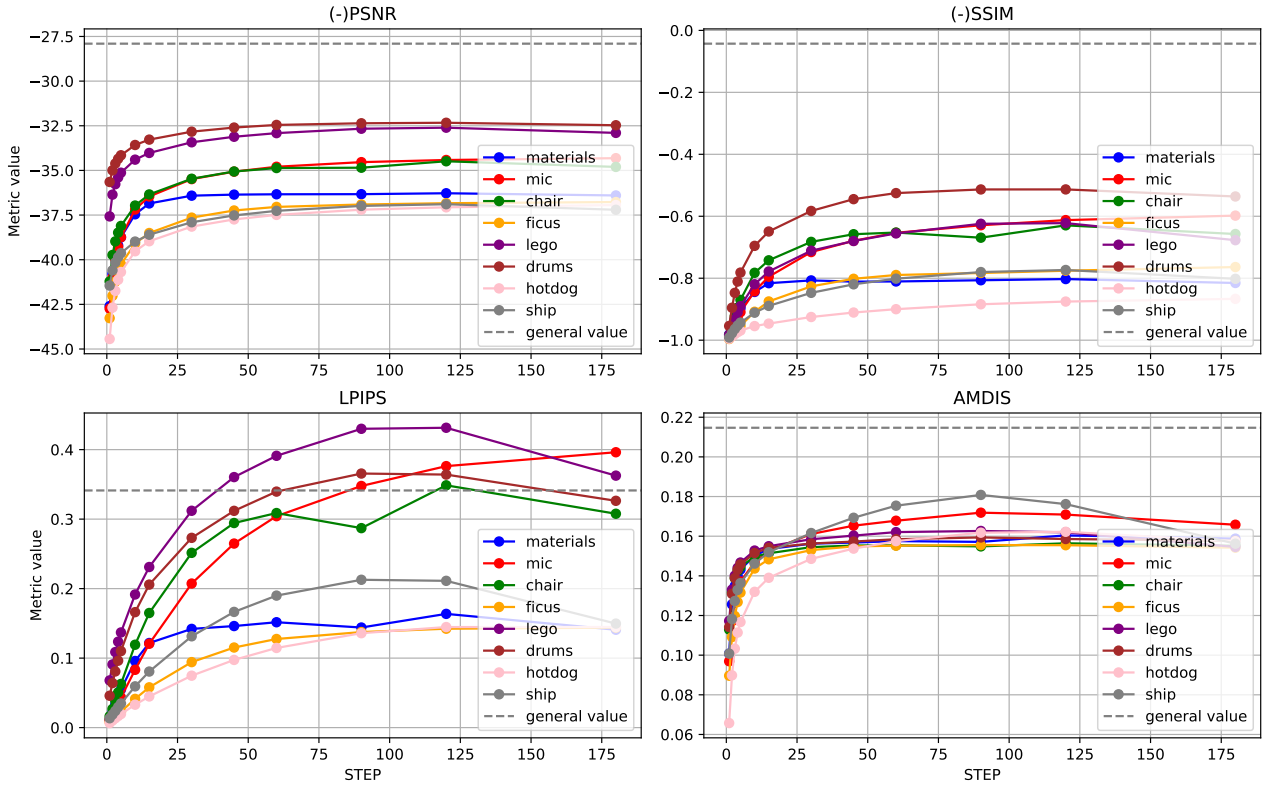


Fig. 5 Viewpoint-to-viewpoint evaluation scores of each evaluation metric. For the realistic synthetic dataset, a total of 13 evaluation scores were calculated using PSNR, SSIM, LPIPS, and the proposed AMDIS while shifting two different images at a STEP by 1° to obtain the average score. The PSNR and SSIM graphs are multiplied by a negative number (-) for an easy comparison. The general score is the average of 10,000 evaluations of two randomly selected images from the NeRF Realistic Compositing Dataset [3].

Table 1 Absolute score of the correlation coefficient between the metrics of RR settings and the PSNR of FR settings. The bold and underlined marks present the best and the second performances, respectively.

