

Supplementary Material for: New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts

Martin Čadík* Robert Herzog* Rafał Mantiuk[◇] Karol Myszkowski* Hans-Peter Seidel*
*MPI Informatik Saarbrücken, Germany [◇]Bangor University, United Kingdom

1 Introduction

In this supplementary material we provide additional images, tables, experiment instruction and other results that were left out from the main publication for brevity. Please refer also to the HTML supplement, that is best suited for online viewing¹.

Image #	with-reference $\bar{u}; \bar{u}_{mask}$	no-reference $\bar{u}; \bar{u}_{mask}$
#1 – apartment	0.980; 0.637	0.890; 0.526
#2 – cg figures	0.933; 0.499	0.830; 0.251
#3 – disney	0.867; 0.367	0.842; 0.458
#4 – kitchen	0.698; 0.386	0.714; 0.378
#5 – red kitchen	0.857; 0.370	0.852; 0.448
#6 – sponza above tree	0.896; 0.188	0.856; 0.526
#7 – sponza arches	0.914; 0.390	0.905; 0.465
#8 – sponza atrium	0.928; 0.314	0.818; 0.425
#9 – sponza tree shadows	0.809; 0.209	0.832; 0.457
#10 – sponza trees	0.974; 0.570	0.933; 0.666
Averages	0.886; 0.393	0.847; 0.460
#11 – apartment II	0.823; 0.229	0.860; 0.478
#12 – atrium	0.701; 0.333	0.759; 0.458
#13 – bathroom	0.446; 0.300	0.408; 0.257
#14 – buddha	0.770; 0.347	0.926; 0.618
#15 – chairs	0.856; 0.357	0.775; 0.397
#16 – city-d	0.895; 0.725	0.796; 0.628
#17 – city-u	0.558; 0.346	0.477; 0.408
#18 – cornell	0.586; 0.291	0.809; 0.419
#19 – dragons	0.596; 0.376	0.463; 0.176
#20 – hall	0.995; 0.759	0.821; 0.612
#21 – icido	0.911; 0.405	0.886; 0.451
#22 – kitchen II	0.841; 0.464	0.872; 0.519
#23 – livingroom	0.645; 0.407	0.525; 0.406
#24 – mpii	0.952; 0.315	0.892; 0.599
#25 – plants-d	0.834; 0.330	0.801; 0.596
#26 – plants-u	0.702; 0.230	0.682; 0.278
#27 – room	0.800; 0.439	0.758; 0.574
#28 – sala	0.535; 0.477	0.640; 0.465
#29 – sanmiguel	0.927; 0.532	0.880; 0.595
#30 – sanmiguel cam3	0.925; 0.423	0.865; 0.585
#31 – sanmiguel cam4	0.779; 0.466	0.829; 0.429
#32 – sibenik	0.853; 0.626	0.951; 0.706
#33 – sponza	0.915; 0.499	0.848; 0.671
#34 – tt	0.786; 0.341	0.822; 0.418
#35 – villa cam1	0.849; 0.373	0.809; 0.456
#36 – villa cam2	0.663; 0.386	0.749; 0.482
#37 – villa cam3	0.865; 0.321	0.873; 0.500
Averages	0.778; 0.411	0.769; 0.488

Table 2: Kendall’s u coefficients of agreement between observers. The range of values is $[-0.11, 1]$ and $[-0.04, 1]$ for EG’12 and the new dataset, respectively.

*e-mail: mcadik@mpi-inf.mpg.de

¹The complete dataset is available at <http://www.mpi-inf.mpg.de/resources/hdr/1qm-evaluation/>

