New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts

Martin Čadík, Robert Herzog, Rafał Mantiuk, Karol Myszkowski, Hans-Peter Seidel

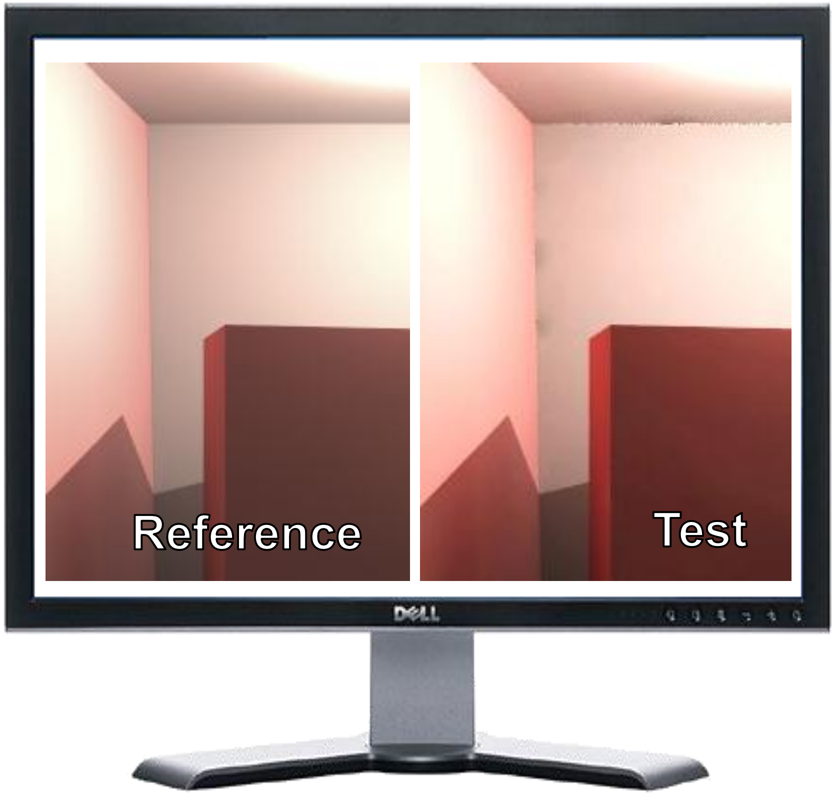
Sponsored by ACM SIGGRAPH

SIGGRAPH ASIA 2012

Outline

- Full-reference Image Quality Metrics (IQM)
- Datasets, experiments – localized distortions
- Evaluation of state-of-the-art IQ metrics
- Analysis of IQM failures
- Conclusions and future work

FR Image Quality Assessment

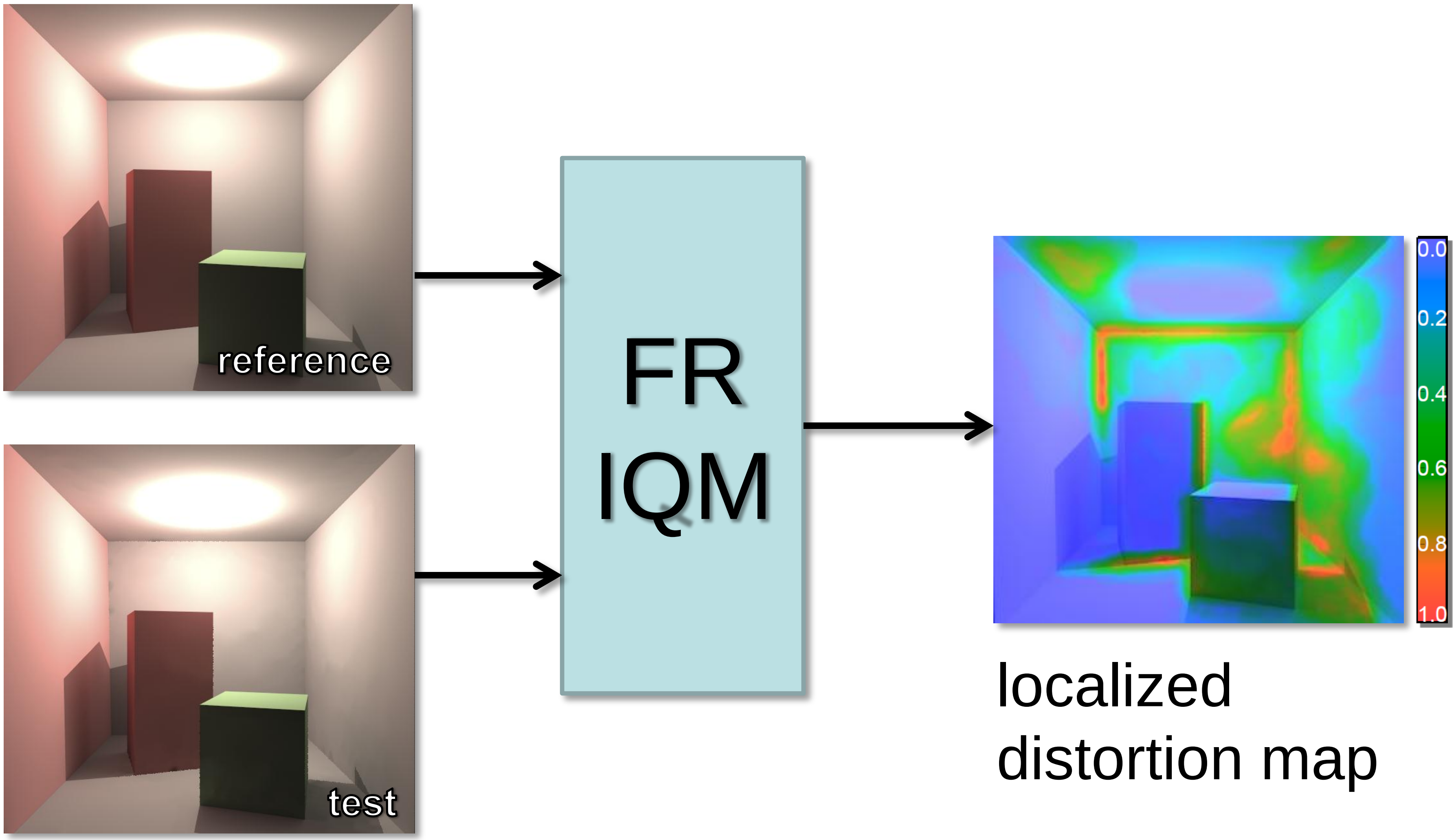


Rate the Quality/ Visibility of Artifacts

Subjective Experiments: + Reliable

– High Cost

Full-Reference Image Quality Metrics



Full-Reference Metrics

- What are they good for?
  - Quality assessment scenarios in compression/transmission, etc.
  - Algorithm analysis/validation/evaluation
  - Guiding/ parameter estimation of renderers
  - Stopping criterions
  - Speed/ quality enhancements
- Are they reliable?

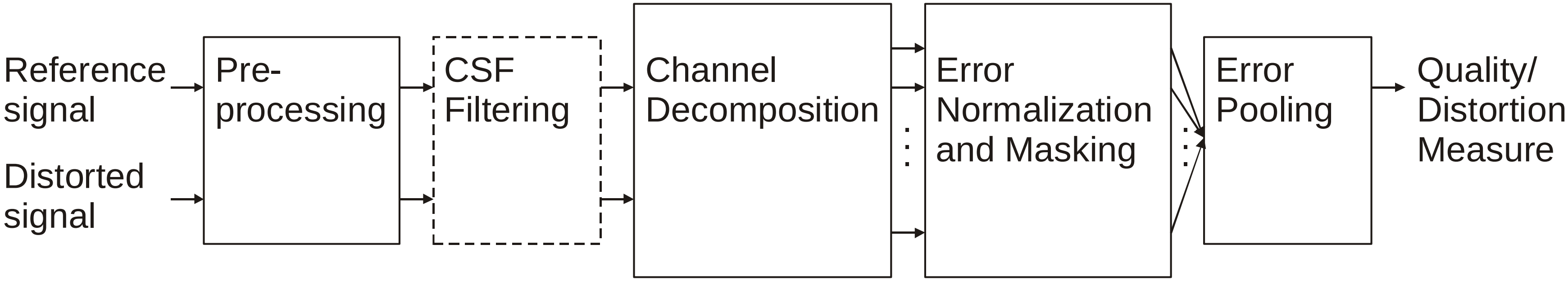
Mathematically Based Metrics

- AD $M = |ref - test|$
- (R)MSE $M = (ref - test)^2$ $MSE = \frac{1}{n}\sum_{i=1}^{n}(ref_i - test_i)^2$
- PSNR $PSNR = 10\log_{10}\frac{MAX^2}{MSE}$
- sCORREL $M = SRCC(ref,\ test)$

  (Spearman's rank correlation coefficient, per 8x8 block)

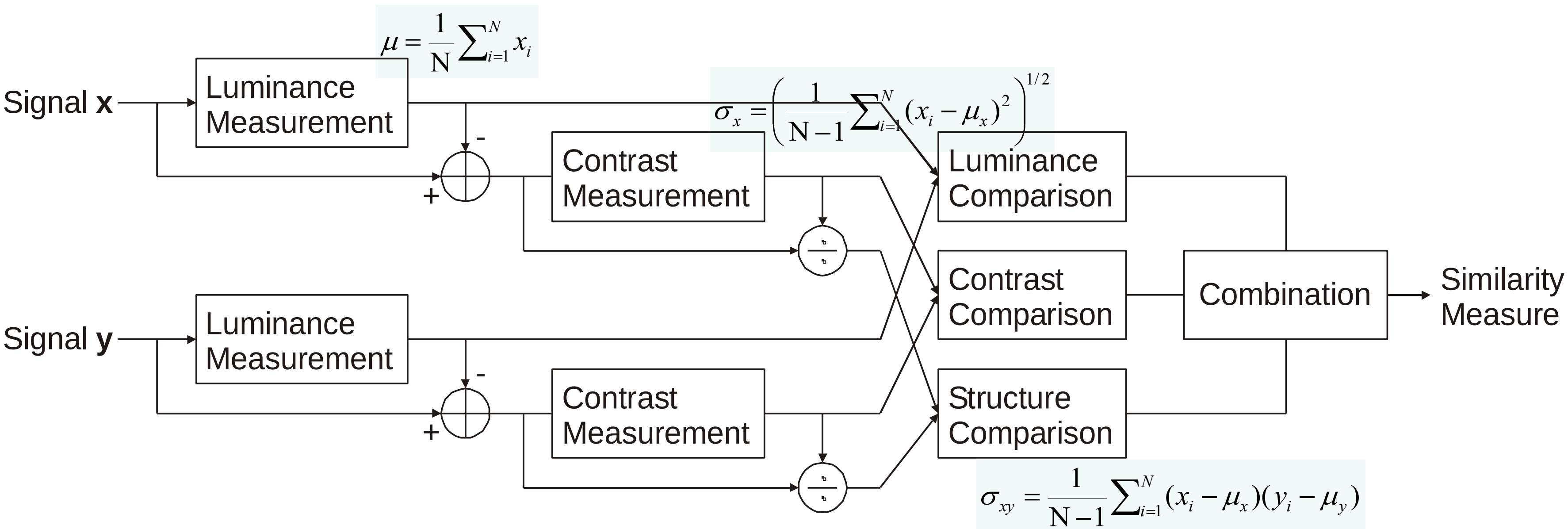
Error Sensitivity-Based Approaches

- General framework



- Visible Differences Predictor [Daly93]
- Perceptual Distortion Measure [Teo, Heeger 94]
- Visual Discrimination Model [Lubin 95]
- Gabor pyramid model [Taylor et al. 97]
- WVDP [Bradley 99]
- HDR-VDP-2 [Mantiuk et al. 05, Mantiuk et al. 11]

Structural Similarity-Based Approaches



- UQI [Wang 02]
- SSIM [Wang 04]
- M-SSIM [Wang et al. 04]
- Multidimensional Quality Measure Using SVD [Shnayderman 04]

Other Metrics

- sCIE-Lab [Zhang and Wandell 98]
  - Spatial extension of CIE Delta E
  - Luminance and color contrast sensitivity
- VSNR [Chandler and Hemami 07]
  - Visual Signal to Noise Ratio
  - Wavelet-based SNR
  - Masking model
- VIF [Wang and Bovik 06, Ch. 3.3]
  - Information-theoretic approach (mutual information)
  - Exploits natural scene statistics

Evaluation of STAR FR-IQM

- 6 IQMs: AD (PSNR, MSE), sCIE-Lab, sCORREL, SSIM, MS-SSIM, HDRVDP-2

- How good are IQMs in **localizing** artifacts?
- Evaluation of distortion **maps** (not just mean-opinion-scores, i.e. one number per image)
- Computer graphics-generated contents and artifacts
- Two subjective tasks: given reference image and with no reference image

Evaluation of STAR FR-IQM (cont.)

