Supplemental Material for DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Statistics of Natural Videos

ANTON S. KAPLANYAN, ANTON SOCHENOV, THOMAS LEIMKÜHLER*, MIKHAIL OKUNEV, TODD GOODALL, and GIZEM RUFO, Facebook Reality Labs

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Perception**; **Virtual reality**; *Image compression.*

Additional Key Words and Phrases: generative networks, perceptual rendering, foveated rendering, deep learning, virtual reality, gaze-contingent rendering, video compression, video generation

ACM Reference Format:
Anton S. Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. Supplemental Material for DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Statistics of Natural Videos. *ACM Trans. Graph.* 38, 4, Article 212 (July 2019), 5 pages. https://doi.org/10.1145/3355089.3356557

Fig. 1. User study setup. Left: display setup; right: HMD setup

1 COMPRESSION DENSITY FOR A DISPLAY

From the CSF, a function of maximum perceptible frequency $f_m(e)$ (cycles/degree) vs. eccentricity can be derived by equating to maximum contrast and solving for frequency. This function, parameterized by fixation distance e, is defined as

$$f_m(e) = \frac{e_2 \ln(1/CT_0)}{\alpha(e + e_2)} \tag{1}$$

where e_2, the half-resolution eccentricity distance, is 2.3°, CT_0, the minimum contrast threshold, is 1/64, and α, a sensitivity falloff parameter, is 0.106.

The pixel distance of a point $\mathbf{x}$ from the point of fixation $\mathbf{x}^f$ is

$$d(p) = ||\mathbf{x}^f - \mathbf{x}||_2. \tag{2}$$

The minimum angular displacement between adjacent pixels at point $\mathbf{x}$ informs critical display frequency and is provided by

$$\theta(\mathbf{x}) = \min\left[cos^{-1}\frac{\mathbf{AB}}{||\mathbf{A}||_2||\mathbf{B}||_2}, cos^{-1}\frac{\mathbf{AC}}{||\mathbf{A}||_2||\mathbf{C}||_2}\right],$$

where

$$A(\mathbf{x}) = \langle (x_0 - x_c) * q, (x_1 - y_c) * q, D\rangle$$
$$B(\mathbf{x}) = \langle (x_0 - x_c + 1) * q, (x_1 - y_c) * q, D\rangle$$
$$C(\mathbf{x}) = \langle (x_0 - x_c) * q, (x_1 - y_c + 1) * q, D\rangle,$$

q is the pixel pitch estimated by dividing physical display width by horizontal resolution, D is the distance between observer and display, and x_c and y_c are the horizontal and vertical center pixel coordinates respectively. Finally, the minimum angular pixel size at particular eccentricity is f_d (cycles/degree) is

$$f_d(\mathbf{x}) = \frac{1}{2|\theta(\mathbf{x})|}.$$

The ratio of f_m to f_d describes the amount of resolution that the eye can pick up vs what the display can deliver. The number of pixels that are needed according to the perceptual falloff follows the sampling rate R

$$R(\mathbf{x}) = \min\left[1.0, \frac{p_r f_m(m(\mathbf{x}))}{f_d(\mathbf{x})}\right],$$

where $m(\mathbf{x})$ is the angle between a point on the display $\langle x_0, x_1, D\rangle$ and the fixation point $\langle x_0^f, x_1^f, D\rangle$, and p_r is the subsampling rate used for controlling sampling rate. Note that R is continuous and bounded on $[0, 1]$. Under this formulation, the number of samples V_{pix} needed to fully cover the retina for a given screen of NxM resolution is provided by

$$V_{pix} = \frac{1}{MN}\sum_{\mathbf{x}\in\mathbb{X}}\left[1 - R(\mathbf{x})\right],$$

where $\mathbb{X}$ is the set of pixel locations, M is vertical resolution, and N is horizontal resolution. Note that p_r is monotonic with sampling rate, allowing us to map desired sampling rate to p_r, giving direct control over the average sampling rate computed over an entire frame.

*Joint affiliation: Facebook Reality Labs, MPI Informatik.

Authors' address: Anton S. Kaplanyan, kaplanyan@fb.com; Anton Sochenov, anton.sochenov@oculus.com; Thomas Leimkühler, tleimkueh@mpi-inf.mpg.de; Mikhail Okunev, mokunev@fb.com; Todd Goodall, todd.goodall@oculus.com; Gizem Rufo, Facebook Reality Labs, gizem.rufo@oculus.com.

