CrossScore: Towards Multi-View Image Evaluation and Scoring

Zirui Wang Wenjing Bian Victor Adrian Prisacariu

University of Oxford
{ryan, wenjing, victor}@robots.ox.ac.uk
https://crossscore.active.vision

Fig. 1: *We propose a novel cross-reference (CR) image quality assessment (IQA) scheme, which evaluates a query image using multiple unregistered reference images that are captured from different viewpoints. This approach sets a new research trajectory apart from conventional IQA schemes such as full-reference (FR), general-reference (GR), no-reference (NR), and multi-modal-reference (MMR).*

Abstract. We introduce a novel *cross-reference* image quality assessment method that effectively fills the gap in the image assessment landscape, complementing the array of established evaluation schemes – ranging from *full-reference* metrics like SSIM [59], *no-reference* metrics such as NIQE [32], to *general-reference* metrics including FID [17], and *Multi-modal-reference* metrics, *e.g.* CLIPScore [16]. Utilising a neural network with the cross-attention mechanism and a unique data collection pipeline

from NVS optimisation, our method enables accurate image quality assessment without requiring ground truth references. By comparing a query image against multiple views of the same scene, our method addresses the limitations of existing metrics in novel view synthesis (NVS) and similar tasks where direct reference images are unavailable. Experimental results show that our method is closely correlated to the full-reference metric SSIM, while not requiring ground truth references.

Keywords: Cross-Reference Image Assessment · Novel View Synthesis

1 Introduction

Accurate image quality evaluation is critical for enhancing the performance of computer vision tasks, including image processing, image generation, and novel view synthesis (NVS).

Image quality assessment (IQA) methods can be categorised based on the type of referencing. The most popular group, *full-reference* (FR) metric, such as PSNR, SSIM [59], LPIPS [70], evaluates the differences between a query image and a reference image in terms of pixels and perceptual quality. These metrics are essential for tasks such as super-resolution, denoising, and compression, and, importantly, assume the availability of an oracle reference image.

Ground truth images, however, may not always be available, for instance, in image generation tasks. Consequently, multiple attempts have been made to alleviate the dependency on ground truth oracles. For instance, FID [17] is a *general-reference* (GR) scheme that evaluates the discrepancy of data distribution between image sets. Alternatively, *multi-modal-reference* (MMR) approaches, such as CLIPScore [16], examine the image-text similarity, whereas *no-reference* (NR) metrics, such as NIQE [32] and PIQE [55] evaluate single-image statistics without referencing.

Although the aforementioned non-ground-truth metrics are widely employed in tasks like image compression, denoising, and generation, these methods generally rely on high-level statistics and global context, consequently lacking the capacity for detailed analysis. This deficiency renders them inadequate for NVS, which requires pixel-level assessment of novel view images with scene-specific context.

The established approach to assess NVS performance involves selecting a subset of test images from an existing camera trajectory, which cannot be used in training, rendering images using their camera parameters, and computing pixel-level FR scores by comparing these rendered images with the original captured test images. While generally simple and effective, this subsampling approach exhibits two primary issues: 1) balancing the number of images between training and evaluation can affect the statistical relevance of the assessment and the effectiveness of the training; and 2) relying on FR metrics precludes the ability to evaluate renderings using true novel trajectories, as ground truth images are not available for true novel views.

These challenges motivate us to develop a novel IQA scheme, which evaluates the quality of a query image using multiple reference views, each observing the same content but from different viewpoints. The key intuition is to leverage multi-view images as a substitute for a ground truth image, enabling a 'perspective'-version FR evaluation. We term this process as *cross-reference* (CR) evaluation and the resulting score as `CrossScore`. Our method differs from prior works in two essential ways: 1) unlike FR-IQA, our approach eliminates the need for aligned reference images; and 2) in contrast to GR-, NR-, and MMR-IQA, our method offers detailed evaluation via multi-view reasoning.

Specifically, we propose to find a cross-reference function that predicts a full-reference metric, *i.e.* SSIM, of a distorted image, by comparing it with multiple unregistered reference images. We implement this function with a neural network that employs the cross-attention mechanism [54].

Alongside the model formulation, a key additional challenge lies in gathering training samples. Our solution involves rendering images throughout the NVS optimisation process, covering a broad spectrum of distortion varieties and intensities. By comparing these images against their original, undistorted versions, we obtain pixel-level SSIM scores that serve as a training objective. This self-supervised data collection approach allows us to build a rich dataset and enables per-pixel supervision for our network training.

In summary, our contribution is threefold. *First*, we unveil CR-IQA, a new image evaluation regime tailored for multi-view scenarios. *Second*, we actualise this concept through a neural network based on cross-attention mechanisms, enabling detailed per-pixel evaluation in the absence of ground truth images. *Third*, we develop a self-supervised data collection scheme that utilises existing NVS algorithms to produce a wide variety of distorted images along with their SSIM score maps, serving as our training samples. Our findings demonstrate that the `CrossScore` aligns closely with the full-reference SSIM score, while eliminating the need for ground truth reference images.