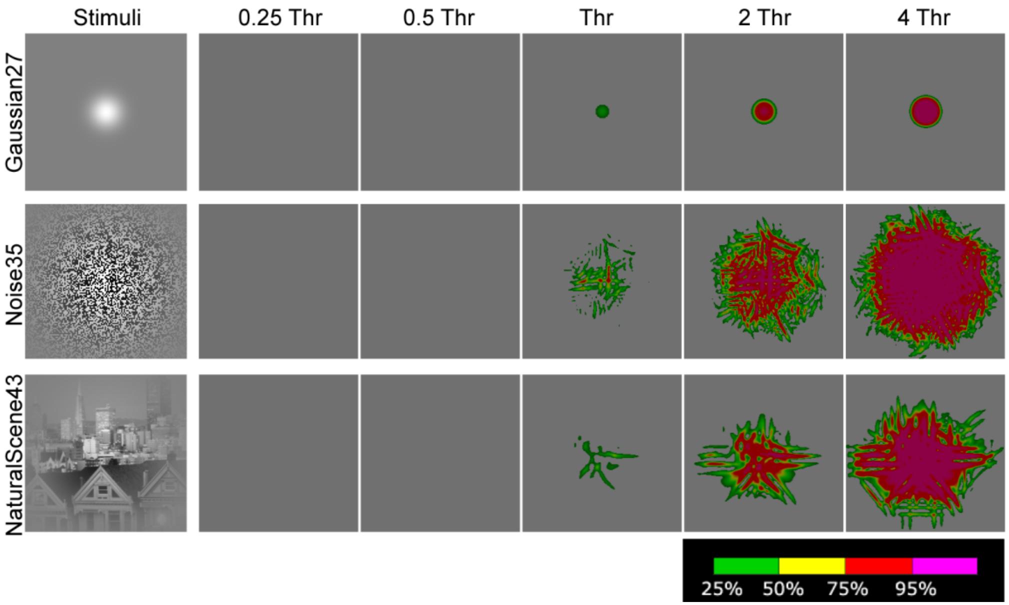
- Input data + Subjective responses = **dataset**
- Datasets
  - Simpler evaluations
  - Reproducible evaluations

  - Should comprise typical artifacts
  - Should be publicly available

http://www.mpi-inf.mpg.de/resources/hdr/iqm-evaluation/

Available Datasets

- IMAGES
  - Modelfest [Watson 99]
  - LIVE image db [Sheikh et al. 06]
  - TID (Tampere Image Database) [Ponomarenko et al. 09]



- VIDEOS
  - VQEG FRTV Phase 1 [VQEG '00]
  - LIVE video db [Seshadrinathan et al. 09]

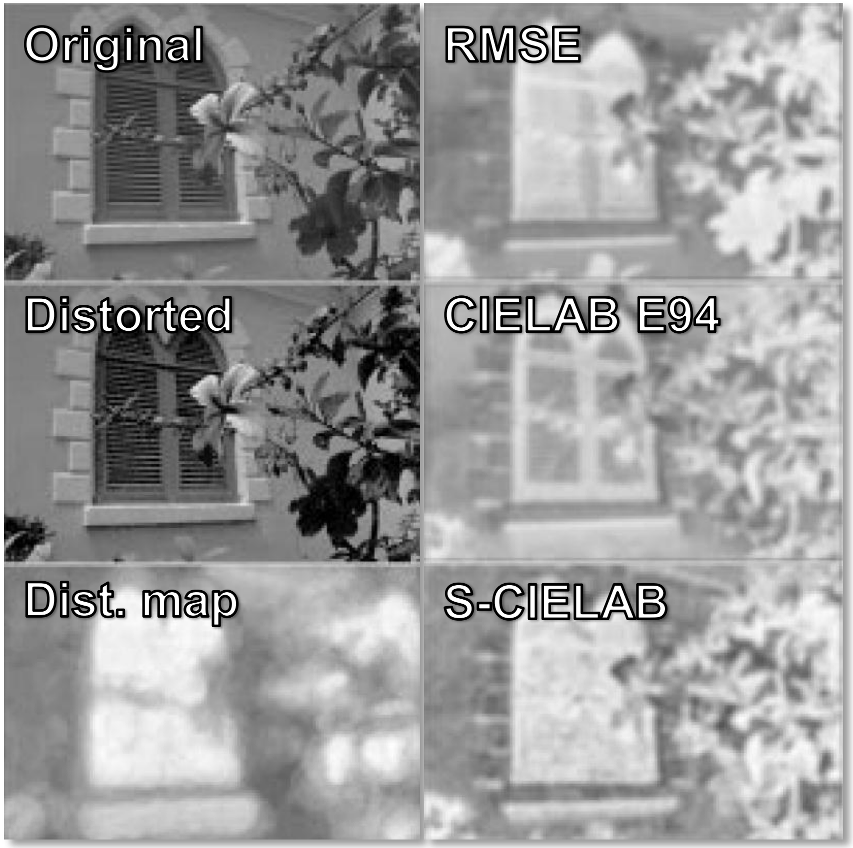
Available Datasets (cont.)

- Mostly only photos/real videos
- Focus on compression/transmission related artifacts
- Subjective responses: only overall quality (MOS)

Mean Opinion Score (MOS)		
MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

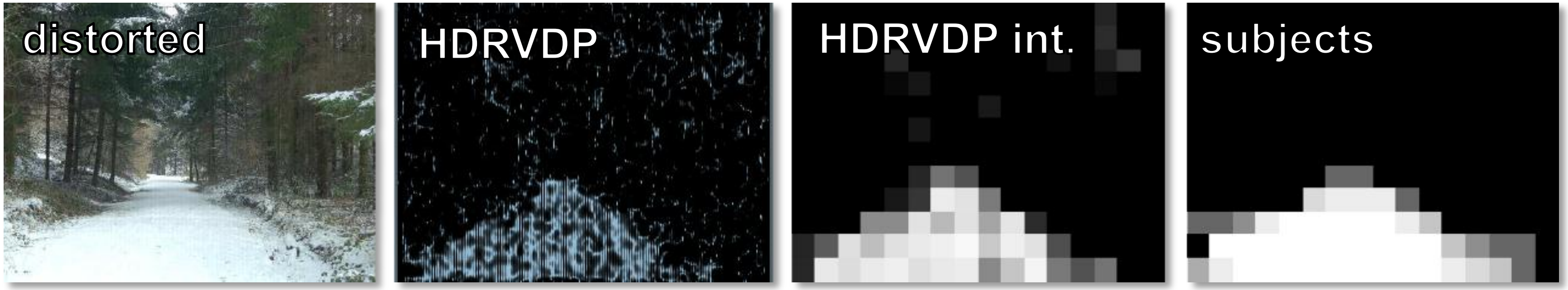
Previous Work

- [Zhang et al., CIC97, SP98]
  - Image distortion maps
  - JPEG compression, half-toning
  - RMSE, CIELAB E94, S-CIELAB



Previous Work (cont.)

- [Mantiuk et al., SPIE05]
  – for calibration of HDRVDP1



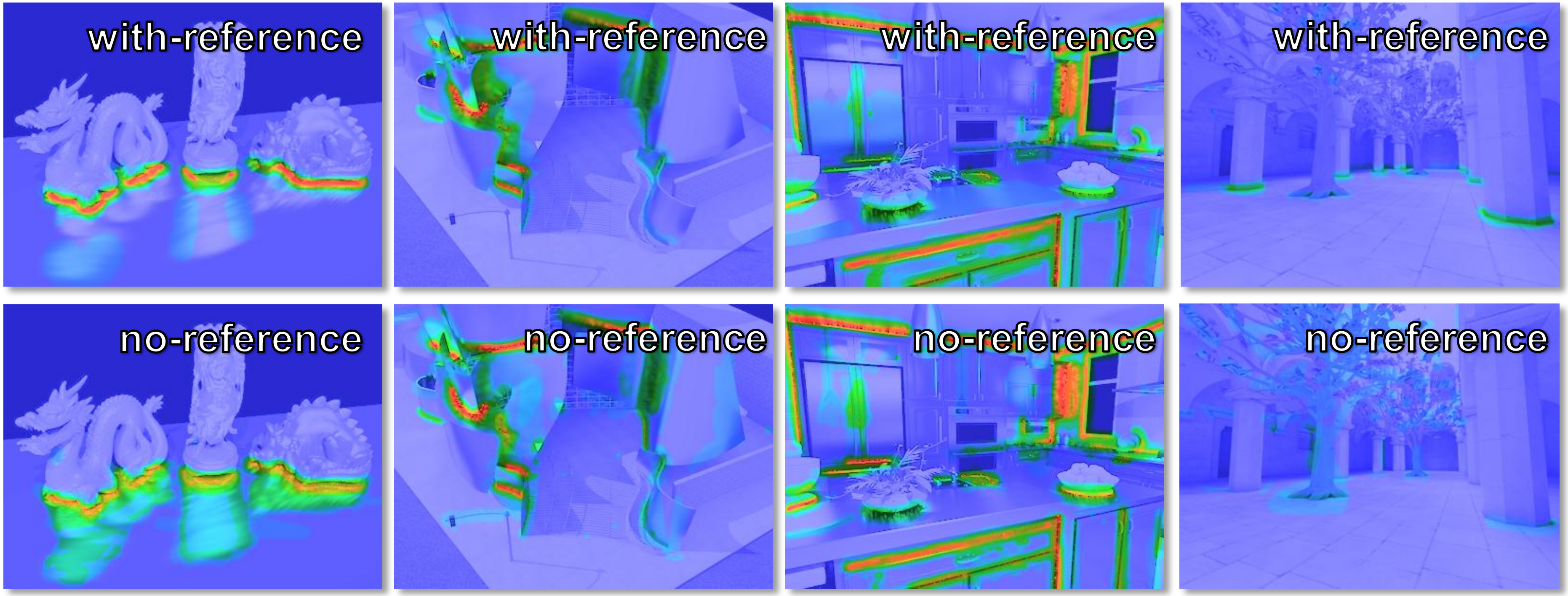
- [Čadík et al., SPIE11]
  – for validation of DRIVQM

Previous Work (cont.)

- Main purpose: to calibrate/validate existing models
- No IQM evaluation
- No CG content
- Simple distortions
  - Pattern noise
  - Blur
  - Random noise
  - Compression artifacts
  - Transmission artifacts

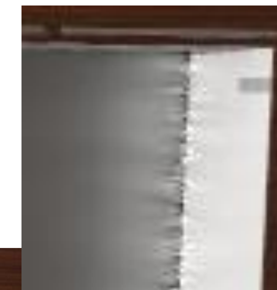
Previous Work (cont.)