0730-0301/2019/7-ART212
https://doi.org/10.1145/3355089.3356557

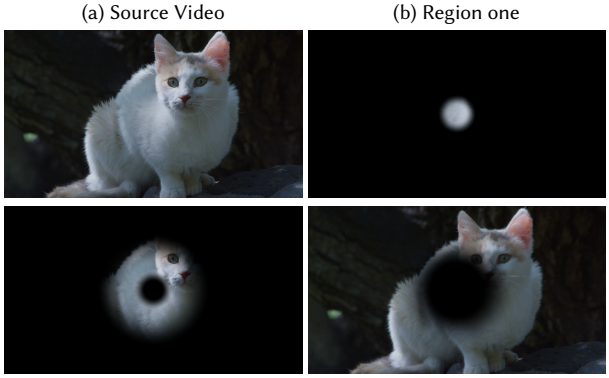

Fig. 2. Regions of interest for both multiscale and H.265 methods.

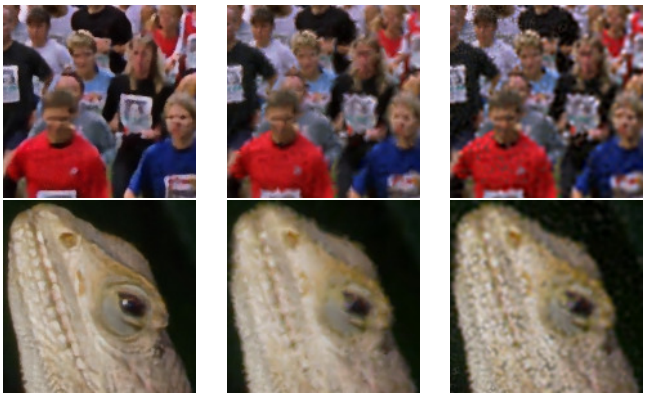

Fig. 3. Comparison among DeepFovea reconstruction, Delaunay, and RBF inpainting from identical sparse samples, using unseen content.

2 CONCENTRIC H.265 PERCEPTUAL BIT-BUDGET ALLOCATION

The number of bits per second is capped at

$$B \leq c_r H W b_{pp} f_r,$$

where H and W are the vertical and horizontal resolution respectively, b_{pp}, bits per pixel, is set to 12 throughout, c_r is the compression rate, and f_r is frame rate. We have three concentric regions that must be allocated from this shared $B/1024$ kbits, according to $B = w_1 B_1 + w_2 B_2 + w_3 B_3$, where w_1, w_2 and w_3 are proportional weights according to region size one, two, and three respectively. For each video and compression rate used in H.265, we use the same region radii, defined as the same multiresolution radii. A visual demonstration of these radii is provided in Fig. 2.

Given fixed radii and fixed bitrate, we must finally decide on the bit allocation for each spatial region. To remain comparable to DeepFovea, we use the ganglion cell density function to develop the relative weights corresponding to region size and retinotopic locations. In other words, we use the normalized midget ganglion cell density map to produce a perceptual importance weighting, which is a simple strategy for bit allocation by keeping the allocation decision in terms of receptor density. This density of cells in the retina correlates highly with the size of brain regions dedicated to each eccentric region [Duncan and Boynton 2003]. The same strategy is repeated when creating the compressed videos for the HMD experiment.

3 DEEPFOVEA RECONSTRUCTION VS INPAINTING

Recent work demonstrates that sparsely sampled scenes can be reconstructed using radial basis functions (RBF) interpolation to achieve a foveated image. In Figure 3, we compare Delaunay-based interpolation, RBF interpolation, and DeepFovea as suggested in [Sun et al. 2017] using identical sampling conditions. For RBF and Delaunay, sampled video frames were reconstructed on a per-frame basis. We found that RBF and Delaunay-based interpolation methods introduce high levels of flicker and spatial noise throughout the whole video, including the foveal regions. Higher-quality spatio-temporal reconstruction can be achieved from both in-hallucination and temporal accumulation using DeepFovea.