2 Related Work

This section offers an overview of image quality assessment (IQA) metrics, sorted by reference image availability and nature, and reviews the current evaluation framework for novel view synthesis (NVS).

2.1 Image Quality Assessment Metrics

Full-Reference Metrics Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) [59] are key metrics in the FR-IQA group for comparing a query image with the ground truth due to their simplicity and accuracy. Further, several variants are proposed to improve performance by evaluating at multi-scale [61], utilising handcrafted features [66, 69] and extend to specific applications, such as image stitching [46], and high dynamic range (HDR) images [28]. Recently, deep neural networks

have advanced FR-IQA towards aligning assessments more closely with human visual perception [13,70]. Overall, FR-IQA offers detailed evaluation at the cost of requiring ground truth images.

Reduced-Reference Metrics RR-IQA methods are designed to address situations where only partial information about the original reference image is accessible. Wang *et al*. [60,62] uses generalised Gaussian density model parameters that model natural images in the wavelet transform domain as RR features. Reduced-reference structural similarity index (RR-SSIM) [41] approximates FR-SSIM by using image statistical properties in the divisive normalisation transform domain. Redi *et al*. [40] uses descriptors based on colour correlogram that describes the spatial correlation of colours as RR data. RR metrics are commonly used for reducing transmission costs or accelerating processing, with a trade-off in information. Since the reference features still come from the ground truth image, the RR group shares the same limitation as FR-IQA in many applications.

No-Reference Metrics NR-IQA methods provide an alternative to evaluating image quality based on the input only when ground truth references are unavailable. Classical approaches like DIVINE [33], BRISQUE [31], NIQE [32] PIQE [55] are designed with handcrafted features and natural scene statistic models to capture image distortions and estimate quality. Kang *et al*. [21] first applied CNN to NR-IQA. TRIQ [67] applied a transformer encoder to features extracted by CNN to predict image quality. MUSIQ [23] addressed the CNN size constraint with a patch-based transformer. NR metrics are primarily tailored to measure distortions, including compression artefacts, noise, and blur. However, their ability to provide a comprehensive analysis of image content is limited by the lack of reference images, rendering them less suitable for multi-view scenarios.

General-Reference Metrics To assess the quality of generated images, commonly used metrics include Inception Score (IS) [3,44], FID [17] and KID [7,65]. These metrics evaluate the overall performance of generative models rather than scoring individual images. For instance, FID measures the squared Fréchet distance between the distributions of the reference image set and the generated image set. These metrics focus on global statistics, making them well-suited for assessing image generation models [14,18] while infeasible for novel view synthesis tasks [29,30].

Multi-Modal-Reference Metrics Cross-modal models facilitate the assessment of alignment between images and text, enabling the development of an MMR-IQA scheme. CLIPScore [16] directly applies the CLIP model [39] to the image captioning task by computing the adjusted cosine similarity between the image and candidate text as a score. LIQE [71] employs a multi-task learning model leveraging vision-language correspondences to estimate the quality score,

scene category and distortion type. CLIP-IQA [57] applies CLIP to IQA with simple antonym prompts to access image qualities such as brightness and noisiness. By associating semantics between text and vision, these metrics are commonly used in text-to-image generation and editing [22,43] and image captioning tasks [34], yet they lack the capability for detailed evaluation.

2.2 Image Quality Assessment in NVS Systems

Common evaluation metrics for NVS tasks include Full-Reference (FR) metrics such as PSNR, SSIM [59], and LPIPS [70], which produce detailed similarity assessment between rendered and ground truth images. Enhancements to the evaluation process have been proposed, including introducing explicit representation [2], simplifying evaluation through metric summarisation [4], incorporating additional robustness metrics [56], and benchmarking with more effective camera coverage [11].

As NVS rapidly evolves to address more complex tasks, conventional FR-style evaluations struggle, particularly with novel views lacking ground truth camera data, as seen in tasks like joint camera parameter and NeRF optimisation [6, 9, 20, 25, 37, 52, 63]. Additionally, large-scale scenes and dramatic camera movements, such as in city-scale [42, 50, 64] and egocentric setups [10, 12, 15, 36, 47, 49, 53], render the subsample-then-compare strategy inadequate. These issues underscore the need for a metric better suited to multi-view evaluation while ground truth is not available.

3 Method

Our goal is to evaluate the quality of a query image $\tilde{I}_q$, using a set of reference images $\mathcal{I}_r = \{I_r^i | i = 1...N_{\text{ref}}\}$ that capture the same scene as the query image but from other viewpoints, where i denotes the i^{th} image in a reference set with N_{ref} images. We refer to this reference set as the *cross-reference* (CR) set. From the NVS application perspective, the query image $\tilde{I}_q$ is often a rendered image with artefacts, and the reference set consists of the real captured images.

To achieve this goal, we propose a simple but effective strategy, by finding a function that predicts a well-established FR score, *e.g.* SSIM, for a query image. Unlike the SSIM function, which takes the input of a pre-aligned ground truth image, our new function takes multi-view images as input.

