Puzzle Similarity: A Perceptually-guided Cross-Reference Metric for Artifact Detection in 3D Scene Reconstructions

Nicolai Hermann[1,2] Jorge Condor[1,2] Piotr Didyk[1,2]

[1]USI, Lugano, Switzerland
[2]IDSIA, Switzerland
{nicolai.hermann, jorge.condor, piotr.didyk}@usi.ch

Abstract

Modern reconstruction techniques can effectively model complex 3D scenes from sparse 2D views. However, automatically assessing the quality of novel views and identifying artifacts is challenging due to the lack of ground truth images and the limitations of no-reference image metrics in predicting reliable artifact maps. The absence of such metrics hinders assessment of the quality of novel views and limits the adoption of post-processing techniques, such as inpainting, to enhance reconstruction quality. To tackle this, recent work has established a new category of metrics (cross-reference), predicting image quality solely by leveraging context from alternate viewpoint captures [47]. In this work, we propose a new cross-reference metric, Puzzle Similarity, which is designed to localize artifacts in novel views. Our approach utilizes image patch statistics from the training views to establish a scene-specific distribution, later used to identify poorly reconstructed regions in the novel views. Given the lack of good measures to evaluate cross-reference methods in the context of 3D reconstruction, we collected a novel human-labeled dataset of artifact and distortion maps in unseen reconstructed views. Through this dataset, we demonstrate that our method achieves state-of-the-art localization of artifacts in novel views, correlating with human assessment, even without aligned references. We can leverage our new metric to enhance applications like automatic image restoration, guided acquisition, or 3D reconstruction from sparse inputs. Find the project page at https://nihermann.github.io/puzzlesim/.

1. Introduction

Image-based rendering and 3D reconstruction from a sparse set of 2D views has received ample attention in recent years, both for pure geometry reconstruction and radiance-field modeling. Classical approaches such as structure from motion (SfM) use simple triangulation and epipolar geometry to produce sparse point clouds of diffuse color [33]. Densifying these representations can be done explicitly [15]. Alternatively, one can learn continuous, implicit representations [3, 24, 28], normally modeled through multi-layer perceptrons. A tangential problem to these efforts is the collection of 2D data and the handling of corrupted, distorted, or simply incomplete sets of images from an object or scene we would like to model. Learning representations from very sparse inputs has been a widely studied topic, where normally learned priors from large datasets are leveraged to enforce 3D consistency to ensure that the resulting reconstructions follow natural statistics [4, 5, 48, 57]. However, quantifying the quality of novel views from reconstructions is still problematic. These views can contain artifacts due to the sparsity of the training dataset, and automatically identifying them helps with restoration (e.g., masking for image-based inpainters [37]) or simply to guide future data acquisition to fill the gaps [17]. Recent works have followed a Bayesian approach to quantify the uncertainty of whether an area belongs to a reconstructed scene or not [8], which could potentially be leveraged for simple artifact detection. However, they require implicit models, with fundamental changes to the scene model, and are not practical for more general applications that require visual artifact identification outside of scene reconstruction, and are incapable of detecting artifacts not arising from lack of coverage.

To tackle this, we propose a novel approach for artifact detection that can be leveraged on any set of images without an encoded explicit or implicit model of the scene or object they depict. Unlike visual difference predictors (VDPs) [20] (which require references) and no-reference quality metrics [25, 26] (which typically do not provide maps, but rather produce single values of overall quality), our approach provides visual artifact maps with no direct references. We leverage learned perceptual patch statistics

from small, clean datasets and compare them to the embedded statistics of new images from a similar distribution (i.e., novel reconstructed views from a 3DGS [15] representation) to obtain artifact maps without aligned references. We test our generated maps through a human experiment where we ask participants to manually identify artifacts and distortions in images to generate ground-truth data of visual artifacts. Our results show that our method agrees with human assessment, correlating better than no-reference, full-reference, and state-of-the-art cross-reference metrics. To summarize, our contributions are the following:

- A novel cross-reference visual artifact identification metric, particularly tailored for 3D reconstruction,
- a novel dataset of human-labeled artifact and distortion maps to fill the gap of validation benchmarks for cross-reference metrics,
- and applications on image restoration and 3D reconstruction enhancement that showcase our approach's utility.

2. Related Work

Our metric is specifically designed for applications in 3D scene reconstruction and image-based rendering. Consequently, this section discusses work on 3D reconstruction first and then on image metrics.

2.1. 3D Reconstruction and Image-based Rendering

Reconstructing 3D objects or scenes from sparse sets of 2D observations is a fundamental problem in vision [20]. Particularly, in the context of novel view synthesis, the objective is to approximate the radiance field (i.e., 5D function encoding spatially varying radiance emission) of specific objects or scenes. Most methods differ either in the model used to encode the function or the rendering procedure. Implicit approaches model the radiance field as a continuous function, approximated by a multi-layer perceptron [24, 36]. Rendering is usually done via sampling the implicit volume using ray-marching [42], which provides spatially varying values of density and anisotropic color emission modeled through Spherical Harmonics. Improvements over this formula have tackled performance limitations, either by using more efficient sampling techniques [9, 28, 30] or by distilling the implicit space into explicit density and anisotropic appearance volumes [50, 56]. On the other hand, purely explicit models do not require any pre-training using implicit functions, and were originally Eulerian in nature [55]. Explicit models are easier to optimize, usually faster, and more interpretable, which can help in different tasks such as scene editing or animation. More recently, anisotropic Lagrangian approaches have found tremendous success [15]. However, these explicit methods have introduced some limitations of their own along the way. Methods like 3D Gaussian Splatting [15] can only model areas that are directly supervised, and degrade less gracefully than implicit counterparts when querying viewpoints substantially outside the training set coverage. Detecting artifacts arising from the lack of coverage is difficult due to the lack of reference images. Our method produces these masks via supervision on the training data solely, which can enable unsupervised restoration (automatic inpainting of the artifacts based on available context [7]) or simply automatically guide further image acquisition to complete the dataset efficiently [17].