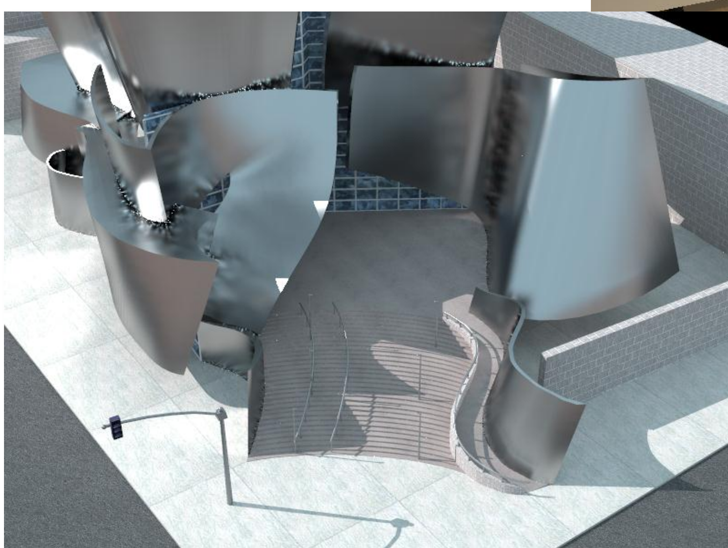
- [Herzog et al., EG12]
  - With-reference and no-reference experiments
  - 10 Supra-threshold CG stimuli



Our Dataset: Example Rendering Artifacts

- e.g., low-freq. noise from glossy instant radiosity or photon density estimation





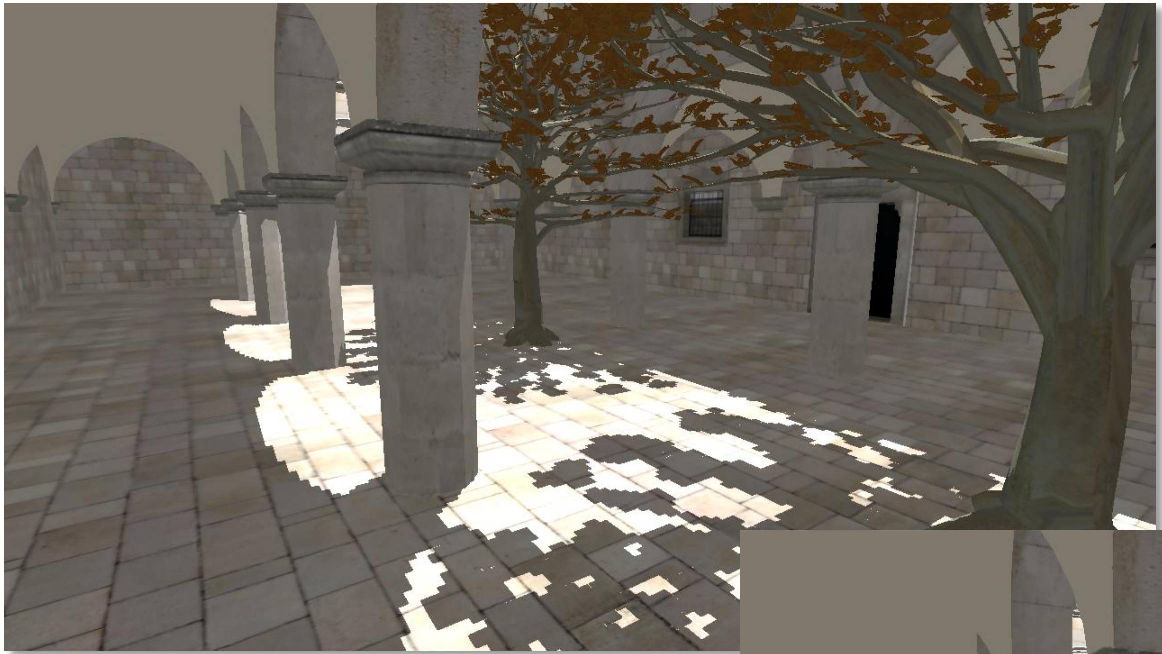
Example Rendering Artifacts

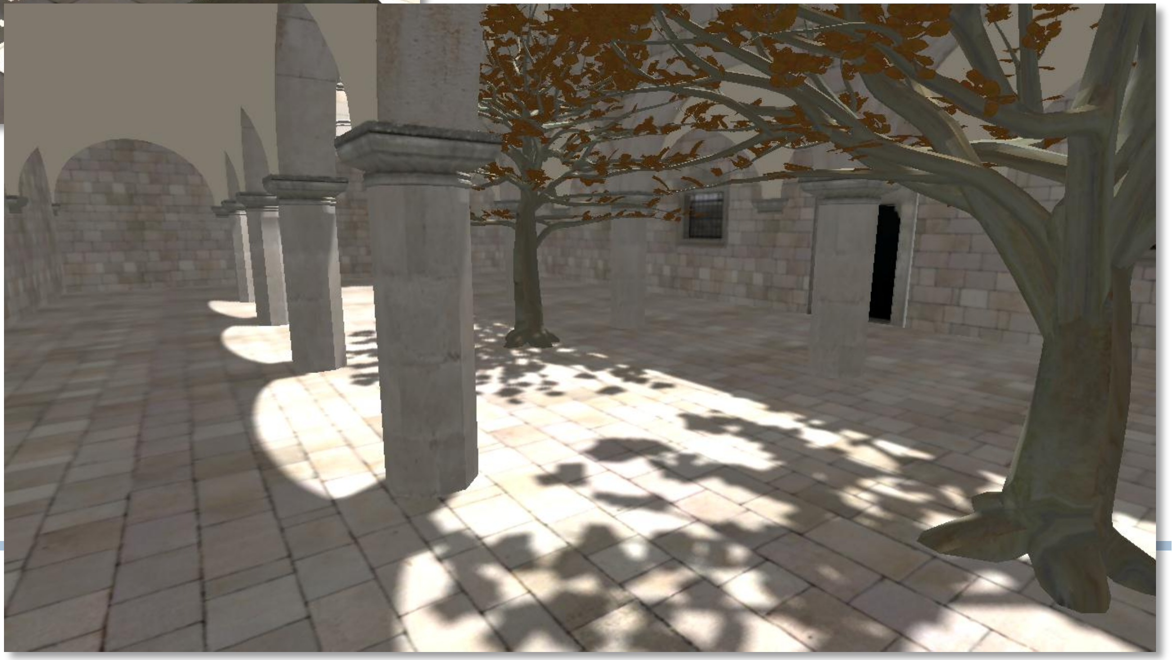
- Clamping Bias (darkening in corners)

Example Rendering Artifacts

- Irradiance caching
  - interpolation errors
  - leaking

Example Rendering Artifacts

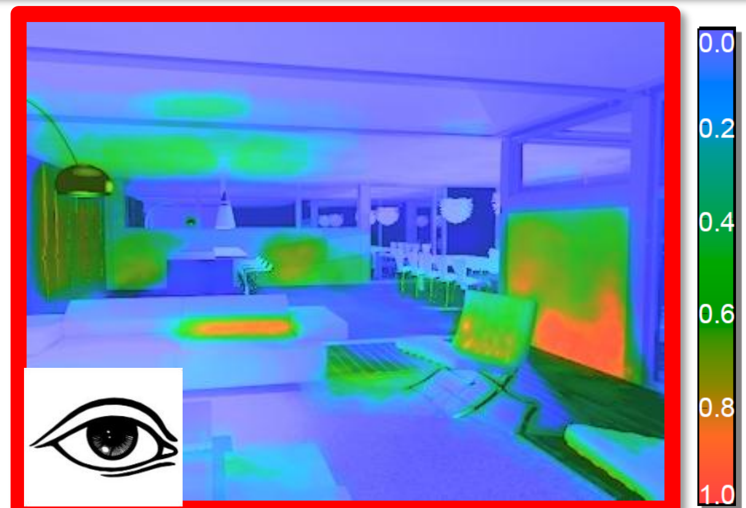




- Shadow Mapping
  (easy to generate large sample set)

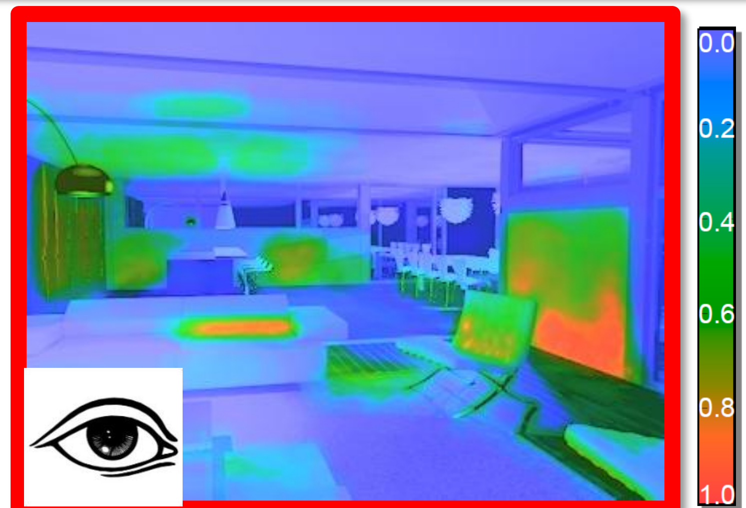
User Experiment - Mean Distortion Maps

- 37 test images
- 35 subjects (expert and non experts)
- Localization of artifacts
- Scribbling interface