Image #	AD	HDR-VDP-2	MS-SSIM	SSIM	sCIE-Lab	sCorrel
#1 – apartment	NaN	NaN	NaN	NaN	NaN	NaN
#2 – cg figures	0.972	0.978	0.979	0.960	0.958	0.869
#3 – disney	0.853	0.959	0.943	0.910	0.907	0.794
#4 – kitchen	0.755	0.764	0.689	0.746	0.744	0.588
#5 – red kitchen	0.605	0.835	0.628	0.681	0.557	0.707
#6 – sponza above tree	0.896	0.966	0.973	0.950	0.963	0.880
#7 – sponza arches	0.808	0.859	0.859	0.876	0.892	0.810
#8 – sponza atrium	0.880	0.950	0.963	0.947	0.979	0.851
#9 – sponza tree shadows	0.827	0.881	0.920	0.884	0.916	0.822
#10 – sponza trees	0.995	0.988	0.992	0.995	0.995	0.815
Overall AUC	0.834	0.891	0.871	0.873	0.867	0.782
#11 – apartment II	0.977	0.938	0.961	0.970	0.980	0.982
#12 – atrium	0.956	0.928	0.894	0.933	0.951	0.931
#13 – bathroom	0.930	0.858	0.931	0.906	0.933	0.946
#14 – buddha	0.822	0.794	0.743	0.782	0.819	0.825
#15 – chairs	0.975	0.924	0.965	0.936	0.965	0.964
#16 – city-d	NaN	NaN	NaN	NaN	NaN	NaN
#17 – city-u	0.911	0.803	0.852	0.851	0.839	0.891
#18 – cornell	0.820	0.879	0.833	0.846	0.818	0.781
#19 – dragons	0.770	0.596	0.541	0.677	0.676	0.840
#20 – hall	NaN	NaN	NaN	NaN	NaN	NaN
#21 – icido	0.951	0.972	0.894	0.961	0.953	0.941
#22 – kitchen II	0.921	0.813	0.896	0.884	0.936	0.882
#23 – livingroom	0.967	0.979	0.979	0.983	0.943	0.818
#24 – mpii	0.983	0.974	0.985	0.968	0.984	0.967
#25 – plants-d	0.901	0.839	0.879	0.895	0.892	0.887
#26 – plants-u	0.941	0.903	0.947	0.888	0.944	0.924
#27 – room	0.770	0.806	0.727	0.787	0.792	0.703
#28 – sala	0.390	0.293	0.435	0.450	0.757	0.896
#29 – sanmiguel	NaN	NaN	NaN	NaN	NaN	NaN
#30 – sanmiguel cam3	0.977	0.921	0.987	0.894	0.986	0.983
#31 – sanmiguel cam4	0.774	0.653	0.729	0.742	0.770	0.946
#32 – sibenik	NaN	NaN	NaN	NaN	NaN	NaN
#33 – sponza	0.964	0.930	0.988	0.994	0.989	0.886
#34 – tt	0.928	0.866	0.933	0.940	0.939	0.816
#35 – villa cam1	0.952	0.873	0.844	0.926	0.935	0.953
#36 – villa cam2	0.576	0.608	0.586	0.635	0.573	0.804
#37 – villa cam3	0.938	0.875	0.928	0.937	0.939	0.947
Overall AUC	0.817	0.751	0.723	0.782	0.793	0.886

Table 3: Areas under ROC curves (at threshold 0.5) in with-ref experiment for HDR-VDP-2, SSIM, MS-SSIM, sCIE-Lab, sCorrel and Absolute Difference. The last row shows the overall AUC over each test set. In case there is no region marked by more than 50% of observers, the AUC is not available (NaN).

2 Experiment instruction

In this section we show the experiment instruction sheets presented to the observers prior to the experiment, as well as the questionnaire filled by them after the experiment.

No-reference Experiment: “Thank you for agreeing to participate in the experiment. The experiment will take approximately 30min to complete. Please, wear your prescription glasses, if you would normally wear them to work with a computer or if they improve your vision. Please read the following instructions carefully before starting the experiment. The purpose of the experiment is to measure the visibility of computer graphics distortions. These distortions reveal itself as blurry, blotchy or noisy areas in an image. They could be also shadows or lighting that looks unrealistically. During the experiment, you will see 27 images containing different types of distortion, some distortions may be less and some more noticeable. There may be various types of distortions present in one image, and also the severity of the distortion may vary. During the experiment, you will see one image on the screen at a time. Your task is to mark (paint) all the areas in the image where you can see the distortion using the brush. The brush is controlled with a mouse

