Puzzle Similarity: A Perceptually-guided No-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 detailed artifact maps. The absence of such quality metrics hinders accurate predictions of the quality of generated views and limits the adoption of post-processing techniques, such as inpainting, to enhance reconstruction quality. In this work, we propose a new no-reference metric, Puzzle Similarity, which is designed to localize artifacts in novel views. Our approach utilizes image patch statistics from the input views to establish a scene-specific distribution that is later used to identify poorly reconstructed regions in the novel views. We test and evaluate our method in the context of 3D reconstruction; to this end, we collected a novel dataset of human quality assessment in unseen reconstructed views. Through this dataset, we demonstrate that our method can not only successfully localize artifacts in novel views, correlating with human assessment, but do so without direct references. Surprisingly, our metric outperforms both no-reference metrics and popular full-reference image metrics. We can leverage our new metric to enhance applications like automatic image restoration, guided acquisition, or 3D reconstruction from sparse inputs.

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 using simple triangulation and epipolar geometry through methods such as structure from motion (SfM) to produce sparse point-clouds of diffuse color [28]. Densifying these representations can be done explicitly [13]. Alternatively, one can learn continuous, implicit representations [3, 22, 25], normally modeled through some kind of multi-layer perceptron (MLP). 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 as well [4, 5, 38, 44]; they normally leverage learned priors on large datasets, helping to fill in the gaps by enforcing 3D consistency and that resulting reconstructions follow natural image statistics. However, quantifying the quality of novel views from these reconstructions is still problematic. These views can contain artifacts due to the sparsity of the dataset their model was constructed from, and automatically identifying them helps with restoration (e.g. masking for image-based inpainters [32]) or simply to guide future data acquisition to fill the gaps [15]. Recent works have followed a bayesian approach to quantification of the uncertainty of an area belonging or not to a reconstructed model or scene [7], which could potentially be leveraged for simple artifact detection; however, they require implicit representations of the scene, with fundamental changes to the scene model, and are not practical for more general applications that require visual artifact identification outside of scene reconstruction, as well as 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. As opposed to visual difference predictors (VDPs) [18] (which require references) and no-reference quality metrics [23, 24] (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 embdedded statistics of new images from a similar distribution (i.e. novel reconstructed views from a 3DGS [13] representation) to obtain artifact maps without direct 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 both no-reference and full-reference metrics. To summarize, our contributions are the following:

- A novel no-reference visual quality/artifact identification metric particularly tailored for image-based rendering,
- a novel dataset of human-assessed artifact and distortion identification to validate our metric,
- and an application on image restoration and 3D reconstruction enhancement that showcases the utility of our approach.

2. Related Work

3D reconstruction and Image-based rendering Reconstructing 3D objects or scenes from sparse sets of 2D observations is a fundamental problem in vision [18]. 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 however, differ either on 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-layered perceptron (MLP) [22, 31]. Rendering is usually done via sampling the implicit volume using ray-marching [36], 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 [8, 25, 27] or by distilling the implicit space into explicit density and anisotropic appearance volumes [39, 43]. On the other hand, purely explicit models do not require any pre-training using implicit functions, and were originally Eulerian in nature [42]. 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 [13]. However, these explicit methods have introduced some limitations of their own along the way. Methods like 3D Gaussian Splatting [13] 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 [6]) or simply automatically guide further image acquisition to complete the dataset efficiently [15].

Image Quality Metrics (IQMs) We can roughly categorize IQMs on their reliance on reference images, or the lack of it. Full-reference techniques traditionally input a reference image and its distorted counterpart and compute pixel or patch wise differences, which can either be pooled across an image to provide a single quality value [37, 49] or in some cases, kept as a pixel-wise map to visualize the location or source of the distortions. Classic examples also include mean-absolute error (MAE) or mean-squared error (MSE). Pixel-map metrics share similarities with visual difference predictors (VDPs), with the distinction usually being that VDPs leverage some explicit or implicit model of the human visual system (HVS) to align distortion predictions with human assessment. In some cases, it is also possible to transform usually pooled metrics like SSIM [37], FSIM [46] and LPIPS [49] into pixel maps, and we show how in Section 4.1.