User Experiment – With Reference

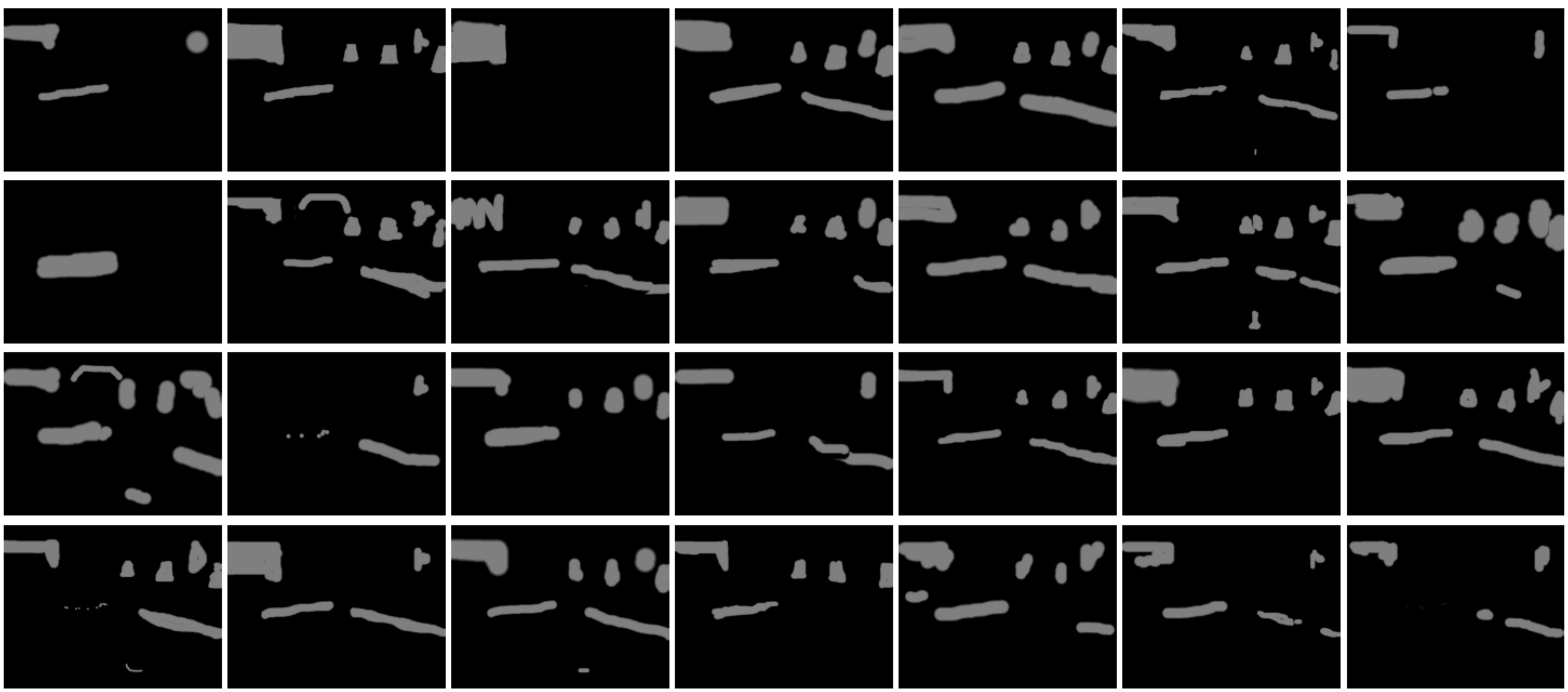
- Noticeable distortions



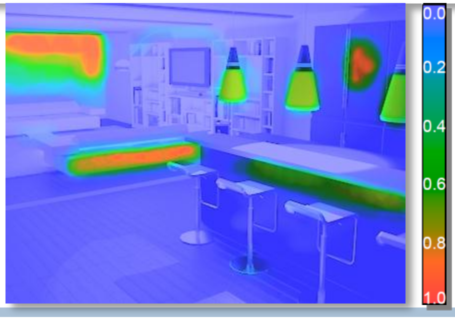
User Experiment – No Reference

- Objectionable distortions

Example User Responses



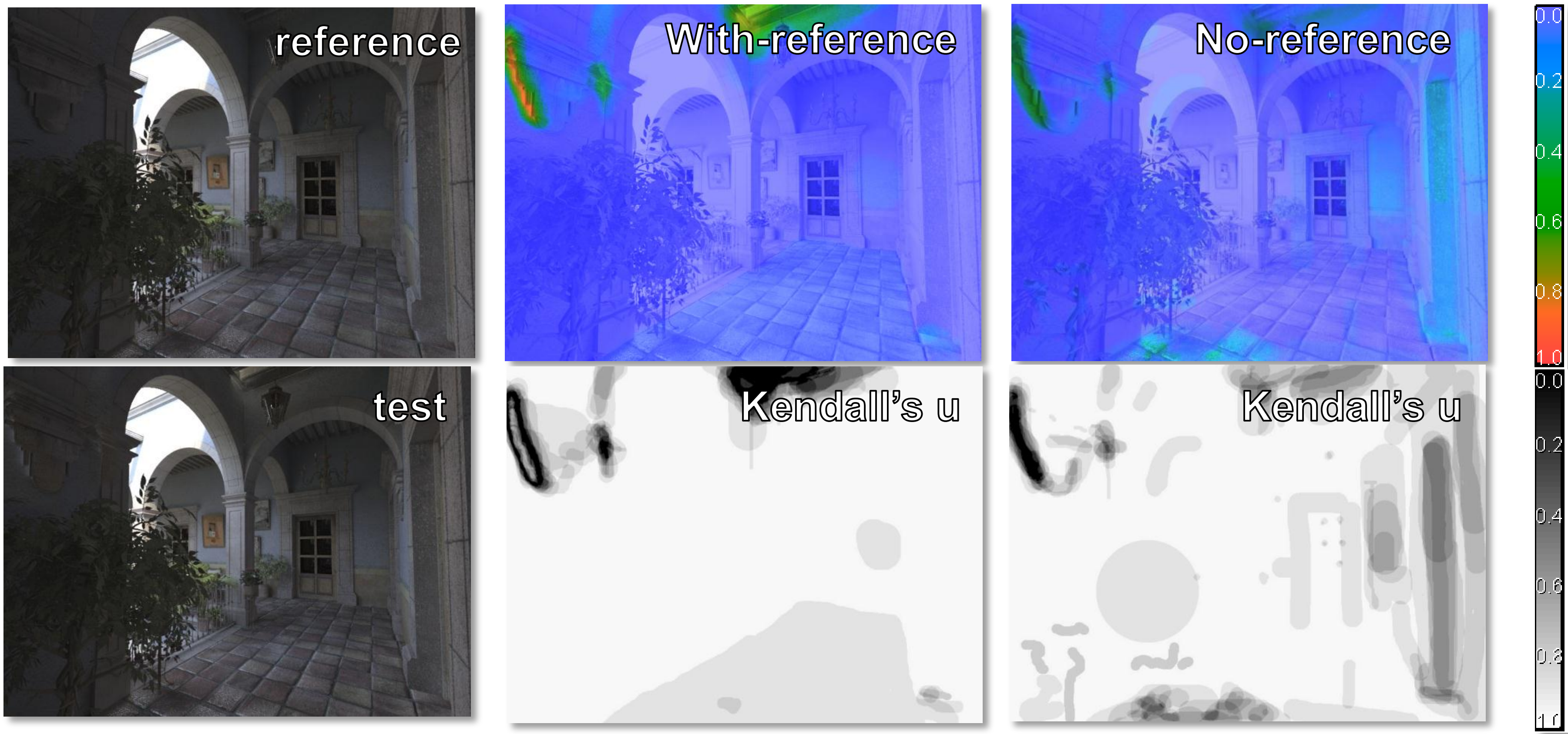
- Probability of detection



Inter-Observer Agreement

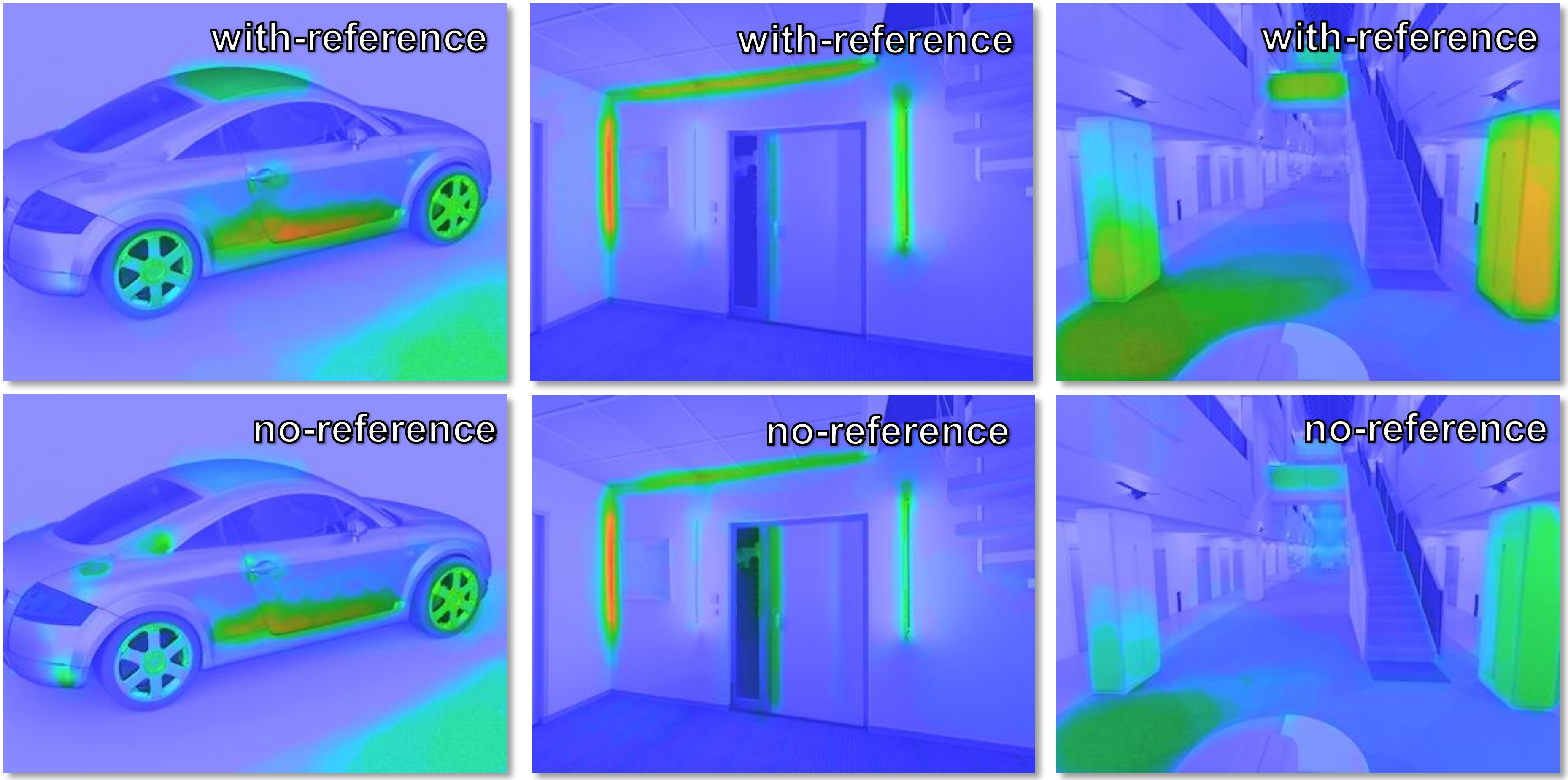
- Kendall's coefficient of agreement u

$\overline{u_{with-ref}}$ =0.78 $\overline{u_{no-ref}}$ =0.77



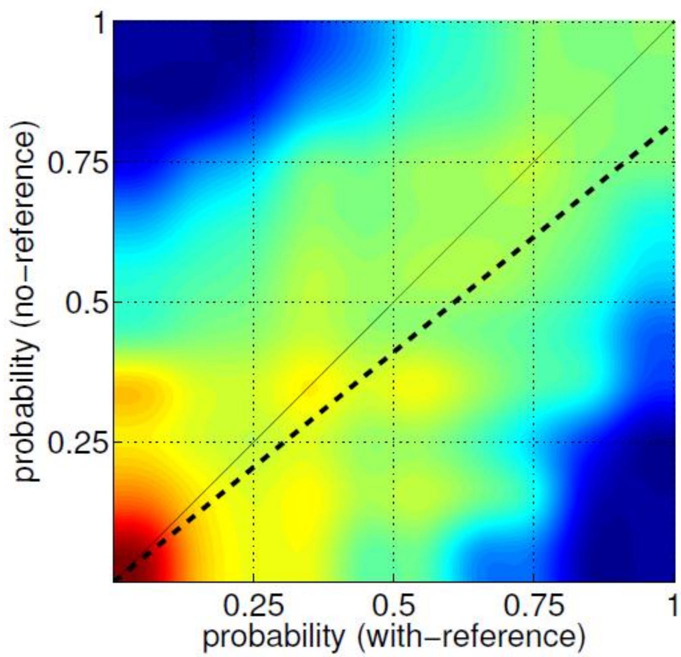
With-reference vs. No-reference

- Results rather similar

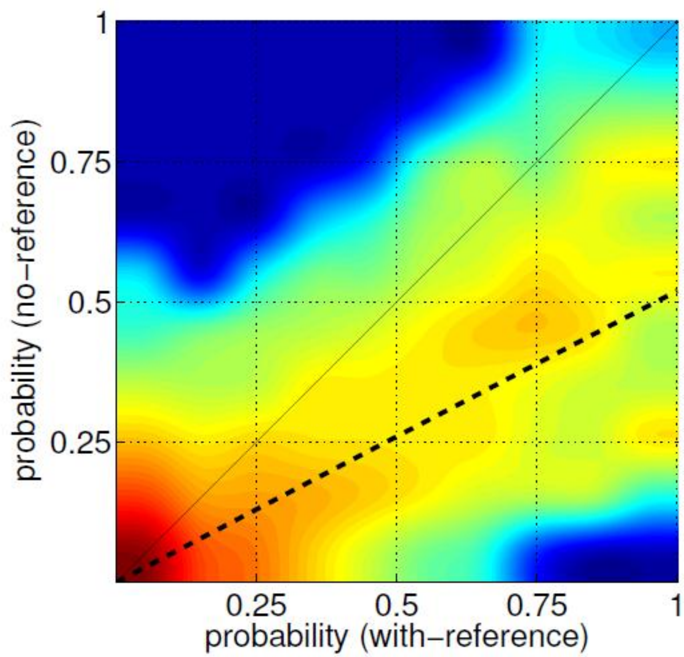


With-reference vs. No-reference (cont.)

- Strong correlation
  - (perhaps people do not need the reference)
- SRCC=0.88 SRCC=0.85