2.2. Image Metrics

Image metrics are classified by the type of prediction they make and their input. Image Quality Metrics (IQMs) typically predict a single number, corresponding to overall image quality, and are often trained on Mean Opinion Score (MOS) datasets. Visibility Metrics (VMs) produce maps corresponding to the perceptibility of distortions. They often rely on models of the human visual system and predict the probability of detecting local artifacts by an observer. Most image metrics are full-reference, necessitating a reference to assess the quality of a test image. In contrast, no-reference metrics predict the quality or visible distortions based solely on the test image. Others use additional information, e.g., partial reference, and are referred to as reduced-reference or cross-reference metrics.

Classical examples of full-reference IQMs include mean-absolute error (MAE), mean-squared error (MSE), peak signal-to-noise ratio (PSNR), SSIM [46], FSIM [59], MS-SSIM [45], and LPIPS [62]. These methods begin by computing local differences between test and reference images and aggregate them into a single quality score as the final step. By omitting this step, it is possible to create a local distortion map, which is a common output of VMs widely adopted in rendering to localize poorly rendered areas [2]. Typically, VMs differ from IQMs in their more explicit modeling of the human visual system [6], which enables the prediction of perceived visibility or the strength of visual differences between images. Improvements over the original framework extended their applicability to high dynamic range imagery [22], making them eccentricity and motion-aware [21, 40], and integrating perceived color [23]. Such metrics have been used in many perceptual optimizations, such as foveated rendering [41] and perceptually aware tone mappers [39].

No-reference metrics eliminate the need for a reference and are most commonly learned from human quality assessment datasets [13, 14, 53], supervised on extracted features from natural image statistics [25–27, 51], or even synthetic scores [52]. Modern deep learning approaches can utilize Transformer architectures [54], and multi-scale Transformers are employed to alleviate the resolution constraints [14]. Hybrid approaches use multi-modal architectures in conjunction with text templates to query image features such

as noisiness, sharpness, or contrast, which can be translated into MOS [44, 63]. The above no-reference metrics rely on global image features and, therefore, are not suitable for obtaining distortion maps. However, similarly to our work, some of these IQMs are capable of producing visual maps. For example, PIQE [29] measures distortions in an image patch based on extracted local features. CNNIQA [13], on the other hand, was one of the first no-reference metrics to employ convolutional neural networks (CNN) to regress mean-opinion scores. More recently, PaQ-2-PiQ [53] uses region proposals to select quality-determining patches. PAL4VST [61] localizes specific artifacts that emerge from image synthesis tasks through binary segmentation. In contrast, our work leverages the latent space of a model pre-trained on natural images to measure the cosine similarity in feature space of candidate image patches to a limited set of image patches from a similar distribution (i.e., images from the same scene in the context of scene reconstruction), rendering a higher level of alignment with human assessment.

In some applications, even though an image metric does not have access to a reference image, it may have access to other information useful for making a prediction. In the context of novel view synthesis, cross-reference metrics utilize a set of unaligned reference images to assess the quality of a single query image belonging to the same scene. This type of metric was recently established by Wang et al. [47] and their metric CrossScore serves as the main baseline for our evaluation. The metric relies on a cross-attention module [43] to correlate a test image with unaligned multiview images to predict quality maps. The maps predict the quality of 14x14 patches of the input image and are trained to mimic unpooled SSIM maps. However, SSIM has been repeatedly shown to be poorly aligned with human quality assessment and perception [31, 32, 62], which fundamentally limits the potential of CrossScore.

In comparison to previous work, our method is a cross-reference metric for novel view synthesis tasks. It predicts local similarities for a synthesized view given a set of unaligned reference views. Its effectiveness in localizing artifacts is achieved by assessing similarity in feature space. In contrast to many IQMs, our method is not designed to predict a single score, but rather a map corresponding to the strength of the visible artifacts.

3. Our Method

Let us establish an analogy for our method: pretend each reference image is a puzzle with many puzzle pieces. To test if a new image is similar to our unaligned references, we would simply shuffle all pieces from all puzzles from our references and try to reassemble the test image only using those pieces. If the new image is very similar to the references, we should have enough puzzle pieces to compose the other image confidently. However, if the image holds regions very different from what we saw in the reference images, we would lack puzzle pieces to assemble this area, effectively leaving holes in the newly assembled puzzle (image). An overview of our approach through this analogy can be seen in Fig. 1.

In our work, the puzzle pieces correspond to embedded image patches. In order to assess patch similarity, an obvious approach would involve computing the dot product between all patches; best-matching pieces would be recorded to create a similarity map. This simplistic approach, however, would hardly align with human assessment if unprocessed patches were used. Inspired by the close correlation between human quality judgment and latent CNN feature maps [38, 62], we employ a pre-trained CNN [11, 18, 35] to embed all the references, computing similarity in the latent feature space. Note that comparing feature map "pixels" in a CNN is similar to comparing individual patches in the input domain; this is due to the locality of the sliding kernels when convolving. The patch size is dependent on the receptive field (showcased in Fig. 2).

Choice of layers Choosing the right layers for embedding is essential to maximize the quality of the predicted spatial maps. While early layers feature small receptive fields and capture fine details, deeper layers have larger receptive fields and capture coarser features. This can be observed in Fig. 3, where we showcase different `VGG` layers. It is essentially a trade-off between prediction granularity, accuracy and speed. We identified that combining multiple layers into our metric computation incorporates the various levels of abstraction and scales in a robust manner. We thus compute the weighted average of the three layers; we empirically found that halving the image resolution more than three times did not significantly improve our results, as the scale becomes too small and the pool of reference vectors too little and specific to find good correspondences among novel images, even for well-reconstructed areas. This observation suggested that features from the layers before the third down-sampling were most useful for our cause.

