Supplemental Document

Optics Letters

Neural compression for hologram images and videos: supplement

LIANG SHI,[1,2] RICHARD WEBB,[1] LEI XIAO,[1] CHANGIL KIM,[1] AND CHANGWON JANG[1,*]

[1]*Meta Reality Lab, 9845 Willows Road NE, Redmond, Washington 98052, USA*
[2]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, USA*
[*]*Corresponding author: Changwon.Jang@fb.com*

Supplement DOI: https://doi.org/10.6084/m9.figshare.21400797

Parent Article DOI: https://doi.org/10.1364/OL.472962

Neural Compression for Hologram Images and Videos: supplemental document

1. TRAINING AND EVALUATION OF HIFIHC

The dynamic focal stack loss is defined as

$$d_{fs}(x,x') = \sum_{d_t \in D_t^{\text{Rand}}} \Big(\Big|\Big| |\text{ASM}(x,d_t)| - |\text{ASM}(x',d_t)| \Big|\Big|_1 + \Big|\Big| \nabla|\text{ASM}(x,d_t)| - \nabla|\text{ASM}(x',d_t)| \Big|\Big|_1 \Big) \tag{S1}$$

where $|\cdot|$ is the element-wise absolute value operator, t denotes the training iteration, D_t^{Rand} is a set of random offsets from the hologram plane to the locations in the volume of 3D content that varies per training iteration, ASM denotes the angular spectrum method for free-space propagation,

$$\text{ASM}(x,d_t) = \mathcal{F}^{-1}\left(\mathcal{F}(x) \odot \exp\left(i2\pi d_t \sqrt{\lambda^{-2} - \mu^2 - v^2} \right) \right), \tag{S2}$$

where $\mu \in \mathbb{R}^{R_x \times R_y}$ and $v \in \mathbb{R}^{R_x \times R_y}$ are the spatial frequencies along the x and y directions, ∇ is the gradient operator, $\odot$ denotes Hadamard element-wise product, $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the 2D Fourier transform and inverse Fourier transform operator.

We train image and video versions of HiFiHC with $w_r = 2^3$ when the bpp is greater than the target bpp, $w_r = 2^{-5}$ when the bpp is less than the target bpp, $w_{holo} = 19.125 \times 2^{-5}$, $w_{fs} = 19.125 \times 2^{-5}$, $w_D = 0.015$. For both versions, D_t^{Rand} consists of 15 random depths. Following HiFIC, We use a learning of 10^{-4} for the first half million iterations and decay it to 10^{-5} for the second half million iterations.

At evaluation, we use mv-extractor [1] (https://github.com/LukasBommes/mv-extractor) to extract the motion vector (see Fig. S1 for a visualization). Because mv-extractor is only designed for the H.264 codec, we encode a separate H.264 video along with the H.265 video only for extracting the motion vectors. We defer motion vector extraction of H.265 video to future engineering as this does not affect the method performance. We optionally adopt the translated macroblock during motion compensation of the residual image if the compensated region yields a smaller mean square error (MSE) than the MSE before compensation. This error check is necessary because the reverse case could happen when the motion vectors connect two blocks with a close appearance but do not correspond to the same area across the frames (i.e., most of the macroblocks in Frame 4 of Fig. S1). In such cases, the translated residual rarely reduces the MSE error. The acceptance of the macroblock is recorded by a binary flag and entropy coded as side information.

2. ADDITIONAL RESULTS

We demonstrate additional results (see Fig. S2, S3, and S4) on hologram image and video compression results for various captured and computer-rendered scenes. These results illustrate the consistently improved performance available with HIFIHC over conventional image and video compression codecs, particularly with the improved resolution of fine details in the refocus results (marked by the white box).