No-reference visual quality metrics relinquish the necessity of direct references, and are most commonly learned, often relying on mean-opinion based supervised learning. Alternatively to leveraging datasets on human quality assessment [12, 24], they can also be supervised on extracted features from natural image statistics [23, 40], or even synthetic scores [41]. They are not usually capable of producing spatial quality maps, which makes it impossible to determine the source of reduced quality or localize the potential artifacts.

Some IQMs however, are capable of producing visual maps as well [12, 26, 40]; this makes them the closest to our work. The most relevant work in this space, PIQE [26] similarly attempts to quantify visual artifacts without relying on any supervision; rather, it extracts local features from image patches and quantifies quality as a measure of distortions in the patch, in accordance with a low-level human vision model. In contrast, our work does not make any assumptions on the HVS, rather leveraging the latent space of a model pretrained on natural images to measure the euclidean distance 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). This enables us to obtain a higher level of alignment with human assessment.

Visual Difference Predictors (VDPs) VDPs usually differ from IQMs in the more explicit integration of models of human vision, leveraging the naturally compressive nature of the human visual system [18] to predict perceived differences between images. Improvements over the original framework have increased its robustness by extending their applicability to high dynamic range imagery [20]; making them eccentricity and motion-aware [19]; and integrating perceived color [21]. VDP frameworks have seen use in foveated rendering applications [35] and perceptually-aware tone mappers [34]. In contrast to VDPs, our model relinquishes the necessity of direct references, while performing similarly to state of the art VDPs such as FovVideoVDP [19].

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 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 Figure 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. Inspired by the close correlation between human quality judgment and latent CNN feature maps [33, 49], we employ a pre-trained convolutional network [10, 16, 30] to embed all the references, computing similarity in the latent feature space. Note that comparing feature map ”pixels” in a convolutional network 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 Figure 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 Figure 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, so we will describe the steps once for one layer ℓ. To find the similarity $s_\ell(x, y)$ of 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:

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

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 S_ℓ.

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

where $\mathcal{I}$ is the test image. We repeat this for a set of layers and combine them into a final quality 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:

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

with $\sum_\ell w_\ell = 1$ and reference images $\mathcal{I}^{1:N}_{\text{ref}}$. 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:

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

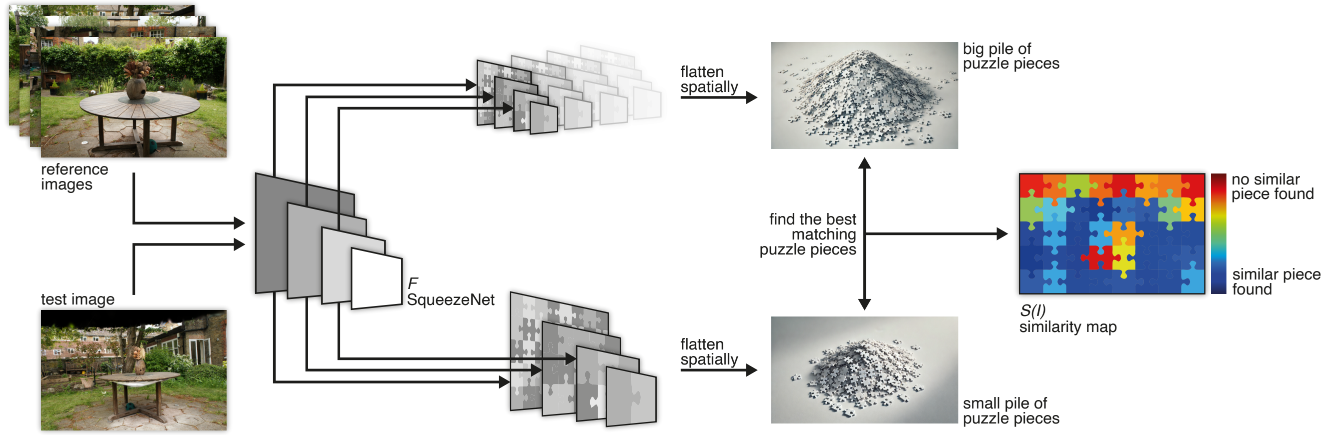


Figure 1. Schematic representation of our Puzzle Similarity metric.