Computing patch similarity To compute the similarity map of a test image, we feed all references and the test image through a pre-trained network $\mathcal{F}$ to obtain the embeddings. We repeat the exact computation for each network layer; thus, we will describe the steps once for one layer ℓ. To find the similarity $s_\ell(x, y)$ to the best matching puzzle piece for a pixel of the embedded test image at some spatial location (x, y), we compute the cosine similarity based on the feature vector $\mathcal{F}_\ell(x, y)$ and all other feature vectors of all N reference images of the same layer ℓ and select the correspondence with the highest similarity:

$$s_\ell(x, y) = \max_{n,x',y'} \hat{\mathcal{F}}_\ell(x, y) \cdot \hat{\mathcal{F}}_\ell^{(n)}(x', y') \qquad (1)$$

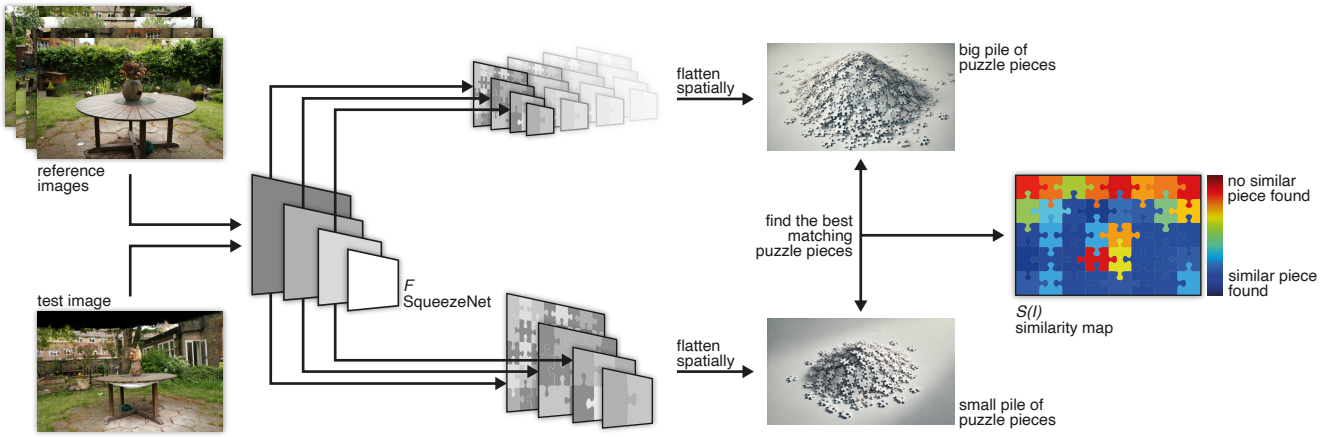


Figure 1. Schematic representation of our Puzzle Similarity metric.

where $\hat{\mathcal{F}}$ denotes a feature vectors scaled to unit length $\hat{\mathcal{F}}_\ell(x,y) = \frac{\mathcal{F}_\ell(x,y)}{||\mathcal{F}_\ell(x,y)||_2} \in \mathbb{R}^{C_\ell}$ and $\cdot$ is the dot product. Note that we compute the cosine similarity with any feature vector of the same layer from all references, not just those at the same spatial position. This relinquishes spatial relations and makes the method robust to simple camera movements that only shift the image horizontally or vertically. We iterate this maximum search for all pixels of the test image's embedding to construct the similarity mask $\mathcal{S}_\ell$.

$$\mathcal{S}_\ell(\mathcal{I}) = \begin{bmatrix} s_\ell(1,1) & s_\ell(1,2) & \cdots & s_\ell(1,W_\ell) \\ s_\ell(2,1) & s_\ell(2,2) & \cdots & s_\ell(2,W_\ell) \\ \vdots & \vdots & \ddots & \vdots \\ s_\ell(H_\ell,1) & s_\ell(H_\ell,2) & \cdots & s_\ell(H_\ell,W_\ell) \end{bmatrix} \tag{2}$$

where $\mathcal{I}$ is the test image. We repeat this for a set of layers and combine them into a final similarity map. To match the spatial dimensions of each layer, we bilinearly upsample each map to the original image size and combine them with an affine combination:

$$\mathcal{S}(\mathcal{I}) = \sum_\ell w_\ell \text{ Upsample } \left(\mathcal{S}_\ell(\mathcal{I},\ \mathcal{I}_{\text{ref}}^{1:N})\right) \tag{3}$$

with $\sum_\ell w_\ell = 1$ and reference images $\mathcal{I}_{\text{ref}}^{1:N}$. To utilize optimized hardware, please note how the computation of S_ℓ can also be expressed as an outer product between the spatially flattened embeddings:

$$\begin{aligned} &\hat{\mathcal{F}}_\ell(\mathcal{I}^{1:N}) \in \mathbb{R}^{N\times H_\ell\times W_\ell\times C_\ell} \\ &\tilde{\mathcal{F}}_\ell(\mathcal{I}^{1:N}) = \text{flatten}\left(\hat{\mathcal{F}}_\ell(\mathcal{I}^{1:N})\right) \in \mathbb{R}^{NH_\ell W_\ell\times C_\ell} \\ &\mathcal{S}_\ell(\mathcal{I}) = \overbrace{\text{rowmax}\ \underbrace{\tilde{\mathcal{F}}_\ell(\mathcal{I}_{\text{ref}}^{1:N}) \otimes \tilde{\mathcal{F}}_\ell(\mathcal{I})}_{\in \mathbb{R}^{NH_\ell W_\ell\times H_\ell W_\ell}}}^{\in \mathbb{R}^{H_\ell W_\ell}} \end{aligned} \tag{4}$$

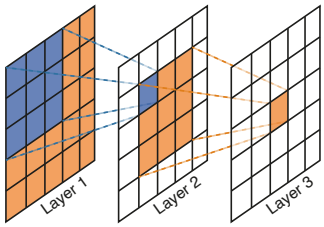
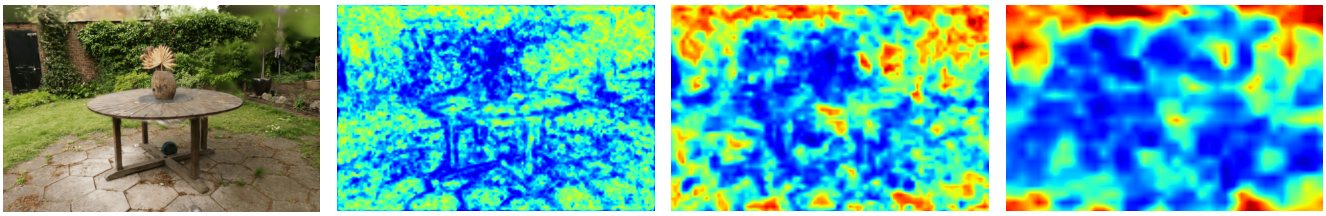