Image #	no-ref _{avg}	AD	VDP2	MS-SSIM	SSIM	sCIE-Lab	sCorrel
#1 – apartment	0.219	0.363	-0.013	0.321	0.354	0.421	-0.042
#2 – cg figures	0.784	0.582	0.670	0.640	0.553	0.558	0.130
#3 – disney	0.739	0.389	0.630	0.471	0.437	0.410	0.253
#4 – kitchen	0.835	0.332	0.324	0.291	0.336	0.348	0.021
#5 – red kitchen	0.836	0.134	0.267	0.200	0.209	0.101	0.276
#6 – sponza above tree	0.558	0.497	0.567	0.631	0.550	0.588	0.429
#7 – sponza arches	0.768	0.255	0.302	0.285	0.357	0.346	0.230
#8 – sponza atrium	0.780	0.480	0.447	0.474	0.586	0.650	0.440
#9 – sponza tree shadows	0.669	0.435	0.204	0.409	0.482	0.499	0.343
#10 – sponza trees	0.122	0.499	0.564	0.587	0.539	0.534	0.428
Overall	0.637	0.294	0.387	0.371	0.373	0.355	0.194
#11 – apartment II	0.737	0.556	0.187	0.397	0.452	0.611	0.651
#12 – atrium	0.866	0.755	0.653	0.593	0.697	0.719	0.616
#13 – bathroom	0.820	0.699	0.527	0.683	0.674	0.686	0.759
#14 – buddha	0.102	0.325	-0.074	0.339	0.312	0.304	0.207
#15 – chairs	0.828	0.653	0.377	0.606	0.577	0.571	0.531
#16 – city-d	0.073	0.090	0.041	0.059	0.067	0.057	0.082
#17 – city-u	0.872	0.738	0.472	0.631	0.613	0.603	0.707
#18 – cornell	0.852	0.471	0.560	0.493	0.529	0.466	0.438
#19 – dragons	0.924	0.302	-0.153	-0.361	0.012	-0.030	0.430
#20 – hall	-0.116	0.119	0.121	-0.187	0.063	0.115	0.123
#21 – icido	0.791	0.307	0.549	0.098	0.367	0.323	0.381
#22 – kitchen II	0.638	0.469	0.181	0.395	0.372	0.537	0.387
#23 – livingroom	0.848	0.806	0.711	0.728	0.795	0.724	0.392
#24 – mpii	0.872	0.601	0.383	0.602	0.515	0.620	0.330
#25 – plants-d	0.628	0.669	0.620	0.682	0.632	0.661	0.663
#26 – plants-u	0.841	0.607	0.670	0.748	0.642	0.707	0.626
#27 – room	0.380	0.184	0.292	0.095	0.224	0.200	-0.182
#28 – sala	0.665	-0.215	-0.029	-0.247	-0.127	0.201	0.556
#29 – sanmiguel	0.609	0.393	-0.159	0.217	0.235	0.236	0.478
#30 – sanmiguel cam3	0.666	0.661	0.189	0.720	0.517	0.701	0.625
#31 – sanmiguel cam4	0.841	-0.179	-0.035	0.112	0.120	0.175	0.589
#32 – sibenik	-0.086	-0.101	0.089	-0.203	-0.078	-0.103	0.115
#33 – sponza	0.581	0.522	0.506	0.670	0.687	0.667	0.383
#34 – tt	0.836	0.440	0.179	0.102	0.276	0.295	0.320
#35 – villa cam1	0.848	0.501	0.226	0.267	0.413	0.407	0.619
#36 – villa cam2	0.715	0.097	0.032	0.123	0.181	0.097	0.489
#37 – villa cam3	0.829	0.362	0.099	0.364	0.400	0.369	0.431
Overall	0.633	0.254	0.128	0.077	0.170	0.186	0.453

Table 4: Spearman correlations of subjective responses in with-reference experiment with subjective responses in no-reference experiment and with the predictions of HDR-VDP-2, SSIM, MS-SSIM, sCIE-Lab, sCorrel and Absolute Difference. The last row shows the overall correlations over the whole test set. The best correlations (excluding the no-reference subjective experiment) for each stimulus are printed in bold.

2.1 Post-experiment questionnaire

Localized quality assessment experiment

Date: March, April 2012

Experiment: no-reference with-reference

Participant code:

Age:

Gender:

Rate your experience with computer graphics methods (tick your answer)

no experience Moderate experience Very experienced
 I use computer graphics I research computer
 rendering software graphics methods

Comments:

scene ID:

comment:

Thank you!

3 Additional description of stimuli

In the following we will summarize the different distortions we have encountered and briefly describe their source and nature with respect to the corresponding rendering algorithms. However, note that an extensive analysis of rendering-specific artifacts is out of scope here and that we restrict ourselves to typical global illumina-

Image #	AD	HDR-VDP-2	MS-SSIM	SSIM	sCIE-Lab	sCorrel
#1 – apartment	0.000	0.000	0.000	0.000	0.000	0.000
#2 – cg figures	0.611	0.561	0.641	0.568	0.552	0.267
#3 – disney	0.263	0.580	0.478	0.445	0.410	0.201
#4 – kitchen	0.231	0.254	0.141	0.204	0.213	0.105
#5 – red kitchen	0.065	0.222	0.090	0.100	0.040	0.134
#6 – sponza atrium	0.426	0.519	0.546	0.467	0.475	0.288
#7 – sponza above tree	0.233	0.235	0.239	0.287	0.269	0.193
#8 – sponza arches	0.485	0.495	0.491	0.599	0.609	0.254
#9 – sponza tree shadows	0.368	0.334	0.409	0.411	0.415	0.278
#10 – sponza trees	0.119	0.065	0.083	0.115	0.121	0.012
Overall MCORR	0.249	0.308	0.248	0.277	0.257	0.187
#11 – apartment II	0.733	0.579	0.661	0.704	0.746	0.758
#12 – atrium	0.679	0.589	0.559	0.671	0.653	0.595
#13 – bathroom	0.670	0.547	0.689	0.673	0.676	0.702
#14 – buddha	0.131	0.117	0.105	0.135	0.128	0.109
#15 – chairs	0.752	0.541	0.692	0.723	0.693	0.680
#16 – city-d	0.000	0.000	0.000	0.000	0.000	0.000
#17 – city-u	0.514	0.343	0.428	0.447	0.405	0.481
#18 – cornell	0.356	0.455	0.380	0.395	0.346	0.355
#19 – dragons	0.508	0.227	0.239	0.393	0.365	0.557
#20 – hall	0.000	0.000	0.000	0.000	0.000	0.000
#21 – icido	0.357	0.496	0.256	0.411	0.367	0.306
#22 – kitchen II	0.392	0.232	0.373	0.375	0.421	0.288
#23 – livingroom	0.780	0.771	0.762	0.785	0.687	0.488
#24 – mpii	0.498	0.407	0.511	0.519	0.508	0.402
#25 – plants-d	0.265	0.206	0.233	0.263	0.248	0.248
#26 – plants-u	0.646	0.516	0.642	0.613	0.616	0.588
#27 – room	0.177	0.209	0.139	0.185	0.176	0.146
#28 – sala	0.003	-0.190	0.053	0.031	0.285	0.630
#29 – sanmiguel	0.000	0.000	0.000	0.000	0.000	0.000
#30 – sanmiguel cam3	0.453	0.233	0.485	0.476	0.516	0.500
#31 – sanmiguel cam4	0.285	0.198	0.250	0.278	0.286	0.587
#32 – sibenik	0.000	0.000	0.000	0.000	0.000	0.000
#33 – sponza	0.521	0.274	0.580	0.680	0.603	0.217
#34 – tt	0.274	0.236	0.348	0.319	0.314	0.179
#35 – villa cam1	0.592	0.409	0.321	0.533	0.518	0.601
#36 – villa cam2	0.052	0.071	0.086	0.118	0.058	0.299
#37 – villa cam3	0.441	0.298	0.433	0.450	0.438	0.455
Overall MCORR	0.350	0.268	0.218	0.314	0.296	0.439