For a query image $\tilde{I}_q \in \mathbb{R}^{H \times W \times 3}$, the SSIM function compares it with its ground truth image I_q in a sliding window fashion, and outputs a score map $\mathbf{S}_{\text{ssim}} \in \mathbb{R}^{H \times W}$:

$$f(\tilde{I}_q, I_q) \mapsto \mathbf{S}_{\text{ssim}}, \tag{1}$$

where $f(\cdot)$ denotes the SSIM function.[1]

[1] The raw SSIM score map shares the same dimensions as a query image, *i.e.* $\mathbf{S}_{\text{ssim}} \in \mathbb{R}^{H \times W \times 3}$. For simplicity, we follow a standard practice that averages the SSIM scores across colour channels, yielding a single-channel score map $\mathbf{S}_{\text{ssim}} \in \mathbb{R}^{H \times W}$.

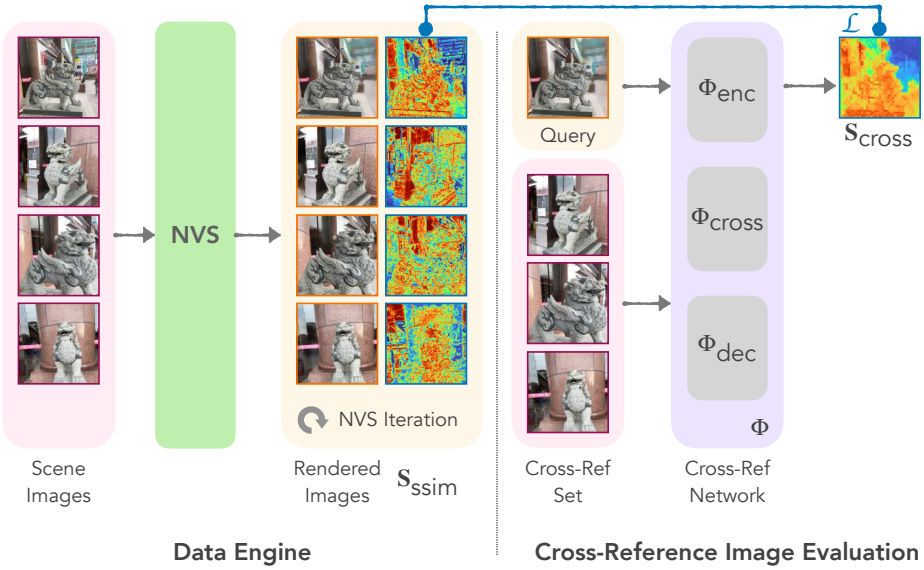

Fig. 2: Data Generation and Training Pipeline. *We employ existing NVS models to generate pairs of rendered images and SSIM maps for training purposes. As the NVS model iterates, rendered images at various optimisation stages are used as the query image for input into our model. Together with a set of reference images from the same scene, our model predicts a score map, supervised by the corresponding SSIM map. More details see Secs. 3.1 and 3.2.*

The aim of our work is to predict a score map $\mathbf{S}_{\text{cross}} \in \mathbb{R}^{H \times W}$, which is highly correlated with $\mathbf{S}_{\text{ssim}}$, by comparing a query image with the CR set $\mathcal{I}_r$, instead of with its fully aligned ground truth image I_q. In other words, we seek a function $g(\cdot)$ that approximates the SSIM function $f(\cdot)$, but making use of the CR set $\mathcal{I}_r$:

$$g(\tilde{I}_q, \mathcal{I}_r) \mapsto \mathbf{S}_{\text{cross}} \approx \mathbf{S}_{\text{ssim}}. \tag{2}$$

The intuition here is to approximate a 'perspective SSIM' function by replacing the ground image with a set of unregistered multi-view images.

We parameterise the cross-reference function $g(\cdot)$ with a neural network Φ. We elaborate on the network design and training strategy in the subsequent sections.

3.1 Network Design

As shown in Fig. 2, our network Φ consists of three parts, i) an image encoder Φ_{enc}, which extracts feature maps from input images; ii) a cross-reference module Φ_{cross}, which associates a query image $\tilde{I}_q$ with images in a CR set $\mathcal{I}_{\text{ref}}$ and produces a latent score map; and iii) a score regression head Φ_{dec} that decodes the latent score map to the final score map $\mathbf{S}_{\text{cross}}$.

Image Encoder Φ_{enc} We adapt a pre-trained DINOv2 [35] network as our image encoder Φ_{enc}, which takes an image I as input and outputs a feature map $\mathbf{F} = \Phi_{\text{enc}}(I)$. This image encoder is applied to all images including query and reference images, and produces feature maps $\mathbf{F}_q$ and $\mathbf{F}_r^i$. We adopt the same patch-wise positional encoding scheme as DINOv2, with each small patch being assigned a positional embedding. Since our cross-reference function takes a set of unordered reference images, image-wise encoding is not applied.