Figure 2. Receptive field of a multi-layer CNN. Note how one pixel in the last layer is an embedding of a patch of the input space.

where the test image $\mathcal{I}$ is a special case with $N = 1$, $\otimes$ is the outer product, and rowmax applies the max over the first dimension. While a naïve implementation of this outer product would require substantial amounts of memory for larger N, H, W, we provide an efficient implementation through blockwise tiling with intermediate max-reduction, which we detail in the Supplemental.

Pre-trained Model Choice The choice of pre-trained neural network, through which the embeddings will be created, is a key component of our work. We primarily considered classic models including `VGG-16`, `VGG-19`, `AlexNet`, and `SqueezeNet` [11, 18, 35]. Some of the critical considerations are model complexity and memory requirements, which we summarized in the Supplemental, as well as their specifically tested performance on our human alignment task. Beyond quality performance, reducing the memory footprint and computational complexity is key as it may impact the possibility of downstream applications of our metric, which, given its differentiability, could be leveraged in optimization procedures.

We empirically found that while `VGG` produces the most fine-grained maps, `AlexNet` and `SqueezeNet` still managed to perform similarly, while doing so at a substantially reduced computational cost. We opted for `SqueezeNet`



(a) Rendering (b) 2nd VGG layer (c) 3rd VGG layer (d) 4th VGG layer

Figure 3. Puzzle Similarity computed on a single VGG layer. Note how the second layer has a finer resolution and mostly cold colors, while the fourth layer is much smoother and features a wider range of values. Warm colors indicate artifacts or poor reconstruction quality.

as it aligned best with our test examples, specifically using layers $\ell \in \{2, 3, 4\}$ with the weights $w_2 = 0.67$, $w_3 = 0.2$, and $w_4 = 0.13$, which we found heuristically.

4. Results

We will now analyze how our method compares against competing approaches for both no-reference and cross-reference visual map prediction in the context of reconstruction and image-based rendering. To quantify the correlation between all these maps and human assessment, we present a novel dataset on human artifact identification, which we manually collected and can be found here[1] to facilitate future research on the topic. As for our method, for each different scene, we compute embeddings on their respective training dataset to compute similarity maps on test views, as explained in Sec. 3. CrossScore leverages the same set of training views for its map predictions.

4.1. A Novel Artifact Identification Dataset

We created a dataset of human-perceived artifacts in 3D reconstructed views with corresponding ground truths collected through a user study. To generate images exhibiting typical reconstruction artifacts, we apply 3D Gaussian Splatting [15] to twelve scenes from the Mip-NeRF360 [3], Tanks and Temples [16], and Deep Blending [10] datasets. We use default parameters but intentionally withhold a significant portion of training images, which we later utilize for additional validation. By omitting these views during training, we increase the likelihood of artifacts appearing in the withheld views. For each dataset, we selected three renderings that demonstrated a mix of well-reconstructed areas, strong artifacts, and subtle artifacts, resulting in 36 samples across 12 datasets.

Experiment details We asked 22 participants to segment visible artifacts in each of the 36 sample images under controlled viewing conditions using the tool developed by Wolski et al. [49], which the authors kindly provided. We include details on the participants' self-reported gender and age distributions in the Supplementary, as well as detailed viewing and display conditions. During the experiment, users had no undistorted, artifact-free references at their disposal and thus had to judge individual images at face value. They would then mark areas found to be unnatural or unappealing, creating a binary mask. With the dataset, we can evaluate the agreement between human judgment and any metric output by simply averaging all binary masks to estimate the probability of each pixel being marked as an artifact. Fig. 4 shows example renderings (a) alongside metric predictions (b)-(d) and their average human-produced mask (e).

We evaluate our method against both no-reference and cross-reference metrics. To assess their alignment with human perception, we correlate their maps with the human ratings from our dataset, as described in Sec. 4.1. NR and CR metrics are the only metrics that can detect artifacts without a direct reference, but the way we collected our dataset gives us access to a ground truth that is normally unavailable. This enables us to assess FR metrics and current VDPs too, although they are otherwise not suitable for the objective in question. We include extensive results in the Supplemental and show that, unlike any other competing metric, our metric even outperforms the best FR metrics on this benchmark, proving the general superiority of our method.