Table 5: Matthews correlations in with-reference experiment with the predictions of HDR-VDP-2, SSIM, MS-SSIM, sCIE-Lab, sCorrel and Absolute Difference. The last row shows the overall correlations over the whole test set. In case there is no region marked by more than 50% of observers, the Matthews correlation is not available (0).

	EG'12 dataset	New dataset
Sample size (number of subjective responses)	2×10	2×28
Number of images	10	27
Angular resolution	unrestricted, 30 ppp (pixels per one visual degree) on average	restricted by a chinrest to 60 ppp
Types of distortion	3	12
Strength of distortions	supra-threshold	near- and supra-threshold
Distortion distribution	uniform	uniform and manually blended
Observers	mostly experts	both experts and naïve
Repetition	no repetitions	21 out of total 35 observers completed no-reference followed by with-reference set

Table 6: The comparison of both our datasets used for testing quality metrics. The new dataset contains a wider variety of distortions, more scenes, observers, and was measured in rigorous conditions.

nation related distortions. Hence, we do not cover artifacts such as banding, tessellation, shadow bias or other more specific artifacts that arise mostly in real-time rendering techniques.

High-frequency noise. Probably the most common distortion encountered in rendered images is noise due to random sampling of high-dimensional integrals when simulating the light transport in a synthetic 3D scene. Our test images with noise were generated using (bi-directional) *path-tracing* or *progressive photon mapping*.

Structured noise and interpolation artifacts. Many rendering artifacts are more complex in appearance and can take on any shape of different spatial frequency. A common example is the (chromatic) noise of varying frequency arising in *photon density estimation* and in the *instant radiosity* algorithm which is the result of an insuffi-

cient local density of photons and virtual point light (VPL) sources, respectively.

Another example is the *irradiance* or *radiance cache* algorithm, which computes the illumination in a 3D scene sparsely and then interpolates the result on the surfaces, thereby, introducing errors either due to an insufficient number of caches or due to noise in the cache samples. In addition to low-frequency noise (ir-)radiance caching artifacts may exhibit step-function-like extrapolation seams, which strongly depends on the image traversal pattern. Although VPL and (ir-)radiance cache artifacts may be similar in appearance, their location in the image differs. Whereas VPL noise appears mostly in corners and on glossy surfaces, (ir-)radiance caching artifacts can occur anywhere in the presence of strong lighting gradients.

VPL clamping and light leaking. In order to prevent VPL noise in corners, clamping can be introduced that dampens the contribution of individual VPLs if their relative contribution to a pixel is above a user threshold. The result is a loss in energy and darkening near corners. The counterpart of this effect is the gradual brightening in corners due to light leaking and over-estimation of photon density which results from neglecting the local visibility and under-estimating the surface-area. Occasionally, (ir-)radiance caching algorithms may also suffer from light-leaking artifacts if the global error parameter is set too large (e.g., *sanmiguel* scene).

Brightness change. Another distortion we have noticed is a consistent change in brightness in large regions of an image, which usually indicates bias in the rendering algorithm. Reasons for this can be of systematic nature (e.g., wrong normalization, incorrect material usage) or approximative nature (e.g., missing light-(sub)paths if, for instance, only one bounce of indirect light is computed, or certain light transport is neglected (e.g., caustics) or approximated).

Aliasing is the result of insufficient super-sampling or missing pre-filtering during rendering. Our examples comprise two prominent instances of aliasing in rendering: shadow maps and image aliasing, which are similar in nature. However, shadow map aliasing mostly appears as “jaggy” edges. Our test images for image aliasing were generated by first downsampling the resolution of the reference images followed by upsampling to the original resolution using the nearest neighbors approach. Upsampling can also introduce blur in the image if using for example the popular Lanczos filter.