	NeRF [1]				NeRF [1]+time				Yoon [30]			
METRIC	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating
PSNR	0.503	0.861	0.498	0.313	**0.936**	0.694	0.016	**0.599**	0.583	0.874	0.215	**0.147**
SSIM [7]	0.008	0.621	0.579	0.052	0.523	0.498	0.006	0.123	0.310	0.644	0.113	0.145
LPIPS [9]	0.503	0.865	0.360	**0.479**	0.919	**0.966**	0.435	0.119	0.455	**0.941**	0.470	0.121
AMDIS	**0.696**	**0.968**	**0.662**	0.045	0.851	0.755	**0.753**	0.365	**0.776**	0.900	**0.779**	0.031
	NR [32]				NSFF [31]				DynamicNeRF [18]			
METRIC	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating
PSNR	0.356	0.681	0.141	0.235	**0.767**	0.877	0.129	0.397	0.863	0.857	0.245	0.420
SSIM [7]	0.084	0.278	0.164	0.168	0.118	0.582	0.031	0.316	0.513	0.605	0.084	0.139
LPIPS [9]	0.421	**0.851**	0.520	**0.518**	0.613	**0.939**	0.403	0.684	0.788	**0.949**	0.577	0.552
AMDIS	**0.638**	0.422	**0.529**	0.344	0.688	0.919	**0.651**	**0.742**	**0.897**	0.934	**0.710**	**0.651**

of the metric in the RR and FR settings. Instead of calculating the correlation coefficients every 12 images as shown in Tables 1, 2, and 3, we calculate them every 288 images in Table 4. This approach helps determine whether they are valid for not only one part of the scene or model but when considering them as a whole. The proposed AMDIS demonstrated the strongest correlations with two of the three FR ground truth metrics, indicating its capability to evaluate NeRF models comparably to FR metrics. These results confirm the effectiveness of the proposed AMDIS evaluation metrics regardless of the scene, NeRF model, or ground truth metric.

5.3 Discussion

The proposed AMDIS assumes that the difference between the two viewpoints in a monocular moving image consists of parallel shifts along the x- and the y-axis, and that the phase

Table 2 Absolute score of the correlation coefficient between the metrics of the RR settings and the SSIM of the FR settings. The bold and underlined marks represent the best and second performances, respectively.

	NeRF [1]				NeRF [1]+time				Yoon [30]			
METRIC	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating
PSNR	0.518	0.852	0.421	0.338	0.890	0.710	0.007	0.342	0.555	0.858	0.014	0.429
SSIM [7]	0.003	0.612	0.531	0.397	0.604	0.637	0.027	0.288	0.410	0.646	0.068	0.052
LPIPS [9]	0.510	0.858	0.248	**0.724**	**0.933**	**0.920**	0.465	0.182	0.478	**0.896**	0.251	0.330
AMDIS	**0.774**	**0.969**	**0.694**	0.430	0.821	0.800	**0.802**	**0.466**	**0.828**	0.895	**0.805**	**0.453**
	NR [32]				NSFF [31]				DynamicNeRF [18]			
METRIC	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating
PSNR	0.356	0.875	0.054	0.452	**0.866**	0.892	0.057	0.456	0.863	0.805	0.042	0.493
SSIM [7]	0.068	0.595	0.008	0.284	0.296	0.629	0.196	0.349	0.605	0.552	0.102	0.097
LPIPS [9]	0.417	**0.906**	0.265	**0.749**	0.772	**0.931**	0.192	0.775	0.811	**0.931**	0.302	0.606
AMDIS	**0.698**	0.733	**0.599**	0.496	0.796	0.928	**0.675**	**0.796**	**0.914**	0.898	**0.796**	**0.720**

Table 3 Absolute score of the correlation coefficient between the metrics of the RR settings and the LPIPS of the FR settings. The bold and underlined marks represent the best and second performances, respectively.

	NeRF [1]				NeRF [1]+time				Yoon [30]			
METRIC	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating
PSNR	0.278	0.808	0.646	0.474	0.860	0.263	0.064	0.041	0.469	0.722	0.130	0.075
SSIM [7]	0.217	0.652	**0.731**	0.292	0.642	0.145	0.087	0.369	0.660	0.517	0.197	**0.283**
LPIPS [9]	0.273	0.771	0.529	**0.769**	**0.902**	**0.925**	0.520	**0.931**	0.589	**0.828**	0.143	0.158
AMDIS	**0.515**	**0.969**	0.547	0.514	0.821	0.408	**0.808**	0.738	**0.673**	0.763	**0.821**	0.063
	NR [32]				NSFF [31]				DynamicNeRF [18]			
METRIC	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating	Balloon2	Jumping	Playground	Skating
PSNR	0.119	0.416	0.118	0.226	0.870	0.797	0.328	0.342	0.828	0.685	0.494	0.689
SSIM [7]	0.328	0.032	0.164	0.390	0.734	0.517	0.144	0.413	0.713	0.438	0.350	0.175
LPIPS [9]	0.194	**0.669**	**0.530**	**0.590**	**0.954**	**0.936**	0.548	0.593	0.753	**0.895**	**0.739**	0.700
AMDIS	**0.438**	0.112	0.466	0.524	0.831	0.820	**0.655**	**0.675**	**0.969**	0.811	0.690	**0.735**

Table 4 Average score of the correlation coefficient between the metrics of the RR setting and FR setting.