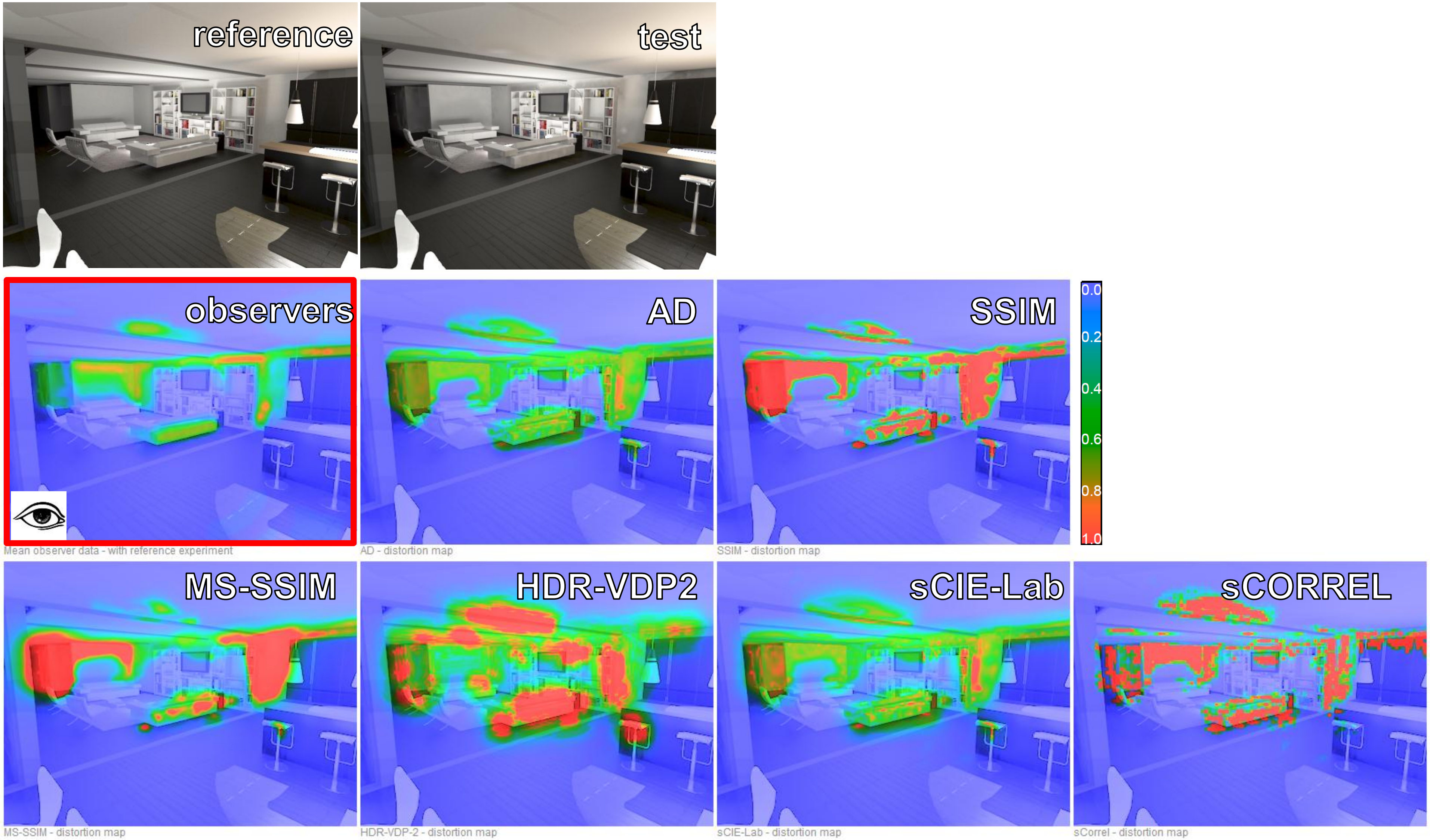


EG'12 dataset

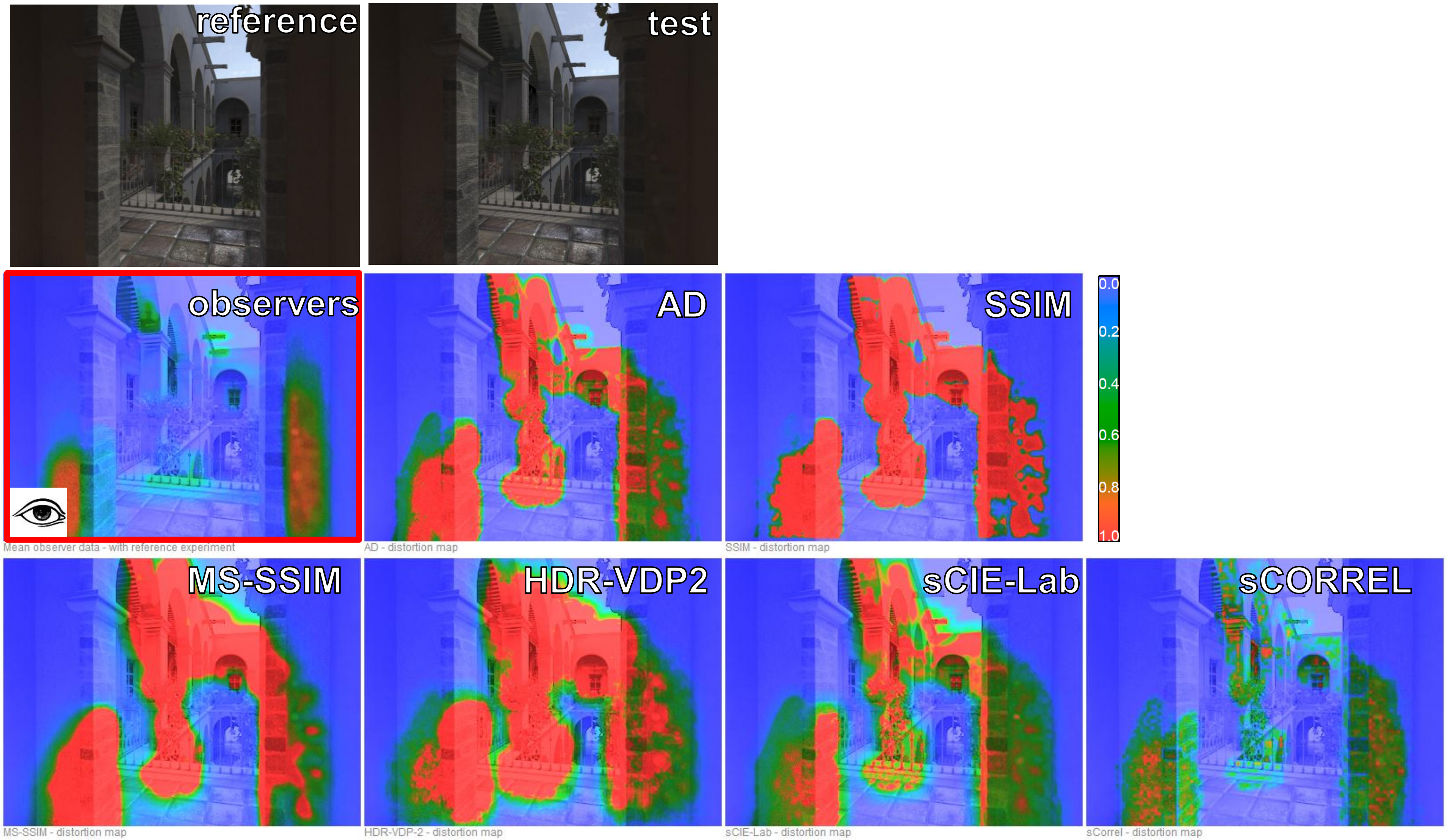


new dataset

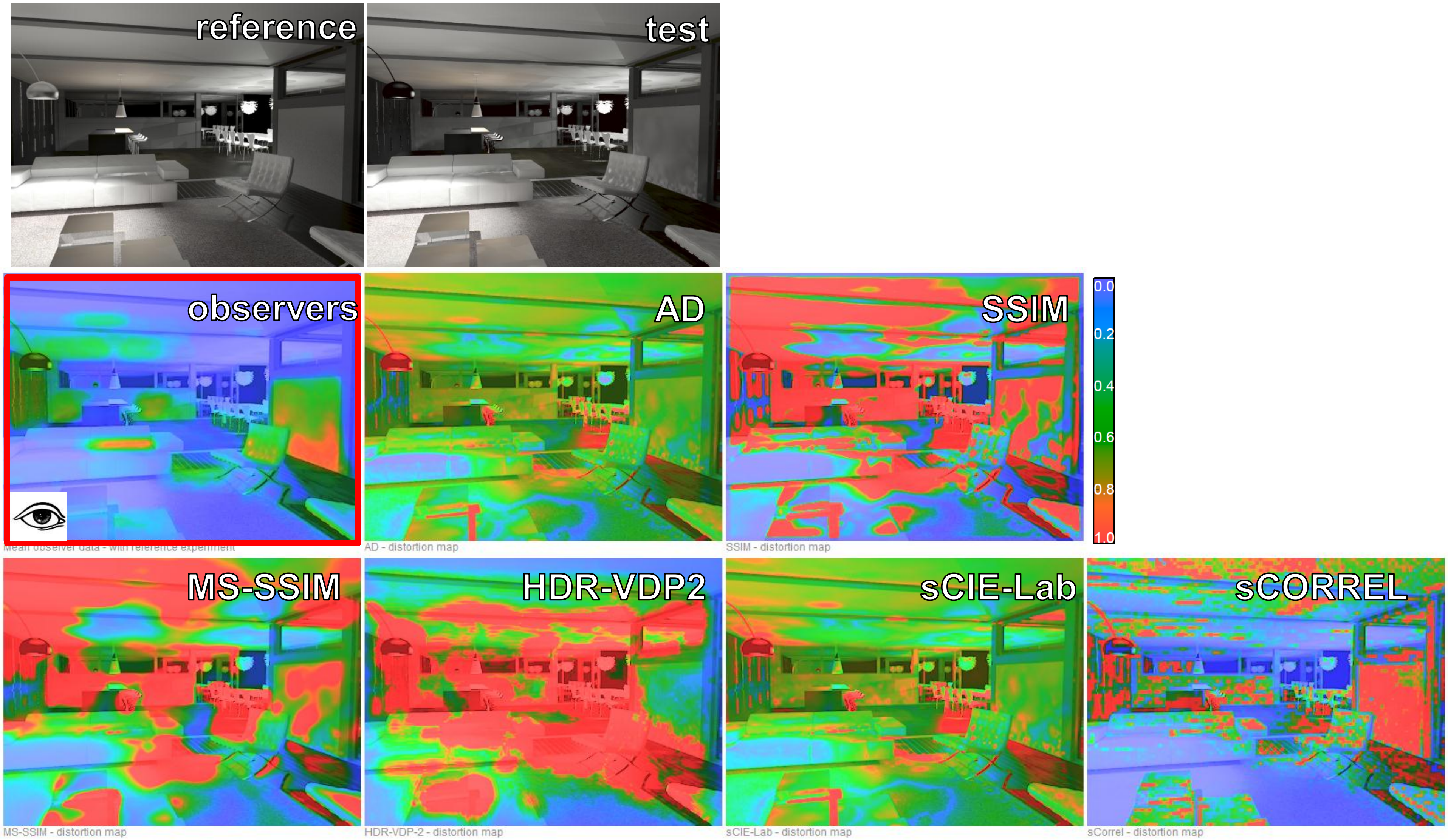
Results – Example of Metric Predictions



Results – Example of Metric Predictions

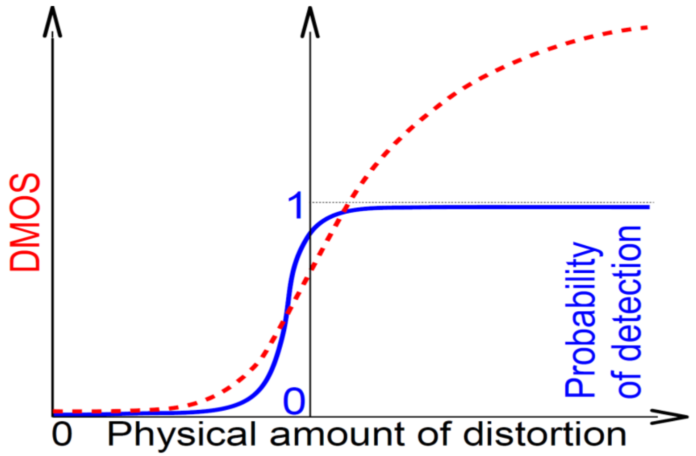


Results – Example of Metric Predictions

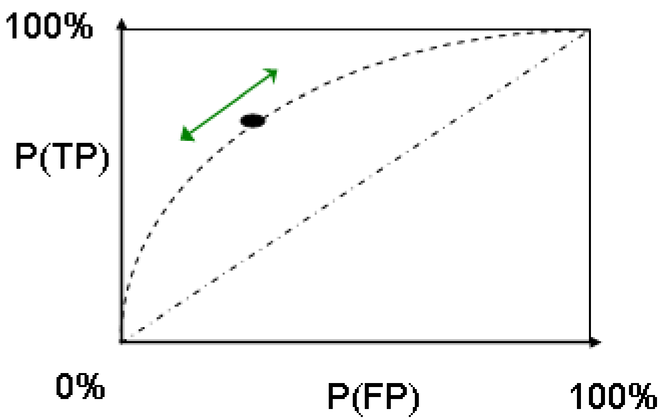


Measures of Metric Performance

- Previous experiments
  - MOS/DMOS {1,2,3,4,5}
- No easy way to capture MOS locally
  - Probability of detection [0,1]