Tone mapping can introduce visually disturbing artifacts as well, in particular if local tone mapping operators (TMO) are applied. Potentially the most apparent artifacts are produced by gradient-based techniques. Therefore, we included examples of two popular gradient TMOs into our test set: typical halo artifacts appear in the *buddha* scene and characteristic gradient “leaking” is demonstrated in the *mpii* scene. Apart from that, every image in the test set was tone mapped for display purpose using a global tone mapping curve (gamma curve) for which we did not notice any visible distortions.

4 Additional data analysis

In the following we show all the stimuli (reference and distorted image pairs) for both datasets (Fig. 2), as well as mean observer data for both no- and with-reference experiments (Fig. 3). The difference between the two datasets is summarized in Table 6.

Descriptive statistics reported in Table 1 enable the analysis of the distortion maps produced by image quality metrics. The agreement between observers is reported using Kendall’s coefficients u both on a per-pixel basis (Fig. 4) and averaged per scene (Table 2). An example of the subjective data we are dealing with in our experiment is given in Fig. 6.

For the sake of brevity, we excluded an analysis of the correlation between with-reference and no-reference experiment (i.e. noticeable and objectionable distortions) for each scene and only show the plots in Fig. 5. The values for the area under the ROC curve (AUC) are provided in Table 3 and the Matthews correlation values are given in Table 5 for each scene.

4.1 Computer graphics experts vs. non-experts

The observers filled in a post-experimental questionnaire (shown in Section 2.1), where they indicated their expertise in computer graphics techniques. Thanks to that we can assess the effect of this knowledge on the gathered results. For that purpose we run another Kruskal-Wallis test separately for each dataset and experiment. The results indicate that the expertise of an observer did not play an important role in our experiment and we average the subjective results for all the observers in the further analysis.

However, there are few interesting exceptions to look at. In particular, the darkening artifact due to VPL clamping in the *sibenik* (#32) scene is only detected by experienced observers in the with-reference experiment, see Fig. 1. Experts also detected the aliasing for the wheels in the *tt* (#34) scene. The difference between experts and non-experts gets slightly higher in the no-reference experiment, as non-experts sometimes tend to mark false-positives (especially in EG’12 dataset). Furthermore, non-experts sometimes overlooked distortions that are familiar to experts.