Cross-Reference Module Φ_{cross} We leverage a Transformer Decoder [54] in our cross-reference module Φ_{cross}. Given a feature map of a query image $\mathbf{F}_q$, the cross-reference module Φ_{cross} outputs a latent score map $\mathbf{M} = \Phi_{\text{cross}}(\mathbf{F}_q, \mathcal{F}_r)$, by comparing $\mathbf{F}_q$ with the set of reference feature maps $\mathcal{F}_r = \{\mathbf{F}_r^i | i = 1...N_{\text{ref}}\}$. Specifically, this cross-reference is conducted by the cross-attention mechanism, where the feature map of query image $\mathbf{F}_q$ is the *query* of the cross-attention, and the set of feature maps of the reference images $\mathcal{F}_r$ serve as *key* and *value* in the cross attention.

Score Regression Head Φ_{dec} With a latent score map predicted from the cross-reference module Φ_{cross}, a small regression head Φ_{dec} is applied to finally predict `CrossScore` $\mathbf{S}_{\text{cross}}$. We use a shallow Multi-layer Perceptron (MLP) to interpret a latent score map to a per-pixel score map. Since DINOv2 encodes images by patches, a latent score vector contains the quality estimation for the entire patch. In order to predict `CrossScore` in pixel level, we use the last layer of the MLP to interpret each latent score estimation to a 196-dimension vector, which is then reshaped to a 14×14 patch that corresponds to a small image patch encoded by DINOv2.

3.2 Training Strategy

Self-supervised Training Data Collection We leverage existing NVS systems and abundant multi-view datasets to generate SSIM maps for our training. Specifically, we select Neural Radiance Field (NeRF)-style NVS systems as our data engine. Given a set of images, a NeRF recovers a neural representation of a scene by iteratively reconstructing the given image set with photometric losses. By rendering images with the camera parameters from the original captured image set at multiple training checkpoints, we generate a large number of images that contain various types of artefacts at various levels. From which, we compute SSIM maps $\mathbf{S}_{\text{ssim}}$ between rendered images and corresponding real captured images, which serve as our training objectives.

Supervision This data collection scheme enables a self-supervised training scheme for our network Φ. We consider each rendered image as a query image $\tilde{I}_q$, and we randomly sample real captured images (exclude the real image of

the query) to form our reference set $\mathcal{I}_r$. The SSIM map of the rendered image $\mathbf{S}_{\text{ssim}}$ is then used to supervise our network to predict a $\mathbf{S}_{\text{cross}}$ with an L_1 loss:

$$\mathcal{L} = |\mathbf{S}_{\text{ssim}} - \mathbf{S}_{\text{cross}}|. \tag{3}$$

4 Experiments

We start this section by outlining our experimental setup in Sec. 4.1, followed by assessing the correlation between our score and SSIM through both qualitative and quantitative analyses in Sec. 4.2. We then demonstrate the application of our CR-IQA in two scenarios: benchmarking unseen NeRF algorithms (Sec. 4.3) and evaluating images rendered from novel trajectories in NVS without ground truth (Sec. 4.4). Additionally, we examine the effectiveness of our cross-reference module via a visualisation of its attention maps in Sec. 4.5 and an ablation study in Sec. 4.6. Lastly, Sec. 4.7 concludes our experiments with a discussion on limitations and future research directions.

4.1 Experimental Setup

Dataset We utilise three datasets in our primary experiments. First, the Map-free Relocalisation (MFR) [1] dataset, initially designed for camera parameter estimation benchmarks, has been adapted for our data collection and network training. This dataset features 460 outdoor videos of objects and buildings in a resolution of 540×960. Second, we utilise the Mip360 [5] dataset, consisting of 9 videos that capture 360-degree scans of diverse scenes, both outdoor and indoor. The original resolution of the images is ∼4K. To facilitate DINOv2 [35] image encoding, we downscale all images by a factor of 4. Third, we randomly select 10 videos from the RealEstate10K (RE10K) [72] dataset, originally in 1920×1080 resolution, which are downscaled to 960×540. Training and evaluation: Our network is solely trained on the MFR dataset, from which 348 and 14 videos are randomly selected as training and evaluation split respectively. In addition to evaluating on MFR evaluation split, we further assess the performance of our method using Mip360 and RE10K datasets.

Metrics To evaluate the effectiveness of our method in predicting scores closely aligned with SSIM values, we use correlation coefficients as our primary evaluative metric. The Pearson correlation coefficient [38] is utilised across the majority of our analyses, complemented by Spearman's rank correlation coefficient [48] for studies involving new camera trajectories. These coefficients, ranging in $[-1, 1]$, measure the strength of association between `CrossScore` and SSIM, with a larger magnitude indicating a stronger correlation.

Baselines We choose five well-established IQA methods as baselines. Two from FR-IQA family: SSIM [59] and PSNR, and three from NR-IQA family: BRISQUE [31], NIQE [32], and PIQE [55].