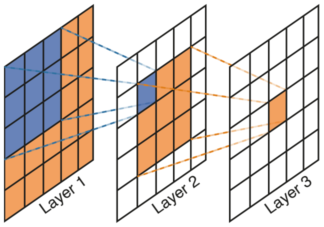


Figure 2. Receptive field of a multi-layer CNN. Note how one pixel in the last layer can be seen as an embedded patch of the input space.

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` [10, 16, 30]. Some of the critical considerations are model complexity and memory requirements, which we summarized in the Supplemental, as well as their specific tested performance in our human assessment 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, with the key advantage of doing so at a substantially reduced computational cost. We opted for `SqueezeNet` 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 stacks against competing approaches for both full-reference and no-reference visual map prediction in the context of reconstruction and image-based rendering.

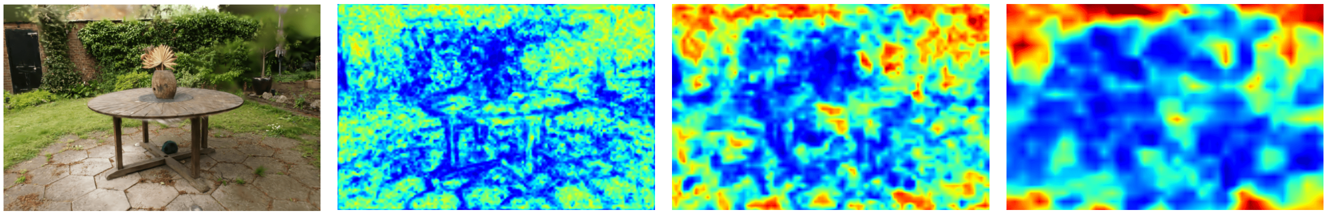
In order to quantify the correlation between all these maps and human assessment, we present a novel dataset on human artifact identification, which we manually collected and can be found here [1]to facilitate future research on the topic.

As for our method, for each different scene, we compute embeddings on their respective training dataset and leverage each of them respectively to compute visual quality maps on validation and artifact-ridden views, as explained in Section 3.

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. In order to generate images with typical reconstruction artifacts, we run 3D Gaussian Splatting [13] on twelve scenes from Mip-NeRF360 [3], Tank and Temples [14] and Deep Blending [9] datasets, using default parameters but withholding substantial amounts of training images. Omitting training views increases the chance of artifacts appearing in those withheld views, while still allowing us to have clean references. For each dataset,

[1] www.placeholder_fake_link_for_dataset.com



(a) Rendering

(b) 2nd VGG layer

(c) 3rd VGG layer

(d) 4th VGG layer

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

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 23 participants to segment visible artifacts in each of the 36 sample images under controlled viewing conditions using the tool developed by Wolski et al. [1], which the authors kindly provided. We include details on the participants' self-reported gender and age distributions in Supplementary material, 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. Figure 4 shows example renderings (a) alongside metric predictions (b)-(d) and their average human-produced mask (e).

We evaluate our method against both full-reference and no-reference IQMs, and the state-of-the-art VDP. To assess their alignment with human perception, we will correlate their maps with the human ratings from our dataset, which we described in Section 4.1.

4.2. Comparison with Full-Reference Metrics

Adapting pooled metrics to produce spatial quality maps In terms of single-valued quality metrics, we selected commonly used reconstruction quality metrics (L1, L2, SSIM [37], LPIPS [49]). We adapt L1, L2 and SSIM by simply removing the pooling step. For LPIPS, after computing the distance in the embedding space, the metric already upsamples each feature map back to the original image resolution. We can then pool across maps but not along 2D image dimensions, allowing us to retain the map. We compare LPIPS with three base models: `VGG`, `AlexNet`, and `SqueezeNet`. We further compare with ꟻLIP [2], a commonly found metric in quality assessment, particularly in the field of physically-based rendering; it already produces spatial quality maps by default. The no-reference metric CNNIQA [12] was applied on patches as described in their paper. We applied padding to avoid cropping the borders and upsampled the final map. PIQE [26] already produces three different kinds of maps that we averaged.