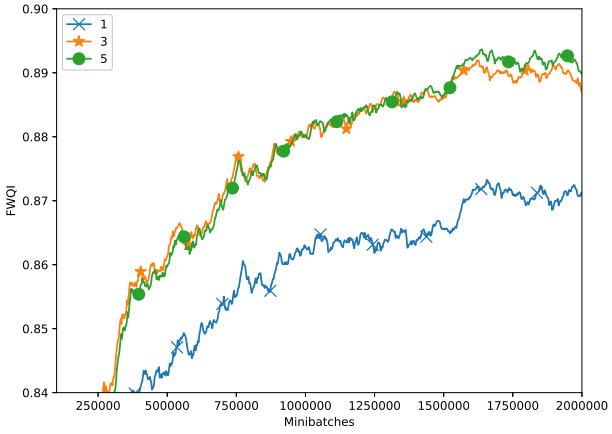

Fig. 4. FWQI graph for networks with different depth of the U-net (number of encoder blocks).

4 ABLATION STUDY ADDITIONAL PLOTS

Figure 4 provides an analysis of DeepFovea performance for different numbers of encoding blocks. We use the FWQI foveated frame-quality metric to measure performance, and we find that the reconstruction quality greatly improves when the number of encoding layers is increased from 1 to 3, but not much when increased from 3 to 5.

5 CORRELATION WITH FWQI AND FA-SSIM

We computed mean gaze from the recorded gaze locations in both screen and hmd subjective studies for each content, method, and sampling rate. First, we treated each gaze location from each subject

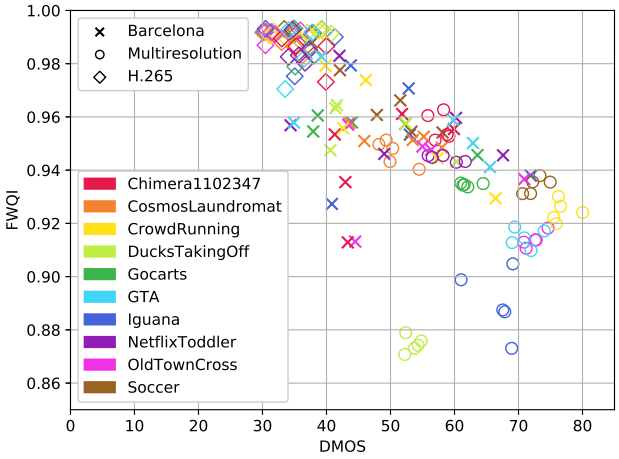

Fig. 5. Scatterplot between FWQI and ground truth DMOS for DeepFovea, Multiresolution, and H.265 methods. Different colors indicate different content.

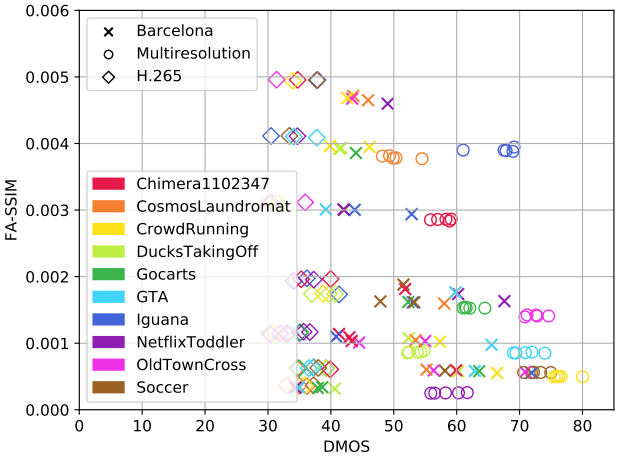

Fig. 6. Scatterplot between FA-SSIM and ground truth DMOS for DeepFovea, Multiresolution, and H.265 methods. Different colors indicate different content.

as a unit impulse located at pixel locations in a frame. Second, we convolved these impulses with a Gaussian window with standard deviation 100, corresponding to 3.34° [Rai et al. 2017]. Finally, we picked the maximum point, which isolates the "most seen" point amongst observers.