	PSNR (FR)	SSIM (FR) [7]	LPIPS (FR) [9]
PSNR (RR)	0.448	0.333	0.102
SSIM (RR) [7]	0.379	0.318	0.146
LPIPS (RR) [9]	0.356	0.279	**0.371**
AMDIS	**0.535**	**0.421**	0.117

component absorbs the discrepancies caused by these shifts.

This study has several limitations. First, in more realistic settings, differences between viewpoints may not only result from parallel shift but also involve rotations and changes in proximity, which can significantly affect the amplitude component. Secondly, it is also be necessary to examine the differences with semantic information-based evaluation metrics that have been proposed in recent years. In this study, we introduced a new metric from a signal processing perspective. However, other methods have developed to evaluate novel viewpoints based on the preservation of semantic information [33], [34]. For example, contrastive language-image pre-training similarity [35], which compares the embedded multimodal semantic representations from deep learning models between the novel-view and corresponding reference images, is often used for the evaluation. Additionally, some studies have evaluated novel viewpoints using large language models [36], [37]. Although the proposed AMDIS metric offers distinct advantages, it is believed that it would be beneficial to use it in conjunction with these other evaluation methods. Finally, as the next step of our study, it is essential to evaluate Gaussian splatting (GS)-based methods [38], [39]. Recent reports suggest that 3DGS methods outperform NeRF-based methods in many aspects, contributing to the rapid advancements in the field of 3D reconstruction. Therefore, it is necessary to continue assessing whether these new reconstruction methods can be appropriately evaluated.

6. Conclusion

In this paper, we proposed a novel RR IQA metric for NeRF, AMDIS. The proposed AMDIS reduces the need for ground truth viewpoint images captured by multiple synchronized cameras. Additionally, we found that the corresponding evaluation score remained stable when calculated for viewpoint angles with different degrees. In future work, we plan to extend the proposed metric to account for not only simple parallel shifts but also variations in camera movement angles and zoom levels.

Appendix

NeRFs are defined as multilayer perceptron (MLP) F_Θ, which takes the 3D position $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \phi)$ as inputs and outputs color $\mathbf{c} = (r, g, b)$ and density σ. The density depends only on the 3D position $\mathbf{x}$, whereas the color $\mathbf{c}(\mathbf{x}, \mathbf{d})$ depends on the 3D position $\mathbf{x}$ and viewing direction $\mathbf{d}$, enabling highly accurate representation in response to changes in the viewing direction. In volume rendering with radiance fields, the expected color $C(\mathbf{r})$ at

a given camera ray $\mathbf{r}(\tau) = \mathbf{o} + \tau\mathbf{d}$ in the near plane τ_n and the far plane τ_f is calculated. The implementation samples N points from the integral interval $[\tau_n, \tau_f]$ and processes them discretely. The sampling is not at regular intervals, but rather random sampling based on a uniform distribution $\mathcal{U}$ from the range defined below to determine the sampling point t_i. By outputting and integrating the density σ_i and the color $\mathbf{c}_i$ for the set of N, the color $\hat{C}(\mathbf{r})$ of a single pixel is calculated. Although the above mentioned implementation theoretically allows rendering, it is inefficient because it not only blurs high-frequency components but also samples a fixed number of objects along each camera ray, regardless of the coarseness or fineness of the objects in the camera ray. To address these issues, Mildenhall et al. [3] proposed the use of positional encoding and hierarchical volume sampling. Hierarchical volume sampling considers the density distribution in the 3D scene, allowing it to skip sampling sparse regions, thereby improving efficiency. Instead of using a single neural field to represent the 3D scene, two neural fields with the same structure are prepared: corse and fine networks. This dual-network approach strategically determines the sampling targets. For evaluating one pixel (one ray) of the rendered image, first skip the ray to the coarse network and sample N_c points to output the color and density. These outputs then serve as cues to identify regions of higher density (i.e., areas with more information) along the same ray in the fine network and sample N_f points. The coarse network is used to achieve this. The evaluation (discrete integration) of one ray in the coarse network to achieve this is defined by the following formula rewritten as a weighted sum of the colors c_i by the weights ω_i.

$$\hat{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} \omega_i c_i, \tag{3}$$

$$\omega_i = T_i(1 - \exp(-\sigma_i \delta_i)). \tag{4}$$

By normalizing the weights ω_i with $\hat{\omega} = \frac{\omega_i}{\sum_{i=j}^{N_c} \omega_j}$, we obtain a probability density function that is constant for each segment along the ray. The distribution obtained from the coarse network generates N_f sampling points, and the final pixel color is obtained by evaluating the fine network using N_c+N_f sampling points combined with N_c sampling points. This allows for efficient rendering according to the density of the target.