4.2. Evaluation

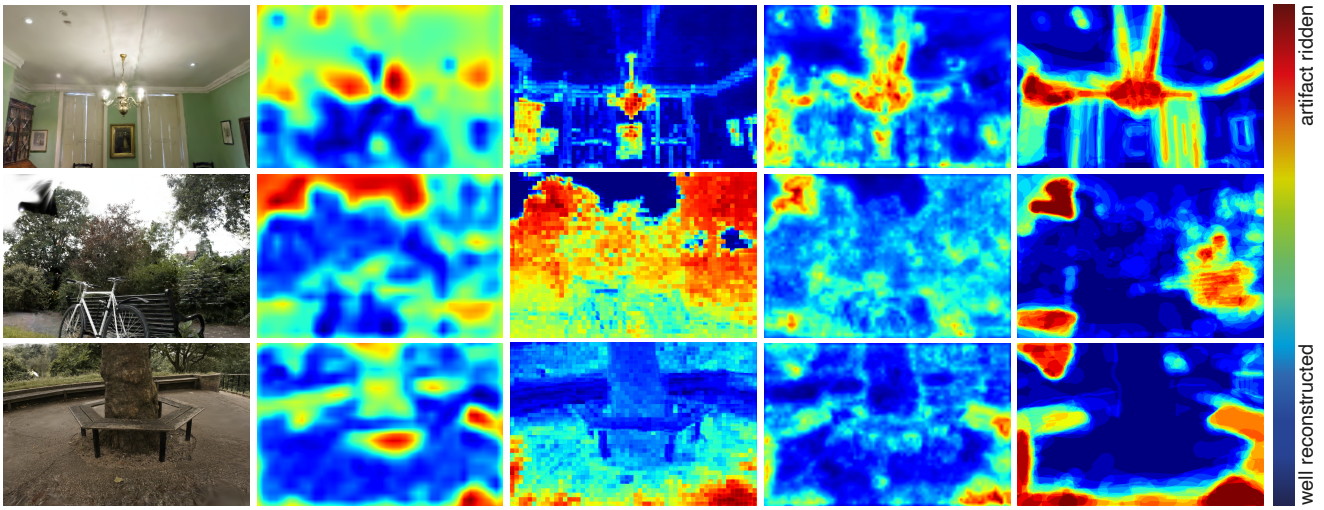
To correlate metric outputs to our human segmentations, we first compute each metric map for each rendering and then compute the Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC). To account for the different domains of the compared metrics and possibly non-linear relations, we fit a 5-parameter logistic curve for a fair comparison as suggested by [1, 19, 34, 60]:

$$q(x) = a_1 \left\{ \frac{1}{2} - \frac{1}{1 + \exp\left(a_2(x - a_3)\right)} \right\} a_4 x + a_5 \quad (5)$$

where x is an individual pixel score and $a_{1\ldots5}$ are optimized through gradient ascent to maximize PCC or SRCC.

Results Tab. 1 reports the average Pearson and Spearman correlations for each scene. Our dataset includes three images per scene, with varying artifact types per scene. For

[1] https://huggingface.co/datasets/nihermann/annotated-3DGS-artifacts



(a) Artifact-ridden view (b) PaQ-2-PiQ map (c) CrossScore map (d) PuzzleSim map (ours) (e) Human labeling

Figure 4. Selection of image quality maps for artifact-ridden renderings from various scenes. The last column shows ground-truth human assessments from our collected dataset. Note that our metric provides the finest resolution, enabling better artifact segregation.

Table 1. Pearson and Spearman Correlation between NR, and CR metrics and Human Perception per Dataset. The dashed line separates NR (above) from CR (below) metrics.

		bicycle	bonsai	counter	drjohnson	flowers	garden	kitchen	playroom	stump	train	treehill	truck
Pearson	PAL4VST [61]	0.139	0.088	0.062	0.153	0.002	0.005	0.068	0.197	0.000	0.104	0.000	0.119
	PaQ-2-PiQ [53]	0.194	0.420	0.432	0.138	0.305	0.428	0.613	0.452	0.534	0.667	0.220	0.424
	PIQE [29]	0.266	0.267	0.255	-0.089	0.583	0.441	0.490	0.091	0.526	0.200	0.376	0.101
	CNNIQA [13]	0.027	0.037	0.064	-0.068	-0.053	0.409	0.324	0.345	0.367	0.400	-0.133	0.005
	CrossScore [47]	0.338	0.331	0.493	0.390	0.476	0.748	0.663	0.603	0.585	0.565	0.300	0.630
	PuzzleSim (ours)	0.594	0.565	0.618	0.461	0.609	0.675	0.768	0.636	0.505	0.642	0.717	0.593
Spearman	PAL4VST [61]	0.083	0.080	0.027	0.111	0.004	-0.003	0.069	0.169	0.000	0.098	0.000	0.108
	PaQ-2-PiQ [53]	0.214	0.495	0.435	0.009	0.261	0.176	0.616	0.329	0.476	0.696	0.152	0.329
	PIQE [29]	0.209	0.409	0.291	-0.130	0.460	0.229	0.620	0.079	0.375	0.224	0.279	0.174
	CNNIQA [13]	-0.085	0.020	0.166	0.157	-0.130	0.255	0.375	0.380	0.316	0.395	-0.219	-0.066
	CrossScore [47]	0.299	0.030	0.243	0.508	0.365	0.590	0.315	0.534	0.490	0.494	0.240	0.431
	PuzzleSim (ours)	0.468	0.393	0.382	0.499	0.428	0.428	0.658	0.601	0.307	0.540	0.548	0.440

example, the *garden* scene has prominent black regions due to holes in the reconstruction, making artifact detection straightforward and leading to high correlation scores for most metrics. However, scenes like *treehill*, *stump*, and *flowers* exhibit artifacts in the form of blurry or unnatural textures while preserving similar color distributions to the ground truth. Puzzle Similarity consistently achieves high correlation with human-perceived artifacts across all datasets, retaining smaller variance across datasets than all competitive methods (See Tab. 2). The higher variance in the averaged results is mainly due to scenes having different types of artifacts, where some are harder to identify than others (e.g., it is easier to identify black areas than small diffuse blobs). Keeping a small variance implies robust performance across various artifact types. We include extended quantitative and qualitative results in our Supplemental.

4.3. Comparison with No-Reference Metrics

We compare with other no-reference metrics capable of producing spatial visibility maps [13, 47, 53, 58, 61]. CNNIQA [13] was applied on patches as described in their paper. We applied padding to avoid cropping the borders and bilinearly upsampled the final map. PIQE [29] already produces three different kinds of maps that we averaged. PAL4VST [61], PaQ-2-PiQ [53] produce maps and were applied as described in their papers, but bilinearly upsam-

pled to match the human maps' resolution.

Puzzle Similarity demonstrates superior accuracy in artifact localization, as shown by the correlation values in Tab. 1. PAL4VST and CNNIQA performed poorly, as expected, given their focus on detecting specific types of distortions that are not necessarily similar to reconstruction artifacts. While PIQE and PaQ-2-PiQ performed well in certain scenes, their overall correlation with human opinion was generally lower in others, reflecting a less robust alignment with human assessment. However, while our method relies on a small subset of images from a similar distribution to the target image (e.g., the training dataset on novel view synthesis of a specific scene), NR metrics do not require any extra images and attempt to generalize to any input.