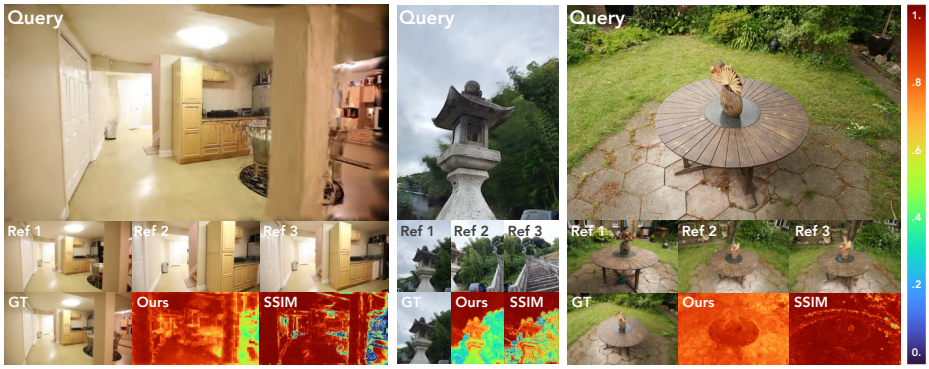

Fig. 3: Qualitative results of* CrossScore *and SSIM on various datasets. *We present examples for test results on each dataset (from left to right: RE10K, MFR, Mip360). We show our score maps have a strong correlation with SSIM score, demonstrating the generalisation capability of our approach across diverse datasets.*

Network Training Architecture: We adopt a pre-trained DINOv2-small network as the image encoder Φ_{enc}, which encodes images with a patch size 14×14 and produces features in 384 channels. The `CLS` token is ignored. The cross-reference module Φ_{cross} incorporates 2 transformer decoder layers with hidden dimension 384, and the decoder Φ_{dec} is equipped with a 2-layer MLP. Pre-processing: During training, we randomly crop 518×518 images from both query and reference images, whilst during inference, our model supports inputs at arbitrary resolutions. Notably, raw SSIM maps may occasionally present values below zero, and we found clamping raw SSIM maps to the range $[0, 1]$ leads to a slightly more stable training process. For the cross-reference set selection, we randomly choose $N_{\text{ref}} = 5$ real images from the same scene as the query image. Optimisation: We apply a constant learning rate of 5e-4 with an Adam-W [26] optimiser, training on 2× NVIDIA A5000 24GB GPUs for 160,000 iterations in 60 hours, with a per-GPU batch size of 24.

Training Data Generation We optimise Gaussian-Splatting (GS) [24], Nerfacto [51], and TensoRF [8] on the MFR [1] dataset for 15,000 iterations, saving checkpoints every 1,000 iterations up to 10,000, and a final one at 15,000 iterations. Images rendered from these checkpoints are compared against ground truths to produce SSIM maps. To reduce the cost of this process, we temporally subsample the MFR dataset by a factor of 8. The entire data processing spanned approximately two weeks, utilising 4× NVIDIA A5000 GPUs. The generated images and SSIM maps take about 1.5TB of storage. Our selection of GS, Nerfacto, and TensoRF was based on their efficiency and output quality. Each method employs a distinct NVS approach: GS models scenes with point clouds, Nerfacto utilises voxel grids, and TensoRF decomposes a 3D scene to planes, ensuring a diverse and high-quality image rendering process. As a result,

Table 1: Correlation between various metrics and SSIM on various datasets. ***FR****: full-reference.* ***NR****: no-reference.* ***CR****: cross-reference. We show* `CrossScore` *is highly correlated with SSIM score on various datasets, while only being trained with the MFR dataset.*

Datasets	FR		NR			CR
	SSIM↑	PSNR↑	BRISQUE↓	NIQE↓	PIQE↓	Ours↑
RE10K	1.00	0.92	0.46	0.32	0.27	**0.99**
Mip360	1.00	0.91	0.19	0.61	0.69	**0.95**
MFR	1.00	0.92	0.23	-0.30	-0.11	**0.83**

Table 2: Evaluating Few-shot NeRFs with Various Metrics. *We show that when comparing two few-shot NeRF models IBRNet [58] and PixelNeRF [68],* `CrossScore` *is consistent with full-reference metrics such as SSIM and PSNR. In this case, all metrics shows that IBRNet performs better than PixelNeRF on MFR dataset.*

NVS	FR		NR			CR
	SSIM↑	PSNR↑	BRISQUE↓	NIQE↓	PIQE↓	Ours↑
PixelNeRF	0.26	9.17	35.46	5.44	35.96	0.40
IBRNet	**0.44**	**18.51**	**23.47**	**2.68**	**23.35**	**0.71**

this approach balances data generation cost while producing a wide variety of distorted images and accurate corresponding SSIM maps.

4.2 Correlation with SSIM

We evaluate `CrossScore` by comparing it to SSIM using Pearson Correlation [38], with results shown in Tab. 1 alongside other baselines. Our approach demonstrates a strong correlation with the full-reference SSIM without using ground truth. Moreover, trained solely on the MFR dataset, our method successfully generalises to various settings, including indoor, outdoor, and 360-degree scanning environments, highlighting its versatile applicability. Fig. 3 provides qualitative results supporting our findings.

4.3 Application: Evaluating Few-shot NeRFs

This experiment demonstrates the application of `CrossScore` for evaluating few-shot NeRF methods, specifically comparing IBRNet [58] and PixelNeRF [68] using official checkpoints. Tab. 2 shows that both SSIM, PSNR, and `CrossScore` suggest IBRNet performs better on the MFR dataset. Note that the aim here is to highlight the ability of `CrossScore` to discern performance differences between methods rather than to benchmark them comprehensively.