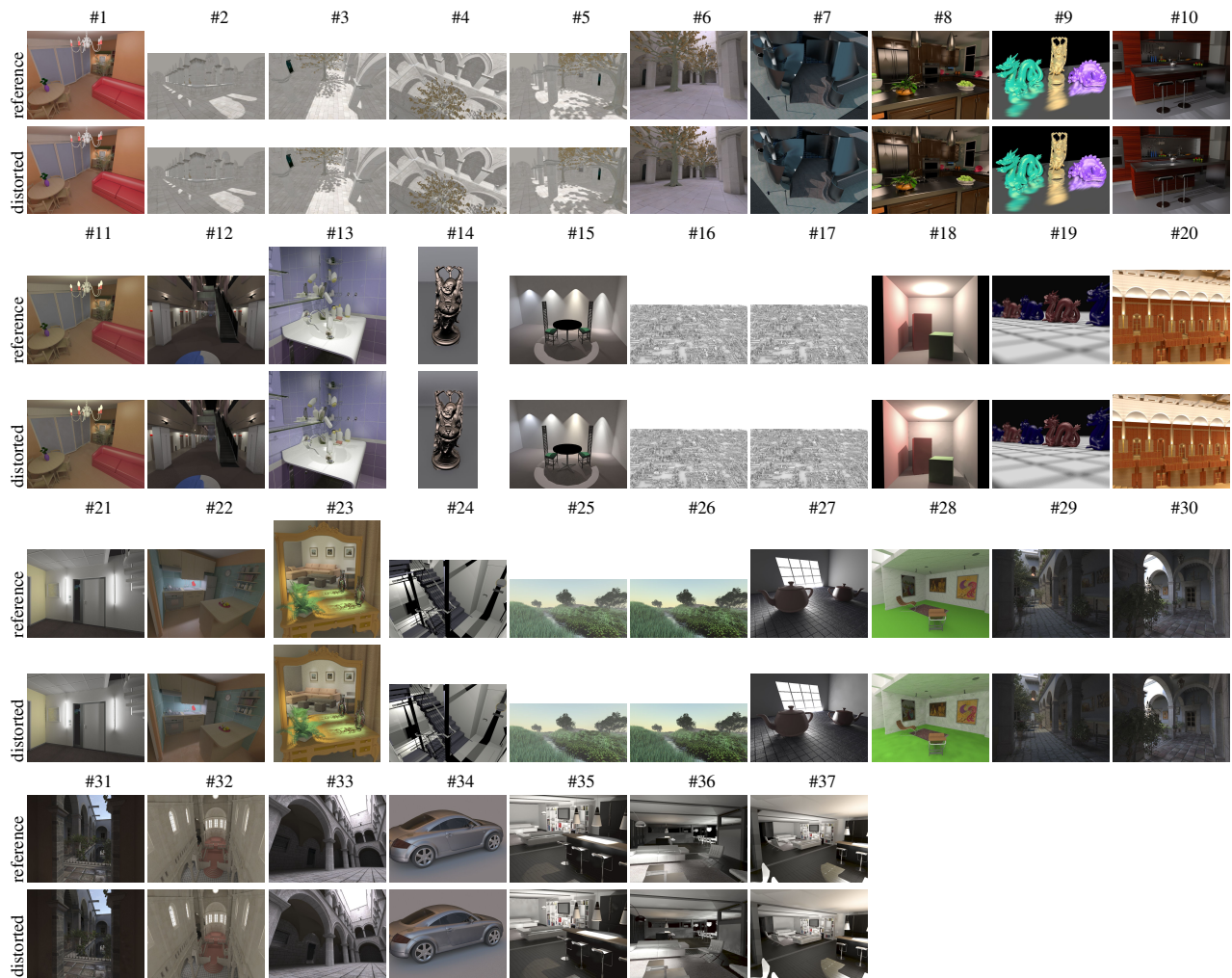


Figure 2: Both experimental datasets. Images #1 – #10 belong to EG'12, images #11 – #37 were used in the new dataset. Top rows: reference images, bottom rows: images containing rendering artifacts. (Best viewed in electronic version.)