Feeding these mean gaze points into FWQI [Wang et al. 2001] and FA-SSIM [Rimac-Drlje et al. 2011] allowed us to compute each metric based on the gaze of an average observer on each frame of each video. Figure 5 compares FWQI with mean gaze to the ground truth DMOS for both DeepFovea and Multiresolution. We find a significant correlation between the metric and our subjective scores. For FA-SSIM, we observe a poor correlation with our computed

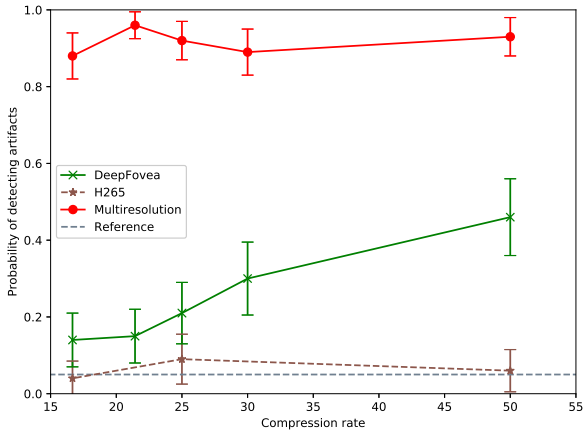

Fig. 7. A summary of detectability results from HMD experiment. Green shows mean detectability for five compression rates measured for DeepFovea. Red shows Multiresolution. Dashed brown line shows H.265 and the dashed black line represents reference videos. The x-axis represents compression rate. Error bars represent bootstrapped 95% confidence intervals.

DMOS, as depicted in Fig. 6. We also observe that the scores and FWQI both independently demonstrate a systematic difference between DeepFovea and Multiresolution. Since FWQI does not model temporal information or masking, it mispredicts on content with high motion and details, such as "DucksTakingOff."

6 HMD DETECTABILITY EXPERIMENT RESULTS SUMMARY

From Fig. 7, we find that DeepFovea is much less detectable than Multiresolution. Also, for lower compression rates, we find that DeepFovea is statistically indistiguishable from H.265 and reference up until 25x compression, which corresponds to a sampling rate of 4%.

7 CORRELATION WITH FWQI AND FA-SSIM PER VIDEO

Figures 8 and 9 depict a per-content breakdown of DMOS results with 95% confidence intervals for the screen and hmd studies respectively. Across the contents and display types, we noticed that "cosmoslaundromat" and "duckstakingoff" had the highest degree of overlap in confidence intervals amongst the methods. In the case of "cosmoslaundromat," most of the content away from the center of the image is part of the out-of-focus background, which can easily be captured by each approach. For "duckstakingoff," most of the content contains rippling water waves, which tends to mask spatial-temporal artifacts.

8 SCREEN EXPERIMENT: DETECTABILITY RESULTS FOR EACH VIDEO

Figure 10 analyzes detectability performance of DeepFovea compared to H.265 and Multiresolution for each video in the screen study. We find that the contents "NetflixToddler" and "DucksTakingOff"

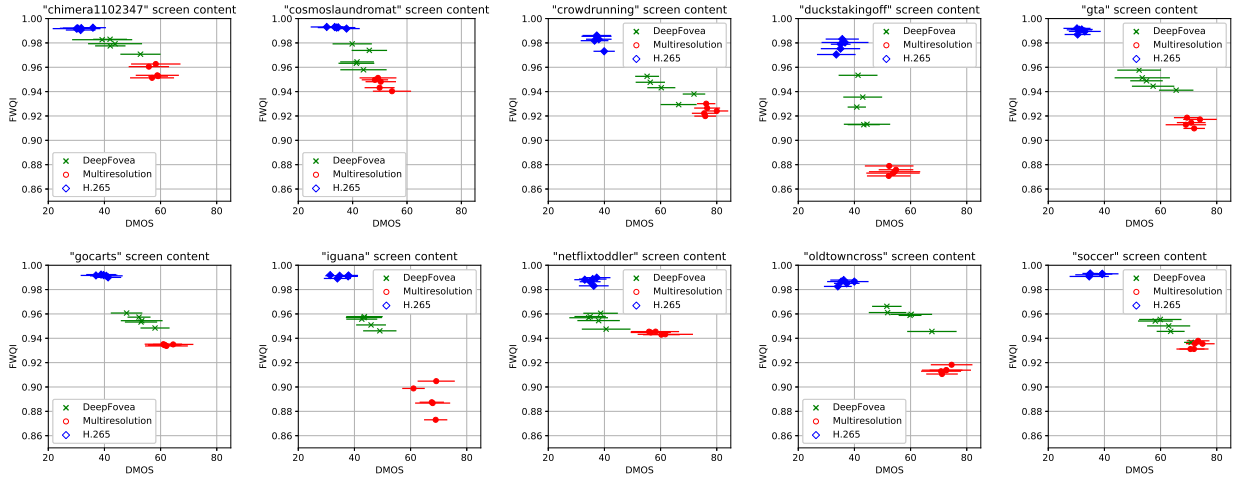

Fig. 8. Scatterplots between FWQI and ground truth DMOS for DeepFovea, Multiresolution, and H.265 methods. Different colors indicate different content. Error bars indicate 95% confidence intervals.