To correlate them to our human ratings, we first compute each metric map for each sample 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, 17, 29, 47]:

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

where x is an individual quality score and $a_{1\dots5}$ are the tunable parameters optimized using gradient ascent to maximize either PCC or SRCC.

Results Puzzle Similarity consistently achieves high correlation with human-perceived artifacts across all datasets, outperforming well-established full-reference (FR) metrics such as SSIM, L1, and L2 norms, *even without access to direct ground-truth correspondences*. Tables 1 & 2 report the average Pearson and Spearman correlations for each scene. Our dataset includes three images per scene, with artifact types varying per scene. For 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 maintaining similar color distributions to the ground truth. In these more subtle cases, Puzzle Similarity significantly outperforms FR metrics. We include extended results and artifact examples in our Supplemental.

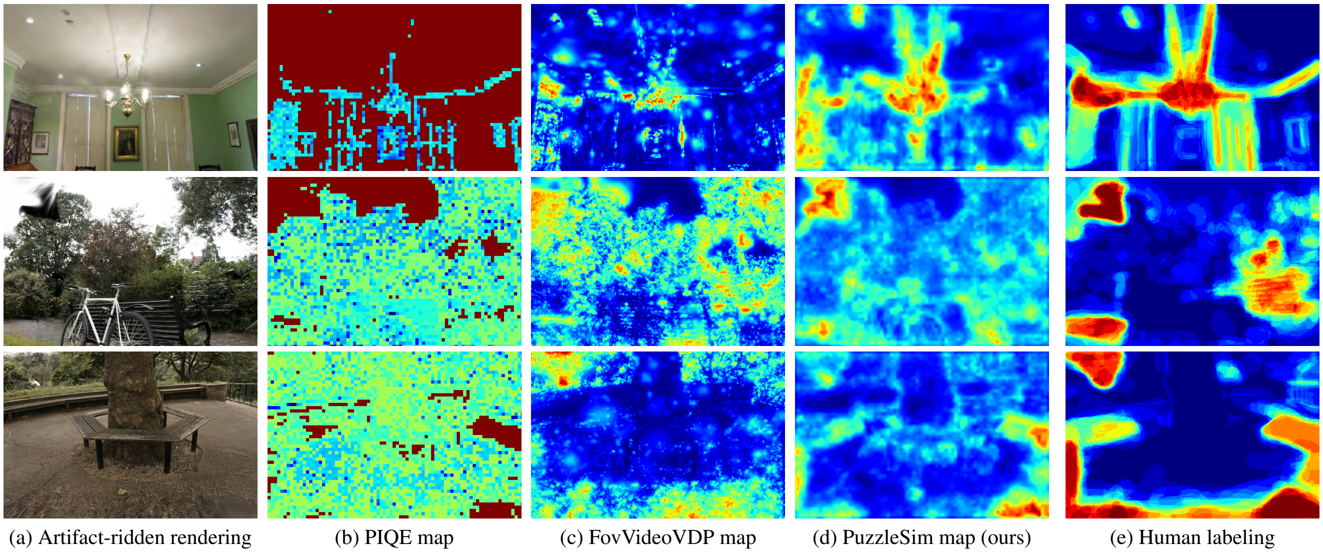


Figure 4. Selection of image quality maps for artifact-ridden renderings from various scenes. The last column shows ground-truth human assessments from our collected dataset.