The aim of NeRF training is to optimize the MLP output color and density, obtaining the parameter Θ that minimizes the difference between the training data image and the image rendered from the same viewpoint. The loss function for this optimization is given by

$$\mathcal{L} = \sum_{\mathbf{r}\in\mathcal{R}} [\|\hat{C}_c(\mathbf{r}) - C(\mathbf{r})\|_2^2 + \|\hat{C}_f(\mathbf{r}) - C(\mathbf{r})\|_2^2], \tag{5}$$

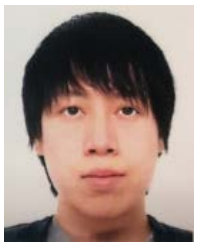
where $\mathcal{R}$ is set of rays, $\hat{C}_c(\mathbf{r})$ is coarse network pixel color, $\hat{C}_f(\mathbf{r})$ is fine network pixel color, and $C(\mathbf{r})$ is training data pixel color.

References

[1] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," Proceedings of the European Conference on Computer Vision, pp.405–421, 2020.
[2] L. Wu, J.Y. Lee, A. Bhattad, Y.X. Wang, and D. Forsyth, "Diver: Real-time and accurate neural radiance fields with deterministic integration for volume rendering," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16200–16209, 2022.
[3] B. Mildenhall, P. Hedman, R. Martin-Brualla, P.P. Srinivasan, and J.T. Barron, "NeRF in the dark: High dynamic range view synthesis from noisy raw images," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16190–16199, 2022.
[4] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural radiance fields for dynamic scenes," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10318–10327, 2021.
[5] Y. Era, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "Content-based image retrieval using effective synthesized images from different camera views via pixelnerf," IEEE 11th Global Conference on Consumer Electronics (GCCE), pp.404–405, IEEE, 2022.
[6] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, "EfficientNeRF efficient neural radiance fields," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12902–12911, 2022.
[7] 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, vol.13, no.4, pp.600–612, 2004.
[8] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, pp.1398–1402, 2003.
[9] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.586–595, 2018.
[10] L. Wang, "A survey on iqa," arXiv preprint arXiv:2109.00347, 2021.
[11] T. Huang, J. Burnett, and A. Deczky, "The importance of phase in image processing filters," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.23, no.6, pp.529–542, 1975.
[12] A.V. Oppenheim and J.S. Lim, "The importance of phase in signals," Proceedings of the IEEE, vol.69, no.5, pp.529–541, 1981.
[13] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, "EfficientNeRF efficient neural radiance fields," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12902–12911, 2022.
[14] L. Wang, J. Zhang, X. Liu, F. Zhao, Y. Zhang, Y. Zhang, M. Wu, J. Yu, and L. Xu, "Fourier plenoctrees for dynamic radiance field rendering in real-time," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.13524–13534, 2022.
[15] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, "Clip-NeRF: Text-and-image driven manipulation of neural radiance fields," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3835–3844, 2022.
[16] K. Kania, K.M. Yi, M. Kowalski, T. Trzciński, and A. Tagliasacchi, "CoNeRF: Controllable neural radiance fields," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18623–18632, 2022.
[17] Y.J. Yuan, Y.T. Sun, Y.K. Lai, Y. Ma, R. Jia, and L. Gao, "NeRF-editing: Geometry editing of neural radiance fields," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18353–18364, 2022.
[18] C. Gao, A. Saraf, J. Kopf, and J.B. Huang, "Dynamic view synthesis from dynamic monocular video," Proceedings of the IEEE/CVF