4.4. Comparison with Cross-Reference Metrics

CrossScore is, to our knowledge, the only other CR metric besides ours that is also reliant on the training views. We also bilinearly upsampled its output to match the human maps' resolution. We show comparisons to CrossScore in Tab. 1 and Tab. 2, in Fig. 4 and extended results in the Supplementary. We outperform CrossScore on most datasets and show better performance both on average and in terms of consistency (with a substantially smaller standard deviation among results). While the expensive domain-specific pretraining of CrossScore should, in theory, be superior to our feature-space patch matching leveraging general models pre-trained on collections of natural images, their reliance on SSIM as its target quality assessment metric limits its potential to accurately model human quality assessment, due to the well-known limitations of the metric in this regard [31, 32, 62]. Their DINOv2 encoder limits map resolution to 14x14 blocks, reducing artifact localization fidelity. Furthermore, our approach is notably simpler: we can leverage any CNN as a feature encoder, allowing seamless adaptation to specific domains simply by swapping out the backbone. No retraining or distillation of any framework component is required.

4.5. Comparison with Full-Reference Metrics

Although full-reference metrics, unlike our method, require a direct reference image for detecting artifacts, we also provide an extensive comparison to them based on our dataset, which includes reference images. On average, we outperform all FR metrics, while CrossScore falls behind FovVideoVDP. Thus, in the context of 3D reconstruction, our metric performs better than all tested NR, CR, and FR metrics. Investigating why even the most advanced VDPs, rooted in complex models of low-level human visual processing, quantitatively fall behind remains a fascinating avenue for future work.

Table 2. Aggregated correlation between Image Metrics and Human Perception with mean and standard deviation across all datasets. Above the dashed line, we list NR, below CR metrics.

Metric	Pearson ↑	Spearman ↑
PAL4VST [61]	$0.078_{\pm 0.112}$	$0.062_{\pm 0.085}$
CNNIQA [13]	$0.144_{\pm 0.247}$	$0.130_{\pm 0.253}$
PIQE [29]	$0.292_{\pm 0.222}$	$0.268_{\pm 0.221}$
PaQ-2-PiQ [53]	$0.402_{\pm 0.178}$	$0.349_{\pm 0.225}$
CrossScore [47]	$0.510_{\pm 0.204}$	$0.378_{\pm 0.209}$
PuzzleSim (ours)	$0.615_{\pm 0.120}$	$0.474_{\pm 0.137}$

5. Application: Progressive Inpainting

Finally, we showcase an application of our metric in automatic restoration of novel views from a reconstructed scene. Whenever it is possible to establish a visual distribution (e.g., we have a training dataset available), we can recursively use our metric to automatically identify visual outliers in novel views and remove them through inpainting. In the Supplementary, we present a quantitative ablation study to show that our metric performs best in this application.

Our Framework We can take a new image $\mathcal{I}$ and employ our *PuzzleSim* metric to obtain the similarity map $\mathcal{S}$.

$$\mathcal{S} = \text{PuzzleSim}(\mathcal{I}) \in \mathbb{R}^{H_\mathcal{I} \times W_\mathcal{I}} \tag{6}$$

To apply neural inpainting, we first need to create a binary mask from the similarity map $\mathcal{S}$, indicating the areas to be inpainted. This involves finding an optimal threshold τ that clearly distinguishes outlier regions. The effectiveness of inpainting depends on carefully setting this mask. If the mask is too large, the inpainting may inadvertently remove clean parts of the scene. If the mask is too small, artifacts might be left untouched. In order to automatically find a balanced threshold, we use a conservative, iterative approach to refine the test image based on the assumption that artifacts have below-average similarity scores. For an initial threshold, we select $N = 50$ candidate values, uniformly spaced between the lowest and mean similarity scores that we use to threshold the similarity map.

$$\begin{aligned} \tau_i &= \min(\mathcal{S}) + \frac{i}{N-1}(\text{mean}(\mathcal{S}) - \min(\mathcal{S})) \\ M_i^{(h,w)} &= \begin{cases} 1 & \text{if } \mathcal{S}^{(h,w)} \leq \tau_i \\ 0 & \text{if } \mathcal{S}^{(h,w)} > \tau_i \end{cases} \end{aligned} \tag{7}$$

with i, h, and w representing indices where $i = 0, \ldots, N-1$, $h = 1, \ldots, H_\mathcal{I}$, and $w = 1, \ldots, W_\mathcal{I}$. With LaMa (big) [37] we generate inpaintings using all N masks and recompute *PuzzleSim* for each option. The quality of each inpainted candidate is evaluated by calculating the average

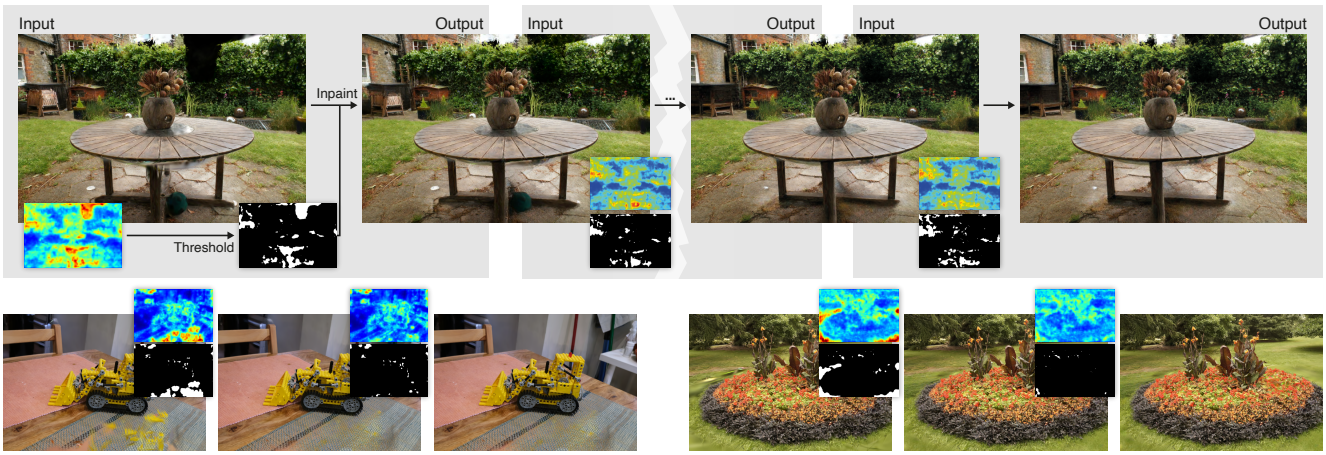


Figure 5. Example showcase of our iterative inpainting application to enhance new views that lack ground-truth correspondences.