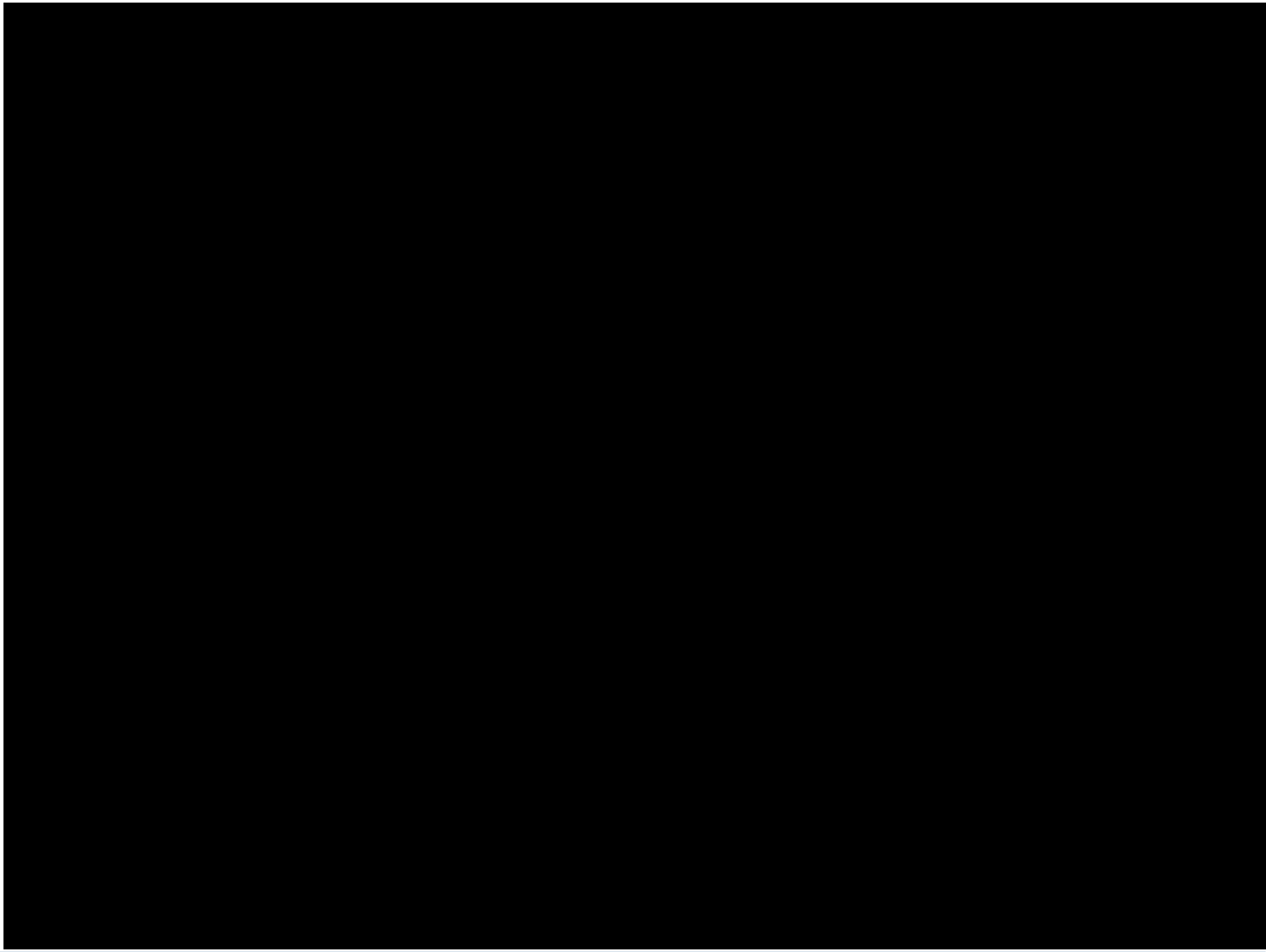


- Receiver operating characteristic (ROC)
  - Area under curve (AUC)
  - Thresholds (25%, 50%, 75%)



Measures of Metric Performance (cont.)

- ROC
  - **TP**
  - **FP**
  - **TN**
  - **FN**

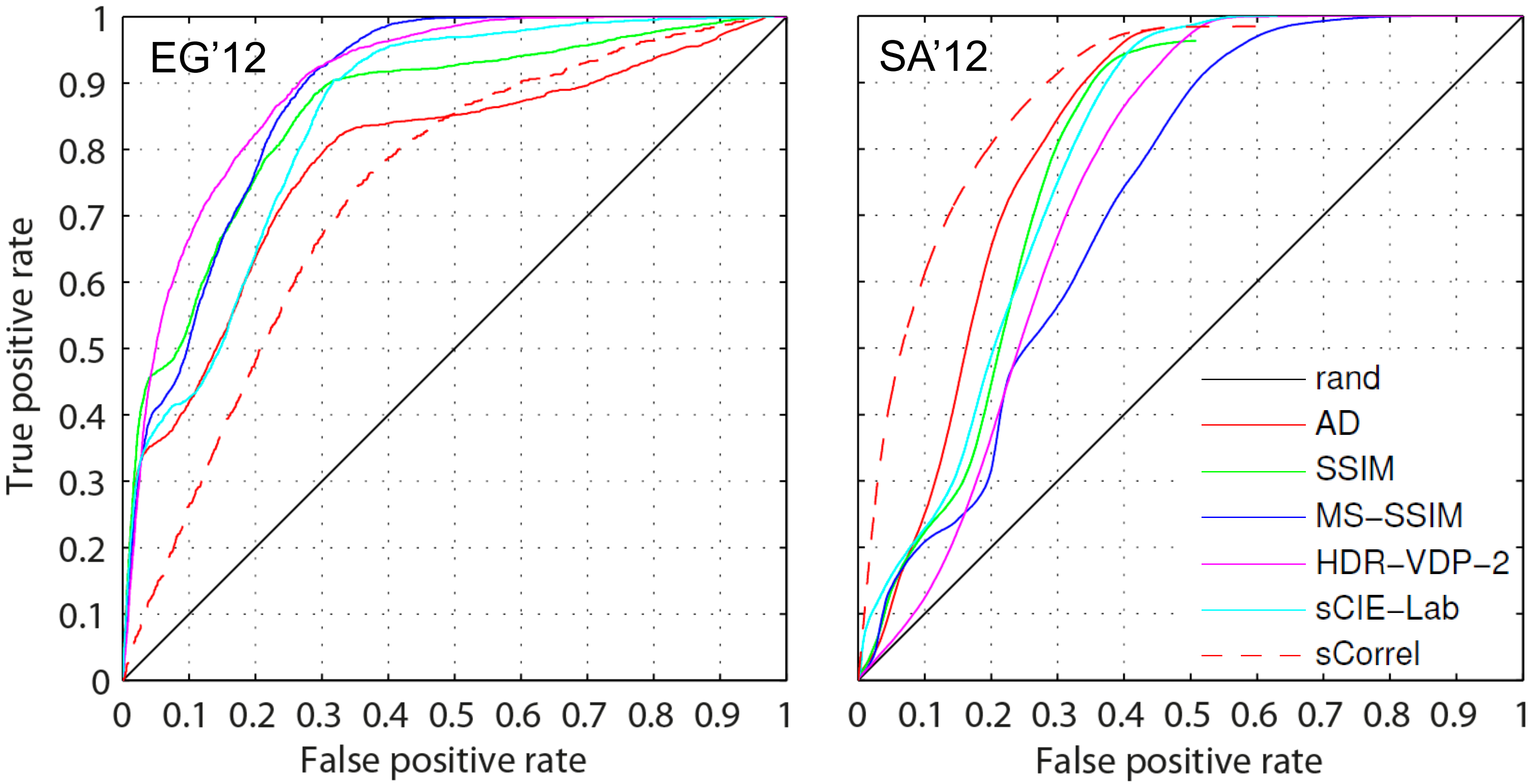


Measures of Metric Performance (cont.)

- Matthews correlation coefficient (MCC)
  - Robust to unbalanced data
  - [-1, +1]
    - 1 – perfect prediction
    - 0 – not better than random
    - -1 – total disagreement

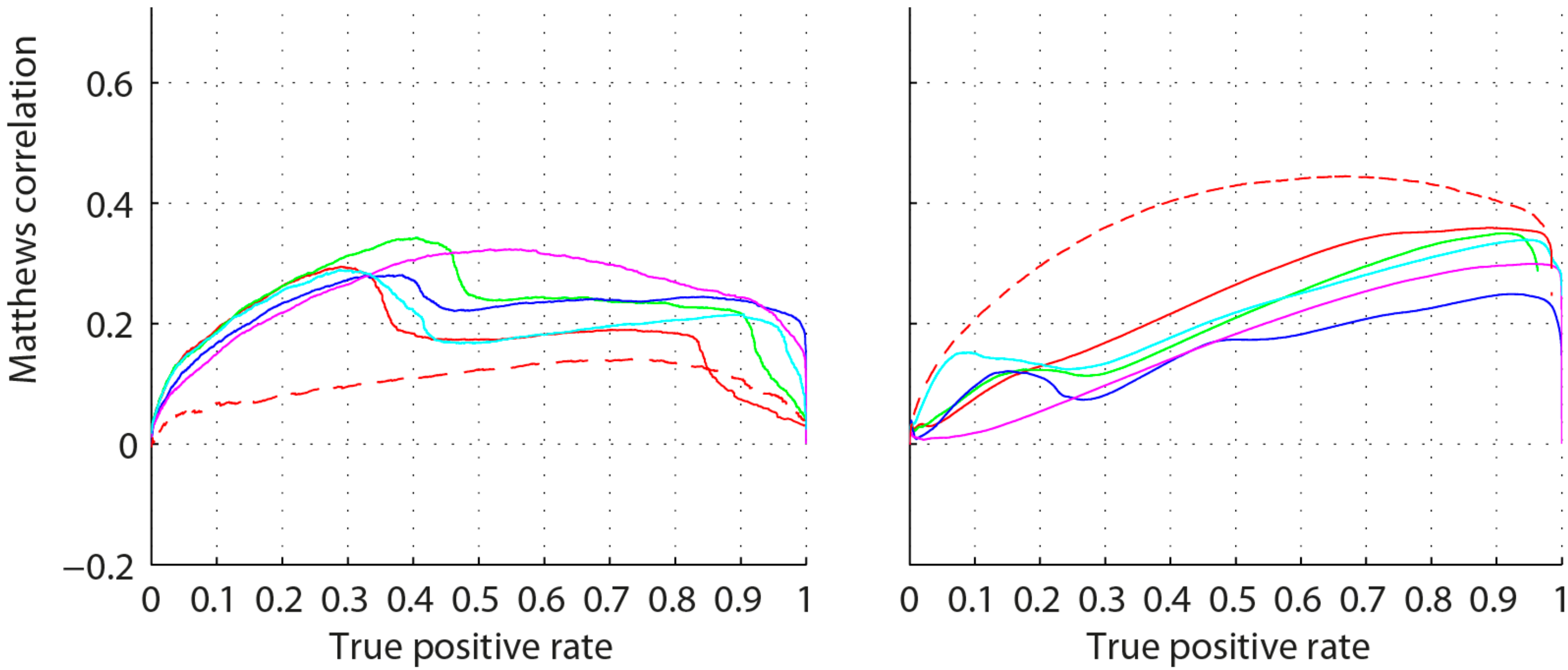
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Metric Performance Comparison – ROC

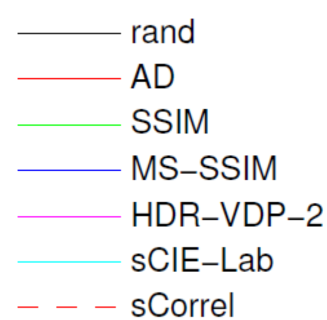


- With-reference experiment results (see paper for no-ref.)