Table 1. Pearson Correlation between Image Metrics (full-reference and no-reference (NR)) and Human Perception per Dataset

		bicycle	bonsai	counter	drjohnson	flowers	garden	kitchen	playroom	stump	train	treehill	truck
Full-Reference	L1	0.199	0.325	0.384	0.280	0.078	0.630	0.533	0.556	0.141	0.314	0.114	0.314
	L2	0.163	0.293	0.232	0.166	0.085	0.546	0.396	0.444	0.123	0.220	0.081	0.223
	SSIM [37]	0.188	0.325	0.370	0.342	0.251	0.603	0.596	0.536	0.231	0.468	0.343	0.490
	ꟻLIP [2]	0.268	0.403	0.452	0.366	0.194	0.719	0.684	0.652	0.239	0.319	0.142	0.402
	LPIPS (vgg)) [49]	0.112	0.190	0.119	0.275	0.216	0.307	0.212	0.183	0.073	0.264	0.111	0.230
	LPIPS (alex)) [49]	0.059	0.122	0.383	0.103	0.134	0.528	0.247	0.177	0.102	0.322	0.221	0.189
	LPIPS (squeeze)) [49]	0.108	0.289	0.256	0.263	0.169	0.211	0.304	0.395	0.158	0.132	0.153	0.139
	FovVideoVDP [19]	0.370	0.457	0.561	0.470	0.310	0.659	0.826	0.752	0.370	0.584	0.339	0.594
NR	CNNIQA [12]	0.075	0.167	0.239	0.068	0.158	0.408	0.322	0.345	0.366	0.400	0.131	0.014
	PIQE [26]	0.265	0.266	0.254	0.116	0.586	0.443	0.490	0.151	0.527	0.202	0.375	0.100
	PuzzleSim (ours)	0.584	0.527	0.567	0.483	0.590	0.637	0.745	0.662	0.460	0.597	0.702	0.555

4.3. Comparison with Visual Difference Predictors

The explicit model of low-level human vision in VDP models usually produces strong correlations between the metric and human assessment. We compare against the state-of-the-art FovVideoVDP [19]; while the metric can take into account higher-order perceptual cues like motion and eccentricity, we disable them as we are 1) analyzing single static frames and 2) not assuming specific viewing conditions such as display size or distance to the screen. Puzzle Similarity not only matches but often surpasses FovVideoVDP, particularly in texture-rich scenes like *treehill*, *stump*, and *flowers*, again while relinquishing direct references.

4.4. Comparison with No-Reference Metrics

Closer to our work, we can compare with other no-reference quality metrics capable of producing spatial quality maps [12, 45]. Puzzle Similarity demonstrates superior accuracy in artifact localization, as shown by the correlation values in Table 3. While PIQE performed well in certain scenes, its overall correlation with human perception was generally lower in others, reflecting a less robust alignment with perceived artifact localization. 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), metrics like PIQE do not require any extra images and attempt to generalize to any input. Still, in the context of 3D reconstruction, our

Table 2. Spearman Correlation between Image Metrics (full-reference and no-reference (NR)) and Human Perception per Dataset.

		bicycle	bonsai	counter	drjohnson	flowers	garden	kitchen	playroom	stump	train	treehill	truck
Full-Reference	L1	0.151	0.237	0.231	0.209	0.051	0.391	0.347	0.501	0.118	0.298	0.191	0.288
	L2	0.155	0.249	0.234	0.218	0.054	0.402	0.353	0.513	0.122	0.296	0.192	0.292
	SSIM [37]	0.160	0.120	0.252	0.413	0.244	0.540	0.273	0.496	0.279	0.468	0.402	0.435
	ꟻLIP [2]	0.205	0.335	0.250	0.272	0.203	0.466	0.438	0.564	0.210	0.294	0.216	0.365
	LPIPS (vgg)) [49]	0.093	0.097	0.081	0.281	0.207	0.155	0.138	0.159	0.051	0.236	0.099	0.171
	LPIPS (alex)) [49]	0.027	0.093	0.169	0.091	0.106	0.238	0.168	0.094	0.048	0.240	0.233	0.132
	LPIPS (squeeze)) [49]	0.057	0.180	0.158	0.233	0.137	0.065	0.162	0.250	0.167	0.101	0.133	0.064
	FovVideoVDP [19]	0.342	0.312	0.454	0.422	0.295	0.534	0.646	0.735	0.382	0.555	0.364	0.475
NR	CNNIQA [12]	0.085	0.205	0.221	0.163	0.131	0.254	0.373	0.379	0.316	0.397	0.221	0.095
	PIQE [26]	0.208	0.409	0.292	0.182	0.459	0.229	0.620	0.195	0.376	0.225	0.277	0.173
	PuzzleSim (ours)	0.468	0.392	0.382	0.501	0.428	0.428	0.658	0.603	0.306	0.541	0.548	0.441