3. ADDITIONAL NOTES ON PROPAGATION DISTANCE VS. HIFIHC PERFORMANCE

When constructing practical holographic head-mounted displays (HMDs), the 3D image formed by the hologram locates approximately one focal length behind the eyepiece, whereas the SLM is typically placed closer to reduce the display form factor. This induces a propagation distance from the hologram on the SLM to the volume of the 3D image, and this distance can vary based on different design choices and targeted applications of the HMDs. Here, We evaluate

HiFiHC performance at various propagation distances. Specifically, we train HiFiHC models at 5mm, 10mm, 15mm, and 20mm propagation distances from the hologram to the 3D image. Figure S5 illustrates the compression performance of HiFiHC versus the conventional codecs. For both image and video compression, HiFiHC maintains the performance gain across all the tested propagation distances. In image compression, HiFiHC shows a slower performance drop when the propagation distance is prolonged. This is because conventional codecs are only optimized for images with natural statistics, and the hologram statistics differ more from it as the propagation distance prolongs. In contrast, the independently-trained HiFiHC model is less sensitive to the hologram statistics changes. For video compression, our hybrid approach of using HiFiHC to compute residual for a high CRF video has to suffer the performance drop of conventional codec. However, it still consistently outperforms a lower CRF video by 1dB or more.

4. ADDITIONAL NOTES ON RATE-DISTORTION CURVES

Figure S6 visualizes the rate-distortion(RD) curve for both image and video compression evaluated at the hologram plane and a stack of object planes, for later of which five layers are evenly sampled through the object space, and the mean PSNR is reported. The image RD curves exhibit very different statistics from the video RD curves. For the image RD curves, HiFiHC achieves a $\sim$3dB improvement over BPG and HEIC at the hologram plane and a +4dB improvement at the object plane. The relative improvement difference at the hologram plane and the object planes is minor. In contrast, video-version HiFiHC only achieves a $\sim$0.2dB improvement at the hologram plane but a +1.4dB improvement at the object plane, resulting in a big relative improvement difference. This is caused by the fact that, in video compression, the CNN is only responsible for reconstructing the residual, which contributes little to the overall intensity of the image. However, the residual often presents the high-frequency information critical to the refocusing power of the hologram. The focal stack loss encourages the reconstructed residual to preserve the most essential details for refocusing, leading to a significant performance gap at the object planes that requires more than a 0.3bpp increase to match. We note that the PSNRs reported for images and videos are not directly comparable as the test scenes have no overlap, and readers should solely focus on the performance difference within each task category. The image compression method may be adopted in extreme cases where directly encoding each frame as an image outperforms video-based HiFiHC.

5. ADDITIONAL NOTES ON PHASE INITIALIZATION OF GROUND TRUTH CGH

In this work, we train and evaluate our method using smooth-phase ground truth holograms as a lot of recent studies that demonstrate high-quality display results choose smooth-phase initialization [2, 3]. As described in our letter, a smooth-phase hologram eliminates the speckle noise and also facilitates the use of the double phase methods. However, we note that it is still a debatable topic as sometimes random phase hologram demonstrates advantages over smooth-phase hologram depending on the system configuration. For example, in direct view systems without an eye-piece lens, a random phase hologram could better utilize the display's space-bandwidth and provide parallax views with a close-to-uniform spectrum. On the other hand, in a pupil-forming system, a smooth-phase hologram tends to have uneven energy distribution in the eyebox with a high peak at the center. This would sometimes result in vignetting or unnatural blur effects, while random phase holograms can facilitate a more natural blur pattern. Nevertheless, recent studies have demonstrated that temporal multiplexing with smooth-phase hologram could overcome such issues [4]. Since in near-eye displays, the eye moves around with the headset; thus, at one moment in time, only one view is technically needed, whereas the motion-induced view change can be alternatively supported by rendering a new hologram optimized for the new view. Thus, it might be more important to reduce speckles instead of supporting view-dependent effects in near-eye display applications. That said, the pros and cons of both smooth and random initialization methods are still being explored yet, and we leave the extensive study of the effect of the phase initialization method as future work. We further stress that our training data are mainly designed to help CNN reproduce consumer graphics content (e.g., games, movies, natural images), but to make the CNN a general-purpose compresser contents from such as domain-specific/scientific applications (e.g., holographic microscope or tomography measurement).