Metric Performance Comparison – MCC

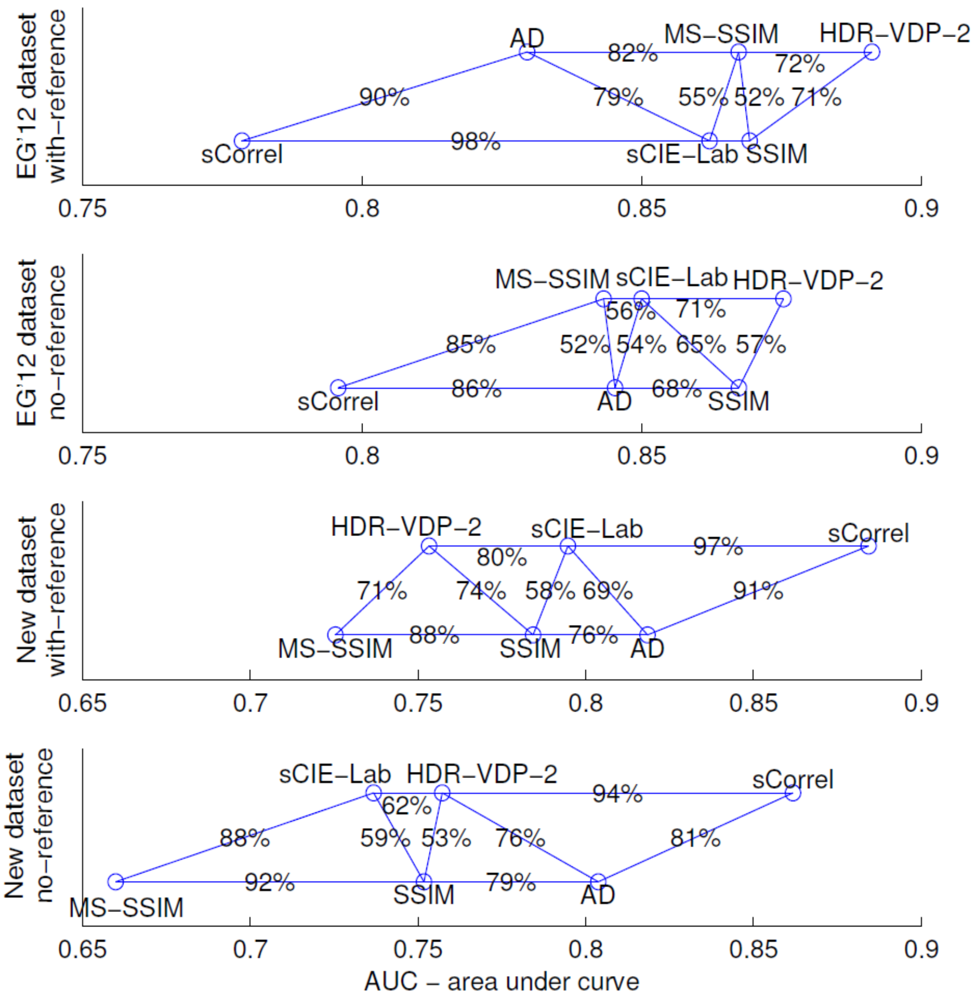


- Rather poor performance
- No champion
- Simple metrics comparable to complex ones

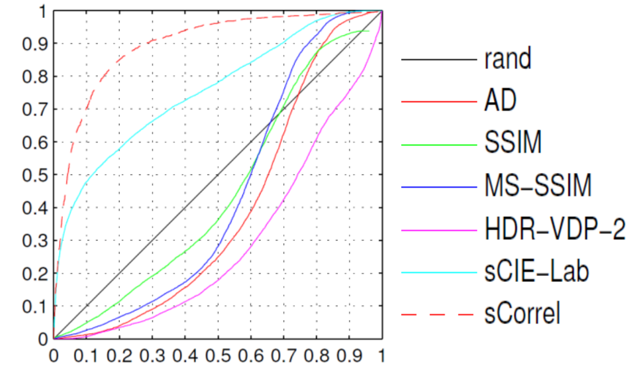


Metric Performance Comparison (cont.)

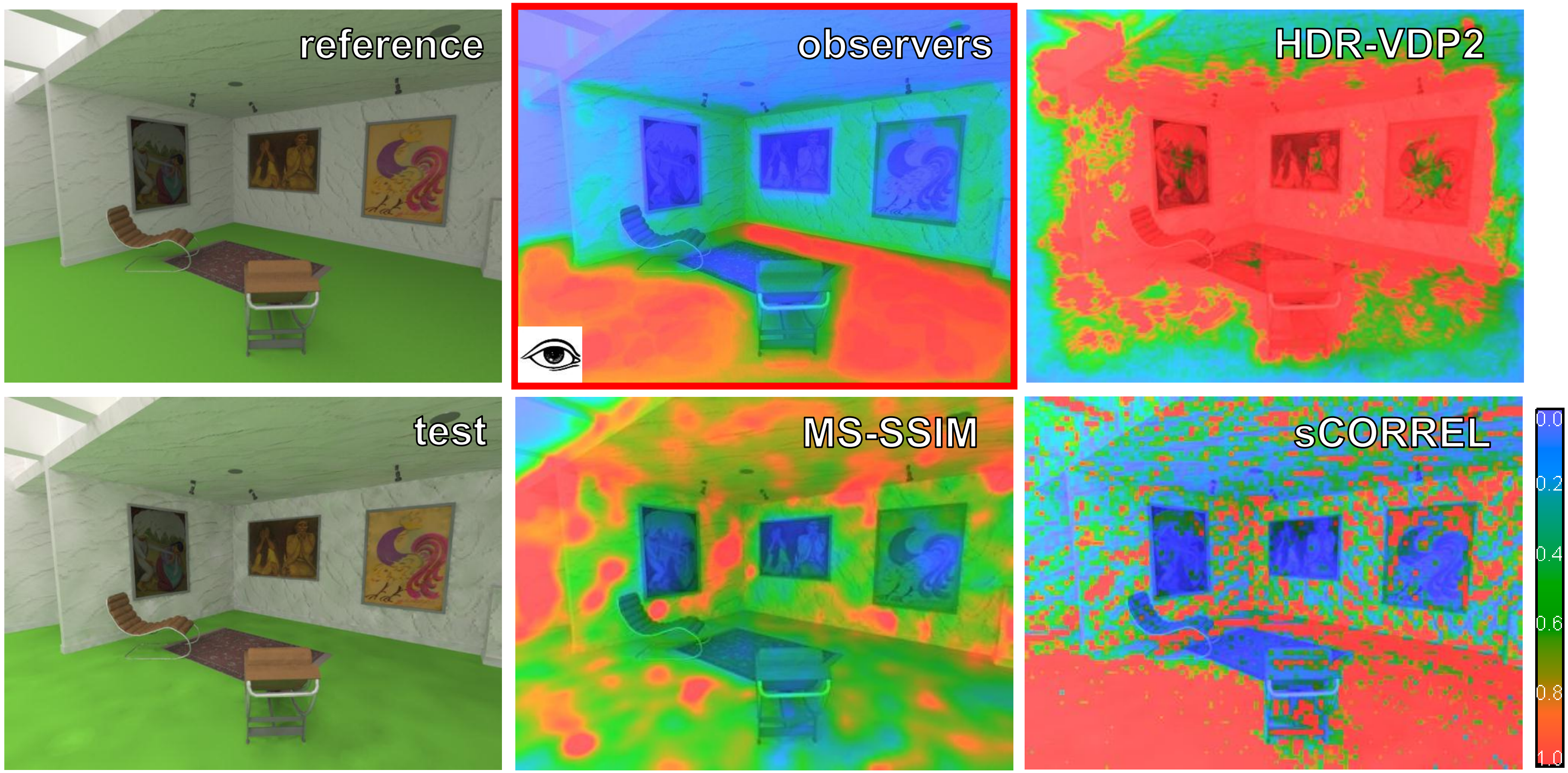
- Bootstrapping (randomization with repetitions 500x)
  - Bonferroni correction
- **No** statistically significant difference between IQMs
- Performance differs significantly per scene