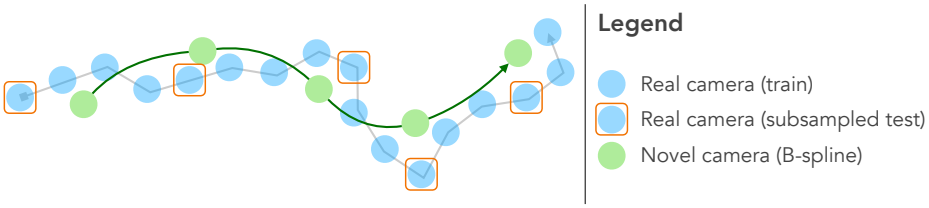

Fig. 4: Illustration of two IQA approaches in NVS: 1) with subsampled test views and 2) with true novel views. *The first approach relies on full-reference metrics that requires ground truth images, precluding test views in training (blue circles enclosed in orange boxes). In contrast, our cross-reference approach bypasses the need for ground truth views, allowing NVS evaluation from true novel views (green circles) and enabling NVS modelling to utilise the entire captured image set.*

Table 3: IQA on Images Renderings From Novel Trajectories. *We evaluate each sequence in two ways: 1) computing SSIM on images rendered from the standard subsampled test split with ground truth images, and 2) computing* `CrossScore` *on images rendered from a novel trajectory, with a cross-reference set randomly sampled from training images. We show that our method can evaluate the quality of Gaussian-Splatting from a novel trajectory without requiring aligned ground truth images. Our cross-reference style score is highly correlated with the full-reference SSIM score, and ranking video quality using* `CrossScore` *is similar to ranking with SSIM.* ***Top****: SSIM and* `CrossScore`*. Higher is better.* ***Bottom****: quality ranking using SSIM and* `CrossScore` *respectively. Lower is better. 'Corr' denotes Pearson correlation for scores and Spearman's rank correlation for rankings.*

	Scene	426	34	10	135	238	284	103	441	345	311	175	244	82	4	**Corr**
Score ↑	SSIM	0.74	0.66	0.64	0.64	0.61	0.61	0.59	0.58	0.56	0.55	0.51	0.50	0.44	0.40	**0.84**
	Ours	0.80	0.78	0.77	0.78	0.66	0.61	0.73	0.75	0.73	0.72	0.62	0.58	0.55	0.53	
Rank ↓	SSIM	0	1	2	3	4	5	6	7	8	9	10	11	12	13	**0.85**
	Ours	0	2	3	1	8	10	6	4	5	7	9	11	12	13	

4.4 Application: IQA on Images Rendered From a Novel Trajectory

In this experiment, we demonstrate that our cross-reference method enables true novel view rendering evaluation. Specifically, given an NVS-reconstructed scene, we evaluated this scene in two distinct ways, as illustrated in Fig. 4. First, we follow the conventional test split, which considers every 8th image as a test image, and compute the SSIM score between the rendered image and ground truth. Second, we evaluate true novel view renderings that are rendered from a novel trajectory[2] with `CrossScore` without ground truth. Tab. 3 indicate a close correlation between `CrossScore` evaluations of novel views and traditional SSIM scores. Additionally, the rankings of rendering quality for these scenes, determined using both SSIM and `CrossScore`, are also closely aligned.

[2]Novel trajectories are generated by interpolating training poses with a B-spline function (degree of 10), creating 20 novel poses per scene.