6. ADDITIONAL NOTES ON DATA PREPROCESSING

Our hologram is initially represented as an amplitude map and a phase map. For natural-image-like 3D scenes, the resulting amplitude rarely goes beyond 1. In practice, we empirically set the max as $\sqrt{2}$ as it upper bounds the maximum amplitude found in the dataset holograms, and if any value beyond exists, it is clipped. The range of phases is bounded by the periodicity and always clipped into $[0, 2\pi]$. The amplitude and phase are converted to real and imaginary components and fed to the network. For all test holograms, we didn't see any influence of the clipping. Again, this assumption is valid for the smooth phase hologram we intended for and may not hold for arbitrary holograms.

7. ADDITIONAL NOTES ON ALTERNATIVE COMPRESSING STRATEGIES

In light of recent works on efficient hologram computation, an alternative solution for compression is compressing the hologram's input and relying on edge devices to directly compute holograms in real-time. A viable input that can readily leverage existing codecs is the RGB+D input (e.g., using 3D-HEVC). However, many works have shown that more complex representations such as layered depth images (LDI), light fields, or full point clouds are necessary to incorporate wavefront diffracted by non-line-of-sight parts in the scene [5, 6]. Such information is critical to producing a natural depth boundary. While which representation is optimal is still in debate, RGB-D has proved insufficient. Efficient codecs for these more advanced representations are either under development or haven't received enough attention yet. Therefore, at this moment in time, transmitting/decoding one of these new representations and then running a real-time CGH algorithm may not be more computational/power-efficient than directly decoding the hologram. Nevertheless, this is another promising path of solution that is worth further investigation.

REFERENCES

1. L. Bommes, X. Lin, and J. Zhou, "Mvmed: Fast multi-object tracking in the compressed domain," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA),* (2020), pp. 1419–1424.
2. L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik, "Towards real-time photorealistic 3D holography with deep neural networks," (2021).
3. A. Maimone, A. Georgiou, and J. S. Kollin, "Holographic near-eye displays for virtual and augmented reality," ACM Trans. Graph. **36**, 1–16 (2017).
4. S. Choi, M. Gopakumar, Y. Peng, J. Kim, M. O'Toole, and G. Wetzstein, "Time-multiplexed neural holography: A flexible framework for holographic near-eye displays with fast heavily-quantized spatial light modulators," in *Proceedings of the ACM SIGGRAPH,* (2022), p. 1–8.
5. L. Shi, B. Li, and W. Matusik, "End-to-end learning of 3D phase-only holograms for holographic display," Light. Sci Appl **11**, 247 (2022).
6. P. Chakravarthula, E. Tseng, H. Fuchs, and F. Heide, "Hogel-free holography," ACM Trans. Graph. (2022).
7. C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," ACM Trans. Graph. **32**, 1–12 (2013).
8. L. Xiao, S. Nouri, M. Chapman, A. Fix, D. Lanman, and others, "Neural supersampling for real-time rendering," ACM Transactions on (2020).
9. B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: practical view synthesis with prescriptive sampling guidelines," ACM Trans. Graph. **38**, 1–14 (2019).
10. D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Computer Vision – ECCV 2012,* (Springer Berlin Heidelberg, 2012), pp. 611–625.

Fig. S1. Visualization of motion vector extracted by mv-extractor, only ones with non-zero magnitude are shown. Readers are encouraged to zoom in and checkout for Big Bunny Bunny Frame 2-4. The rectangle box denotes the macroblock, the processing unit in conventional image and video compression codecs based on linear block transforms (i.e., DCT). The arrow line denotes the motion vector, where it starts from the center of the macroblock in the reference frame and ends at the center of the macroblock in the current frame. The Big Buck Bunny scene represents a type of scene where some content undergoes a non-translation motion, whereas the rest remains close to stationary. The Orchids represent another type of scene where the whole scene undergoes a translation motion with perspective projection.