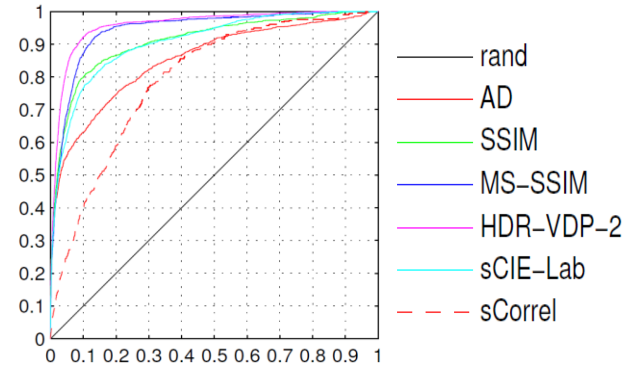
Analysis of Metric Failures



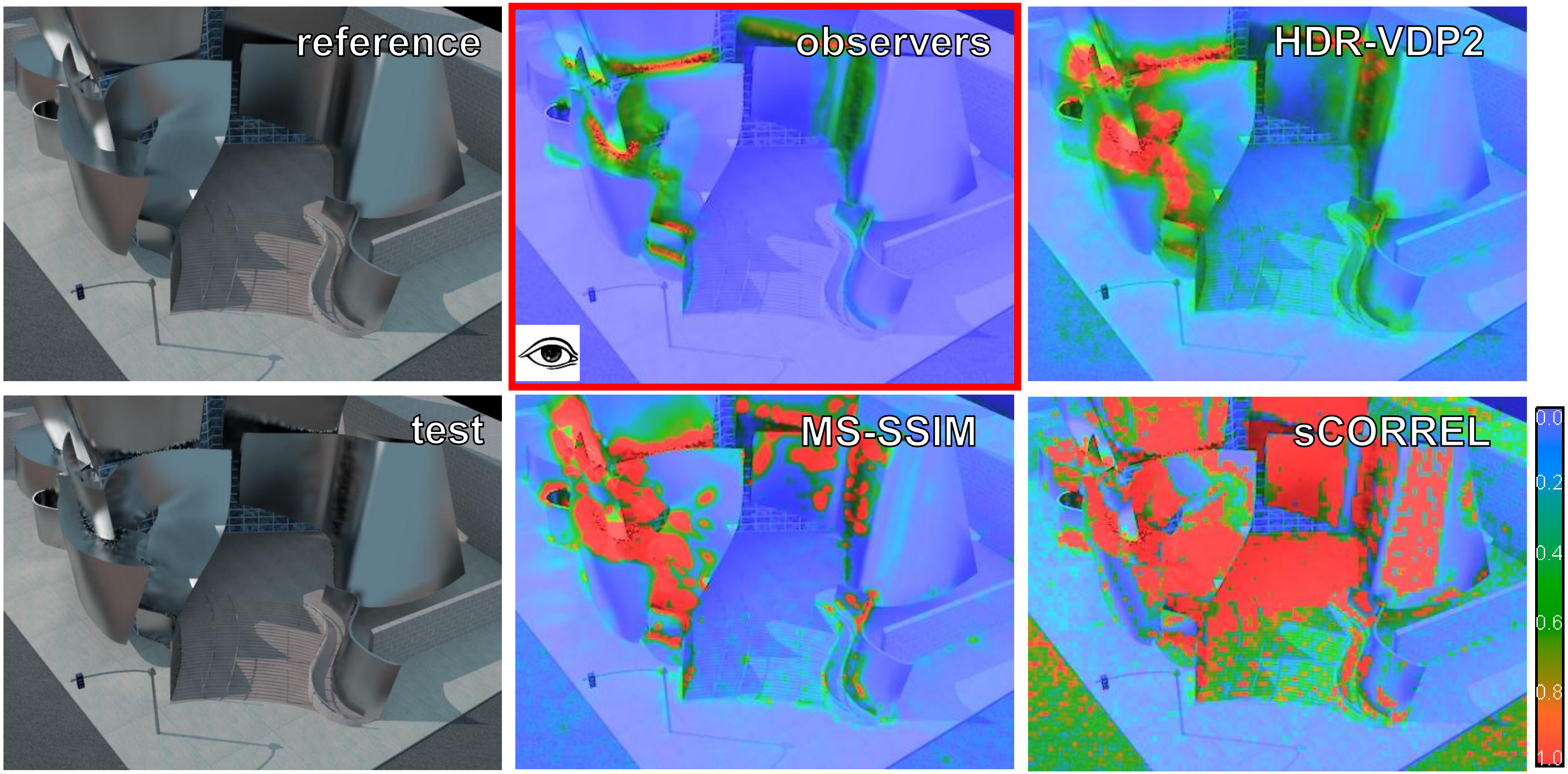
Brightness and contrast change



Analysis of Metric Failures

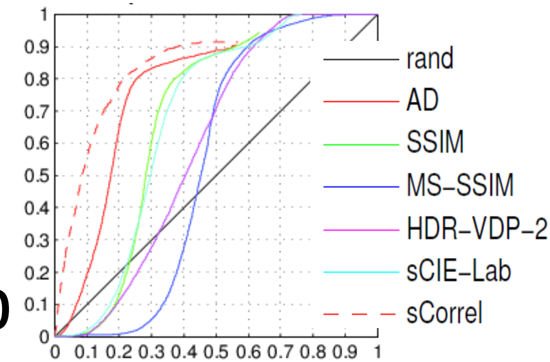


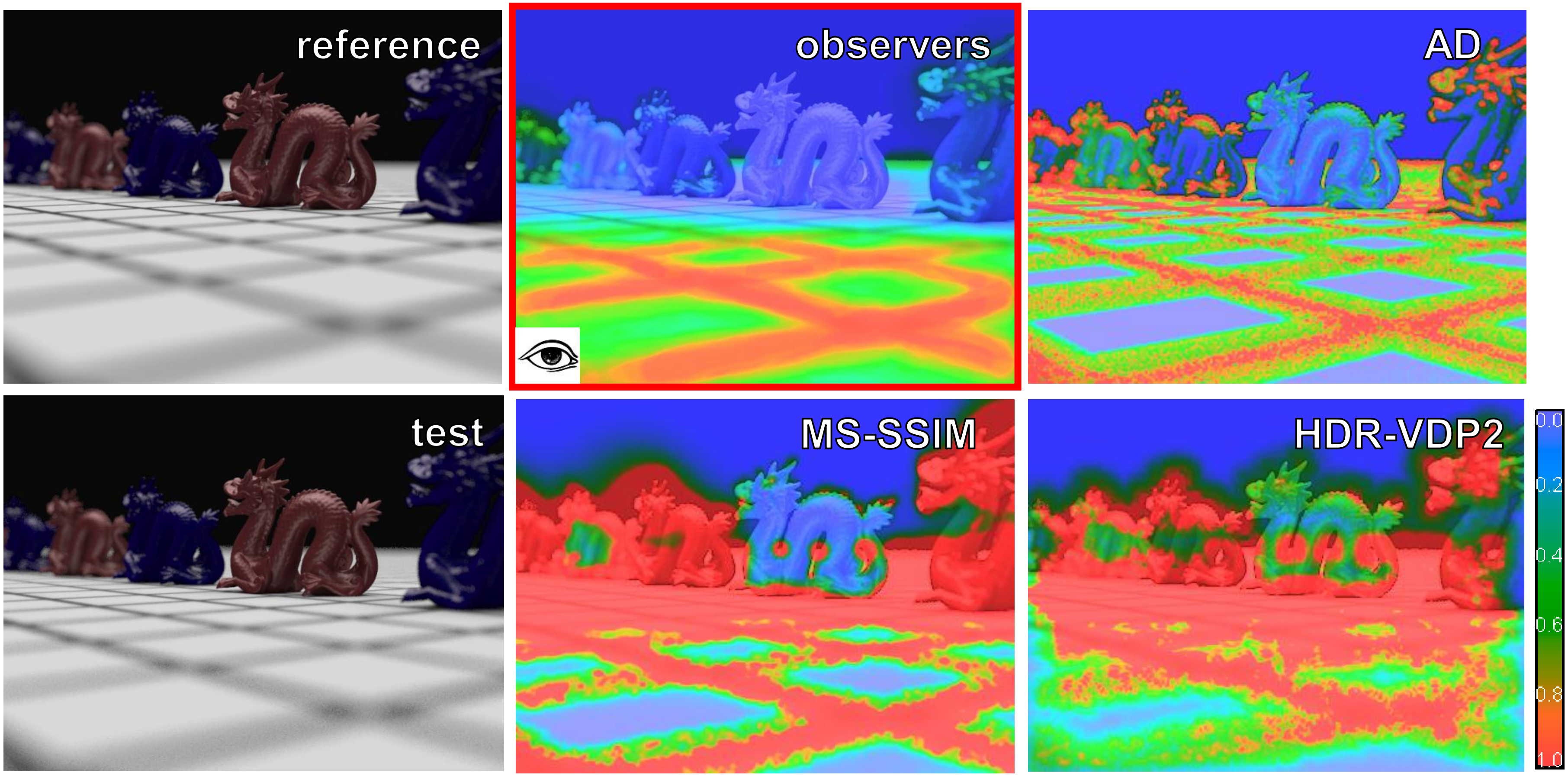
Visibility of low-contrast differences



Analysis of Metric Failures

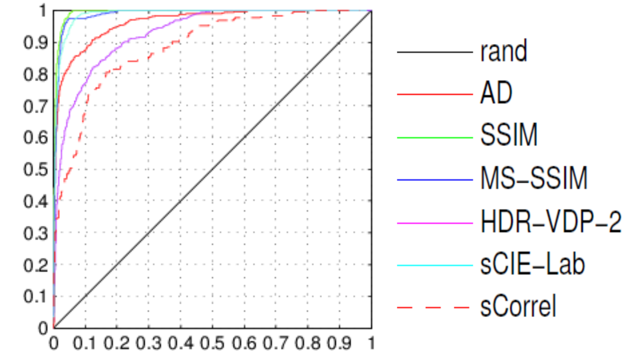
Spatial accuracy of the prediction map





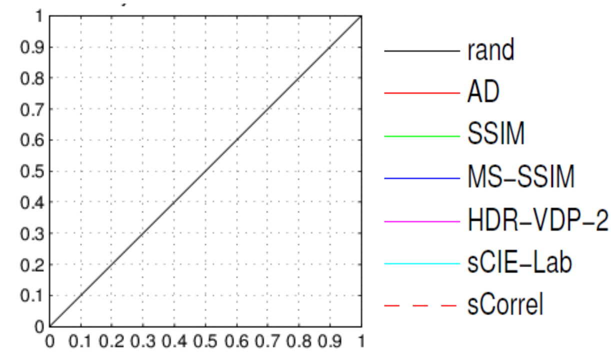
Analysis of Metric Failures

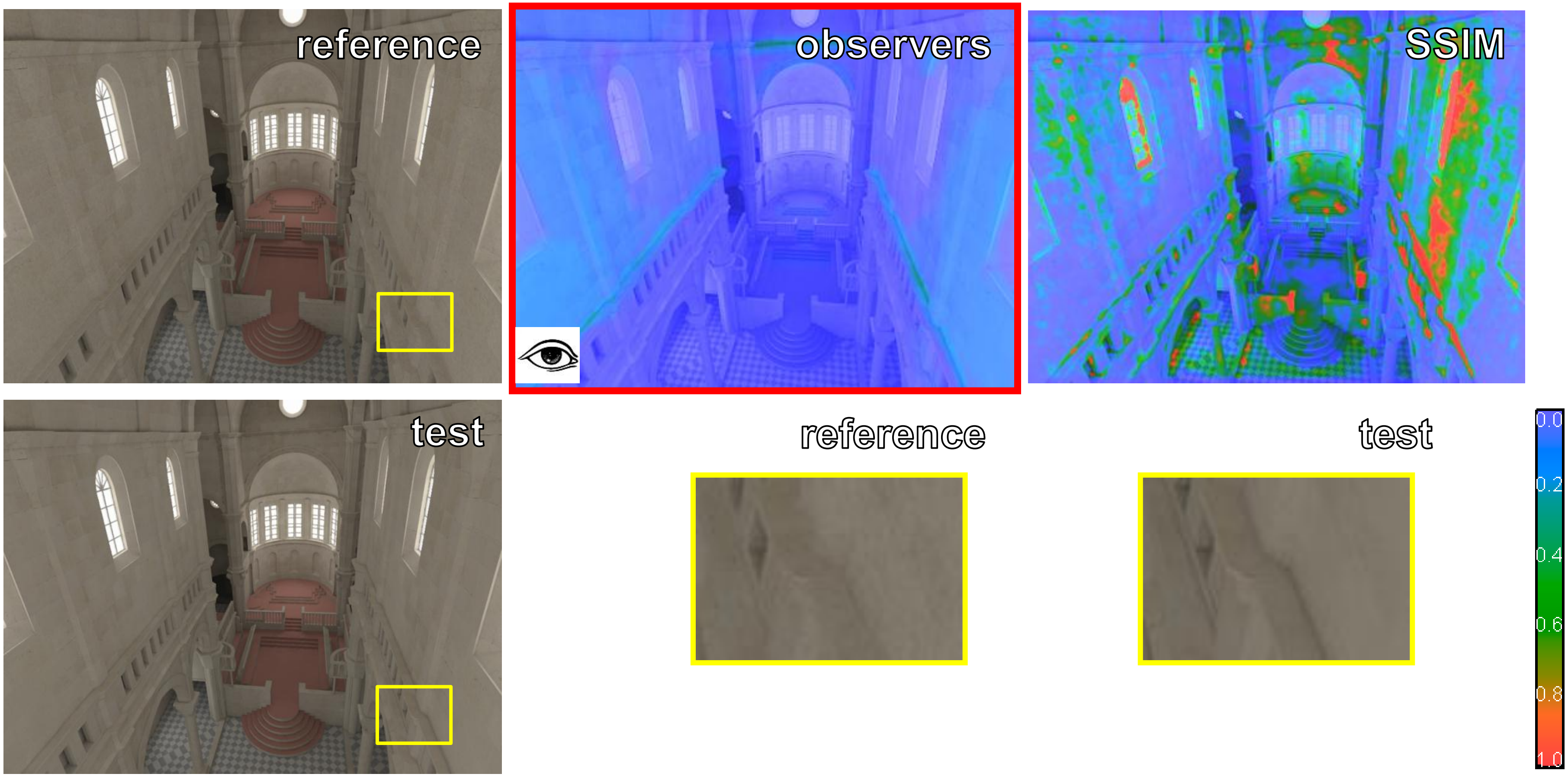
Plausibility of shading



Analysis of Metric Failures

Plausibility of shading (cont.)



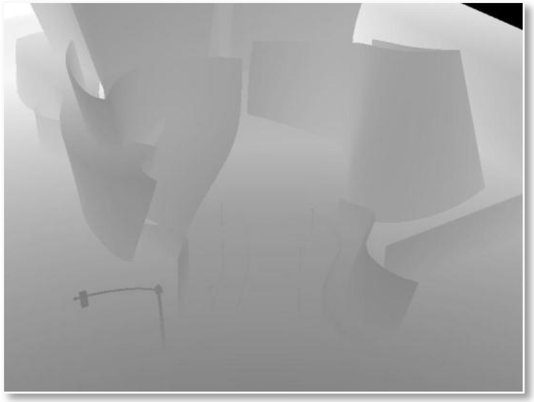


Conclusions

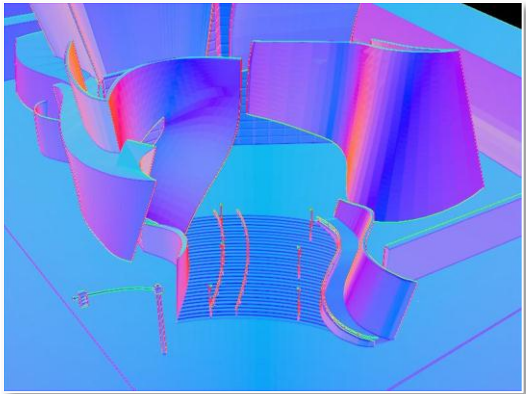
- Rendering datasets for IQM evaluation with subjective localized distortion maps
- With reference ≈ no-reference experiments
- State-of-the-art IQMs far from subjective ground-truths
- No universally reliable metric exists
- Large space for improvements

FW: How to Improve Existing Metrics?

- Data-driven approaches (machine learning)
- Edge-stopping decompositions
- Utilize more information if possible (CG)
  - Similarly to NoRM [Herzog et al. EG'12]



depth buffer



normals

diffuse texture buffer

Future Work (cont.)

- Datasets – more uses possible
  - Development and evaluation of future metrics
  - Visual saliency of rendering artifacts
  - Vision science (real, not “laboratory” stimuli)

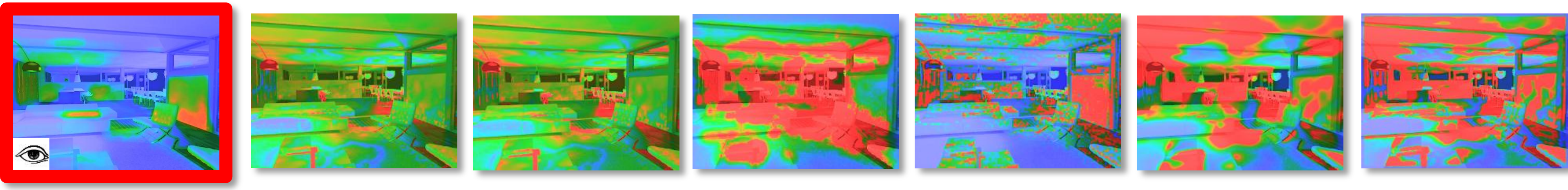
- Effects of visual attention, inattentional blindness, etc.

Thank You For Your Attention

Martin Čadík, Robert Herzog, Rafał Mantiuk,
Karol Myszkowski, Hans-Peter Seidel

http://www.mpi-inf.mpg.de/resources/hdr/iqm-evaluation/

mcadik@mpi-inf.mpg.de



SIGGRAPH ASIA 2012