International Conference on Computer Vision, pp.5712–5721, 2021.
[19] K. Park, U. Sinha, J.T. Barron, S. Bouaziz, D.B. Goldman, S.M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.5865–5874, 2021.
[20] K. Park, U. Sinha, P. Hedman, J.T. Barron, S. Bouaziz, D.B. Goldman, R. Martin-Brualla, and S.M. Seitz, "HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields," arXiv preprint arXiv:2106.13228, 2021.
[21] M. Kawai, R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, "Free-viewpoint sports video generation based on dynamic NeRF considering time series," Proceedings of the 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), pp.408–409, 2022.
[22] A.M. Demirtas, A.R. Reibman, and H. Jafarkhani, "Full-reference quality estimation for images with different spatial resolutions," IEEE Transactions on Image Processing, vol.23, no.5, pp.2069–2080, 2014.
[23] Z. Wang, A.C. Bovik, and B.L. Evan, "Blind measurement of blocking artifacts in images," Proceedings of the International Conference on Image Processing, pp.981–984, 2000.
[24] C. Liu, W.T. Freeman, R. Szeliski, and S.B. Kang, "Noise estimation from a single image," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.901–908, 2006.
[25] A. Mittal, A.K. Moorthy, and A.C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Transactions on Image Processing, vol.21, no.12, pp.4695–4708, 2012.
[26] J.A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color distribution information for the reduced-reference assessment of perceived image quality," IEEE Transactions on Circuits and Systems for Video Technology, vol.20, no.12, pp.1757–1769, 2010.
[27] X. Gao, W. Lu, X. Li, and D. Tao, "Wavelet-based contourlet in quality evaluation of digital images," Neurocomputing, vol.72, no.1-3, pp.378–385, 2008.
[28] L. Ma, S. Li, F. Zhang, and K.N. Ngan, "Reduced-reference image quality assessment using reorganized dct-based image representation," IEEE Transactions on Multimedia, vol.13, no.4, pp.824–829, 2011.
[29] J.S. Yoon, K. Kim, O. Gallo, H.S. Park, and J. Kautz, "Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5336–5345, 2020.
[30] J.S. Yoon, K. Kim, O. Gallo, H.S. Park, and J. Kautz, "Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5336–5345, 2020.
[31] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6498–6508, 2021.
[32] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.12959–12970, 2021.
[33] A. Jain, M. Tancik, and P. Abbeel, "Putting nerf on a diet: Semantically consistent few-shot view synthesis," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.5885–5894, 2021.
[34] Y. Uchida, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "An evaluation metric for single image-to-3D models based on object detection perspective," SIGGRAPH Asia 2024 Technical Communications, pp.1–4, 2024.
[35] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," Proceedings of the International Conference on Machine Learning (ICML), pp.8748–8763, 2021.
[36] Y. Hirakawa, T. Wada, K. Morishita, and R. Shimizu, "An empirical analysis of gpt-4v's performance on fashion aesthetic evaluation," SIGGRAPH Asia 2024 Technical Communications, pp.1–4, 2024.
[37] D. Haraguchi, N. Inoue, W. Shimoda, H. Mitani, S. Uchida, and K. Yamaguchi, "Can GPTs evaluate graphic design based on design principles?," SIGGRAPH Asia 2024 Technical Communications, pp.1–4, 2024.
[38] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," ACM Transactions on Graphics, vol.42, no.4, pp.1–14, 2023.
[39] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3D content creation," Proceedings of the International Conference on Learning Representations (ICLR), pp.1–18, 2024.

Ren Togo received the B.S. degree in Health Sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is also a Radiological Technologist. He is currently a Specially Appointed Assistant Professor with the Laboratory of Media Dynamics, Faculty of Information Science and Technology, Hokkaido University. His research interests include machine learning and its applications. He is a member of the IEEE, ACM, AAAI, and IEICE.



Rintaro Yanagi received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2019, and the M.S. degree from the Graduate School of Information Science and Technology, Hokkaido University, in 2021, and Ph.D. degree the Graduate School of Information Science and Technology, Hokkaido University, in 2024. He is a researcher at National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interest includes machine learning and its applications. He is a member of ACM.

Masato Kawai received the M.S. degree from the Graduate School of Information Science and Technology, Hokkaido University, in 2024. His research interests include sports analytics and 3D analysis.

Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He joined Graduate School of Information Science and Technology, Hokkaido University in 2008. He is currently a professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests are AI, IoT and big data analysis for multimedia signal processing and its applications. He was a special session chair of IEEE ISCE2009, a Doctoral Symposium Chair of ACM ICMR2018, an organized session chair of IEEE GCCE2017-2019, a TPC Vice Chair of IEEE GCCE2018, a Conference Chair of IEEE GCCE2019, etc. He has been also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE and ITE.

Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 1995 to 1996. She is currently a professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of ITE Transactions on Media Technology and Applications, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE). She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).