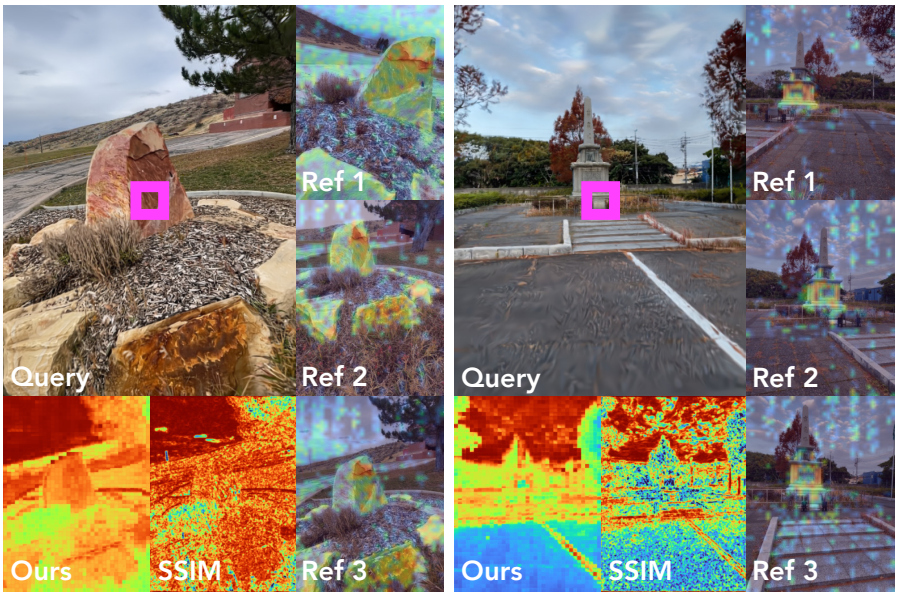

***Fig. 5: Visualisation of attention weights from the cross-reference module $\Phi_{\texttt{cross}}$. Top left**: a query image with a region of interest (centre of image) highlighted with a magenta box. **Right column**: We show 3 reference images from our cross-reference set with attention maps overlaid. The attention maps illustrate the attention that is paid to predicting image quality at the query region. Red and blue denote high and low attention weights respectively. Note that we use $N_{\text{ref}} = 5$ but only 3 is shown due to space constraint. **Bottom**: Predicted `CrossScore` map and SSIM map. Red and blue denote high and low quality image regions respectively.*

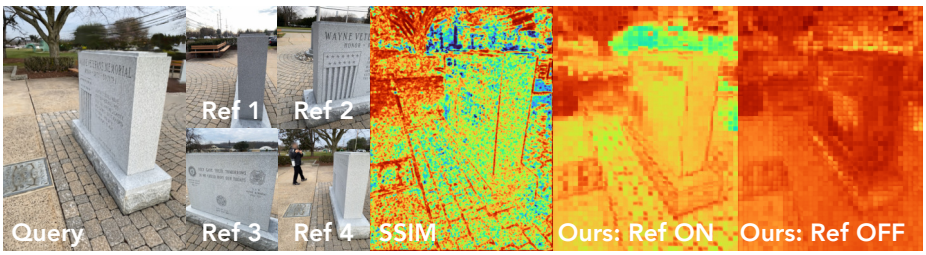

***Fig. 6: Ablation study: reference set enabled (on) and disabled (off).** We show that with reference images enabled, the score map predicted by our method contains more details. When the reference images are disabled, the model tends to assign everything a high score. This is also evidenced by quantitative results in Tab. 4.*

4.5 Visualising Attention Weights

To delve deeper into our cross-reference method, Fig. 5 visualises the attention weights in the cross-attention layer for the central patch of a query image. This

Table 4: Ablation study: reference set enabled (✓) and disabled (✗). *Our method performs closer to SSIM when reference images are provided. Note that when reference images are disabled, the predicted scores still show a certain level of correlation, as certain noise patterns can be identified from local image statistics. In this case our method degrades to a no-reference-style image evaluation.*

Scene	4	10	34	82	103	135	175	238	244	284	311	345	426	441	Avg	**Corr**
SSIM	0.40	0.64	0.66	0.44	0.59	0.64	0.51	0.61	0.50	0.61	0.55	0.56	0.74	0.58	0.57	**1.00**
Ours ✓	0.46	0.72	0.72	0.48	0.64	0.75	0.56	0.61	0.66	0.58	0.65	0.72	0.82	0.71	0.65	**0.83**
Ours ✗	0.71	0.81	0.83	0.73	0.80	0.84	0.80	0.79	0.86	0.80	0.74	0.86	0.89	0.85	0.81	**0.68**

illustration confirms that the cross-attention mechanism effectively focuses on similar content from the cross-reference set, thereby providing insight into the results of the ablation study in Sec. 4.6.

4.6 Ablation Study: Enable and Disable Reference Views

This experiment demonstrates that our cross-reference module effectively uses the cross-reference set for quality prediction. When provided with reference images, the module offers detailed and accurate evaluations, as shown in Fig. 6, in contrast to the high scores predicted across almost all regions when reference images are disabled. Note that, in this context, we disable reference images by setting all pixels in reference images to zero. Quantitative support for these findings is presented in Tab. 4.

4.7 Limitations and Future Work

We outline two future research directions: First, enhancing the sharpness of our score maps to match the clarity of full-reference SSIM, possibly by integrating pixel-level positional encoding or super-resolution methods to mitigate the blurring from patch-wise encoding of ViT models. Second, tackling the issue with unconventional images, such as those from fish-eye lenses that lead to inaccurate predictions, as illustrated in Fig. 7.

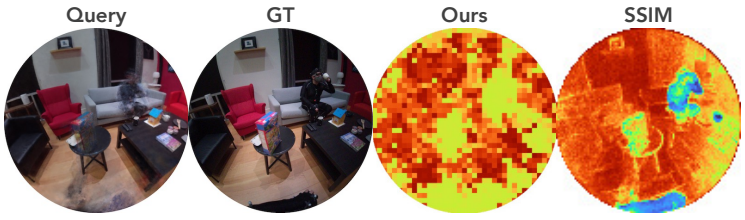

Fig. 7: Limitations*: evaluating a fish-eye-style query image.*

5 Conclusion

In summary, we introduce a novel Cross-Reference Image Quality Assessment (CR-IQA) scheme, filling a critical gap in existing IQA schemes. By leveraging a neural network with cross-attention mechanisms and a unique NVS-enabled data collection pipeline, we demonstrate the feasibility of accurately evaluating the quality of an image by comparing it with other views of the same scene. Our experimental results indicate that our predictions closely align with ground-truth-dependent metrics.