Table 3. Aggregated correlation between Image Metrics and Human Perception with mean and standard deviation across all datasets.

	Metric	Pearson ↑	Spearman ↑
Full-Reference	L1	0.322$_{\pm 0.203}$	0.251$_{\pm 0.157}$
	L2	0.248$_{\pm 0.175}$	0.257$_{\pm 0.158}$
	SSIM [37]	0.395$_{\pm 0.166}$	0.340$_{\pm 0.168}$
	ꟻLIP [2]	0.403$_{\pm 0.216}$	0.318$_{\pm 0.150}$
	LPIPS (vgg) [49]	0.191$_{\pm 0.147}$	0.147$_{\pm 0.114}$
	LPIPS (alex) [49]	0.216$_{\pm 0.169}$	0.137$_{\pm 0.107}$
	LPIPS (squeeze) [49]	0.215$_{\pm 0.148}$	0.142$_{\pm 0.097}$
	FovVideoVDP [19]	0.524$_{\pm 0.191}$	0.460$_{\pm 0.165}$
NR	CNNIQA [12]	0.224$_{\pm 0.174}$	0.237$_{\pm 0.153}$
	PIQE [26]	0.314$_{\pm 0.188}$	0.304$_{\pm 0.167}$
	PuzzleSim (ours)	0.592$_{\pm 0.119}$	0.475$_{\pm 0.137}$

metric performs better than all tested no-reference metrics.

5. Application: Progressive Automatic Artifact Inpainting

Finally, we will showcase a possible application of our metric in automatic restoration of novel images 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.

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

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

To apply neural inpainting, we first need to create a binary mask from the quality map $\mathcal{Q}$, indicating the areas to be inpainted. This involves finding an optimal threshold τ that clearly distinguishes artifact regions. The effectiveness of inpainting depends on setting this mask carefully. 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 quality scores. For an initial threshold, we select $N = 50$ candidate values, uniformly spaced between the lowest and mean quality scores that we use to threshold the quality map.

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

with i, h, and w representing indices where $i = 0, \ldots, N-1$, $h = 1, \ldots, H_{\mathcal{I}}$, and $w = 1, \ldots, W_{\mathcal{I}}$. 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 quality difference within the inpainted region, 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 drawing new thresholds close to the previous one. For each new threshold, we repeat the selection