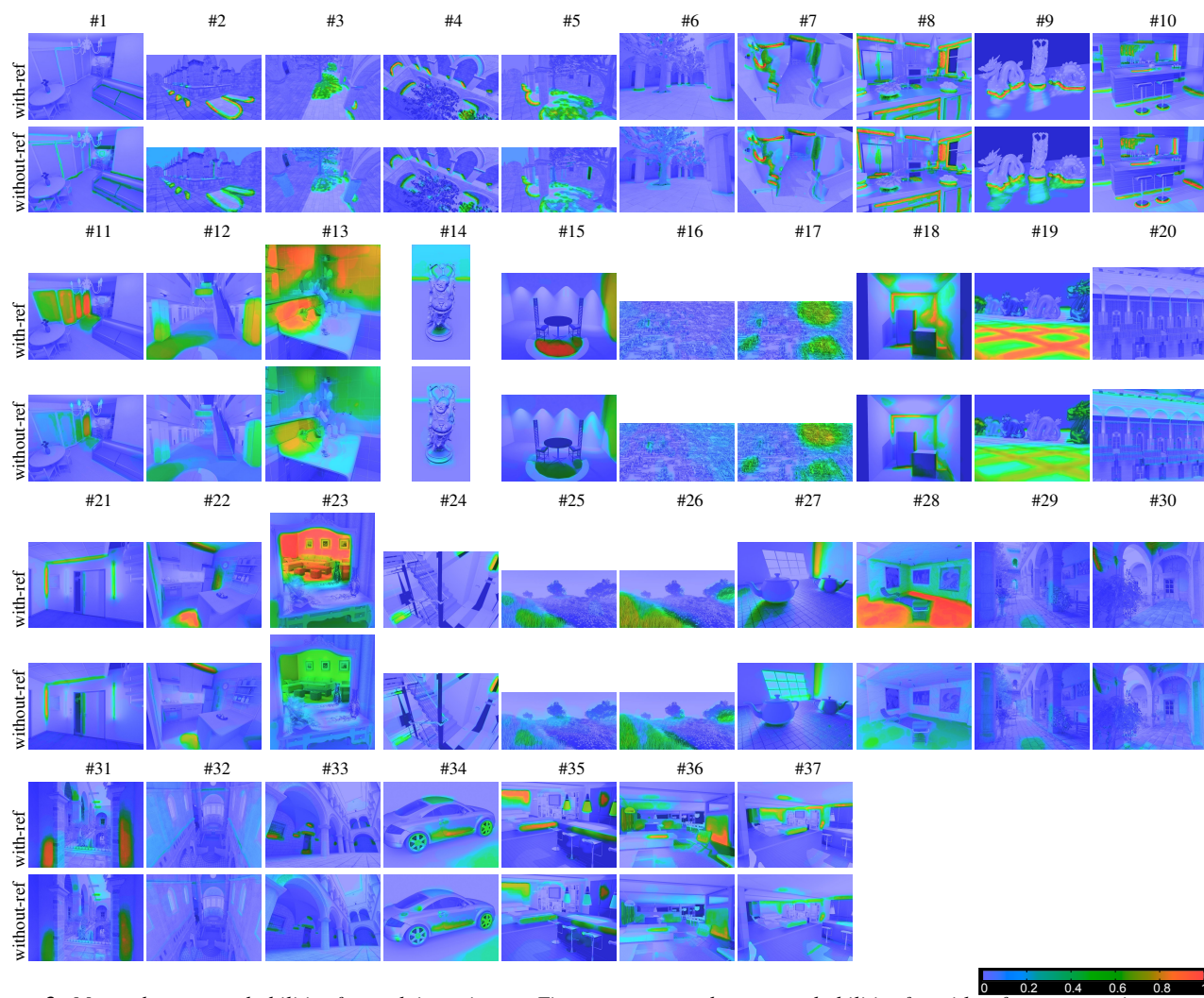


Figure 3: Mean observer probabilities for each input image. First rows: mean observer probabilities for with-reference experiment, second rows: mean observer probabilities for no-reference subjective experiment. (Best viewed in electronic version.)

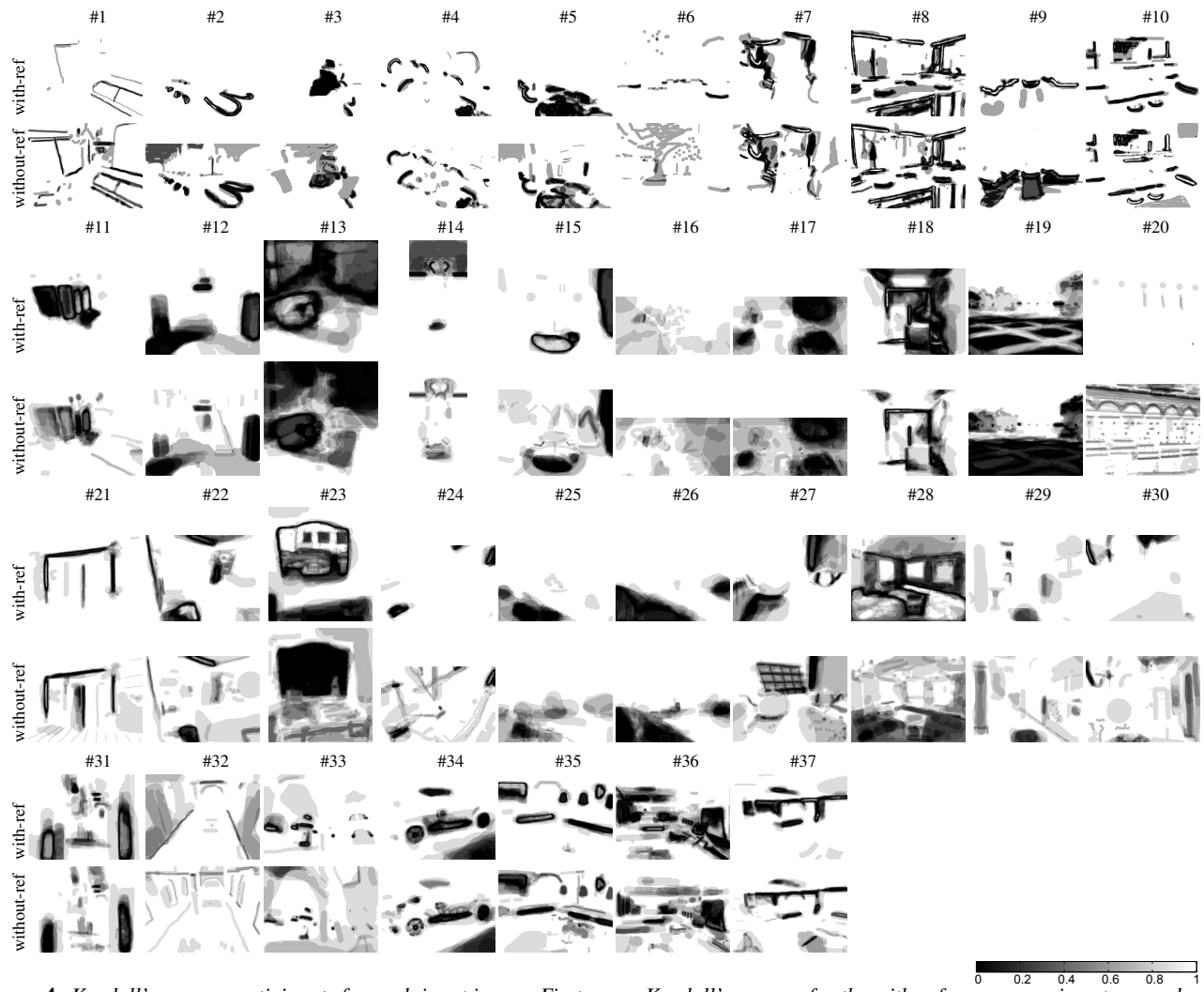


Figure 4: Kendall's u over participants for each input image. First rows: Kendall's u maps for the with-reference experiment, second rows: Kendall's u maps for the no-reference subjective experiment. (Best viewed in the electronic version.)