Fig. S2. Additional comparison of HiFiHC, HEIC, and BPG performance on hologram images. Readers are encouraged to zoom in and examine details. The second and the third row in each label mark the peak signal to noise ratio (PSNR) and structure similarity index (SSIM) for the hologram amplitude (first number) and the refocused DoF image (second number). Source images: Couch (top left) from Kim et al. [7], Mr. Elephant by Glenn Melenhorst, Robot (bottom left), and Home (bottom right) scene from Xiao et al. [8]

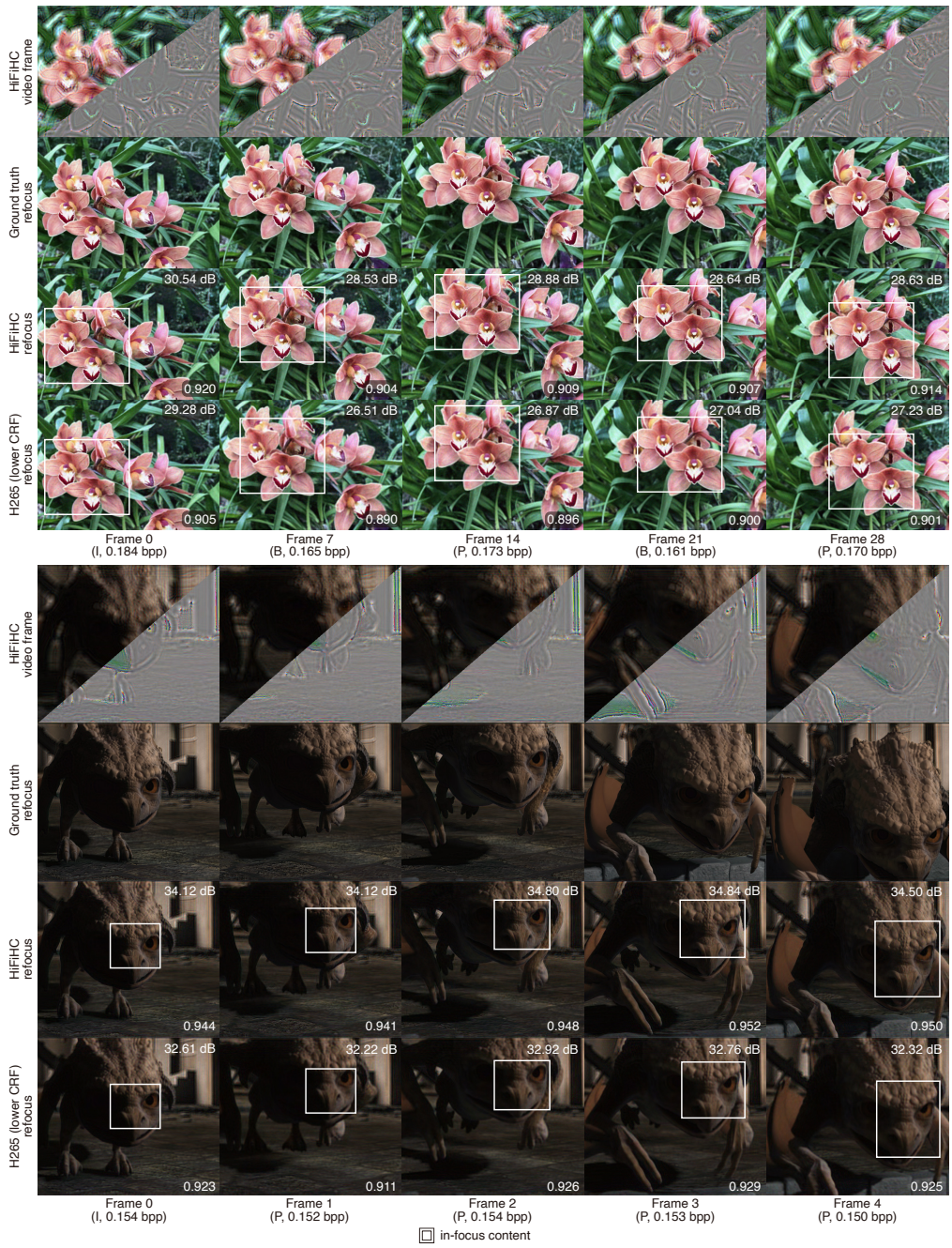

Fig. S3. Additional comparison of HiFiHC and H.265 (at lower CRF) performance on hologram videos. Readers are encouraged to zoom in and examine details. The top and bottom numbers in each inset mark the PSNR and SSIM for the refocused DoF image. The second row in the frame label marks the frame type and the bpp of the HiFiHC latent code. Source images: Orchids (top) from [9], and Market_5 (bottom) from MPI Sintel dataset [10]. The H265 (lower CRF) results in the use CRF of 18 and 19 for Orchids and Market_5, respectively, both of which yield a similar amount of additional bpp compared to the HiFiHC.