Acknowledgement

This research is supported by an ARIA research gift grant from Meta Reality Lab. We gratefully thank Shangzhe Wu, Tengda Han, and Zihang Lai for insightful discussions, and Michael Hobley for proofreading.

References

1. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free visual relocalization: Metric pose relative to a single image. In: ECCV (2022)
2. Azzarelli, A., Anantrasirichai, N., Bull, D.R.: Towards a robust framework for nerf evaluation. arXiv preprint arXiv:2305.18079 (2023)
3. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
4. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
5. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR (2022)
6. Bian, J.W., Bian, W., Prisacariu, V.A., Torr, P.: Porf: Pose residual field for accurate neural surface reconstruction. In: ICLR (2023)
7. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: ICLR (2018)
8. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: ECCV (2022)
9. Chen, Y., Lee, G.H.: Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In: CVPR (2023)
10. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: The epic-kitchens dataset: Collection, challenges and baselines. TPAMI (2021)
11. De Luigi, L., Bolognini, D., Domeniconi, F., De Gregorio, D., Poggi, M., Di Stefano, L.: Scannerf: a scalable benchmark for neural radiance fields. In: WACV (2023)
12. Deng, N., He, Z., Ye, J., Duinkharjav, B., Chakravarthula, P., Yang, X., Sun, Q.: Fov-nerf: Foveated neural radiance fields for virtual reality. IEEE Transactions on Visualization and Computer Graphics (2022)
13. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. TPAMI (2020)

14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
15. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Empirical Methods in Natural Language Processing (EMNLP) (2021)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
19. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: CVPR (2024)
20. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: ICCV (2021)
21. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: CVPR (2014)
22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023)
23. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021)
24. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG (2023)
25. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: ICCV (2021)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
27. Lu, F., Lan, X., Zhang, L., Jiang, D., Wang, Y., Yuan, C.: Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In: CVPR (2024)
28. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. TOG (2011)
29. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. TOG (2019)
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021)
31. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing (2012)
32. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters (2012)
33. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing (2011)
34. Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L.: Improving multimodal datasets with image captioning. In: NeurIPS (2024)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. TMLR (2023)

36. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: ICCV (2023)
37. Park, K., Henzler, P., Mildenhall, B., Barron, J.T., Martin-Brualla, R.: Camp: Camera preconditioning for neural radiance fields. TOG (2023)
38. Pearson, K.: Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London (1896)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
40. Redi, J.A., Gastaldo, P., Heynderickx, I., Zunino, R.: Color distribution information for the reduced-reference assessment of perceived image quality. IEEE Transactions on Circuits and Systems for Video Technology (2010)
41. Rehman, A., Wang, Z.: Reduced-reference image quality assessment by structural similarity estimation. IEEE Transactions on Image Processing (2012)
42. Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban radiance fields. In: CVPR (2022)
43. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
44. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016)
45. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
46. Solh, M., AlRegib, G.: Miqm: A novel multi-view images quality measure. In: International Workshop on Quality of Multimedia Experience (2009)
47. Somasundaram, K., Dong, J., Tang, H., Straub, J., Yan, M., Goesele, M., Engel, J.J., De Nardi, R., Newcombe, R.: Project aria: A new tool for egocentric multimodal ai research. arXiv preprint arXiv:2308.13561 (2023)
48. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology (1904)
49. Sun, J., Qiu, J., Zheng, C., Tucker, J., Yu, J., Schwager, M.: Aria-nerf: Multimodal egocentric view synthesis. arXiv preprint arXiv:2311.06455 (2023)
50. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: CVPR (2022)
51. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al.: Nerfstudio: A modular framework for neural radiance field development. In: SIGGRAPH (2023)
52. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: CVPR (2023)
53. Tschernezki, V., Larlus, D., Vedaldi, A.: Neuraldiff: Segmenting 3d objects that move in egocentric videos. In: 3DV (2021)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
55. Venkatanath, N., Praneeth, D., Bh, M.C., Channappayya, S.S., Medasani, S.S.: Blind image quality evaluation using perception based features. In: National Conference on Communications (2015)
56. Wang, C., Wang, A., Li, J., Yuille, A., Xie, C.: Benchmarking robustness in neural radiance fields. arXiv preprint arXiv:2301.04075 (2023)

57. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
58. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
59. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
60. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: Human Vision and Electronic Imaging X (2005)
61. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Asilomar Conference on Signals, Systems & Computers (2003)
62. Wang, Z., Wu, G., Sheikh, H.R., Simoncelli, E.P., Yang, E.H., Bovik, A.C.: Quality-aware images. IEEE Transactions on Image Processing (2006)
63. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: NeRF−−: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
64. Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D.: Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In: ECCV (2022)
65. Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K.: An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755 (2018)
66. Xue, W., Zhang, L., Mou, X., Bovik, A.C.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Transactions on Image Processing (2013)
67. You, J., Korhonen, J.: Transformer for image quality assessment. In: ICIP (2021)
68. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR (2021)
69. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing (2011)
70. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
71. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: CVPR (2023)
72. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification. TOG (2018)