similarity difference before and after inpainting, denoted as δ_i. To discourage overly large masks, we add a regularization term that penalizes them. Further details on the mathematical definitions of δ_i and the regularization term are provided in the Supplementary material. We then select the candidate that maximizes δ. After determining the initial threshold, we iteratively refine the inpainted image by sampling new thresholds close to the previous one. The size of this interval depends on a hyperparameter α and the spread of similarity scores. Keeping this range small ensures stable convergence and prevents excessive, disrupting inpainting. If the upper limit of the interval is below the minimum similiarity value $\min_{h,w} \hat{\mathcal{S}}^{(h,w)}$, we revert to the initial sampling method in Eq. (7) as an empty mask would be meaningless and cause division by zero when computing δ_i. Finally, we terminate the process if no further improvement is achieved (i.e., $\max_i \delta_i \leq 0$), returning the final inpainting result. This framework guarantees a monotonic improvement in *PuzzleSim* similarity. In Fig. 5, we showcase several novel views from the reconstructed scenes *garden*, *kitchen*, and *flowers* using only a fraction of the original training views (20-30%). We process this artifact-ridden new view through the iterative inpainting framework presented above. Our method successfully detects and inpaints artifacts in the original reconstruction, producing high-quality inpainting consistent with the distribution of the original scenes.

6. Limitations and Future Work

While our method demonstrates promising results, there are some limitations to consider. Even with our optimized implementation, finding the maximum similarity for a great number of vectors becomes expensive as the number of reference images and image resolution rise (see Supplemental material for implementation and runtime analysis). Performing approximate maximum search or fitting Gaussian mixture models in the embedding space can improve computational performance [12, 60]. Furthermore, our metric is empirically calibrated, but choosing the weights to combine layers and weighting in the channel dimension in a data-driven manner could advance the metric further. The resolution at which our metric can be utilized is currently bound by the CNN backbone's generalizability to higher resolutions. Currently, we can not support Vision Transformer backbones as it takes special care to maintain a limited receptive field. Although our metric is differentiable, it is unlikely to produce valuable gradients due to the max operation across many vectors. Alternatively, softmax operations could be explored to make the metric more suitable for gradient-based optimization.

7. Conclusion

We proposed Puzzle Similarity, a cross-reference image metric for detecting and localizing artifacts in novel views of 3D scene reconstructions. By leveraging learned patch statistics from input views, our method generates spatial artifact maps without requiring ground-truth references, addressing a key challenge in evaluating reconstructed scenes. To enable the evaluation of cross-reference metrics, we also provide a dataset of human-assessed quality and artifact localization for 3D scene reconstruction.

Our evaluation shows that Puzzle Similarity outperforms all tested full-reference, cross-reference and no-reference metrics in capturing artifacts aligned with human perception, demonstrating robustness across diverse artifact types and texture-rich scenes. Furthermore, we apply our metric to automatic image restoration, illustrating its potential to enhance scene reconstruction quality. Puzzle Similarity provides an effective, perceptually aligned, reference-free solution for artifact localization, with promising applications in few-shot reconstruction and guided acquisition.

Acknowledgements We would like to thank Krzysztof Wolski for making their image segmentation tool available to us, Volodymyr Kyrylov for providing the idea and first prototype of the memory-efficient implementation, and Sophie Kergaßner for designing figures. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement N°804226 PERDY), from the Swiss National Science Foundation (SNSF, Grant 200502) and an academic gift from Meta.

References

[1] Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafal K. Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Towards a Quality Metric for Dense Light Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3720–3729, Honolulu, HI, 2017. IEEE. 5

[2] Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D. Fairchild. FLIP: A Difference Evaluator for Alternating Images. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 3(2):1–23, 2020. 2

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields, 2021. 1, 5

[4] Yihang Chen, Qianyi Wu, Mengyao Li, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Fast Feedforward 3D Gaussian Splatting Compression, 2024. arXiv:2410.08017. 1

[5] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-Regularized Optimization for 3D Gaussian Splatting in Few-Shot Images, 2024. arXiv:2311.13398 [cs]. 1

[6] Scott J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, pages 2–15. SPIE, 1992. 2

[7] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Image inpainting through neural networks hallucinations. In *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, Bordeaux, France, 2016. IEEE. 2

[8] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty Quantification for Neural Radiance Fields, 2023. arXiv:2309.03185 [cs]. 1

[9] Kunal Gupta, Milos Hasan, Zexiang Xu, Fujun Luan, Kalyan Sunkavalli, Xin Sun, Manmohan Chandraker, and Sai Bi. MCNeRF: Monte Carlo Rendering and Denoising for Real-Time NeRFs. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, Sydney NSW Australia, 2023. ACM. 2

[10] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics*, 37(6):1–15, 2018. 5

[11] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, 2016. arXiv:1602.07360 [cs]. 3, 4

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. Conference Name: IEEE Transactions on Big Data. 8

[13] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional Neural Networks for No-Reference Image Quality Assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, Columbus, OH, USA, 2014. IEEE. 2, 3, 6, 7

[14] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, Montreal, QC, Canada, 2021. IEEE. 2