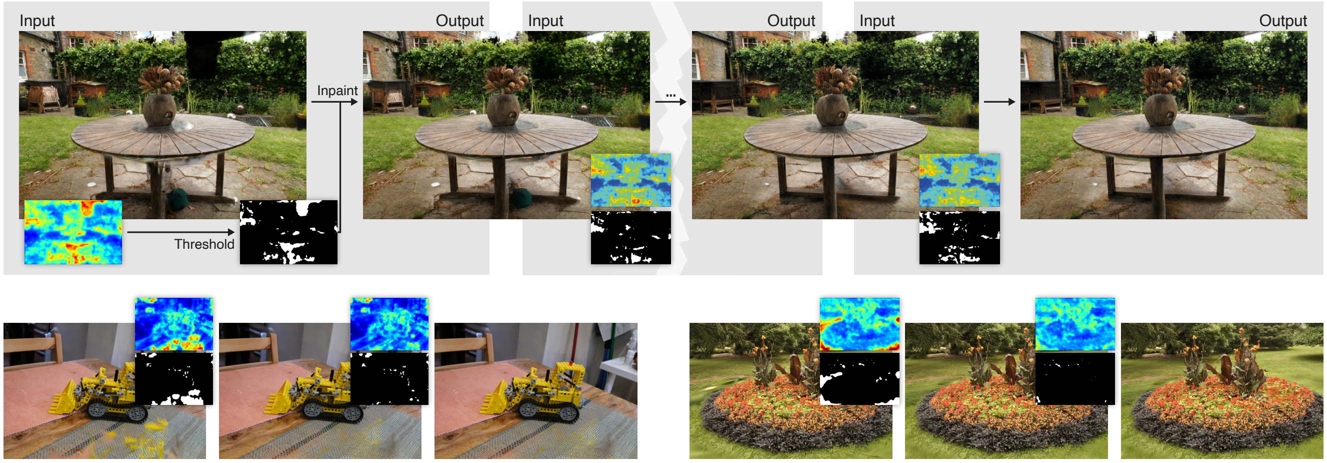


Figure 5. Example showcase of our iterative inpainting application to enhance new views that lack ground-truth correspondences.

procedure, but with fewer candidates sampled within a narrow range around the previous threshold. The size of this interval depends on a hyperparameter α and the spread of quality scores. Keeping this range small ensures stable convergence and prevents excessive inpainting that could disrupt scene geometry.

If the upper limit of the interval is below the minimum quality value $\min_{h,w} \hat{\mathcal{Q}}^{(h,w)}$, we revert to the initial sampling method in Eq. 7 since 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* quality.

In Figure 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 several limitations to consider. Finding the maximum similarity with many other vectors becomes expensive as the number of reference images and image resolution rises. Performing approximate maximum search or fitting Gaussian mixture models in the embedding space can improve computational performance [11, 48]. 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 limited by the CNN backbone's generalizability to higher resolutions. Although it is differentiable, it is unlikely to produce valuable gradients due to the max operation across many vectors. The next step would be to explore softmax alternatives to make the metric more suitable for gradient-based optimization.

7. Conclusion

In this work, we have introduced Puzzle Similarity, a no-reference visual quality metric designed to detect and localize artifacts in novel views generated by 3D scene reconstruction methods. By leveraging learned patch statistics from input views, our method can generate spatial artifact maps without needing ground-truth references, which is a significant advantage for evaluating reconstructed scenes. Furthermore, we have provided a novel dataset of human-assessed quality and artifact detection specifically tailored for 3D scene reconstruction approaches.

Our evaluation demonstrates that Puzzle Similarity outperforms traditional full-reference metrics, such as SSIM and L2 norms, as well as spatial quality no-reference metrics, like PIQE, in capturing artifacts that align with human perception. Compared to sophisticated visual difference predictors like FovVideoVDP, which leverage explicit fitted models of low-level human vision, Puzzle Similarity achieves comparable or superior performance in complex texture-rich scenes, proving its robustness across a range of artifact types and scenarios, while relinquishing the necessity of direct references.

Additionally, We demonstrate our metric applied to the problem of automatic artifact inpainting, highlighting its potential for enhancing quality in scene reconstructions. Overall, Puzzle Similarity offers an effective solution for artifact localization in 3D reconstruction, paving the way for more perceptually aligned, reference-free quality assessment in computer vision and graphics, and we are hopeful it will power many more downstream applications in the future, such as 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. 5, 6, 7

[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, 4

[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] 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

[7] 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

[8] 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

[9] 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. 4

[10] 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

[11] 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

[12] 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, 5, 6, 7

[13] 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, 4

[14] 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. 4

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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. 3, 6, 7

[20] 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. 3

[21] 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. 3

[22] 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

[23] 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

[24] 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, 2

[25] 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

[26] 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. 2, 5, 6, 7

[27] 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

[28] 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

[29] 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

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

[31] 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

[32] 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

[33] 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

[34] 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. 3

[35] 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. 3

[36] 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

[37] 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, 5, 6, 7

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

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

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

[46] 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

[47] 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

[48] 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. 8

[49] 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, 5, 6, 7