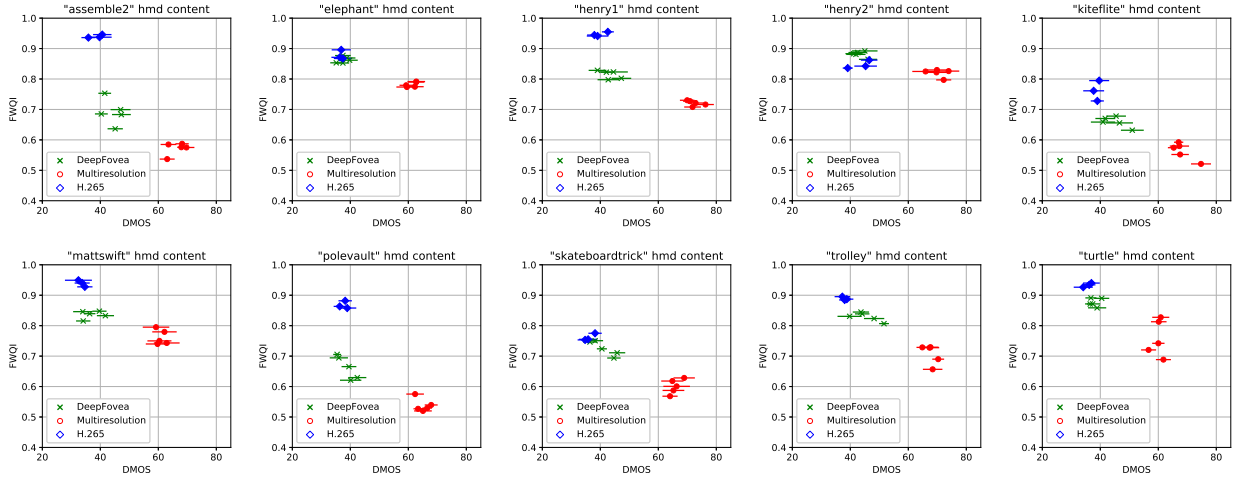

Fig. 9. Scatterplots between FWQI and ground truth DMOS for DeepFovea, Multiresolution, and H.265 methods. Different colors indicate different methods. Error bars indicate 95% confidence intervals.

are tend to mask artifacts, sometimes causing methods to become visually indistinguishable. Contents with high camera and object motion such as “OldTownCross” and “CrowdRunning” are particularly difficult for both Multiresolution and DeepFovea, with relatively higher visibility of artifacts when compared to other content.

9 HMD EXPERIMENT: DETECTABILITY RESULTS FOR EACH VIDEO

Figure 11 analyzes detectability performance of DeepFovea compared to H.265 and Multiresolution for each video in the HMD study. We find that multiple contents are indistinguishable for subjects when comparing DeepFovea with H.265. When looking specifically at “Assemble2” and “Trolley” we noticed that artifact detectability is higher than for other contents. These contents, like before, demonstrate a large degree of motion, which is not represented naturally in the viewed reconstructions.

REFERENCES

Robert O Duncan and Geoffrey M Boynton. 2003. Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron* 38, 4 (2003), 659–671.

Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 205–210.

S. Rimac-Drlje, G. Martinović, and B. Zovko-Cihlar. 2011. Foveation-based content Adaptive Structural Similarity index. *International Conference on Systems, Signals and Image Processing* (2011), 1–4.

Qi Sun, Fu-Chung Huang, Joohwan Kim, Li-Yi Wei, David Luebke, and Arie Kaufman. 2017. Perceptually-guided Foveation for Light Field Displays. *ACM Trans. Graph. (Proc. SIGGRAPH)* 36, 6, Article 192 (2017), 192:1–192:13 pages.