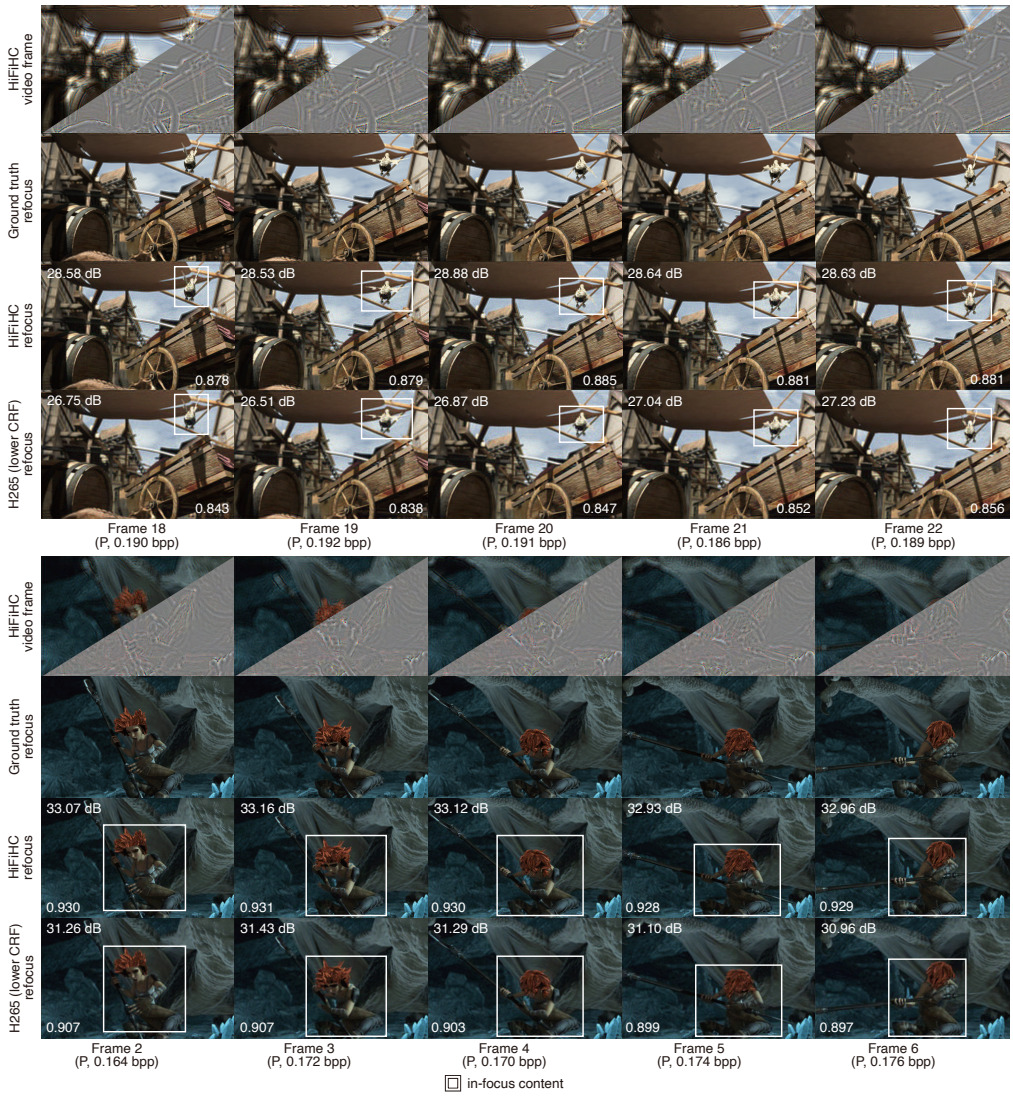

Fig. S4. Additional comparison of HiFiHC and H.265 (at lower CRF) performance on hologram videos. Readers are encouraged to zoom in and examine details. Each inset's top and bottom numbers mark the PSNR and SSIM for the refocused DoF image. The second row in the frame label marks the frame type and the bpp of the HiFiHC latent code. Source images: Market_6 (top) and Cave_4 (bottom) from MPI Sintel dataset [10]. The H265 (lower CRF) results use CRF of 20 and 20 for Market_6 and Cave_4, respectively, both of which yield a similar amount of additional bpp compared to the HiFiHC.