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1, 2, 5

[16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4):1–13, 2017. 5

[17] Georgios Kopanas and George Drettakis. Improving NeRF Quality by Progressive Camera Placement for Unrestricted Navigation in Complex Environments, 2023. arXiv:2309.00014 [cs, eess]. 1, 2

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 3, 4

[19] Yixuan Li, Peilin Chen, Hanwei Zhu, Keyan Ding, Leida Li, and Shiqi Wang. Deep Shape-Texture Statistics for Completely Blind Image Quality Evaluation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, page 3694977, 2024. 5

[20] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Visible difference predicator for high dynamic range images. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, pages 2763–2769 vol.3, 2004. ISSN: 1062-922X. 1, 2

[21] Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. FovVideoVDP: a visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics*, 40(4):1–19, 2021. 2

[22] Rafal K. Mantiuk, Dounia Hammou, and Param Hanji. HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content, 2023. arXiv:2304.13625. 2

[23] Rafal K. Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. ColorVideoVDP: A visual difference predictor for image, video and display distortions, 2024. arXiv:2401.11485. 2

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, 2020. arXiv:2003.08934 [cs]. 1, 2

[25] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. Conference Name: IEEE Transactions on Image Processing. 1, 2

[26] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. Conference Name: IEEE Signal Processing Letters. 1

[27] Anush Krishna Moorthy and Alan Conrad Bovik. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011. Conference Name: IEEE Transactions on Image Processing. 2

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. arXiv:2201.05989 [cs]. 1, 2

[29] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, 2015. 3, 6, 7

[30] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty R. Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks, 2021. arXiv:2103.03231. 2

[31] Jim Nilsson and Tomas Akenine-Möller. Understanding SSIM, 2020. arXiv:2006.13846 [eess]. 3, 7

[32] Jean-François Pambrun and Rita Noumeir. Limitations of the SSIM quality metric in the context of diagnostic imaging. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2960–2963, 2015. 3, 7

[33] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. ISSN: 1063-6919. 1

[34] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15 (11):3440–3451, 2006. Conference Name: IEEE Transactions on Image Processing. 5

[35] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015. arXiv:1409.1556 [cs]. 3, 4

[36] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning Persistent 3D Feature Embeddings, 2019. arXiv:1812.01024 [cs]. 2

[37] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182, 2022. ISSN: 2642-9381. 1, 7

[38] Taimoor Tariq, Okan Tarhan Tursun, Munchurl Kim, and Piotr Didyk. Why Are Deep Representations Good Perceptual Quality Features? In *Computer Vision – ECCV 2020*, pages 445–461. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science. 3

[39] Taimoor Tariq, Nathan Matsuda, Eric Penner, Jerry Jia, Douglas Lanman, Ajit Ninan, and Alexandre Chapiro. Perceptually Adaptive Real-Time Tone Mapping. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, Sydney NSW Australia, 2023. ACM. 2

[40] Cara Tursun and Piotr Didyk. Perceptual Visibility Model for Temporal Contrast Changes in Periphery. *ACM Trans. Graph.*, 42(2):20:1–20:16, 2022. 2

[41] Okan Tarhan Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics*, 38(4): 1–14, 2019. 2

[42] Heang K. Tuy and Lee Tan Tuy. Direct 2-D display of 3-D objects. *IEEE Computer Graphics and Applications*, 4(10): 29–34, 1984. Conference Name: IEEE Computer Graphics and Applications. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. arXiv:1706.03762 [cs]. 3

[44] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, pages 2555–2563. AAAI Press, 2023. 3

[45] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402 Vol.2, 2003. 2

[46] Zhou 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. Conference Name: IEEE Transactions on Image Processing. 2

[47] Zirui Wang, Wenjing Bian, and Victor Adrian Prisacariu. CrossScore: Towards Multi-View Image Evaluation and Scoring. In *Computer Vision – ECCV 2024*, pages 492–510, Cham, 2025. Springer Nature Switzerland. 1, 3, 6, 7

[48] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing Ghostly Artifacts from Casually Captured NeRFs, 2023. arXiv:2304.10532 [cs]. 1

[49] Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter

Seidel, Anthony Steed, and Rafał K. Mantiuk. Dataset and Metrics for Predicting Local Visible Differences. *ACM Trans. Graph.*, 37(5):172:1–172:14, 2018. 5

[50] Muyu Xu, Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Xiaoqin Zhang, Christian Theobalt, Ling Shao, and Shijian Lu. WaveNeRF: Wavelet-based Generalizable Neural Radiance Fields, 2023. arXiv:2308.04826. 2

[51] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without Human Scores for Blind Image Quality Assessment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, Portland, OR, USA, 2013. IEEE. 2

[52] Peng Ye, Jayant Kumar, and David Doermann. Beyond Human Opinion Scores: Blind Image Quality Assessment Based on Synthetic Scores. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4241–4248, 2014. ISSN: 1063-6919. 2

[53] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3572–3582, Seattle, WA, USA, 2020. IEEE. 2, 3, 6, 7

[54] Junyong You and Jari Korhonen. Transformer for Image Quality Assessment, 2021. arXiv:2101.01097 [cs]. 2

[55] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks, 2021. arXiv:2112.05131 [cs]. 2

[56] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for Real-time Rendering of Neural Radiance Fields, 2021. arXiv:2103.14024 [cs]. 2

[57] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images, 2021. arXiv:2012.02190 [cs]. 1

[58] Sergey Zagoruyko and Nikos Komodakis. Learning to Compare Image Patches via Convolutional Neural Networks. pages 4353–4361, 2015. 6

[59] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. Conference Name: IEEE Transactions on Image Processing. 2

[60] Lin Zhang, Lei Zhang, and Alan C. Bovik. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. Conference Name: IEEE Transactions on Image Processing. 5, 8

[61] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual Artifacts Localization for Image Synthesis Tasks. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7545–7556, Paris, France, 2023. IEEE. 3, 6, 7

[62] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. pages 586–595, 2018. 2, 3, 7

[63] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14071–14081, Vancouver, BC, Canada, 2023. IEEE. 3