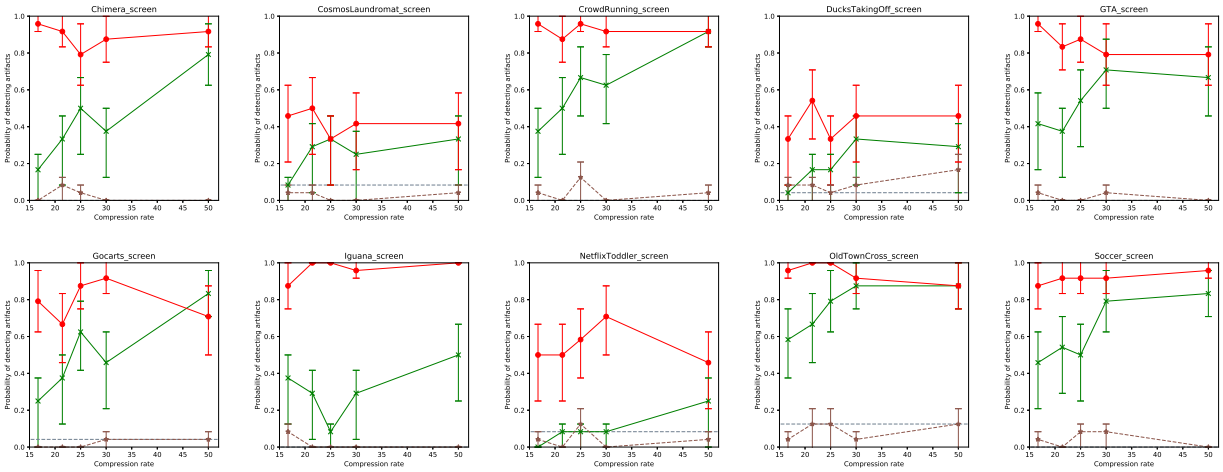

Fig. 10. Per-video results for screen study. Error bars indicate 95% confidence intervals.

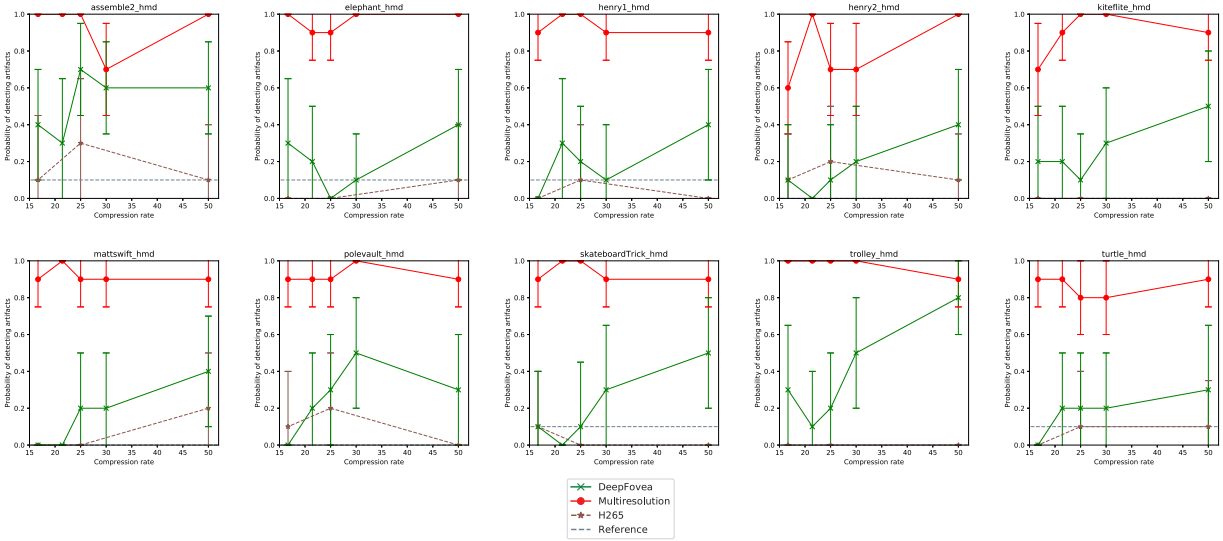

Fig. 11. Per-video results for HMD study. Error bars indicate 95% confidence intervals.

Zhou Wang, Alan Conrad Bovik, Ligang Lu, and Jack L Kouloheris. 2001. Foveated wavelet image quality index. *Proc. SPIE* 4472 (2001), 42–53.