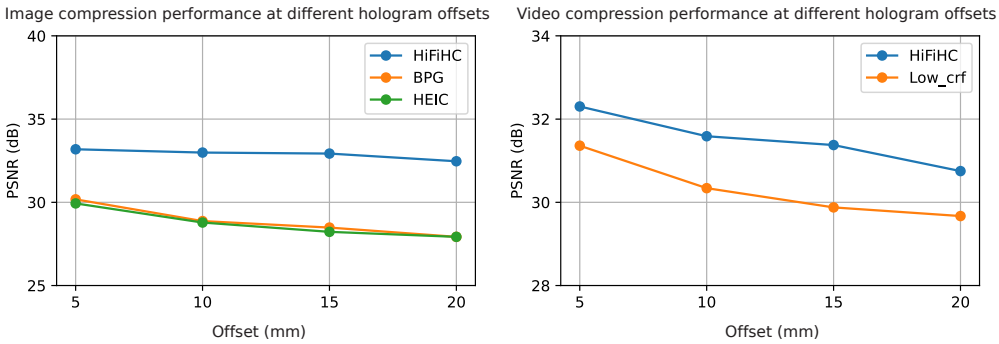

Fig. S5. Performance comparison of hologram image compression (left) and video compression (right) under different offsets between the 3D volume and the hologram. The PSNRs are calculated for the refocused insets in Fig. 2, Fig. S2 (for image compression); Fig. 3, Fig. S3, Fig. S4 (for video compression). For both hologram image and video, the PSNR decreases as the offset increases.

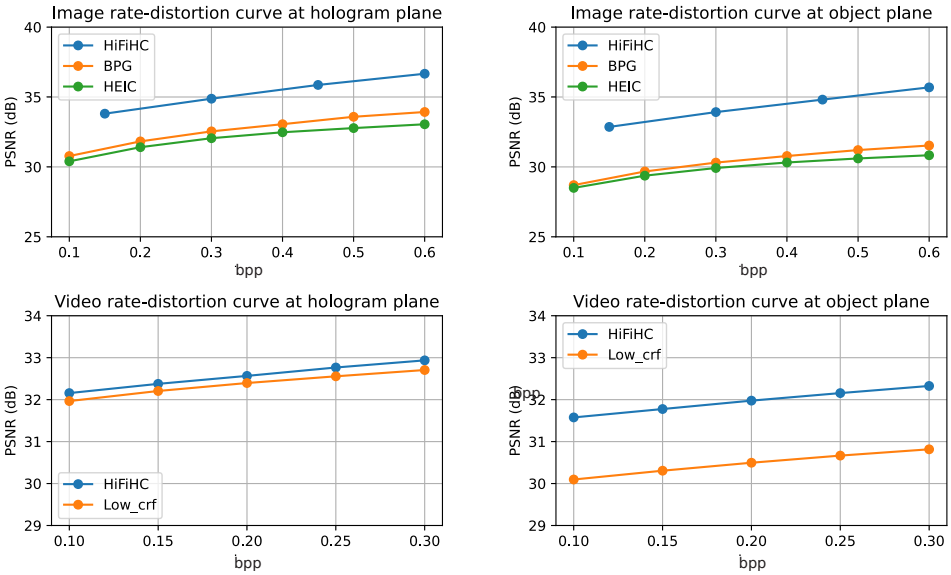

Fig. S6. Rate-distortion curves for image and video compression evaluated at the hologram plane and a stack of object planes. For object planes, five layers are evenly sampled through the object space and the mean PSNR is reported. The PSNRs are calculated for the scenes in Fig. 2, Fig. S2 (for image compression); Fig. 3, Fig. S3, Fig. S4, and additional test scenes from Xiao et al. [8] (for video compression).