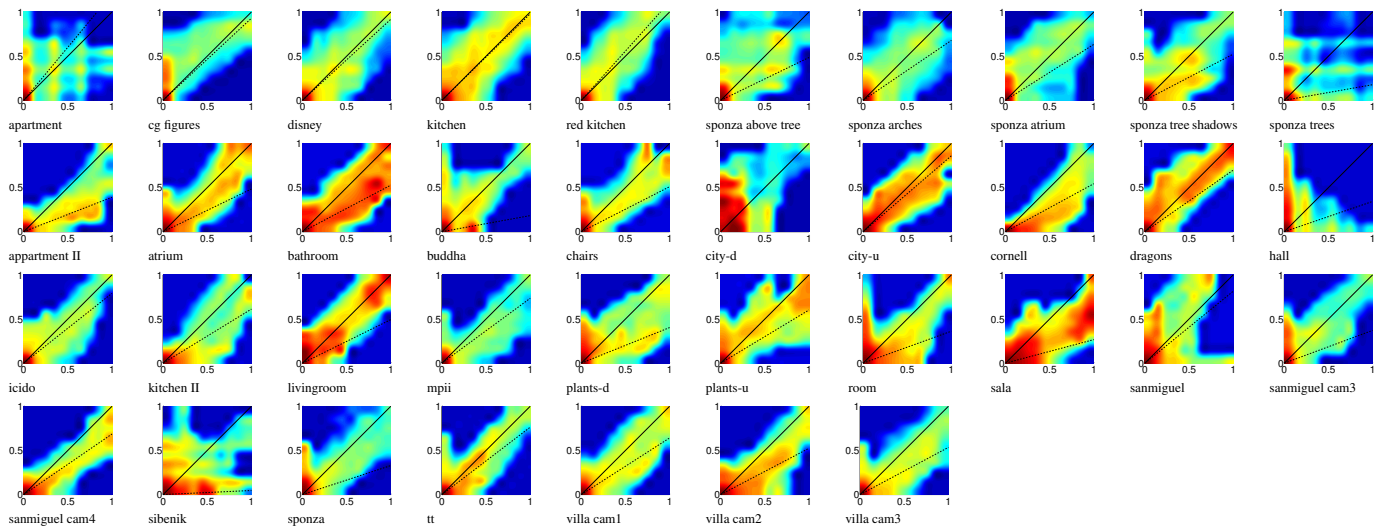


Figure 5: Correlation between the with-reference (x-axis) and no-reference artifact-detection probabilities (y-axis) for each scene.

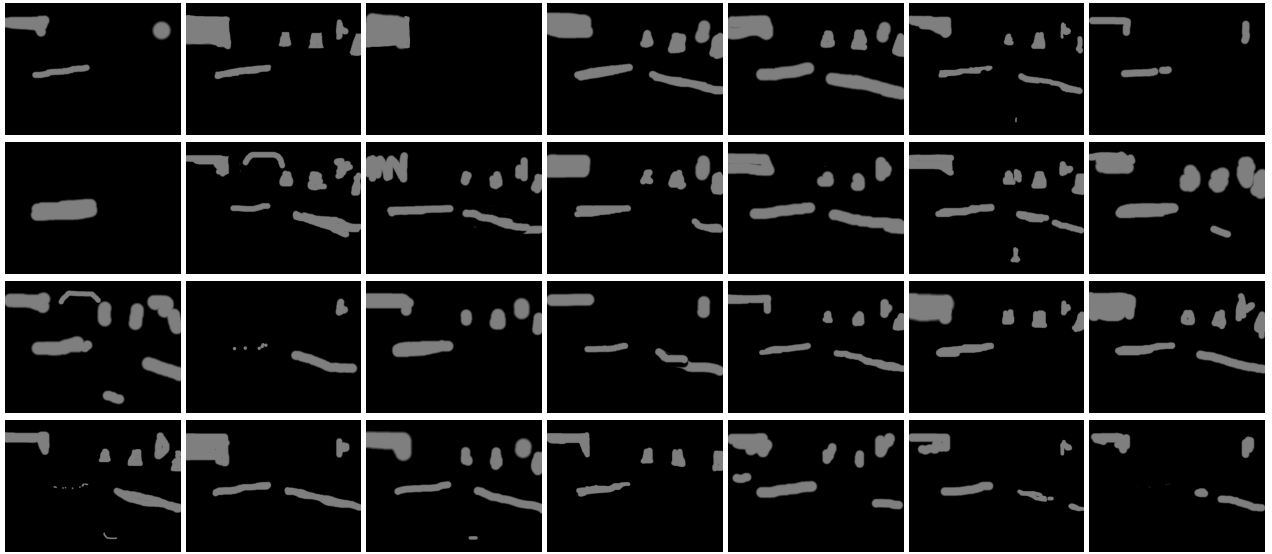


Figure 6: An example of agreement between all 28 observers for the villa cam1 scene in the with-reference experiment, $u = 0.849$, $\bar{u}_{mask} = 0.373$.